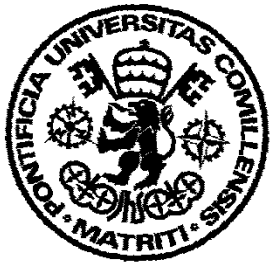
UNIVERSIDAD PONTIFICIA COMILLAS
MADRID

Escuela Técnica Superior de Ingeniería (ICAI)
Departamento de Electrotecnia y Sistemas

Modelos especializados en la detección incipiente de fallos

Rafael Palacios Hielscher

Tesis doctoral



Madrid 1998

Modelos especializados en la detección incipiente de fallos

Universidad Pontificia Comillas de Madrid
Colección Tesis Doctorales: Nº 215 / 1998

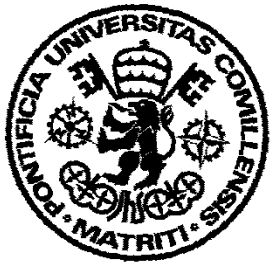
UNIVERSIDAD PONTIFICIA COMILLAS
MADRID

Escuela Técnica Superior de Ingeniería (ICAI)
Departamento de Electrotecnia y Sistemas

Modelos especializados en la detección incipiente de fallos

Rafael Palacios Hielscher

Tesis doctoral



Madrid 1998

A mi mujer, Isabel

porque ha sido mi mejor descubrimiento

Agradecimientos

En la página final de la tesis, que aparece aquí por obligaciones de formato, quiero aprovechar la oportunidad para agradecer la colaboración de todas las personas que han participado en este trabajo.

Después de tantos años de leer, programar, calcular y dibujar gráficas, esta es sin duda la parte más difícil de la tesis, ya que la mayoría de la gente no va a leer nada más que esto.

En primer lugar quiero expresar mi agradecimiento a mi amigo José Ignacio Pérez Arriaga, que por otro lado ya figura en la portada por haber sido, además, mi director de tesis. Tengo que agradecerle una brillante labor de dirección. Por su paciencia durante todos estos años. Por saber escuchar, entender y entonces proponer un nuevo plan de trabajo. Por haberme rescatado de algunos desvíos hacia temas "que son otra tesis" porque sino no habría terminado. Y por todos los comentarios y discusiones (algunos más formales y otros en plena excursión campestre). También agradecer algunas de las aportaciones que han contribuido a dar forma de tesis a unas ideas iniciales, simples pero atrevidas, que sugerían apartarse de las técnicas clásicas de ajuste de parámetros. Finalmente destacar las múltiples revisiones del documento, rápidas pero sobre todo rigurosas.

En segundo lugar quiero agradecer a Miguel Ángel Sanz todo el tiempo que me ha dedicado durante estos años, sus aportaciones, revisiones y comentarios. La tesis se ha visto muy beneficiada de esta segunda dirección. También por ser la persona que me convenció para hacer la tesis y por no haber dejado de animarme a terminarla.

También agradecer a George Verghese sus sugerencias y conversaciones relacionadas principalmente con la estimación recursiva de parámetros y a Juan Luis Zamora la minuciosa revisión de algunos capítulos.

Mis primeros pasos en la instalación de sensores, adquisición de datos, análisis de señales reales y detección de fallos tuvieron lugar hace ahora más de diez años. En aquel tiempo fue Juan Carlos Lavalle quien me enseñó lo fundamental de estos temas y quien me introdujo en el mundo de la investigación. Aunque no ha podido ayudarme durante el desarrollo de la tesis, los primeros pasos en este camino fueron fundamentales. Le envío mi agradecimiento por una vía especial.

También doy las gracias a la Universidad Pontificia Comillas como institución, por la ayuda que supone un ambiente de trabajo agradable y con buenos medios técnicos y humanos. Este trabajo ha sido parcialmente financiado por Iberdrola dentro del plan de Ayudas a la investigación científica y al desarrollo tecnológico.

Agradezco la colaboración de los compañeros del IIT que me han ayudado en mayor o menor medida, ya sea en aspectos técnicos como en apoyo moral (esos momentos en que la tesis se atasca). Afortunadamente se trata de un grupo muy grande, la mayoría buenos amigos, y se llenaría media página si pongo a todos. Pero sin que nadie se sienta ofendido quiero destacar a Lucía Muñoz, Juan Rivier y Rafael Collantes.

Pero lo más importante ha sido la ayuda de mi mujer. Porque sólo a base de tiempo la tesis sale adelante con mucha dificultad; pero cuando te sale bien la vida, luego la tesis viene por añadidura. Ha sido decisivo un esfuerzo conjunto durante los últimos meses para "no apuntarnos a todos los planes y excursiones" y así poder terminar antes.

Mis padres y hermanos también han tenido que sufrir los agobios de esta tesis, que han sido la continuación de los de la carrera. Les agradezco la paciencia y comprensión que han tenido durante tanto tiempo.

En muchas ocasiones esta tesis se ha visto temporalmente interrumpida por largos y difíciles tratamientos médicos. La tesis tampoco habría llegado a su fin de no haber sido por los cuidados recibidos en este tiempo. Muchas personas han intervenido, pero especialmente el Dr. Javier Hornedo de Oncocenter, Madrid y el Dr. George Canellos de Dana Farber Cancer Institute, Boston.

Índice

Capítulo 3
Estructura general del sistema de detección incipiente de fallos

Capítulo 4
Estructura del procedimiento de ajuste

Capítulo 5
Optimización del sistema

Capítulo 6
Detección de fallos mediante estimación de parámetros

Capítulo 7
Ejemplo práctico

Capítulo 8
Conclusiones y futuros desarrollos

Apéndice A
Modelado de sistemas mediante Bond–Graphs

Apéndice B
Bibliografía

Capítulo 1

Introducción

1.1 Motivación

Ningún proceso industrial tiene una fiabilidad total sino que en todos los casos existe una posibilidad de que el proceso no realice satisfactoriamente el trabajo para el cual ha sido diseñado. Diremos que el proceso ha fallado cuando no realiza su trabajo dentro de la tolerancia requerida.

El fallo de un proceso se produce como consecuencia del fallo en alguno de sus componentes. Es frecuente que el fallo en un único componente provoque una parada de todo el proceso, si bien existen sistemas redundantes en los cuales la avería de un componente puede no ser suficiente para provocar un fallo del proceso. En este tipo de procesos una avería localizada disminuye la fiabilidad o las prestaciones del sistema pero no provoca un fallo. Pensemos por ejemplo en el caso de un avión comercial al cual se le avería un motor en pleno vuelo. Todos los aviones comerciales están diseñados para poder volar, maniobrar y aterrizar con un motor

averiado por lo tanto este fallo de un componente no da lugar a un fallo de todo el sistema.

Las consecuencias del fallo de un proceso son muy variadas, pueden ir desde producir una leve incomodidad (sistemas no críticos) hasta poner el peligro vidas humanas (sistema de seguridad), pasando por producir grandes pérdidas económicas. Por lo tanto existe mucho interés por evitar que se produzcan los fallos o al menos en disminuir la gravedad de sus consecuencias. Para ello se trabaja tanto en la fase de diseño como en la fase de explotación.

Durante la fase de diseño se aumentan la fiabilidad de los sistemas aplicando coeficientes de seguridad, sobredimensionando piezas, utilizando equipos redundantes etc. La falta de conocimiento profundo sobre las características de los materiales y sobre la fiabilidad de los componentes lleva a diseños de mayor coste para evitar fallos.

Durante la fase de explotación se realizarán inversiones en mantenimiento para evitar y corregir fallos en el proceso. Pueden aplicarse distintos planes de mantenimiento en función de criterios de seguridad o ponderando los costes de mantenimiento con los costes de la aparición de averías inesperadas.

Las inversiones que se realicen durante la fase de diseño y durante la fase de explotación para evitar fallos en procesos dependerán fundamentalmente de la gravedad de las consecuencias que puedan producir estos fallos. Por esta razón existen muchos equipos y procesos industriales que justifican claramente la realización de inversiones orientadas a evitar o disminuir la aparición de fallos. Por otro lado también ocurre que muchos sistemas trabajan actualmente por debajo de sus posibilidades, con objeto de garantizar ciertos niveles de seguridad. Ante un aumento de las necesidades de producción se plantea la duda de realizar inversiones en nuevos equipos o mejorar la explotación de las instalaciones existentes. Al aplicar sistemas de monitorización y diagnóstico a los equipos existentes, se puede trabajar a mayor rendimiento con igual o mejor nivel de seguridad.

Esta tesis surge con objeto de mejorar la explotación de procesos industriales gracias a la detección incipiente de fallos. La detección de fallos se realiza mediante análisis de los datos del proceso obtenidos en tiempo real. Para este análisis de los datos se utilizan modelos de comportamiento

del proceso y una serie de módulos que forman la estructura general del sistema de detección. El ajuste del sistema de detección debe orientarse a una especialización en la detección incipiente de fallos y el procedimiento de ajuste que se propone permite considerar distintos criterios de evaluación del sistema (seguridad, robustez...) de forma simultánea.

1.2 Planes de mantenimiento

Todos los equipos industriales requieren acciones de mantenimiento a lo largo de su vida operativa. Las acciones de mantenimiento se llevan a cabo tanto por la aparición de averías inesperadas como por sustitución de componentes que sufren desgaste o envejecimiento. Normalmente se habla de tres estrategias de mantenimiento (o tipos de mantenimiento) que definen, para cada equipo, el mejor momento en que deberían realizarse las acciones de mantenimiento. El plan de mantenimiento que se aplica en cada caso se elige en función de criterios de seguridad, técnicos y económicos.

1.2.1 Mantenimiento correctivo

En la estrategia de mantenimiento correctivo sólo se realizan acciones de mantenimiento cuando se produce una avería. Es el plan que realiza menos acciones de mantenimiento, sin embargo, éstas pueden resultar muy caras cuando la reparación de las averías resulta larga y costosa, como en el caso de averías en cadena.

Se aplica en el caso de equipos o procesos en los que una parada de una duración equivalente al tiempo de reparación no supone grandes gastos económicos ni provoca una situación de peligro. También puede aplicarse en el caso de sistemas redundantes donde un fallo simultáneo sea improbable, la reparación sea sencilla y el fallo de uno de los componentes sea fácilmente detectable y no se propague a otros sistemas. Por ejemplo, los tubos fluorescentes para iluminación se montan en mamparas de al menos dos tubos y sólo se sustituye un tubo cuando se observa que se ha fundido, a pesar de conocer de forma estadística cuál es la vida del componente.

Otro caso son los sistemas en que los fallos son realmente aleatorios y no dependen de los cuidados del equipo. Por ejemplo los neumáticos de un vehículo se sustituyen cuando se pinchan, pero no es posible predecir ni evitar el pinchazo, por lo tanto ninguna acción previa de sustitución del neumático estaría justificada.

1.2.2 Mantenimiento preventivo

En la estrategia de mantenimiento preventivo el objetivo es evitar la aparición de fallos. Para ello se recurre a realizar acciones de mantenimiento, tales como la sustitución de componentes afectados por desgaste, con suficiente antelación. Para aplicar este plan de mantenimiento es necesario disponer de ciertos conocimientos sobre la vida media de los componentes para poder definir el mejor momento de la sustitución.

Suele aplicarse en el caso de sistemas de funcionamiento crítico o sistemas de seguridad en los cuales una parada inesperada provoca grandes pérdidas económicas o una situación de peligro. En los sistemas en los que se aplica mantenimiento preventivo se planifican paradas periódicas, que se realizan sin peligro y produciendo las mínimas pérdidas posibles, para realizar la sustitución de componentes.

Algunos ejemplos son la sustitución de aceites y filtros en máquinas y vehículos, las revisiones periódicas de centrales nucleares, las revisiones de aviones cada cierto número de horas de vuelo, etc.

1.2.3 Mantenimiento predictivo

Las acciones de mantenimiento se adelantan o se retrasan en función de un conocimiento preciso del estado de salud de los componentes. Este tipo de mantenimiento evita realizar sustituciones innecesarias de componentes sanos gracias a un mayor conocimiento del equipo. Por otro lado puede evitar las pérdidas asociadas a paradas inesperadas si los sistemas de predicción de fallos son suficientemente buenos.

La aplicación de un plan de mantenimiento predictivo se basa fundamentalmente en la existencia de un sistema de diagnóstico en tiempo real. Este sistema debe conocer el estado de salud de los componentes y debe ser capaz de predecir fallos en el sistema con cierta exactitud.

El ahorro económico que supone el plan de mantenimiento predictivo frente al preventivo puede justificar la inversión en equipos y tecnología para diagnóstico en tiempo real.

1.3 Sistemas de diagnóstico y sistemas de detección de fallos

Los sistemas de detección de fallos se encargan de detectar una anomalía o mal funcionamiento de un proceso. Por ejemplo, un medidor del caudal impulsado por una bomba que detecte una disminución brusca del caudal, estando la bomba conectada, es un sencillo sistema de detección de fallos.

Un sistema de diagnóstico es capaz de analizar todos los síntomas disponibles de manera conjunta y localizar el origen de la avería. Siguiendo con el ejemplo anterior, un sistema de diagnóstico sería capaz de distinguir entre un fallo del motor que acciona la bomba, la aparición de fugas en el circuito o la falta de fluido en la aspiración.

Muchas veces los sistemas de diagnóstico no cuentan con toda la información necesaria para emitir una conclusión definitiva por lo que algunas veces pueden sugerir la realización de medidas complementarias o emitir los diagnósticos con un determinado grado de certeza. Este es el caso del sistema experto de diagnóstico de transformadores TRAFES [Sanz93-1] [Sanz93-2]. Aunque la detección de anomalías requiere utilizar sólo algunos sensores, para el proceso de diagnóstico es útil contar con más medidas [Joussellin95].

En el caso de sistemas de diagnóstico en continuo, que se explican más adelante, se puede conocer en cualquier momento el estado de salud de cada componente. Estos sistemas organizan la información orientada a componentes o modos de fallo concretos.

Es interesante diseñar los sistemas de detección de fallos de forma que faciliten el proceso de diagnóstico. Para ello es posible crear sistemas para la detección de fallos en componentes concretos, y hacer los sistemas de detección poco sensibles al ruido, intentando evitar las falsas alarmas en condiciones especiales previamente identificadas y confirmando la situación de fallo. En el caso de definir un sistema de detección basado en el análisis de relaciones entre variables, pero sin tener en cuenta el origen físico de estas relaciones, es interesante estudiar la conexión entre cada modo de fallo y las anomalías que se detectan. El objetivo es que cada modo de fallo quede establecido por afectar a un subconjunto determinado de residuos, lo que define un código de fallo (*fault signature* [Gertler93]). De esta manera se simplifica el trabajo del sistema de diagnóstico pasando parte del trabajo al diseño de un sistema de detección de fallos más sofisticado.

1.4 Técnicas de detección de fallos

Existen muchas maneras de detectar y diagnosticar fallos en equipos industriales. En la mayoría de los casos no se realiza ningún tipo de comprobación de los equipos ya que se trata de componentes no críticos y donde los fallos son evidentes (por ejemplo bombillas eléctricas). En otras ocasiones los fallos sólo pueden detectarse mediante complejos ensayos de calibración (por ejemplo instrumentos de medida). En función del tipo de equipo se pueden aplicar distintas técnicas de detección de fallos.

1.4.1 Técnicas de control de calidad

En muchos procesos de producción el control de calidad es una operación obligada y la información que en ella se recoge permite detectar fallos en la cadena de producción y puede ayudar a localizarlos. Por ejemplo en la fabricación de piezas, si se detecta un aumento en el número de piezas que se encuentran fuera de tolerancia se habrá detectado un fallo en el proceso de rectificado. Si se conoce además en qué rectificadora se han acabado las piezas se habrá localizado el fallo. Esta técnica no evita la producción de piezas defectuosas, como haría un sistema de detección incipiente de fallos

en la máquina rectificadora [Lavalle92] [Lavalle93], sin embargo evita realizar calibraciones innecesarias en las máquinas herramientas.

En definitiva, las técnicas de control de calidad contribuyen a la detección y diagnóstico de fallos, si bien se trata de un método no preventivo ya que detecta los fallos *a posteriori*.

1.4.2 Técnicas de inspección

Cuando se trata de equipos cuyos fallos afectan a la producción o a la seguridad se realiza algún tipo de control de funcionamiento. Este control puede comprender desde una simple inspección visual hasta un sistema de diagnóstico en continuo.

Las pérdidas económicas o situaciones de peligro derivadas de fallos en máquinas o equipos pueden obligar a tomar medidas preventivas con objeto de evitar la aparición de fallos. Estas medidas pueden comprender sustituciones o inspecciones de componentes que obligan a detener la máquina (mantenimiento preventivo) o inspecciones orientadas a conocer el estado de salud de los componentes (mantenimiento predictivo). Para que estas acciones sean efectivas, la frecuencia de las inspecciones viene marcada por la velocidad de degradación de los componentes, que suele estar relacionada con el grado de utilización del equipo.

Si las constantes de tiempo son largas, las técnicas de inspección manual pueden resultar apropiadas aunque en muchas ocasiones la información obtenida puede ser insuficiente. Este es el caso de sistemas donde la medida de una magnitud depende de las condiciones de trabajo o donde los equipos de medida utilizados para la inspección no recogen suficiente información. Por ejemplo, los niveles de vibración en máquinas rotativas dependen del grado de carga y de la velocidad (por citar los dos parámetros fundamentales). Los operarios de manteninimeinto suelen realizar rondas periódicas con equipos portátiles de adquisición de datos para registrar los niveles de vibración de las máquinas con objeto de detectar fallos en rodamientos, engranajes, motores, etc. Es importante que las condiciones de trabajo en que se realiza la recogida de información sean siempre las mismas para que los datos tengan validez. También es importante tener en cuenta que la mayoría de los equipos de medida utilizados (llamados "vibrómetros") sólo ofrecen una medida global del nivel de vibración que

es una información muy pobre frente a la información de un análisis espectral completo. Los vibrómetros calculan el nivel medio de vibración en un espectro ancho (entre 10 Hz y 1 kHz según la norma ISO 2372) pero no suelen ofrecer información sobre en nivel de vibración a determinadas frecuencias características de una máquina, que sería una información imprescindible para realizar el diagnóstico o una verdadera detección incipiente de fallos.

La detección basada en inspecciones periódicas resulta sin embargo muy económica, especialmente si los equipos portátiles se utilizan para inspeccionar gran número de máquinas (como hacen las empresas dedicadas al mantenimiento de equipos industriales).

1.4.3 Sistemas de monitorización

Gracias a la evolución de la electrónica, cada vez resulta más asequible para la industria montar equipos de monitorización en continuo. Estos equipos están formados por un conjunto de sensores, sendos acondicionadores de señal y un equipo para de adquisición de datos (también llamado concentrador de señales). Hace unos años los sistemas de monitorización en continuo y los sensosres permanentes sólo estaban justificados para máquinas críticas y como algo ocasional [Piety88], hoy en día es habitual que las máquinas se entreguen con los sensores instalados o al menos que en el diseño se tenga en cuenta la ubicación de sensores.

Los sistemas de monitorización permiten adquirir información a mayor frecuencia y con mayor claridad, tanto por el rigor en la carencia de muestreo como por la repetitividad en el punto de medida. Además la instalación permanente de sensores permite obtener datos, durante el funcionamiento normal, de puntos que serían inaccesibles sin detener y desmontar el equipo.

En el caso de utilizar sistemas continuos de monitorización, las situaciones de funcionamiento especiales pueden filtrarse para manejar solamente el conjunto de datos correspondientes a situaciones homogéneas. Otra ventaja fundamental es que los sistemas de monitorización, a diferencia de los equipos portátiles, pueden manejar grandes volúmenes de información, analizando todos los datos aportados por los sensores antes de simplificarlos.

1.4.4 Sistemas de diagnóstico en continuo

Un sistema de diagnóstico en continuo es capaz de deducir el estado de salud del equipo y de sus componentes a partir de datos reales tanto de operación.

Aunque el diagnóstico basado en los datos puntuales de un momento puede ser de gran utilidad, el máximo beneficio se obtiene al aplicar el sistema de diagnóstico en tiempo real. Por ejemplo el mantenimiento de los transformadores eléctricos de potencia se realiza basándose en el resultado de análisis puntuales de las propiedades físico-químicas y de la concentración de gases en el aceite aislante del transformador. El diagnóstico que se puede emitir no es muy preciso, fundamentalmente porque no se suele conocer la temperatura en el instante de muestreo ni la historia de carga del transformador (como se menciona en [Crowley90]). Pero lo más importante es que los análisis están tan espaciados en el tiempo que generalmente no resulta posible detectar los fallos de manera incipiente. La configuración ideal en este caso es utilizar un sistema de diagnóstico en tiempo real, que analice las señales suministradas por una serie de sensores y que sea capaz de recomendar la realización de análisis de laboratorio y de interpretar sus resultados [Sanz93-1] [Sanz93-2].

Un sistema de diagnóstico de fallos en tiempo real debe detectar las anomalías y localizar la causa del problema. En general debe realizar las siguiente tareas:

- Recoger datos del proceso de manera periódica (la frecuencia de muestreo depende fundamentalmente de la dinámica del proceso)
- Analizar los datos para detectar fallos en el proceso
- En caso de fallo, localizar el componente dañado

El proceso de localizar las causas del fallo resulta muy complejo en la mayoría de los casos, por el gran número de componentes y de señales a considerar así como por la falta de conocimiento profundo sobre los modos de fallo, y generalmente se recurre a técnicas de inteligencia artificial para poder abordar el problema. En los sistemas de diagnóstico en tiempo real desarrollados en el área de sistemas inteligentes del Instituto de Investigación Tecnológica se han desarrollado distintos tipo de sistemas

expertos para realizar el proceso de diagnóstico [Lavalle92], [Lavalle93], [Sanz93-1], [Sanz93-2], [Sanz94]. Se distinguen tres módulos para realizar las tareas antes mencionadas, estos módulos son respectivamente el sistema de adquisición de datos, el sistema de detección de anomalías y el sistema experto de diagnóstico.

El **sistema de adquisición de datos** está formado por los sensores —que se instalan en el equipo a monitorizar— los acondicionadores de señales, el convertidor analógico a digital y un programa de comunicaciones. Por lo tanto se trata fundamentalmente de un sistema *hardware*, que en la práctica está bastante bien resuelto (salvo el desarrollo de determinados sensores) y cada vez es más asequible para la industria. Además muchos equipos industriales incorporan sistemas de control y tienen instalados sensores para medir los parámetros más característicos, estos sensores y las bases de datos que hayan generado (si existen) pueden utilizarse para desarrollar un sistema de detección de fallos.

El **módulo de detección de anomalías** se encarga de analizar todos los datos que se obtienen del proceso y en su caso indicar al sistema de diagnóstico si hay una anomalía. En general hay que pensar en el sistema de detección de anomalías como un conjunto de módulos que pueden utilizar diferentes técnicas (ver capítulo 2) y pueden estar organizados por componentes.

El **módulo de diagnóstico** es un sistema experto basado en reglas, que en algunos casos ha combinado técnicas de lógica borrosa [Sanz93-1] [Sanz93-2]. Este módulo trabaja al más alto nivel y utiliza toda la información disponible del proceso y la obtenida por el módulo de detección de anomalías.

Esta tesis se centra en el problema de la detección de fallos y no en el proceso de diagnóstico, por lo tanto este trabajo está dirigido al módulo de detección de anomalías y no al módulo de diagnóstico. Pero como se ha mencionado anteriormente, el módulo de detección de anomalías puede resolver en gran medida el problema de diagnóstico, por ejemplo cuando la detección de fallos se realiza orientada a componentes en lugar de al análisis de las señales de forma independiente. En esta tesis se remarca la necesidad de analizar las variables por grupos, fundamentalmente mediante la utilización de modelos.

1.5 Descripción general de la tesis

En este primer capítulo de introducción se ha descrito el marco de aplicación de la metodología de detección incipientes de fallos que se propone en esta tesis. También se han descrito brevemente los distintos tipos de mantenimiento y las técnicas de detección de fallos que se aplican actualmente en la industria.

En el segundo capítulo se describen aquellas técnicas de detección de fallos que pueden trabajar de forma incipiente. Se expone la necesidad de utilizar modelos dinámicos de los procesos monitorizados para poder realizar detección incipiente sin depender de las condiciones de funcionamiento del proceso. También se describen las dos técnicas fundamentales de detección en continuo: la detección basada en el análisis de residuos y la detección basada en el análisis de la evolución de los parámetros del modelo. Los modelos se utilizan para estimar el valor de determinadas variables características de un proceso. Por comparación entre los valores estimados por el modelo y los valores reales de las variables (medidos con sensores) se obtiene una medida de discrepancia conocida como residuo. Los métodos de detección basados en el análisis de residuos utilizan directamente esta información para decidir si existe fallo o no. Por otro lado, los métodos de detección basados en el análisis de la evolución dinámica de los parámetros, utilizan modelos adaptativos que se reajustan de manera continua con objeto de anular los residuos. En este caso la detección se basa en analizar la evolución de los valores que van tomando los parámetros del modelo en cada instante de muestreo.

En el tercer capítulo (página 35) se describe la estructura del sistema de detección incipiente de fallos que es aportación de esta tesis. Esta estructura es suficientemente general como para considerar detección basada en análisis de residuos o detección basada en el análisis de la evolución de los parámetros. Se trata de una estructura modular que admite utilizar de manera simultánea varios criterios que indiquen la existencia de fallos (a estos criterios se les llama **atributos de fallo**, AF). La decisión final sobre el estado de salud del proceso se lleva a cabo en un módulo independiente llamado **función de detección de fallos** (FDF) que

se encarga de evaluar conjuntamente los diferentes atributos de fallo. La función de detección de fallos puede ser una combinación lineal de los atributos de fallo, una función de tipo lógico o una función basada en lógica borrosa. El esquema general del sistema de detección incipiente de fallos es el siguiente:

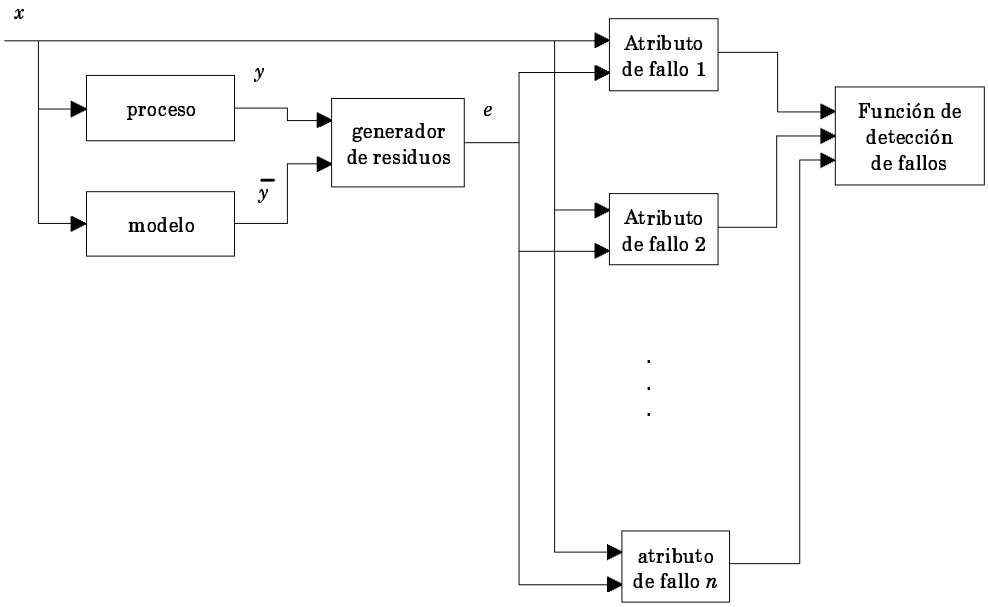


Figura 1.1: Esquema general del sistema de detección de fallos

El cuarto capítulo (página 51) define la estructura necesaria para realizar el ajuste del sistema y es el tema central de la tesis. En este capítulo se hace una comparación de técnicas de ajuste de parámetros sobre un caso ejemplo. Esta comparación pone en duda el procedimiento de ajuste de parámetros habitual que se basa en la utilización de datos de funcionamiento normal y en el método de los mínimos cuadrados. Por el contrario, el procedimiento de ajuste propuesto en esta tesis se basa principalmente en los siguiente puntos:

- La utilización de historias de fallo (conjuntos de datos adquiridos durante procesos reales de degradación).
- El ajuste global de todos los parámetros que intervienen en el sistema de detección incipiente de fallos; es decir, los parámetros del modelo, de los atributos de fallo y de la función de detección de fallos.

- La utilización de técnicas de optimización multi–objetivo para tener en cuenta varios atributos de detección (criterios de valoración de la eficacia del sistema de detección de fallos).

El esquema general del sistema de detección, junto con los módulos utilizados para el ajuste es el siguiente:

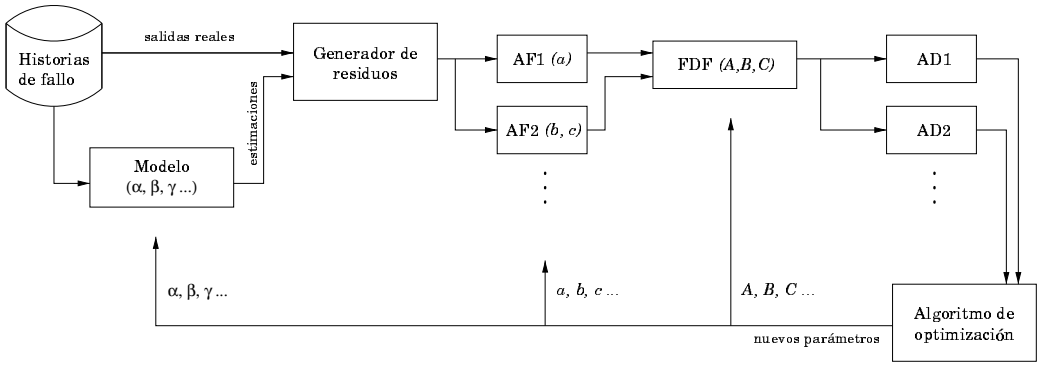


Figura 1.2: Esquema del procedimiento de ajuste

El quinto capítulo (página 81) hace una revisión de los métodos de optimización no lineal, estudiando especialmente los problemas propios del ajuste del sistema de detección propuesto en la tesis. Concretamente se estudian los métodos de optimización por búsqueda directa y se realiza una comparación de la robustez de estos métodos. También se expone la problemática de los mínimos locales en problemas de optimización no lineal y en concreto la problemática intrínseca al sistema de detección de fallos. Para resolver este problema se proponen soluciones generales para el caso unidimensional y multidimensional. Por último se proponen soluciones concretas para el ajuste del sistema de detección incipiente de fallos, como son la mejora de velocidad por optimizaciones parciales y la reducción del efecto de los mínimos locales mediante un filtrado que suaviza la función objetivo.

El capítulo sexto (página 125) analiza el caso particular de la detección de fallos mediante modelos adaptativos, que se basa en el análisis de la evolución de los parámetros. En este capítulo se describen los algoritmos de estimación recursiva de parámetros y se explica cómo se aplican estas técnicas a la detección incipiente de fallos. Sin embargo, la aportación principal del capítulo es el estudio de las limitaciones que presenta esta

técnica y la comparación con la detección incipiente basada en el análisis de los residuos.

El capítulo siete (página 155) muestra un ejemplo práctico de aplicación de toda la metodología. Se trata de un sistema de detección del envejecimiento del amortiguador vertical de un tren. A lo largo del capítulo se describe paso por paso cómo se construye un sistema de detección incipiente de fallos y cómo se realiza el ajuste del mismo. Por último se muestran los resultados del método de detección y se comparan con otras técnicas.

Por último, el capítulo ocho resume las aportaciones fundamentales de la tesis y los resultados obtenidos.

Capítulo 2

Técnicas de detección incipiente de fallos

2.1 Introducción

Este capítulo se centra en la descripción de las principales técnicas de detección incipiente de fallos. Sobre estas técnicas se basa la estructura del sistema de detección de fallos propuesta en esta tesis. La detección de fallos de manera incipiente es un tema de interés para la industria porque produce ahorro en gastos de mantenimiento y en costes de reposición [Piety88].

Ya se ha visto que algunos métodos de detección, como aquéllos basados en control de calidad, no son válidos para realizar detección incipiente. Los únicos métodos válidos se basan en la utilización de datos que se obtienen de forma periódica y en el conocimiento del funcionamiento del proceso. La adquisición de datos es una cuestión que cada vez presenta menos

problemas en la industria, sin embargo el tratamiento que actualmente se realiza con los datos no obtiene el máximo provecho de los mismos.

La mayor parte de los sistemas industriales de detección de fallos son sistemas de alarma basados en el análisis de señales aisladas. Como se analiza a continuación estos sistemas tienen una capacidad de detección incipiente muy limitada. Por otro lado, los sistemas de detección basados en modelos de comportamiento del proceso son sistemas más eficaces pero su utilización en la industria no es generalizada.

2.2 Detección basada en el análisis de señales aisladas

La mayor parte de los sistemas de monitorización incorporan un pequeño módulo que permite definir los límites de funcionamiento normal para cada variable. Estos límites actúan de umbral para disparar sistemas de alarma y constituyen el sistema más sencillo de detección de fallos en tiempo real. Es habitual tener al menos dos niveles de alarma: uno de aviso, que permite realizar modificaciones de las condiciones de trabajo de los equipos y programar una inspección, y otro de parada, que indica la necesidad de detener el proceso inmediatamente o incluso envía automáticamente una señal de parada al sistema de control.

Calificar estos sistemas como "detectores incipientes de fallos" o incluso "sistemas de mantenimiento predictivo", como los llaman frecuentemente sus distribuidores, es algo discutible. No cabe duda que en procesos de funcionamiento muy estable, las variaciones en determinadas variables son un indicativo claro de fallo (por ejemplo la frecuencia de la tensión de una red eléctrica). En los casos en los que esta tesis tiene interés (sistemas que no deben fallar inesperadamente), es importante que la detección de estas variaciones sea anterior al fallo y no una consecuencia del mismo, lo que en cierto modo distingue la detección incipiente de una información útil para el diagnóstico. Si un proceso de degradación es identificable por la variación monótona de una variable, entonces es mejor realizar una estimación de la tendencia de la variable que comprobar si supera un umbral fijo. La información que aporta la tendencia es de mejor calidad por varias razones: al estar calculada a partir de varias muestras supone un filtrado de ruidos

y de cortas situaciones transitorias en la señal; además permite estimar el tiempo que tardará en alcanzarse el umbral de parada.

Como norma general hay que destacar que la detección basada en el análisis de señales aisladas tiene una utilidad muy limitada, ya que ninguna señal interna de un proceso suele ser suficientemente estable como para permitir la definición de unos umbrales estrechos. Lo habitual, sin embargo, es que los cambios en las condiciones de trabajo o simplemente cambios externos al proceso (como cambios atmosféricos) provoquen alteraciones en las variables monitorizadas. Estas alteraciones pueden estar justificadas por las razones mencionadas sin que lleguen a suponer ningún peligro, mientras que en otras ocasiones pueden ser realmente un indicativo de fallo.

Se estudia a continuación, como ejemplo, el caso de un sencillo circuito eléctrico formado por una batería y una resistencia variable (figura 2.1). Las baterías tienden a mantener un nivel de tensión eléctrica constante; sin embargo, durante el funcionamiento del circuito la batería irá agotando su capacidad hasta que en cierto momento aparece una caída brusca de tensión. Este sencillo sistema se va a estudiar como un proceso industrial que funciona de manera estable hasta que en cierto momento sufre una

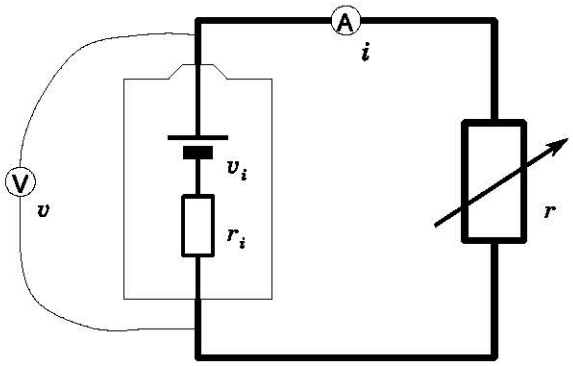


Figura 2.1: Esquema del circuito eléctrico

degradación acelerada que lleva al fallo del sistema. No se trata en realidad de un fallo de la batería, que podría volver a recargarse, pero si se puede considerar un fallo del sistema. En [Smith95] se estudian los modos de fallo de las baterías y de los sistemas de alimentación en corriente continua que se utilizan para el control de los sistemas eléctricos. El sistema de

alimentación en corriente continua de una subestación eléctrica es crítico ya que su fallo inhabilita las protecciones. Este ejemplo también tiene interés hoy en día ya que cada vez se utilizan más equipos alimentados con baterías (ordenadores portátiles, teléfonos móviles, etc) donde un sistema efectivo de detección de descarga es de gran ayuda.

Las dos variables fácilmente medibles en dicho circuito son v (la tensión eléctrica de la pila), e i (la intensidad) pero ninguna de ellas se mantiene constante durante el funcionamiento normal del circuito. Esto se debe a que los cambios en la resistencia variable provocan alteraciones tanto en v como en i. La intensidad depende directamente del valor de la resistencia (r) ya que $i=\frac{v}{r}$, y por otro lado la tensión v depende de la intensidad del circuito (teniendo en cuenta que la pila no es una fuente de tensión infinita, $v=v_i-r_i \cdot i$).

Durante el funcionamiento normal del circuito se pueden producir variaciones de tensión e intensidad, como consecuencia de cambios en la carga. La figura 2.2 muestra como ejemplo una historia de funcionamiento del circuito durante 70 minutos. Tras una hora de funcionamiento se observa que tanto la tensión como la intensidad bajan a niveles mínimos, indicando que se ha agotado la capacidad de la pila. Durante la primera

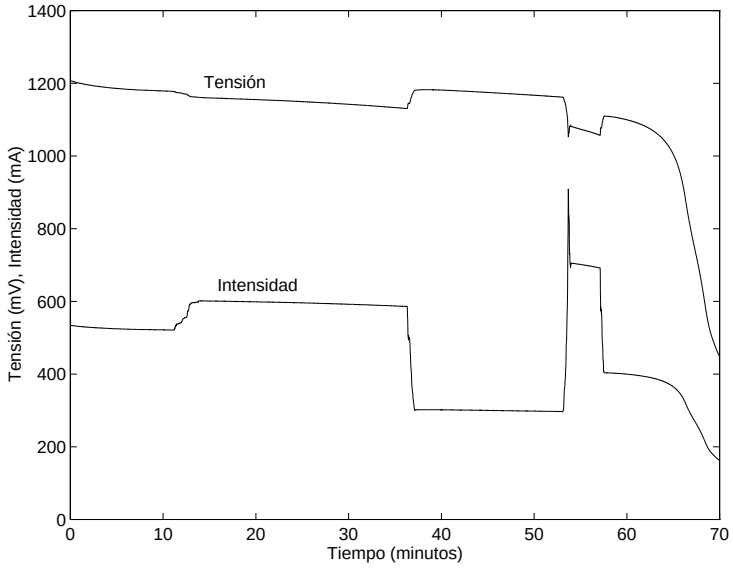


Figura 2.2: Curva de tensión e intensidad del circuito

hora no hay anomalías pero se observan cambios en la tensión y en la

intensidad que, sin embargo, no son indicativos de fallo ni de descarga prematura de la batería.

Para aplicar técnicas de detección basadas en el análisis de señales aisladas habría que definir unos umbrales para cada medida v e i. Dichos umbrales deben ser suficientemente amplios como para no provocar alarmas durante el periodo de funcionamiento normal (v_{min}=1000 mV, v_{max}=1200 mV). En este ejemplo concreto la detección basada en umbrales se produciría muy tarde como para considerarla incipiente. Como puede observarse en la gráfica la detección se produciría a tan solo 3 minutos del colapso total, por lo tanto no parece un método de detección apropiado. Un estudio más profundo de este ejemplo, basado en datos reales de varias curvas de funcionamiento está recogido en [Palacios97]. Este artículo encuentra unos parámetros apropiados para detección incipiente de la descarga de la batería de Ni-Cd utilizada. Estudios comparativos entre el comportamiento y las características de distintos tipos de baterías, incluidas las de Ni-Cd, pueden encontrarse en [Powers95] y [Riezenman95].

Las limitaciones que presenta el análisis de señales aisladas pueden superarse realizando un tratamiento más sofisticado de las señales aportadas por los sensores. Las técnicas avanzadas para el tratamiento de conjuntos de variables se conocen como técnicas de modelado y se describen en el siguiente apartado. Más adelante, en el apartado 2.4.1, se describen las técnicas de detección basadas en el análisis de los residuos y se retoma este ejemplo para mostrar con mayor claridad las limitaciones de la detección basada en análisis de señales aisladas.

2.3 Técnicas de modelado

Las técnicas de modelado surgen con objeto de caracterizar el comportamiento de un sistema. Como se ha visto en el apartado anterior, resulta difícil distinguir si los cambios producidos en una variable son consecuencia de una anomalía o están justificados por la propia dinámica del sistema.

Un método para poder aclarar esta distinción sería contar con un segundo sistema, de iguales características al sistema monitorizado y

sometido a las mismas condiciones; entonces la evolución de las variables de los dos sistemas debería ser idéntica salvo que exista alguna anomalía. Este concepto se aplica por ejemplo en algunos sistemas anti-bloqueo de frenos [Srinivasa80] [Wellstead97], y en sistemas anti-patinamiento de los vehículos actuales[1]. Gracias a la simetría que existe entre las ruedas (igual diámetro, inercia, mecanismo de freno, etc). se puede detectar fácilmente si una rueda está patinando sobre el suelo; basta para ello con disponer de un sensor de velocidad angular en cada rueda. Durante la frenada se conoce la velocidad del vehículo (dada por la mayor velocidad angular) y las ruedas que se bloquean en cada momento, ya que su velocidad angular disminuye drásticamente en comparación con la velocidad del resto de las ruedas. Ocurre lo contrario durante la aceleración, donde un exceso de velocidad en una rueda motriz frente al resto de las ruedas indica que la primera está patinando sobre el suelo. Es decir la detección de bloqueos de frenos y patinamientos de ruedas se realiza únicamente mediante comparaciones de equipos de similares características y que deberían tener el mismo comportamiento en situación normal.

En los procesos industriales no es tan fácil encontrar esta redundancia de equipos, trabajando en las mismas condiciones, que pueda aprovecharse para verificar el comportamiento. Por ello se recurre a crear modelos matemáticos del proceso que estimen el valor de un conjunto de variables (llamadas **variables de salida**) a partir de otro conjunto de variables (llamadas **variables de entrada**) que definen las condiciones de funcionamiento. La utilización de modelos matemáticos del proceso orientada al diagnóstico se llama frecuentemente **redundancia analítica**, ([Gertler91], [Isermann84], [Isermann93], [Frank90], [Rizzoni91], [Chow84]) para diferenciar de la redundancia física.

Los modelos matemáticos no tienen estructura fija, son en general algoritmos que operan matemáticamente con las variables de entrada para obtener una estimación de las variables de salida. Sin embargo pueden clasificarse en dos grandes grupos: modelos de "caja negra" y modelos

[1] Breves descripciones, sin entrar en detalles técnicos, de estos sistemas (*anti-lock braking system* (ABS), *electronic traction systems* (ETS) y otros) pueden encontrarse en `http://www.mercedes.de`

"basados en fundamentos físicos". Ambas técnicas de modelado se explican a continuación.

2.3.1 Modelos matemáticos de caja negra

Son aquéllos que calculan las variables de salida a partir de una serie de operaciones matemáticas con las variables de entrada, sin que dichas operaciones estén basadas en la estructura física o tipo de proceso que modelan. Mediante técnicas de identificación de sistemas es posible encontrar una relación de las variables de entrada que permita calcular las variables de salida, aunque no siempre con la precisión deseada. Cuanto mayor sea el orden o la complejidad del modelo propuesto, mejor será la capacidad de reproducir la señal de salida con fiabilidad.

Dentro de este tipo de modelado se encuentran los modelos estadísticos (AR, ARMA ...) [Basseville93], las redes neuronales [Hammerstrom93] [Sjöberg95], etc. Para crear este tipo de modelos se propone una estructura de modelo y se realiza el mejor ajuste posible. Si el resultado no se considera satisfactorio, lo que suele comprobarse comparando la salida estimada por el modelo y la salida real del proceso, será necesario probar con otra estructura de modelo diferente.

Un caso particular de los modelos de caja negra, que se conocen con el nombre de modelos de "caja gris", son aquéllos que utilizan cierta información cualitativa del proceso que describen. Este tipo de información puede ser el conjunto de variables que afectan al mismo subsistema, o relaciones causa–efecto que permiten definir correctamente qué variables son de entrada y cuáles son de salida, lo cual simplifica el planteamiento del modelo al restringir el número de variables a considerar. Los modelos de caja gris evitan que los métodos de ajuste encuentren relaciones entre variables independientes de entrada y salida debido a artificios matemáticos o a particularidades en los datos utilizados durante la identificación.

En definitiva, toda la información cualitativa que se pueda introducir durante la definición de la estructura del modelo permite obtener modelos más robustos a partir del mismo volumen de datos de entrenamiento.

2.3.2 Modelos matemáticos basados en fundamentos físicos

Cuando el proceso que se quiere modelar tiene un comportamiento dinámico conocido y bien descrito en base a fundamentos físicos, puede plantearse el modelo matemático directamente. Las ecuaciones matemáticas utilizadas en ingeniería para describir el comportamiento de un sistema son modelos matemáticos de dicho sistema.

La mayor diferencia con los modelos de caja negra consiste en que los parámetros de las ecuaciones suelen representar características físicas de los componentes. Algunas de estas características pueden ser conocidas (por ejemplo las dimensiones físicas del sistema, masa, rigidez, etc.) y se convierten en constantes, lo que simplifica el modelo ya que reduce el número de parámetros a estimar durante el proceso de ajuste. La ventaja fundamental es la definición precisa de la estructura del modelo en cuanto a orden, ligaduras entre variables, restricciones en los valores de los parámetros, etc.

Los modelos de caja negra suelen establecer relaciones lineales entre las variables de entrada y de salida, ya que no es posible definir una estructura matemática suficientemente general como para poder modelar adecuadamente cualquier proceso no lineal. Sin embargo existen procesos cuyo comportamiento es claramente no lineal, en estos casos es importante poder disponer de las ecuaciones físicas que describen el comportamiento del sistema para poder definir un modelo representativo. En caso contrario los modelos basados en redes neuronales tienen la capacidad de adaptarse a procesos no lineales y son probablemente la mejor opción cuando no se tiene ninguna información de tipo físico [Muñoz96].

Por último, los parámetros obtenidos en el proceso de ajuste pueden aportar una información muy útil para los técnicos ya que sus valores corresponden a características físicas de los componentes. Por lo tanto el valor de los parámetros de modelos basados en fundamentos físicos es un indicador del estado de salud del componente. Este concepto es importante para el proceso de diagnóstico, ya que la estimación de parámetros en modelos basados en fundamentos físicos equivale, en cierto modo, a medir directamente características de los componentes [Filbert92].

2.4 Detección incipiente de fallos

Entendemos por incipiente que la detección se produce antes de que el fallo se manifieste por completo, es decir cuando aparece un proceso de degradación. Para ello es necesario obtener y analizar datos del sistema con suficiente velocidad. Para cada muestra que se obtiene del sistema debe emitirse un diagnóstico de salud o fallo, si bien los cálculos que se realizan para el análisis pueden utilizar datos del pasado además del dato actual. La utilización de datos del pasado es imprescindible para el cálculo de tendencias y para modelos donde explícitamente se hace referencia a estos datos (generalmente modelos matemáticos que incluyen derivadas). Los programas de diagnóstico deben almacenar en la memoria del ordenador cierta cantidad de información relativa a datos pasados.

Como se ha explicado anteriormente, el análisis basado en señales aisladas es muy limitado por lo que se recurre a la utilización de modelos. Utilizando modelos del proceso que se quiere monitorizar se obtiene la estimación de una o varias señales del proceso. Las señales estimadas son iguales a las señales de salidas del proceso —salvo un cierto nivel de ruido— siempre que el modelo sea representativo del comportamiento del proceso y que la situación general no haya cambiado con respecto a las condiciones en el momento de realizar el ajuste. La situación debe ser la misma tanto en condiciones de trabajo (punto de funcionamiento) como en el estado de salud de los componentes del proceso.

Si cambian las condiciones de trabajo de manera que el punto de funcionamiento sea muy distinto a los utilizados para ajustar los parámetros, es probable que las estimaciones producidas por los modelos no coincidan con las señales reales de salida con la precisión necesaria. Fundamentalmente esto es debido a efectos no lineales que no hayan sido considerados en el modelo y que no se manifiestan mientras no cambie el punto de funcionamiento de manera significativa.

Si cambia el estado de salud de los componentes de manera que el funcionamiento del proceso se vea afectado, las señales de salida no serán las mismas que las estimadas por los modelos, ya que estas últimas reproducen la condición normal de funcionamiento.

Se presentan a continuación dos enfoques diferentes utilizados para detección de anomalías que se fundamentan en la utilización de modelos del

proceso monitorizado. Un estudio más profundo de las diferencias entre los dos métodos se expone más adelante en el capítulo 6.

2.4.1 Detección basada en el análisis de residuos

Los **residuos** miden la discrepancia existente entre las señales reales de salida del proceso y los valores estimados por los modelos. El residuo asociado a cada señal se calcula para cada instante de muestreo utilizando el valor actual de la señal —proporcionado por un sensor— y el valor actual calculado por el modelo. Existen varias maneras de obtener esta discrepancia instantánea, generalmente se utiliza el cuadrado de la diferencia (ecuación 2.1, a) o el valor absoluto de la diferencia (ecuación 2.1, b).

$$\begin{aligned} e &= (y - \bar{y})^2 \quad \text{(a)} \\ e &= |y - \bar{y}| \quad \text{(b)} \end{aligned} \tag{2.1}$$

donde e es el valor del residuo, y es el valor de la señal de salida real, $\bar{y}$ es el valor de la predicción del modelo.

Ambos métodos de cálculo evitan trabajar con residuos negativos y no distinguen si la señal estimada se ha hecho mayor o menor que la real ni tienen en cuenta si la discrepancia es grande o pequeña con relación al valor de la señal. El error cuadrático penaliza más los errores grandes (mayores que la unidad) y subestima los errores pequeños; a la hora de seleccionar el método de cálculo de los residuos es importante tener en cuenta este comportamiento en función de los niveles de discrepancia esperados.

En determinadas situaciones puede ser interesante trabajar con errores porcentuales en lugar de absolutos. Por ejemplo una discrepancia de 2 V en una señal de tensión de 220 V puede ser despreciable frente una diferencia de 2 V en una señal de 10 V. En este caso los residuos pueden calcularse proporcionalmente al valor de la señal según la ecuación:

$$e = \left(\frac{y - \bar{y}}{y} \right)^2 \tag{2.2}$$

La manera de calcular los residuos afecta directamente al sistema de detección de fallos. Normalmente se adopta el mismo criterio que se utilice para realizar el ajuste de los parámetros. En el capítulo 4 se estudia cómo afecta a la detección de fallo el criterio adoptado para el ajuste.

El cálculo de residuos suele representarse gráficamente según el siguiente diagrama (figura 2.3)

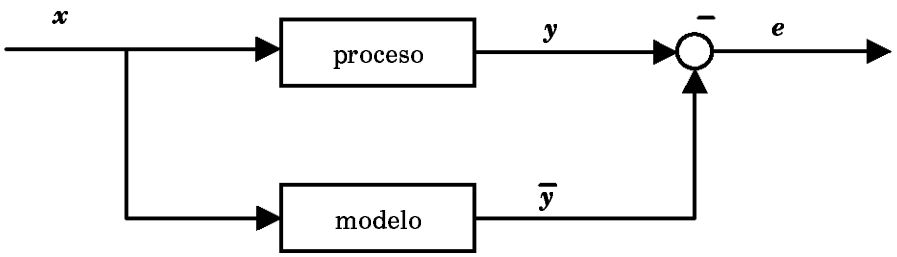


Figura 2.3: Diagrama de cálculo de residuos

donde $\boldsymbol{x}$ son las variables de entrada, $\boldsymbol{y}$ las variables de salida, $\bar{\boldsymbol{y}}$ las estimaciones del modelo y $\boldsymbol{e}$ los residuos correspondientes a cada señal de salida.

Puesto que la manera de calcular los residuos no es necesariamente una diferencia entre las señales de salida real y estimada, resulta más preciso representar el procedimiento según la figura 2.4.

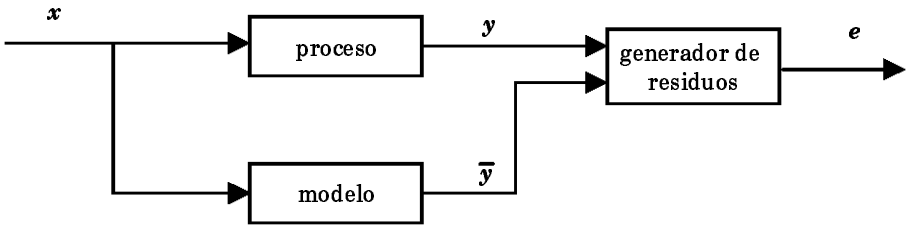


Figura 2.4: Diagrama general de cálculo de residuos

Una manera especial de calcular los residuos es utilizar varias muestras consecutivas y obtener una medida global de la discrepancia de dos señales en lugar de valores que representan discrepancias puntuales. Este método de cálculo es apropiado cuando no se emite un diagnóstico por cada muestra leída. Por ejemplo un sistema de diagnóstico puede tener caracterizado el proceso transitorio de arranque de una máquina herramienta y deducir su

estado de salud comparando la señal transitoria real con la señal generada por un modelo (que depende fundamentalmente de la inercia de la pieza y de la temperatura de los cojinetes). La comparación de las dos señales transitorias se realiza generalmente calculando la suma de los cuadrados de las diferencias, lo que da lugar a un único valor para todos los datos del transitorio de arranque. En este caso el valor del residuo vendría dado por la ecuación 2.3

$$e = \sum_{i=1}^{n} \left(y_i - \bar{y}_i \right)^2 \qquad \textbf{(2.3)}$$

Otra posibilidad es realizar la comparación mediante el cálculo de determinados parámetros característicos de la señal, como por ejemplo las componentes espectrales a determinadas frecuencias. En este tipo de problemas, aunque la frecuencia de muestreo sea rápida durante un corto período de tiempo, todos esos datos se reducen a un pequeño conjunto de valores y no se obtienen nuevos valores hasta que vuelva a lanzarse la adquisición de datos. Por lo tanto la frecuencia de muestreo, desde el punto de vista del diagnóstico, viene marcada por la separación de los tiempos de muestreo y no por la frecuencia del sistema de adquisición de datos.

Análisis de los residuos

El método de detección más sencillo basado en el análisis de los residuos es fijar un umbral para los mismos. Si las discrepancias entre la señal real y la predicción del modelo son suficientemente grandes, los residuos superan dicho umbral y se detecta la anomalía. Otros procedimientos pueden apoyarse en el cálculo de la tendencia de los residuos, en el filtrado de los mismos, etc.

Es importante remarcar la gran diferencia que existe entre esta técnica de detección y la detección basada en el análisis de señales aisladas, que es el método que más se aplica en la actualidad en los procesos industriales. El análisis basado en residuos es mucho más robusto (emite menos falsas alarmas) porque los residuos son poco dependientes de las condiciones de trabajo, a diferencia de las señales producidas por los sensores que generalmente vienen marcadas por el punto de funcionamiento. Retomando el ejemplo de la batería (figura 2.2, página 18) está claro que las variaciones

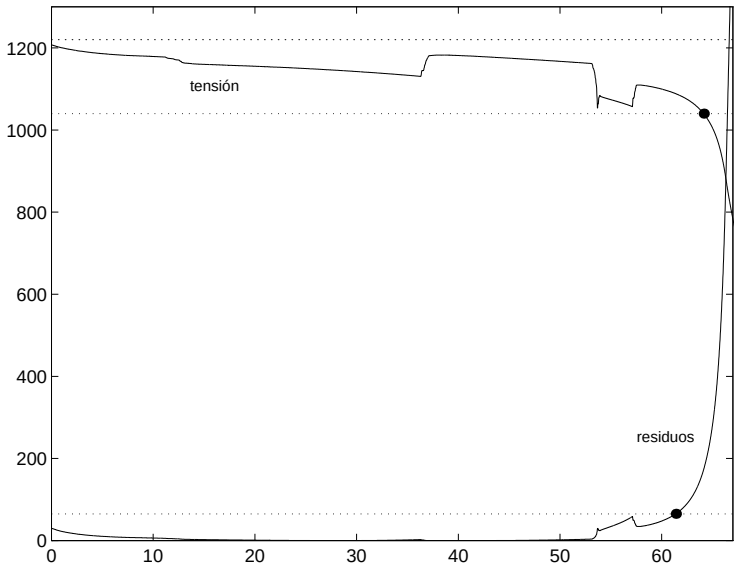


Figura 2.5: Evolución de la tensión y de los residuos

de las condiciones de trabajo provocan grandes variaciones en la tensión de la pila, lo que dificulta la detección de la descarga de la misma basándose en el análisis de esta señal. Sin embargo es sencillo crear un modelo de estimación del valor de la tensión a partir del valor instantáneo de la intensidad, que también refleja los cambios en las condiciones de funcionamiento:

$$\bar{v} = v_i - r_i \cdot i \qquad \textbf{(2.4)}$$

Las variaciones del valor estimado de la tensión son muy parecidas a las variaciones del valor medido, lo que provoca que los residuos sean aproximadamente constantes durante todo el período de funcionamiento normal. La figura 2.5 muestra las variaciones de tensión como consecuencia de los cambios de carga y la curva de los residuos (calculados según la ecuación 2.1, a) que se obtiene utilizando un modelo lineal ajustado por el método de los mínimos cuadrados. Puede observarse que la curva de residuos es más estable (menos sensible a los cambios en las condiciones de trabajo) durante el período de funcionamiento normal, pero experimenta mayores variaciones cuando se degrada la batería.

Siguiendo el criterio más conservador, que intenta envitar la aparición de falsas alarmas durante el período de funcionamiento normal, se pueden definir fácilmente los umbrales de detección. Un sistema de detección

basado en el análisis aislado de la tensión utilizaría como umbrales las dos líneas de puntos que rodean la curva de tensión. Estas líneas se han definido a partir de los valores extremos de la tensión, durante el período de funcionamiento normal, con un pequeño margen de seguridad. De manera análoga, un sistema de detección basado en el análisis de los residuos utilizaría como umbral la línea de puntos de la parte inferior. En este caso el umbral se define a partir del valor máximo que adquieren los residuos durante la condición normal más el pequeño margen de seguridad.

Utilizando estos umbrales se ha calculado, en la zona en la que comienza la degradación de la batería, el instante en que cada método de detección localizaría el fallo. Los puntos (•) marcan en la gráfica el instante de detección para el método basado en el análisis de señales aisladas (tensión) y para el método basado en el análisis de los residuos. Puede observarse que el análisis de residuos permite una detección más incipiente debido fundamentalmente a la conjunción de las dos razones comentadas anteriormente: menor sensibilidad a las variaciones de las condiciones de trabajo durante el funcionamiento normal (lo que se traduce en umbrales más estrechos) y mayor crecimiento cuando empieza la degradación.

Sensibilidad del modelo

Hay que tener en cuenta que en todos los procesos pueden aparecer distintos modos de fallo pero las variaciones que experimentan los residuos pueden ser de diferente tamaño en cada caso. A estas variaciones de los residuos con respecto a la magnitud de un determinado modo de fallo se les llama "sensibilidad del modelo frente al fallo". La sensibilidad depende fundamentalmente de las variables que intervienen en el modelo, de la estructura del modelo y de los parámetros utilizados. Por lo tanto en función del tipo de modelo y de las técnicas empleadas para su ajuste se pueden obtener modelos especialmente sensibles o insensibles a un tipo de fallo. En [Palacios97] se muestran, para un caso concreto, las variaciones de sensibilidad frente al fallo en función de los parámetros utilizados en el modelo.

Un aspecto interesante de la sensibilidad de los residuos es la posibilidad de realizar el diagnóstico directamente [Gertler91] [Gertler93]. Suponiendo que se dispone de varios modelos de un sistema, pero cada modelo es sensible a unos modos de fallo determinados y no a otros, es posible localizar

el tipo de fallo analizando qué modelos han detectado anomalía y cuales no. Para analizar si es posible identificar los diferentes modos de fallo se estudia el subconjunto de residuos que se ve alterado para cada modo de fallo (considerando un único fallo cada vez). Con esta información se construye la "matriz de incidencia" donde cada columna representa un modo de fallo y cada fila el residuo de un modelo (ecuación 2.5). Las columnas correspondientes a cada modo de fallo tendrán un cero cuando el residuo no es sensible al fallo y un uno cuando es detectable. Es evidente que sólo son detectables los modos de fallos cuya columna tenga algún elemento distinto de cero, y que sólo es posible emitir un diagnóstico correcto si todas las columnas son diferentes. La secuencia de ceros y unos que forma cada columna de la matriz de incidencia constituye el código de fallo del modo correspondiente.

Modos de fallo:	1	2	3		1	2	3	
residuo 1	1	1	0		1	0	0	
residuo 2	0	0	1		0	0	1	**(2.5)**
residuo 3	0	1	0		0	1	0	
	caso a				*caso b*			

Es deseable que los códigos correspondientes a cada modo de fallo se diferencien es más de un elemento ya que esto facilita su aislamiento. Esta propiedad se cumple si el número de ceros y de unos es igual para cada modo de fallo y se dice que la matriz de incidencia es fuertemente aislada (ecuación 2.5, caso b). Cuando la matriz no es fuertemente aislada pueden producirse diagnósticos equivocados, sobre todo cuando el fallo es incipiente y no activa todos los residuos instantáneamente. Por ejemplo en el *caso a* (ecuación 2.5), si el segundo modo de fallo no activa el tercer residuo se obtiene el código de fallo [1 0 0], lo que llevaría a establecer un diagnóstico incorrecto, ya que dicho código de fallo corresponde al primer modo.

2.4.2 Detección basada en el análisis de parámetros

Una alternativa a la detección de fallos basada en el análisis de residuos, pero que también utiliza modelos matemáticos del sistema, es la detección basada en el análisis de los parámetros. Esta técnica utiliza la información

inherente a los parámetros que se obtienen tras un proceso de ajuste del modelo.

El procedimiento es válido para detectar problemas en continuo, pero también se puede aplicar para detectar y diagnosticar fallos a partir de un conjunto aislado de datos. Esta capacidad marca una diferencia clara frente al análisis basado en la evolución de los residuos del apartado anterior.

En el caso de utilizar modelos basados en fundamentos físicos, los parámetros corresponden a características de los componentes del sistema y por lo tanto el proceso de estimación de parámetros mediante técnicas de ajuste equivale a medir dichas características de forma indirecta. Hay que tener en cuenta que la mayoría de los ensayos experimentales de componentes se basan en este concepto. Por ejemplo para medir la constante de rigidez de un muelle se realizan mediciones de la fuerza aplicada para distintas variaciones de longitud. El resultado es un conjunto de puntos en el plano Δl–f que definen una recta, salvo pequeñas variaciones debidas a errores de medida (suponiendo que la característica del muelle fuese perfectamente lineal).

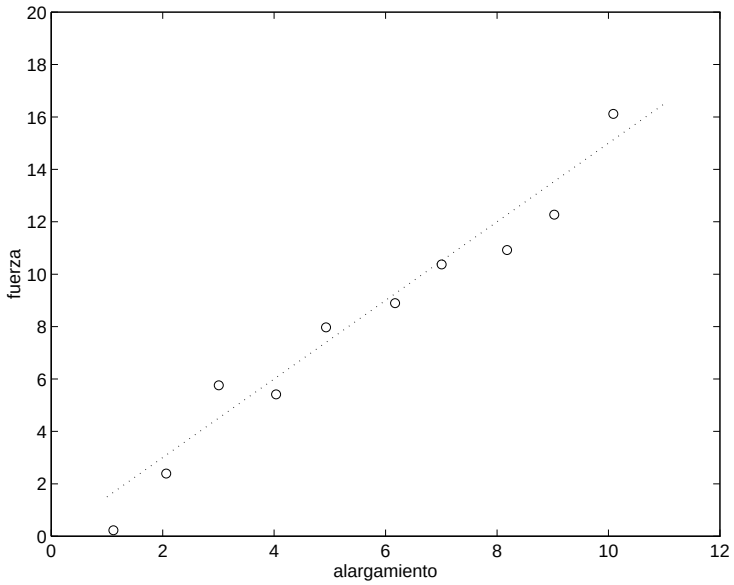


Figura 2.6: Curva simulada de la respuesta de un muelle

Los puntos de medida mantienen aproximadamente la proporción de la constante de rigidez del muelle y por lo tanto cumplen la siguiente relación:

$$f = k \cdot \Delta l + e \qquad \textbf{(2.6)}$$

donde f es la fuerza aplicada en el muelle, Δl es la elongación o variación de longitud, k es la constante de rigidez que se desea obtener y e representa el error de medida de f ó Δl.

Con objeto de minimizar los efectos del ruido, se realiza un ajuste lineal por el método de los mínimos cuadrados. Con ello se obtiene una estimación de la constante del muelle, k. Desde el punto de vista de identificación de sistemas este procedimiento es equivalente al ajuste del modelo lineal:

$$\bar{f} = k \cdot x \tag{2.7}$$

El ajuste se realiza a partir de un conjunto de datos de funcionamiento $(\Delta l, f)$ estimando el valor de la fuerza, $\bar{f}$ y comparando el resultado con el valor medido, f, el parámetro k que se obtiene es el mismo que se obtiene por medio del ajuste lineal.

En el diagnóstico de fallos, si se utiliza un modelo matemático basado en fundamentos físicos, donde los parámetros representan características de los componentes, el proceso de ajuste del modelo para obtener los valores de los parámetros equivale a *medir* el valor de las características de los componentes. Estos valores pueden obtenerse en cualquier momento, a partir de un conjunto de datos de funcionamiento, para compararlos con los valores de diseño. De esta manera contamos con un procedimiento de detección de fallos que además localiza el componentes dañado, es decir un procedimiento de detección y diagnóstico de fallos.

Para aplicar el método de ajuste de parámetros para detección incipiente de fallos es necesario repetir la estimación de los parámetros de forma periódica. Cada vez que se realiza una estimación pueden detectarse fallos comparando con los valores de diseño de las características, en este caso la velocidad de detección viene marcada fundamentalmente por la frecuencia de las estimaciones.

Otra alternativa es realizar un estudio de la dinámica de la evolución de los parámetros con el tiempo. Este tipo de análisis permite conocer los procesos internos de degradación y envejecimiento de los componentes. Esta información unida al cálculo de parámetros dinámicos tales como la tendencia, permiten realizar un diagnostico incipientes de mayor calidad y una estimación de la vida remanente de los componentes. No hay que

olvidar que el cálculo de tendencias sólo es posible si la estimación de parámetros se realiza con suficiente frecuencia. Las técnicas de estimación recursiva de parámetros permiten realizar, de manera eficiente, un reajuste del modelo por cada muestra que se lee del sistema. Pero estas técnicas, que se estudian en profundidad en el capítulo 6, requieren que las condiciones de trabajo del sistema sean suficientemente variadas.

Mediante el análisis de la evolución de los parámetros también es posible realizar detección indipiente cuando se utilizan modelos de caja negra, aunque con modelos basados en fundamentos físicos la interpretación de los valores y de la evolución de los parámetros es más sencilla para los ingenieros de mantenimiento.

En modelos de caja negra es especialmente importante analizar la estabilidad de los parámetros. En algunos casos puede ocurrir que se utilicen variables que no son totalmente independientes, como por ejemplo x_1=T y x_2=2T; si la ecuación del modelo incluye el término $a\ x_1+b{\cdot}x_2$ entonces los parámetros a y b son inestables, ya que existen infinitas combinaciones de a y b que ajustarían el modelo. El mismo problema ocurre con determinadas estructuras de modelo que pueden dar lugar a conjuntos de parámetros que no sean únicos y sin ambigüedad al ajustar el modelo a los datos observados. Este problema se conoce en muchos casos con el nombre de "sobreentrenamiento"o bien *overspecified* y hace que los modelos se especialicen tanto en la representación de los datos observados que luego no son capaces de generalizar a nuevos datos. Como consecuencia el reajuste del modelo cambia significativamente los parámetros sin que el sistema real haya modificado su comportamiento; es decir, se está modelando el ruido.

La estabilidad de los parámetros es un requisito básico para los sistemas de detección basados en el análisis de los parámetros, pero es más fácil de obtener en modelos lineales y en modelos basados en fundamentos físicos, suponiendo que se cuenta con suficiente información de funcionamiento.

La representación gráfica de un sistema de análisis de la evolución de los parámetros es la siguiente:

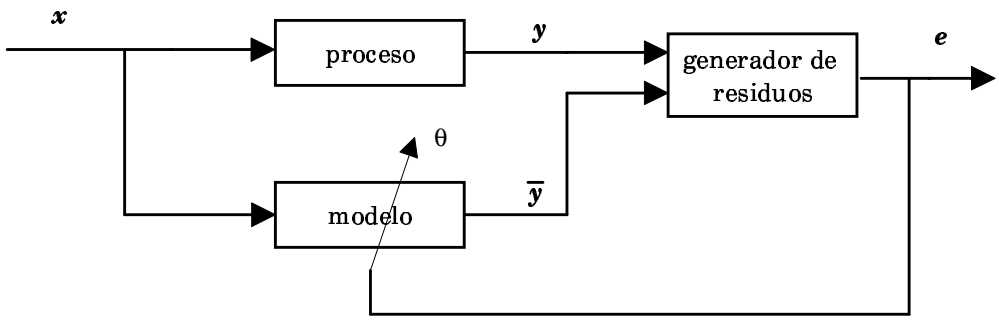


Figura 2.7: Diagrama general del ajuste de parámetros

donde la flecha que atraviesa la caja del modelo indica que los residuos se utilizan para reajustar los parámetros θ. Esta flecha representa el algoritmo de estimación recursiva de los parámetros del modelo, ver capítulo 6.

En el capítulo siguiente se describe una estructura general para definir un sistema de detección incipiente de fallos. Esta estructura permite aplicar detección basada en análisis de residuos y detección basada en análisis de parámetros. Además contempla la posibilidad de utilizar conjuntamente varios algoritmos para analizar los posibles modos de fallo del sistema.

Capítulo 3

Estructura general del sistema de detección incipiente de fallos

3.1 Introducción

Este capítulo está dedicado a la descripción del esquema del sistema de detección de fallos propuesto en la tesis. La estructura del sistema es suficientemente general como para representar distintas técnicas de detección incipiente de fallos. Se trata de una estructura modular donde queda definida la secuencialidad de los distintos tipos de algoritmo y el flujo de datos. La estructura del sistema pretende ser lo más general posible, pero dejando claros los nombres y la función de cada uno de los módulos. Además de la estructura del sistema se ha definido también la manera de ajustar todos sus parámetros de forma que el sistema que finalmente se

obtiene está especializado en detección incipiente de fallos, lo que constituye el objetivo principal de la tesis.

Este capítulo se centra más en la descripción de los módulos que se aplican durante el funcionamiento en continuo del sistema de detección, mientras que el capítulo siguiente trata los aspectos relacionados con el procedimiento de ajuste.

Los módulos que componen el sistema de detección son el modelo matemático, el generador de residuos, los atributos de fallo, y la función de detección de fallos. La figura 3.1 representa el esquema general del sistema de detección incipiente de fallos con todos sus módulos. Estos módulos se describen en los apartados siguientes.

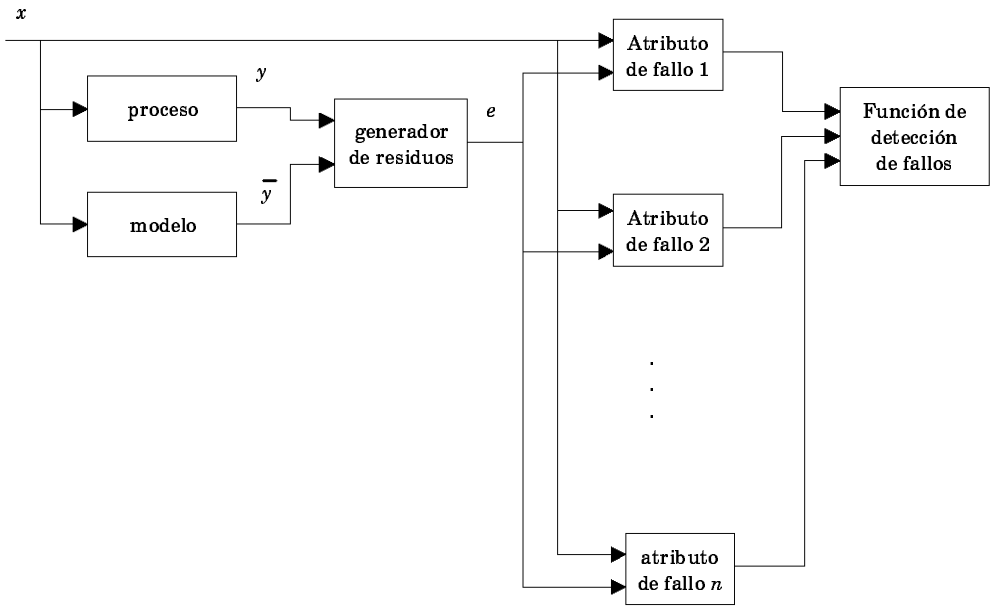


Figura 3.1: Esquema general del sistema de detección de fallos

En el apartado ? se vuelve a mostrar el esquema general y se explica cómo se aplica para el método de detección basado en el análisis de residuos y para el método de detección basado en el análisis de la evolución de los parámetros.

3.2 Modelo matemático del proceso

La primera parte del esquema corresponde a las técnicas de detección incipiente de fallos basados en la utilización de modelos. Se ha explicado en el capítulo anterior la necesidad de utilizar modelos matemáticos del proceso con objeto de obtener un sistema de detección incipiente de fallos de alta calidad, ya que cualquier análisis basado en señales aisladas sufre serias limitaciones.

El modelo matemático es un conjunto de ecuaciones o algoritmos que se resuelven en paralelo con el funcionamiento del proceso que se quiere monitorizar. Este módulo genera la estimación de un conjunto de variables del proceso a partir de otro conjunto de señales que se consideran las variables de entrada. La definición de variables de entrada y salida no es un aspecto crítico para el sistema, ya que el modelo puede plantearse tanto de manera directa como inversa, lo importante es conocer que existe una relación entre variables de forma que si una cambia es origen o consecuencia de la variación de otras. El modelo es un algoritmo que describe numéricamente esta dependencia.

La definición de modelos es por lo tanto bastante flexible en cuanto a las variables que utiliza. En sistemas complejos, donde se utilizan varios modelos, puede ocurrir que una variable que se considera de salida en un modelo se utilice simultáneamente como entrada en otro. Esta situación daría lugar a una representación algo más complicada que no se ha incluido en la figura 3.1 para no crear una confusión innecesaria. La representación completa sería la siguiente:

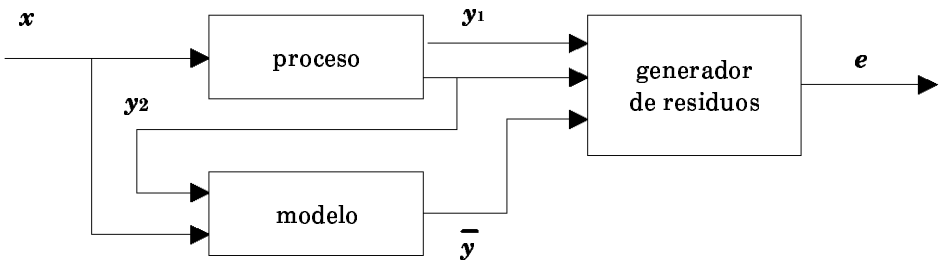


Figura 3.2: Representación completa de la utilización de modelos

donde $\boldsymbol{x}$ representa el conjunto de variables de entrada, $\boldsymbol{y}_1$ son variables de salida que no intervienen en el cálculo de los modelos, $\boldsymbol{y}_2$ son variables de salida pero que también se utilizan como entrada en algún modelo, $\bar{\boldsymbol{y}}$ son las estimaciones de las variables $\boldsymbol{y}_1$ e $\boldsymbol{y}_2$ producidas por los modelos, y finalmente $\boldsymbol{e}$ son los residuos calculados. Hay que tener en cuenta que $\boldsymbol{y}_2$ son el conjunto de variables especiales que además de ser estimadas por algunos modelos, también se utilizan como entrada en otros. Por eso aparecen en el gráfico en la entrada del bloque de modelos y en la entrada del generador de residuos.

Todas estas representaciones del modelo se basan en el concepto de redundancia analítica, descrita anteriormente, que considera el modelo ejecutándose en paralelo con el sistema y utilizando los mismos datos de entrada. Una representación más orientada a una instalación real, pero equivalente a la anterior sería:

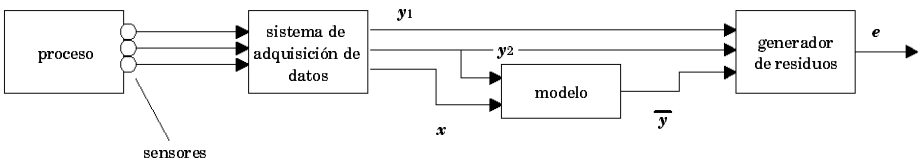


Figura 3.3: Representación orientada a señales

En este gráfico se ha dibujado un conjunto de sensores que recogen los datos del proceso y que están conectados a un sistema de adquisición de datos. El sistema de adquisición de datos transforma las señales eléctricas en valores digitales y pone los datos a disposición de otros módulos, sin hacer ninguna distinción sobre si los datos son de entrada o de salida. El modelo utiliza los subconjuntos de señales $\boldsymbol{y}_2$ y $\boldsymbol{x}$ para estimar las señales $\bar{\boldsymbol{y}}$, mientras el módulo de generación de residuos utiliza las señales $\boldsymbol{y}_1$, $\boldsymbol{y}_2$ e $\bar{\boldsymbol{y}}$ para obtener el vector de residuos $\boldsymbol{e}$. Esta representación está más orientada al diseño informático del sistema de detección de fallos que a los diseños utilizados en control y en identificación de sistemas.

3.3 Módulo generador de residuos

Este módulo mide las discrepancias entre las señales estimadas por el modelo y las señales recogidas por los sensores. De acuerdo con el método de detección de fallos basado en el análisis de residuos (ver apartado 2.4.1, página 24) los residuos aportan información sobre el estado de salud del proceso y serán estudiados en otros componentes del sistema de detección. Sin embargo, al aplicar las técnicas de detección basadas en el análisis dinámico de la evolución de los parámetros del modelo (apartado 2.4.2, página 29) los residuos se utilizarán como información principal para reestimar los parámetros. En este caso el resto de los componentes del sistema de detección se encargarán del análisis de los parámetros mientras que los residuos se mantendrán generalmente próximos a cero.

La manera de obtener el valor de los residuos a partir de dos señales —una real y otra estimada— es muy variada; este módulo está pensado para admitir cualquier algoritmo de cálculo de residuos. En el capítulo anterior se han comentado los dos métodos de cálculo más utilizados, el valor absoluto de la diferencia $|y_i - \bar{y}_i|$ y el cuadrado de la diferencia $(y_i - \bar{y}_i)^2$ pero cualquier otro método también es válido.

Aunque en los diagramas de bloques sólo se dibuje una flecha para representar los residuos, hay que tener en cuenta que se trata de un vector de valores en cada instante de muestreo. Cada variable estimada por los modelos dará lugar a un residuo, por lo tanto a un elemento del vector $\boldsymbol{e}$. (En los gráficos se han representado las variables en negrita indicando que se trata de vectores).

3.4 Atributos de fallo

Conceptualmente los atributos de fallo son indicadores de la existencia de fallos o funcionamientos anómalos del proceso monitorizado. En general se trata de algoritmos que analizan los valores de los residuos o los valores de los parámetros del modelo y cuyo resultado es un indicativo de fallo.

Los atributos de fallo son indicadores más precisos y robustos de la existencia de fallo que los meros valores de los residuos. Los algoritmos

pueden incluir el cálculo de filtrados, tendencias y otras características. También se contempla en la estructura general del sistema de detección la posibilidad de utilizar las variables del proceso en los algoritmos de los atributos de fallo. Con esta posibilidad se permite que los atributos tengan en cuenta las condiciones de trabajo del proceso. Por ejemplo durante periodos transitorios de conexión o desconexión, que son fácilmente detectables observando las variables de entrada, puede anularse temporalmente la capacidad de detección de fallos (tal y como se propone en [Crowley90]).

Se ha pensado que varios atributos de detección puedan evaluarse en paralelo ya que para cada caso concreto puede haber varios criterios de fallo aplicables. La idea es que los atributos de fallo sean algoritmos generales, dentro de lo que sea posible, de manera que se pueda disponer de una pequeña base de datos de atributos de fallo de la cual se seleccionan los que se estiman más convenientes en cada caso[1].

Veamos a continuación algunos ejemplos de atributos de fallo que ayudarán a comprender su funcionamiento. En detección basada en análisis de los residuos el indicador de fallo más sencillo e intuitivo es el sobrepaso de un cierto umbral. Este concepto es equivalente, en el caso de detección basada en el análisis de parámetros, a estudiar si el parámetro se sale de unas bandas de comportamiento normal definidas alrededor del valor teórico del parámetro. El atributo de fallo general para detectar el sobrepaso de un umbral superior es:

$$AF_1 = u_1 - e \tag{3.1}$$

Según esta definición, AF_1 es una función que tiene un sólo parámetro (el umbral u_1) y mide el nivel de sobrepaso de dicho umbral. Si el atributo es menor o igual a cero no hay sobrepaso, pero si es positivo entonces los residuos son mayores que el umbral. Cuanto mayor sea AF_1, mayor será el nivel de sobrepaso.

[1]Siempre es posible incluir todos los atributos de fallo y dejar que el algoritmo de ajuste se encargue de descartar los que resultan poco útiles, pero dificultaría el proceso de optimización.

Otro atributo que ayuda a confirmar la existencia de un fallo es el número de muestras consecutivas que han superado el umbral. Este atributo es una variable entera que normalmente vale cero, pues los residuos suelen mantenerse por debajo del umbral. Cuando los residuos superan el umbral empieza funcionar como un contador y crece linealmente con el número de muestras. Cuanto mayor sea este atributo, mayor seguridad habrá de que se trata de un fallo y no de un ruido o de una perturbación transitoria, pues la condición de fallo es confirmada en cada muestra. Si en cualquier momento los residuos dejan de superar el umbral, se restablece el valor cero en el atributo. La mayor diferencia entre este atributo (AF_2) y el anterior (AF_1) se manifiesta cuando los residuos superan el umbral pero no mantienen un crecimiento apreciable; en esta situación AF_1 se mantendrá constante independientemente de tiempo que lleve activa la indicación de fallo mientras que AF_2 irá creciendo con el tiempo. El interés que puede tener cada uno de estos atributos de fallo es algo que depende claramente del tipo de aplicación. Lo único importante en este punto es hacer notar que puede haber varios criterios que pueden utilizarse de forma conjunta o excluyente y que la estructura del sistema propuesto admite todas las posibilidades y no establece ninguna restricción en este sentido.

Los valores de tendencias de variables son indicadores muy valiosos ya que ofrecen una información fácil de interpretar, filtrada por el propio algoritmo de cálculo y es sencillo obtenerlos. La fórmula que calcula la pendiente del ajuste lineal de un conjunto de n puntos (x_i, y_i) a una recta $(\tilde{y}=Ax+B)$ es la siguiente:

$$A=\frac{n\cdot\sum_{i=1}^{n}x_i y_i-\sum_{i=1}^{n}x_i\cdot\sum_{i=1}^{n}y_i}{n\cdot\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2} \qquad \textbf{(3.2)}$$

En el caso de detección incipiente de fallos no es interesante considerar la tendencia de todos los datos desde el arranque del proceso, sino que interesa analizar la pendiente de los últimos puntos. La pendiente de los datos de

una variable que en funcionamiento normal se encuentra siempre dentro de unos límites fijos, es un atributo que tiende a cero para historias largas de datos. Si se consideran conjuntos de datos demasiado grandes, la pendiente no refleja el comportamiento reciente aunque se produzcan variaciones bruscas; es decir, no vale para realizar detección incipiente de fallos. En definitiva, el único interés estriba en analizar un conjunto pequeño formado por los datos más recientes. Se puede rescribir la ecuación 3.2 teniendo en cuenta que y serán los residuos y x será el tiempo.

$$AF_3 = \frac{n \cdot \sum_{i=T-n+1}^{T} t_i e_i - \sum_{i=T-n+1}^{T} t_i \cdot \sum_{i=T-n+1}^{T} e_i}{n \cdot \sum_{i=T-n+1}^{T} t_i^2 - \left(\sum_{i=T-n+1}^{T} t_i \right)^2} \qquad \textbf{(3.3)}$$

Esta ecuación obtiene el valor de la tendencia de los residuos para el instante de muestreo T considerando un conjunto de n muestras. Los datos considerados son la muestra en el instante T y las $n-1$ muestras de los instantes anteriores (esto hace n muestras en total). Como el cálculo según la ecuación 3.2 es invariante con una translación en x, en la ecuación 3.3 puede trasladarse el origen de tiempos al instante T o al instante $T-n+1$ sin alterar el resultado. Esto permite tener calculados los términos que sólo dependen de t_i con antelación $\left(\sum_{i=T-n+1}^{T} t_i \text{ y } \sum_{i=T-n+1}^{T} t_i^2\right)$ lo que simplifica la ecuación y evita problemas de redondeo —especialmente al resolver el denominador— cuando los tiempos son grandes.

El cálculo de la pendiente es aplicable como indicador de la tendencia de los residuos o para analizar la evolución de los parámetros del modelo por lo que resulta apropiado en los dos métodos de detección. En el esquema general del sistema de detección incipiente de fallos propuesto en esta tesis se considera como un atributo de fallo más. Este atributo sólo depende de un parámetro, el número de muestras n a considerar.

3.5 Función de detección de fallos

La idea de utilizar varios atributos de fallo resulta muy atractiva porque permite definir varios criterios de fallo simultáneamente. Además, la facilidad de incluir o eliminar atributos de fallo (gracias al diseño modular) y la posibilidad de adaptarlos específicamente al problema (ajustando sus parámetros de cálculo) ofrecen toda la flexibilidad necesaria.

El problema que ahora se plantea es la manera de combinar la información de los diferentes atributos de fallo para obtener un diagnóstico único. Este problema se resuelve en el último módulo del esquema, al que se llamará Función de Detección de Fallos (o FDF).

Este módulo utiliza toda la información proporcionada por los atributos de fallo y la combina de manera que se pueda dar una importancia relativa diferente a cada atributo. La función de detección de fallos es una expresión de cualquier tipo, pero lo más sencillo es que se trate de una combinación lineal de los atributos de fallo o una función de tipo lógico.

3.5.1 Función de detección de fallos lineal

Se trata de una suma ponderada de los atributos de fallo. Cada atributo tiene asociado un parámetro que representa el peso o la importancia relativa de dicho atributo frente a los demás. Los pesos son útiles porque permiten sumar atributos de fallo de naturaleza diferente. Los atributos de fallo crecen a un ritmo distinto cuando aparece una anomalía en el proceso y por lo tanto un determinado nivel puede ser indicativo de fallo para un atributo, mientras que el mismo nivel puede ser un valor normal para otro. Los pesos asociados a cada atributo corrigen estas diferencias y en definitiva normalizan los valores de los atributos de fallo, lo que permite realizar una suma de los mismos.

La ecuación general de una FDF lineal es la suma de los atributos de fallo ponderada por los pesos correspondientes:

$$FDF = A_1 AF_1 + A_2 AF_2 + \ldots + A_m AF_m \qquad \textbf{(3.4)}$$

donde AF_i representa los diferentes atributos de fallo y A_i son los correspondientes pesos asociados.

La obtención de los valores de los pesos A_i no es inmediata ya que resulta difícil definir cuál es el ritmo de crecimiento de cada atributo en caso de fallo. La idea que se propone es que los atributos de fallo proporcionen un valor menor que la unidad de forma que durante una situación típica de fallo todos ellos alcancen el valor 1 de manera simultánea. Por ejemplo si se deduce, para un caso concreto, que hay fallo cuando un atributo de fallo (AF_1) supera el valor 10 y esta situación coincide con un valor 0,2 del segundo atributo de fallo (AF_2); pueden redefinirse los siguientes pesos: $A_1=1/10$, $A_2=1/0{,}2$. Entonces la función de detección de fallo tendría la siguiente expresión:

$$FDF = 0{,}1 \cdot AF_1 + 5 \cdot AF_2 \qquad \textbf{(3.5)}$$

En caso de fallo, la función de detección de fallos alcanza un valor igual al número de atributos utilizados. También se puede dividir todos los pesos por el número de atributos y entonces se detectará el fallo cuando la FDF alcance el valor unidad.

Las funciones de detección de fallos lineales no son apropiadas en el caso de atributos de fallo complementarios (aquellos que observan características que se manifiestan ante distintos modos de fallo). En este caso resulta difícil definir los pesos correctamente y además resulta difícil la detección porque sólo uno de los atributos aumenta cada vez. En el caso de complementariedad total no hay que dividir los pesos por el número de atributos de fallo y sin embargo se considera que existe fallo cuando FDF sea mayor o igual a uno.

Otro aspecto de la normalización que producen los pesos asociados a cada atributo de fallo es la posibilidad de comparar la eficacia de los atributos de fallo en situaciones diferentes. Si se considera el ejemplo anterior, se pueden definir los atributos de fallo normalizados (AF_{1N} y AF_{2N}) correspondientes a los atributos de fallo (AF_1 y AF_2) de la siguiente manera: $AF_{1N}=AF_1/10$ y $AF_{2N}=AF_2/0{,}2$. Con esta definición AF_{1N} y AF_{2N} alcanzan el valor unidad de manera simultánea en caso de fallo. En consecuencia, la FDF puede definirse con dos coeficientes iguales a 0,5 para que alcance el valor unidad en caso de fallo:

$$FDF=0{,}5 \cdot AF_{1N} + 0{,}5 \cdot AF_{2N} \qquad \textbf{(3.6)}$$

Suponiendo ahora que tras estudiar otro sistema de detección, que utiliza los mismos atributos normalizados, se llega a la conclusión de que resulta más conveniente utilizar la siguiente función de detección de fallos:

$$FDF = 0{,}7 \cdot AF_{1N} + 0{,}3 \cdot AF_{2N} \qquad \textbf{(3.7)}$$

entonces se puede deducir que en este segundo caso el primer atributo de fallo resulta más interesante que el segundo, comparado con el primer caso. En realidad los coeficientes 0,7 y 0,3 que se están utilizando ahora corrigen los valores de 10 y 0,2 utilizados anteriormente para la normalización; es decir, se está normalizando dos veces. Hay que señalar que ningún procedimiento automático puede encargarse del ajuste simultáneo de los dos pesos, ya que la dependencia que existe entre ellos da lugar a infinitas soluciones equivalentes. Sin embargo el interés que tiene esta manera de utilizar los atributos de fallo y los pesos es lo que permite comparar la importancia relativa de varios atributos en distintas situaciones. Se utilizan siempre los atributos normalizados con unos pesos fijos, que intentan corregir las diferencias entre los atributos de manera general, y luego se adaptan los coeficientes de la FDF a cada situación particular.

3.5.2 Función de detección de fallos de tipo lógico

En este caso se trata de establecer una relación lógica entre la activación de distintos atributos. Puesto que los atributos de fallo calculan un valor indicativo de la existencia de fallos en el proceso, podemos considerar que cada atributo está activo si su valor supera cierto límite. Cada atributo de fallo utilizará su propio límite ya que la velocidad con que aumenta cada uno es diferente y al utilizar límites independientes no es necesario realizar una normalización de los atributos. Como los atributos de fallo se anulan en situación normal de funcionamiento y crecen con la posibilidad de fallo, todas las comparaciones serán del tipo `mayor_que` y todos los límites serán positivos. Por ejemplo, la comparación del atributo de fallo AF_1 con el límite U_1 se formaliza mediante la siguiente relación lógica:

$$AF_1 > U_1 \qquad \textbf{(3.8)}$$

Esta relación es `cierta` si el valor que adquiere el atributo AF_1 es mayor que el límite U_1, que es una constante positiva. Se dice en este caso que el atributo de fallo está activo, lo que significa que este atributo está indicando la existencia de una anomalía.

Las relaciones lógicas entre las activaciones de distintos atributos de fallo pueden ser de tipo **and** o de tipo **or**. En las expresiones tipo **and** deben cumplirse todas las relaciones de manera independiente para que se cumpla la expresión completa, mientras que en las expresiones tipo **or** basta con que se cumpla una relación para que la expresión completa se considere `cierta`. Los dos casos básicos de función de detección de fallos de tipo lógico son los siguientes:

$$FDFand = AF_1 > U_1 \textbf{ and } AF_2 > U_2 \textbf{ and } \cdots \textbf{ and } AF_m > U_m \qquad \textbf{(3.9)}$$

$$FDFor = AF_1 > U_1 \textbf{ or } AF_2 > U_2 \textbf{ or } \cdots \textbf{ or } AF_m > U_m \qquad \textbf{(3.10)}$$

La ecuación 3.9 responde a un criterio más conservador donde *FDFand* es `falso` a menos que todos los *AF* estén activos simultáneamente. Por el contrario en el caso de la función de detección de fallos de la ecuación 3.10 se considera que los atributos son complementarios (observan características que se manifiestan ante distintos modos de fallo) y por lo tanto *FDFor* vale `verdadero` cuando cualquiera de ellos se activa.

La elección de una lógica tipo **and** o tipo **or** depende de la complementariedad de los atributos de fallo y de las especificaciones del sistema en cuanto a necesidad de confirmar la existencia de fallos antes de emitir el diagnóstico definitivo.

3.5.3 Función de detección de fallos basada en lógica borrosa

La FDF de tipo lógico depende mucho de los valores U_i, que requieren una definición precisa. Para relajar esta rigidez en la definición de los límites de activación es apropiado aplicar la teoría de conjuntos borrosa.

La lógica borrosa permite utilizar etiquetas semánticas tales como *alto*, *medio* y *bajo* para la definición de las reglas. En este caso la ecuación 3.9 se transformaría en la siguiente expresión:

$$FDFand = AF_1 \text{ es } alto \textbf{ and } AF_2 \text{ es } alto \textbf{ and } \cdots \textbf{ and } AF_m \text{ es } alto \quad \textbf{(3.11)}$$

La etiqueta *alto* es un concepto menos estricto que la comparación con un límite fijo. Existen varias maneras de describir el concepto de *alto*, pero lo más sencillo es utilizar funciones trapezoidales, que en este caso queda definida por dos valores (U^1 y U^2). Si el valor del atributo de fallo es menor que U^1 se puede afirmar con seguridad que no es *alto*, y por el contrario si fuese mayor que U^2 sería *alto* con certeza. Esta definición deja una zona de incertidumbre (o borrosa) para valores comprendidos entre U^1 y U^2 donde no se puede afirmar con total certeza si el nivel el *alto* o no. Se puede establecer la siguiente gráfica para representar el grado de certeza de la condición AF_1 es *alto* para distintos valores del atributo de fallo:

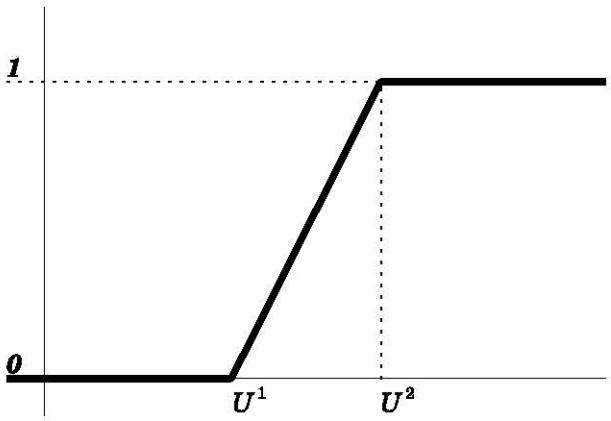


Figura 3.4: Certeza de la etiqueta *alto* en función del valor del atributo de fallo.

Cuando el atributo de fallo adquiere un valor comprendido entre U^1 y U^2 se estima una certeza de la afirmación *AF* es *alto* mediante la proporción representada en la gráfica. La combinación de los diferentes atributos de fallo mediante operaciones lógicas tipo **and** o tipo **or** deben tener en cuenta la certeza de cada término para elaborar el resultado final.

Esta manera de definir los límites de los atributos de fallos requiere dos parámetros en lugar de uno; esto dificulta el proceso de ajuste mediante

procedimientos automáticos. Sin embargo la experiencia demuestra que las personas ven facilitada su tarea de enunciar reglas y de definir límites para las variables. Las reglas se definen de una manera abstracta, utilizando un lenguaje más coloquial, y los límites se definen con mayor flexibilidad y de manera independiente a las reglas enunciadas, sin que la persona se vea forzada a dar un valor del cual no está segura. En definitiva la lógica borrosa en este caso facilita a las personas la tarea de definir la función objetivo pero no es el procedimiento apropiado para realizar un ajuste automático.

3.6 Resumen de la estructura general del sistema

El sistema de detección incipiente de fallos tiene una estructura modular muy flexible. Está basado en la utilización de un **modelo** que simula el comportamiento del proceso, lo que hace que sea poco sensible a variaciones en las condiciones de funcionamiento. Cualquier técnica de modelado puede aplicarse en este módulo.

Un **generador de residuos** se encarga de obtener una medida de discrepancia entre el comportamiento estimado por el modelo y la respuesta real del sistema.

A continuación, los **atributos de fallo** se encargan de buscar indicios de fallo aplicando diferentes técnicas de detección o algoritmos especializados en modos de fallo concretos. Cuando se aplica la detección basada en análisis de residuos, los atributos de fallo utilizan como fuente principal de información el valor instantáneo de los residuos. Sin embargo cuando se aplica la detección basada en el análisis de la evolución de los parámetros se utilizan los valores estimados para los parámetros del modelo. En ambos casos el esquema general (figura 3.1, página 36) es equivalente y sólo cambia el tipo de información que se utiliza como entrada a los atributos de fallo que es consecuencia del tratamiento que se realice con los residuos.

Finalmente un módulo llamado **función de detección de fallos** se encarga de evaluar todos los atributos de fallo de forma conjunta de manera

que se obtenga un diagnóstico único que puede ser de situación normal o de fallo incipiente.

3.7 Procedimiento de ajuste

Cuando se utilizan modelos matemáticos para simular el comportamiento de procesos industriales, es necesario realizar un ajuste de los parámetros de dichos modelos. Generalmente suelen aplicarse técnicas de ajuste por mínimos cuadrados utilizando conjuntos de datos correspondientes a condiciones de funcionamiento normal del proceso. Las técnicas de mínimos cuadrados tienen por objeto encontrar los parámetros que consiguen la máxima similitud entre las predicciones del modelo y las señales reales del proceso, tomando como base un conjunto de datos de funcionamiento normal. Este criterio es completamente independiente de la detección incipiente de fallos. Se ha observado que la calidad de un sistema de detección incipiente de fallos depende en gran medida de los parámetros utilizados (esta afirmación se describe en mayor detalle en el capítulo siguiente). Por lo tanto el procedimiento de ajuste es fundamental para conseguir un buen sistema de detección.

Para completar la descripción del sistema de detección es necesario entrar en aspectos propios del procedimiento de ajuste. La estructura del sistema de detección incipiente de fallos permite distinguir claramente tres grupos de parámetros: los parámetros del modelo, los parámetros de los atributos de fallo y los parámetros de la función de detección de fallos. Tradicionalmente los parámetros del modelo se ajustan de manera independiente por el método de los mínimos cuadrados y utilizando datos de comportamiento normal. Mientras que los valores del resto de los parámetros se fijan en base a la experiencia de la persona que define el sistema de detección y no requiere ningún ajuste. Por el contrario, en esta tesis se propone realizar un ajuste global de todos los parámetros del sistema de detección, bajo unos criterios específicos para detección incipiente de fallos y en base a datos reales de funcionamiento normal y de procesos de degradación (llamados historias de fallo). Para ello se definen unos módulos, llamados **atributos de detección**, que se encargan de

valorar el comportamiento del sistema de detección para las distintas condiciones de fallo.

Puesto que existen diferentes criterios para valorar un sistema de detección, normalmente se utilizarán varios atributos de detección que habrá que considerar de manera conjunta. Se han elegido las técnicas de optimización multi–objetivo como método para valorar simultáneamente varios atributos de detección. El procedimiento de ajuste es un algoritmo iterativo que propone conjuntos de parámetros, mediante métodos clásicos de optimización, hasta encontrar aquellos que proporcionan los mejores resultados. El esquema general del procedimiento de ajuste se muestra en la figura siguiente:

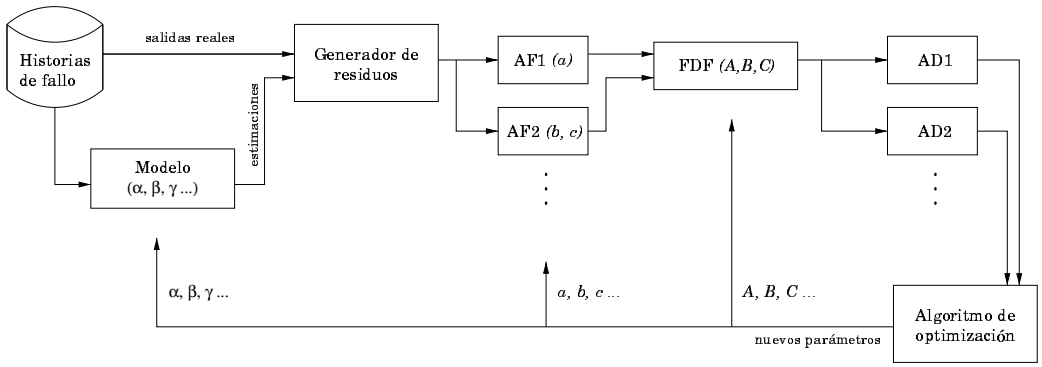


Figura 3.5: Esquema del procedimiento de ajuste

El capítulo 4 está dedicado íntegramente al procedimiento de ajuste, en él se describe el esquema general, se incluyen ejemplos de atributos de detección y se incluye una introducción a la optimización multi–objetivo.

Capítulo 4

Estructura del procedimiento de ajuste

4.1 Introducción

Los modelos matemáticos de cualquier tipo de proceso industrial, mecanismo o sistema dinámico son ecuaciones o algoritmos que dependen en general de varios parámetros. Es necesario obtener el valor de los parámetros para que los modelos queden totalmente definidos. Dos sistemas dinámicos iguales, por ejemplo el mismo diseño de suspensión aplicado a dos vehículos distintos, estarán representados por las mismas ecuaciones de movimiento. Sin embargo estas ecuaciones tendrán como parámetros la masa del vehículo, la rigidez de los muelles y otras características físicas que serán diferentes en cada caso. El modelo que representa adecuadamente el comportamiento dinámico de un vehículo necesita que los valores de sus parámetros estén bien ajustados a las

características de dicho vehículo. Se llama "ajuste" al proceso de obtención de los parámetros de un modelo.

En modelos basados en fundamentos físicos cada parámetro suele representar una característica del sistema. En algunos casos las características de los componentes son conocidas o se pueden medir mediante pequeños ensayos independientes. Sólo en estos casos el ajuste de los modelos puede realizarse directamente, sin utilizar datos experimentales de funcionamiento del sistema real.

En general el ajuste de modelos debe realizarse a partir de datos experimentales del funcionamiento del sistema en distintas situaciones. Esto es frecuente en el caso de modelos basados en fundamentos físicos e imprescindible cuando se utilizan otras técnicas de modelado como redes neuronales, modelos de caja negra o modelos estadísticos. Una de las dificultades típicas es la falta de datos suficientes para poder identificar correctamente todos los parámetros del modelo. Otra dificultad habitual es seleccionar apropiadamente el criterio de comparación entre el comportamiento del modelo y el comportamiento del sistema real. Del tipo de criterio empleado dependerán ligeramente los valores de los parámetros y en consecuencia el funcionamiento del modelo.

En este capítulo se describen algunos de los problemas asociados a la calidad de los datos utilizados en el proceso de ajuste. Se muestra, mediante un ejemplo completo, cómo varían los parámetros y las predicciones de los modelos en función del criterio utilizado durante el ajuste. Este análisis cuestiona la utilidad de los métodos tradicionales de ajuste de parámetros para modelos utilizados en detección incipiente de fallos y da paso a una nuevo planteamiento del procedimiento de ajuste que es una aportación fundamental de esta tesis. En este capítulo se propone un procedimiento de ajuste de los modelos, que luego se extiende a todo el sistema de detección de fallos, basado en datos de situaciones reales de fallo. Este proceso de ajuste permite obtener modelos específicos para la detección incipiente de fallos, que es el tema central de la tesis.

4.2 Comprobación de un sistema de detección de fallos

Los sistemas de detección de fallos que se están tratando en esta tesis deben trabajar en tiempo real para poder realizar la detección de manera incipiente. A partir de los datos obtenidos en cada instante de muestreo, el sistema de detección realizará los cálculos necesarios para decidir si el proceso se encuentra en situación de fallo o en situación normal. Estos cálculos pueden tener en cuenta los últimos datos y datos del pasado, pero el sistema de detección emite un diagnóstico por cada muestra tomada del proceso.

Puesto que el diagnóstico sólo depende de los datos de entrada y salida del proceso, es posible simular el comportamiento del sistema de detección a partir de los registros de funcionamiento del proceso. Por lo tanto, si se dispone de los datos de entrada y salida de un proceso (en formato digital) pueden obtenerse todos los diagnósticos que habría generado el sistema de detección de fallos en cada instante. Conociendo si los datos corresponden a situaciones de funcionamiento normal o a situaciones de fallo, puede comprobarse si el sistema de detección emite correctamente los diagnósticos de salud y de fallo. Este tipo de comprobación es válida para cualquier técnica de detección de fallos, desde la detección basada en umbrales de señales aisladas hasta el esquema de sistema de detección propuesto en esta tesis.

Tanto el ajuste del sistema de detección como la comprobación deben realizarse para distintas condiciones de trabajo, utilizando varios registros de datos. A los conjuntos de datos utilizados para ajustar y comprobar el sistema de detección de fallos los llamaremos conjuntos de entrenamiento (*training sets*) o historias de fallo. Los sistemas de detección de fallos tradicionales sólo se basan en la utilización de conjuntos de datos obtenidos durante condiciones normales de funcionamiento. Por el contrario, el procedimiento de ajuste que se propone en esta tesis requiere también la utilización de datos de procesos completos de degradación. Sin esta información no es posible realizar una comprobación del funcionamiento del sistema de detección. El método de ajuste propuesto es más completo porque utiliza toda la información disponible sobre el funcionamiento del proceso. La calidad de cualquier sistema de detección depende en gran medida de la riqueza de información disponible en el momento de realizar

el ajuste, es decir de la riqueza de información de los conjuntos de entrenamiento.

4.3 Riqueza de datos en el proceso de ajuste

La riqueza de los datos se mide en dos aspectos, la cantidad de datos (que ayuda a disminuir los efectos del ruido) y sobre todo la cantidad de situaciones de funcionamiento diferentes recogidas en los datos.

Si al ajustar un sistema de detección sólo se tiene en cuenta una condición de funcionamiento, es muy probable que se produzca una falsa alarma en cuanto se cambie el punto de trabajo. Veamos como ejemplo el caso de un detector de fallos de motores de combustión interna basado en el análisis del nivel de vibración. Si en el proceso de ajuste sólo se utilizan datos correspondientes a un grado de carga del motor, el nivel de vibración de todos los datos será probablemente bastante uniforme. Un sistema de detección basado en umbrales ajustaría los límites próximos a dicho nivel de vibración. Como consecuencia el sistema de detección emitirá una falsa alarma si cambia el grado de carga del motor y por lo tanto cambia el nivel de vibraciones. Es decir, el cambio de las condiciones de funcionamiento ha llevado al motor a un punto de trabajo diferente (aunque normal) que ha engañado al sistema de detección.

El mismo problema ocurre en el diagnóstico de transformadores eléctricos de potencia. En este caso los métodos tradicionales de detección están basados principalmente en el análisis de muestras de aceite. Sin embargo los muestreos generalmente son infrecuentes y además dependen en gran medida de la temperatura del aceite y de la historia de carga del transformador. Un cambio en las condiciones de trabajo puede modificar el resultado de los análisis sin que exista ninguna anomalía. Los sistemas de detección que no tengan en cuenta las diferentes condiciones de trabajo (o que se ajusten sin utilizar los datos de las diferentes situaciones) tendrán una utilidad muy limitada en la mayoría de los casos. Se puede afirmar que la falta de información es la principal causa de obtención de falsas alarmas.

Se han hecho algunas propuestas, que se describen a continuación, para reducir el número de falsas alarmas cuando no se dispone de suficiente

información. Esta falta de información puede tener lugar en la estructura del modelo o en la riqueza de los datos. Un **modelo** puede producir malas predicciones justo después de un cambio de carga, debido a efectos transitorios desconocidos o no incluidos en el modelo, y sin embargo realizar buenas predicciones el resto del tiempo. Un sistema que ajuste sus límites de detección a la condición transitoria (situación más desfavorable sin anomalía) no sería muy efectivo el resto del tiempo porque los límites serían muy grandes. Sin embargo si se aumenta la sensibilidad del sistema, se obtendrán falsas alarmas durante los transitorios. La solución propuesta en [Crowley90] y [Panossian97] es utilizar umbrales distintos durante los cambios de carga; de esta forma se puede aumentar la sensibilidad del sistema fuera de los periodos transitorios sin obtener falsas alarmas. Para reducir el problema de **falta de datos**, se pueden utilizar límites variables [Muñoz96] [Panossian97]. En este caso los límites varían dinámicamente en función de las condiciones de trabajo. Si el sistema se encuentra en una condición de trabajo de la cual existían suficientes datos durante el proceso de ajuste, los límites serán más estrechos ya que se trata de una situación conocida y suficientemente estudiada. Pero si la condición de trabajo es relativamente infrecuente y durante el proceso de ajuste no se disponía de muchos datos, es posible que el modelo no pueda representarla con la precisión deseada. En este caso los límites deben ser más relajados. La estructura del sistema de detección de fallos propuesta en esta tesis es suficientemente general como para contemplar límites variables, ya que el cálculo dinámico de los límites forma parte de un atributo de fallo.

En definitiva toda condición de funcionamiento no considerada durante el proceso de ajuste representa una incertidumbre, y es difícil predecir el comportamiento de los sistemas de detección en estas condiciones. En algunos casos unas nuevas condiciones de funcionamiento pueden indicar que existe un fallo en otro componente diferente al analizado. Se describe a continuación, como ejemplo, el caso de un motor eléctrico. Supongamos que el motor está alimentado a 220 V y que se han ajustado modelos que relacionan par, velocidad, intensidad y tensión, utilizando la información de una extensa base de datos. En este caso valores de tensión situados en el entorno del 5% del valor nominal habrán sido considerados durante el proceso de ajuste. Si por ejemplo se produce una situación en la cual la tensión es de 30 V, el comportamiento del motor será totalmente anormal.

En esta situación lo más probable es que las predicciones de los modelos no se parezcan a la realidad y por lo tanto que los residuos experimenten un cambio significativo. Como consecuencia se emitirá un aviso de fallo en el motor. Este aviso puede considerarse una falsa alarma ya que el fallo se encuentra seguramente en el sistema de alimentación y no en el motor. Una solución ese identificar primero si los datos de entrada al sistema de detección de anomalías se encuentran dentro de la región de confianza del espacio [Muñoz96]. Dicha región de confianza se define durante el proceso de ajuste en función de los datos utilizados en el mismo. Este tipo de análisis puede facilitar el trabajo del módulo de diagnóstico que en caso contrario puede verse desbordado por una secuencia de avisos de fallo en toda una cadena de componentes en serie.

4.4 Criterios para ajustar los parámetros de modelos

Generalmente el ajuste de modelos se realiza por el método de mínimos cuadrados [Hsia77]. Este método ajusta los parámetros de un modelo intentando que sus estimaciones se asemejen lo más posible al comportamiento del sistema real. El criterio de comparación utilizado es la suma de los cuadrados de las diferencias entre las estimaciones instantáneas del modelo y los correspondientes valores reales del sistema. El método de los mínimos cuadrados obtiene el conjunto de parámetros que hace mínima esta suma.

Se trata de un método sencillo y rápido de aplicar. En el caso de sistemas lineales el problema se reduce a resolver un pequeño sistema de ecuaciones. En el caso no lineal existen métodos específicos para obtener el resultado, como el algoritmo de Levenberg-Marquardt [Marquardt63] [Moré77]. Por estas razones es el método más utilizado en identificación de sistemas.

Sin embargo existen otras maneras de expresar el parecido entre los valores producidos por los modelos y las señales reales, que dan lugar a otros métodos de ajuste. Por ejemplo los métodos de máxima verosimilitud (que equivalen al método de mínimos cuadrados sólo si el ruido de las medidas es de tipo gausiano [Hsia77, Appendix 3A]), o el método minimax (que

minimiza la máxima diferencia entre las dos señales). Cuando el objetivo del modelo es utilizarlo para detección incipiente de fallos no está claro qué método de ajuste producirá los parámetros más beneficiosos.

Para mostrar la diferencia entre los resultados obtenidos con distintas técnicas de ajuste se tomará como ejemplo el comportamiento de un vehículo que baja un bordillo. El movimiento vertical del vehículo en esta situación es semejante a una función sinusoidal amortiguada (ver figura 4.3). El modelo que se utilizará para realizar los ajustes será una función coseno con un amortiguamiento (ecuación 4.1); se trata por lo tanto de un modelo matemático empírico que no se basa en fundamentos físicos del sistema.

El movimiento vertical del vehículo ha sido simulado mediante un modelo simplificado del sistema de suspensión (en [Elbeheiry96] se realiza un estudio comparativo entre distintos sistemas de suspensión en vehículos) que no debe confundirse con el modelo matemático que se utilizará para el ajuste. Se va a suponer que los datos simulados corresponden a un ensayo real y que la función cosenoidal amortiguada es el modelo matemático que se quiere ajustar utilizando distintas técnicas. Se ve más adelante que los parámetros del modelo toman distintos valores en función de la técnica de ajuste utilizada.

4.4.1 Simulación de la respuesta dinámica del vehículo

Se han despreciado los momentos de inercia del vehículo y las deformaciones estructurales. De esta forma puede estudiarse el comportamiento dinámico del vehículo modelando una sola rueda (y suponiendo una masa equivalente a la cuarta parte del total del vehiculo). Esto es lo mismo que considerar un reparto uniforme de peso, los mismos tarados de suspensión en cada rueda y que el escalón actúa simultáneamente sobre todas las ruedas.

Para simplificar el problema del contacto entre la rueda y el bordillo, se ha utilizado una curva de tipo cosenoidal como ecuación del suelo. Esta función indica que la sustentación del suelo no desaparece instantáneamente de debajo de la rueda sino que lo hace de forma progresiva, lo que resulta mucho más parecido a la realidad.

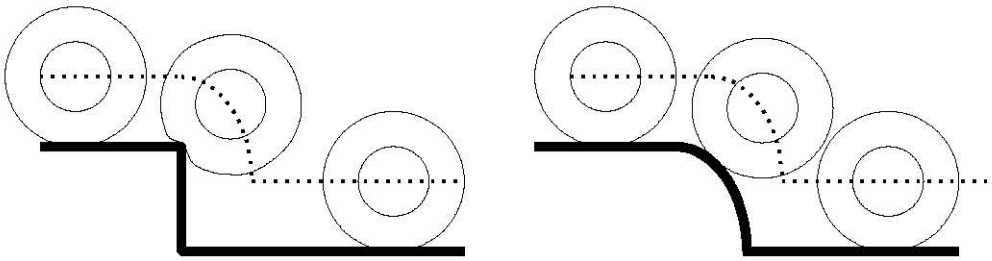


Figura 4.1: simplificación del problema de contacto entre rueda y bordillo

La figura 4.1 compara el comportamiento real con la situación programada para la simulación. Suponiendo que la transición del bordillo mide 20 cm (igual que la altura) y que la velocidad del coche es de 10 m/s (36 km/h), en la función temporal del suelo la transición sólo dura 2 ms.

Se ha modelado el sistema de suspensión como un sistema de dos grados de libertad (ver figura 4.2). La masa principal (Mc) corresponde al propio vehículo y la otra (Mr) representa la masa de la rueda y de otros elementos que se desplazan con la misma como frenos, ejes, etc. Para la simulación se ha supuesto un peso del vehículo de 1600 kg lo que da lugar a una masa $Mc=400$ kg. Como rigidez del neumático se ha tomado el valor $kn=190.000$ N/m y como rigidez de la suspensión $ks=16.000$ N/m, que da lugar a una frecuencia natural de 1 Hz. Estos valores son representativos de un coche normal y han sido tomados de [Elbeheiry96]. Se ha despreciado el amortiguamiento del neumático y se ha considerado un amortiguamiento de la suspensión de 1.500 $\frac{\text{N}}{\text{m/s}}$, que equivale aproximadamente a la tercera parte del valor crítico. También es un valor bastante frecuente, que absorbe la energía de una entrada en forma de escalón en 2 ó 3 ciclos. El esquema del sistema de suspensión, y una representación utilizando las técnicas de Bond-Graph (ver apéndice A) se muestra a continuación:

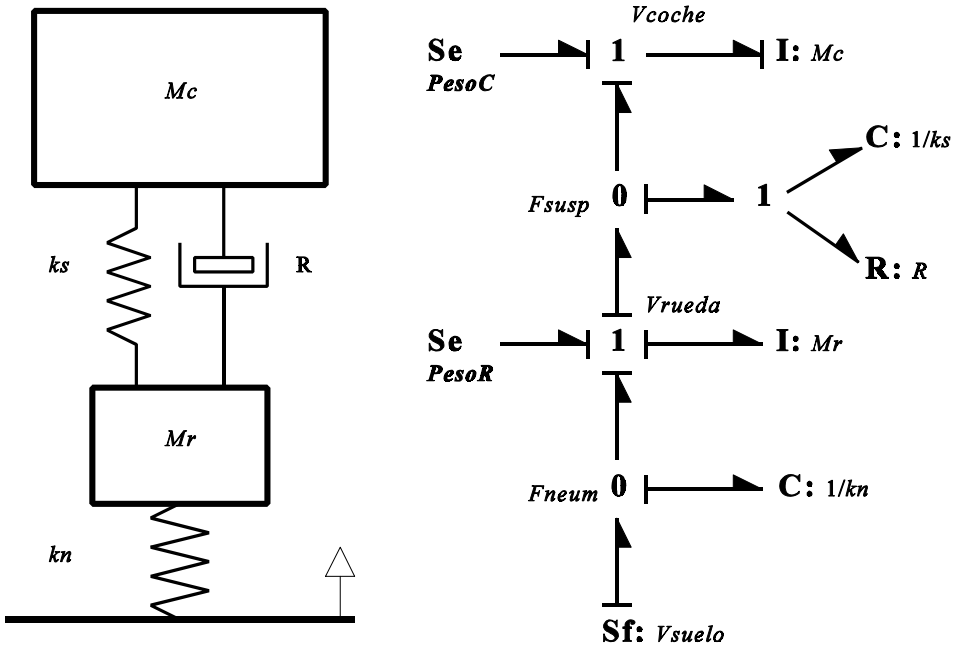


Figura 4.2: Esquema del sistema de suspensión

La siguiente gráfica muestra las señales de desplazamiento del vehículo y de la rueda producidas en la simulación.

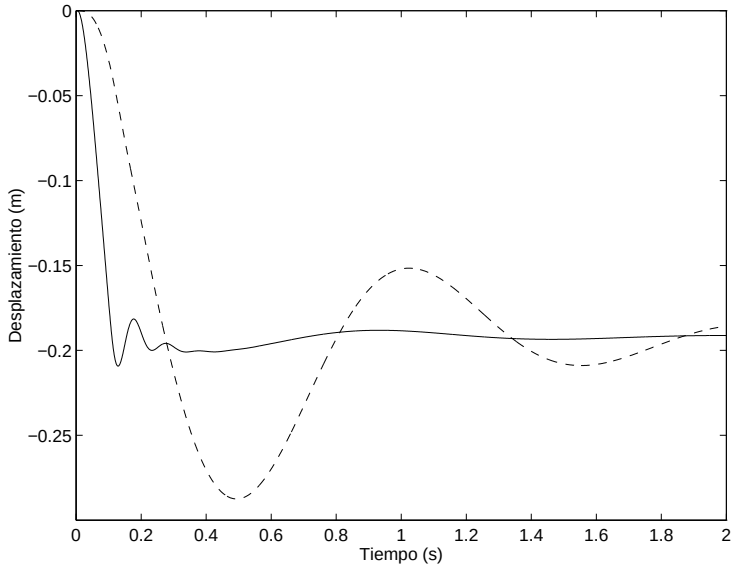


Figura 4.3: Señales de desplazamiento del coche y de la rueda

En ambos casos se llega en régimen permanente a un desplazamiento aproximado de -0.2 m respecto a la posición de equilibrio inicial. Esto es debido a la disminución del nivel del suelo en casi 20 cm (0,192 m exactamente). Se observa que la línea continua, que representa el movimiento de la rueda, sufre oscilaciones a una frecuencia de unos 10 Hz, esto se debe a la poca masa de la rueda y a la gran rigidez del neumático. La línea discontinua representa el movimiento del coche, en este caso las oscilaciones tienen una frecuencia algo superior a 1 Hz. Esta es la respuesta que se quiere modelar.

Por el aspecto de la curva se puede suponer que una función cosenoidal amortiguada es un buen aproximador. Sin embargo esta función no representa exactamente el movimiento real del sistema debido fundamentalmente a efectos no lineales. Entre los efectos no lineales destaca sobre todo el comportamiento del neumático. Se ha considerado que tiene una rigidez constante y por lo tanto produce una fuerza proporcional a la deformación (existen modelos de comportamiento del neumático mucho mejores como los propuestos en [Dugoff70], [Dugoff71], [Ray95], [Padovan94]), pero esta fuerza sólo puede ser de compresión, nunca de tracción, ya que la rueda no puede *agarrarse* al suelo. En la figura 4.4 se muestran las gráficas de movimiento de la rueda durante los primeros instantes. Debido a las inercias del sistema la rueda tarda 0,1 s en alcanzar el suelo desde que sale del bordillo. Esto hace que la rueda pierda el contacto con el suelo durante la primera décima de segundo, como se observa claramente en la curva de fuerza.

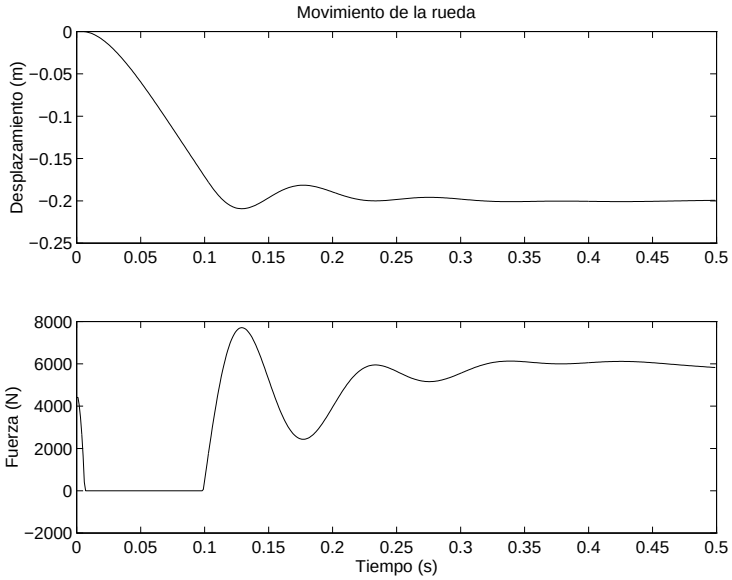


Figura 4.4: Señales de movimiento y fuerza del neumático

Además de estos efectos habría que añadir otros en un caso real. Por ejemplo el ruido en las señales, falta de linealidad y efectos de histéresis en los muelles, falta de proporcionalidad entre esfuerzo y velocidad en el amortiguador etc.

4.4.2 Métodos clásicos de ajuste del modelo

Se ha supuesto en este ejemplo que los datos obtenidos mediante simulación son los datos reales del sistema. A continuación se muestran los resultados de ajustar un modelo a estos datos utilizando distintas técnicas.

El modelo matemático utilizado es el siguiente:

$$\tilde{y} = -A + Ae^{\lambda t} \cdot \cos(2\pi f t) \qquad \textbf{(4.1)}$$

donde A es conocido ya que se trata de la posición final del coche en el ensayo que se quiere estudiar. Los únicos parámetros son λ (el amortiguamiento) y f (la frecuencia). El ajuste se ha realizado durante los primeros 5 s, el tiempo de simulación es t. La única entrada del modelo es

la altura del escalón (A) y la única salida es la posición relativa entre el vehículo y el suelo ($\bar{y}$).

Este modelo matemático es muy sencillo y no vale para estimar la posición relativa del vehículo para cualquier tipo de suelo sino que sólo vale para cambios bruscos de tipo escalón. Sin embargo es un primer paso hacia un sistema para la detección de fallos o de envejecimiento en el sistema de suspensión de un vehículo. Su implantación sería muy sencilla ya que bastaría instalar un sensor de medida de distancia al suelo para obtener todos los datos necesarios, como se muestra en la figura 4.5.

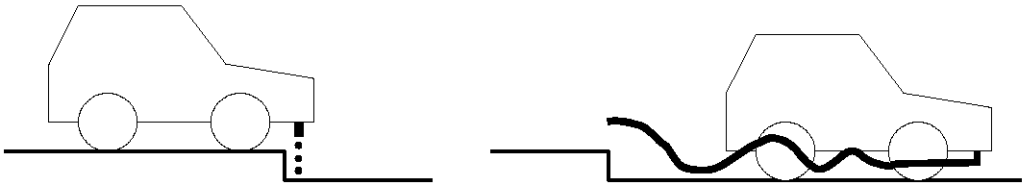


Figura 4.5: Funcionamiento del sensor para medir la distancia al suelo

Se han aplicado tres métodos de ajuste para comparar los resultados. Los métodos utilizados son los siguientes:

- **Suma de errores cuadráticos**. El método consiste en calcular el cuadrado de la diferencia entre los valores reales (o simulados) y los valores estimados por el modelo, para cada muestra entre $t=0$ y $t=5$. Luego se calcula la suma de todos estos valores instantáneos, que es el indicativo del grado de similitud funcional según este método. Minimizando este valor se obtiene el mejor modelo de acuerdo al criterio de mínimos cuadrados.
- **Suma de errores absolutos**. Es parecido al anterior, pero no se calcula el cuadrado de los errores sino el valor absoluto (para evitar que al obtener la suma se compensen los errores positivos con los negativos).
- **Error máximo**. Se utiliza como indicativo de la aproximación funcional el valor máximo de todos los errores instantáneos. El método consiste en hacer mínimo este error máximo. Es un problema complicado porque suele presentar problemas de mínimos locales.

Los resultados obtenidos se muestran en el cuadro siguiente. Los métodos de suma de errores cuadráticos y absolutos producen unos resultados diferentes, a pesar de basarse en un criterio muy parecido que haría pensar que los métodos son equivalentes. El método de error máximo da lugar a unos parámetros bastante diferentes.

Método	λ	f
Suma de errores cuadráticos	1,4282	0,9288
Suma de errores absolutos	1,4977	0,9362
Error máximo	1,1809	0,8954

Estas diferencias en los parámetros hacen que el modelo tenga comportamientos diferentes, y sobre todo que los errores instantáneos sean distintos. En la figura 4.6 se ha representado gráficamente el comportamiento de cada modelo. En la columna de la izquierda se ha dibujado la señal real de movimiento del coche, en trazo fino, y la predicción del modelo, en trazo grueso (sólo en el caso del método de error máximo se aprecia una diferencia clara). En la segunda columna se ha dibujado el error cuadrático en cada instante, para poder comparar el comportamiento de cada método.

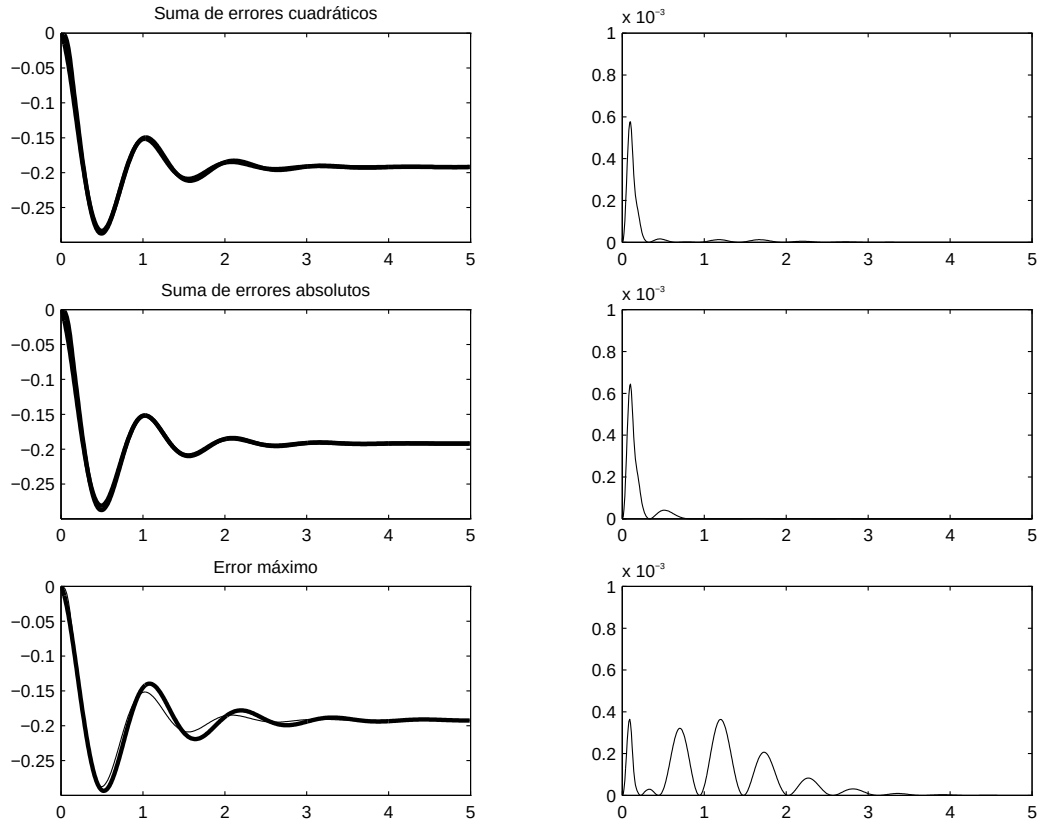


Figura 4.6: Comparación entre distintos criterios de ajuste

Aunque se ha utilizado un modelo matemático muy sencillo, que sólo vale para este tipo de ensayo, puede verse que las diferencias en los resultados son grandes. Se observa que el máximo error se produce en el segundo caso, en el cual se realizó un ajuste que minimiza la suma de errores absolutos.

Si un sistema de detección se basa sólo en comprobar si los residuos sobrepasan un umbral fijo, el valor máximo del error en condiciones normales de funcionamiento es un dato fundamental. Con objeto de evitar falsas alarmas en situación normal, los umbrales se fijan a valores ligeramente superiores a los valores máximos del error. En este ejemplo los dos primeros casos dan lugar a un umbral cercano a 0,7e−3 mientras que en el último caso el umbral puede valer 0,4e−3 (casi la mitad). Sin conocer cómo responde el sistema en situación de fallo incipiente, por ejemplo la respuesta con amortiguadores desgastados, resulta difícil decidir cuál de los métodos de ajuste resulta más apropiado. Sin embargo bajo el supuesto de un aumento progresivo de los residuos en caso de fallo, el último método sería el más apropiado ya que utiliza un valor del umbral menor.

Como conclusión, parece que el ajuste de modelos basado en minimizar el error máximo es el método de ajuste más apropiado –en este caso– para detección incipiente de fallos, aunque no sea el método que mejor ajusta el sistema en condiciones normales. Con este ejemplo queda claro que el ajuste de modelos basado en el criterio de mínimos cuadrados no es siempre el mejor método de ajuste cuando el modelo se quiere utilizar para detección incipiente de fallos.

4.4.3 Ajuste de modelos para detección incipiente de fallos

En el apartado anterior se ha visto que los resultados de las estimaciones de los modelos, y por lo tanto las curvas de residuos, son diferentes en función del método de ajuste utilizado para definir los parámetros de los modelos. Además, tampoco es posible decidir si alguno de los métodos expuestos es mejor que los otros con caracter general, ya que la mejor opción depende de dos factores fundamentales:

Por un lado debe tenerse en cuenta la evolución que sufren los residuos en caso de fallo, lo que anteriormente se ha llamado "sensibilidad del modelo al fallo" (apartado 2.4.1, página 28). Esto indica la necesidad de utilizar datos del proceso de degradación del sistema, además de los datos de funcionamiento normal, para poder realizar un buen ajuste.

Por otro lado también debe considerarse el tipo de sistema de detección de fallos que se va a encargar del análisis de los residuos. No es lo mismo un sistema de detección de fallos basado únicamente en un umbral de los residuos que un sistema de detección que además incluya cálculos de tendencias y otros atributos. Y tampoco es lo mismo exigir la ausencia total de falsas alarmas que ser más flexible en este aspecto en favor de otras características del sistema.

Por lo tanto, esta tesis sostiene que no se puede establecer a priori si un método de ajuste de modelos es mejor que otro sin tener en cuenta el resto del sistema de detección y el tipo de aplicación.

La conclusión final ha sido que el procedimiento de ajuste no debe centrarse sólo en los parámetros del modelo, sino que la aportación fundamental de esta tesis es realizar un ajuste global de todos los parámetros que intervienen en el sistema de detección de fallos. Es decir

los parámetros del modelo, los parámetros de los atributos de fallo y los parámetros de la función de detección de fallos.

Los parámetros asociados a cada atributo de fallo pueden variar el resultado de éstos. La dinámica de los atributos de fallo depende tanto de la dinámica de los residuos como de sus propios parámetros. En último lugar, los parámetros de la función de detección de fallos ponderan la importancia relativa que se da a cada atributo de fallo para emitir el diagnóstico, lo que en definitiva decide el comportamiento del sistema de detección.

4.5 Procedimiento de ajuste del sistema de detección incipiente de fallos

El procedimiento de ajuste que se propone en esta tesis intenta sacar el máximo partido de toda la información disponible, tanto datos de comportamiento normal, como datos obtenidos en situación de fallo. Los datos obtenidos en situación de fallo (en realidad durante un proceso de degradación), son fundamentales para poder comprobar la eficacia de un método de detección. Cuando no existe este tipo de datos, el sistema de detección sólo puede basarse en la suposición de que los residuos cambien en caso de fallo. Sin embargo, varios conjuntos de parámetros de un modelo pueden ser igualmente robustos en situación normal pero pueden tener distinta sensibilidad en caso de fallo, dando lugar a diferentes variaciones en los residuos.

Cuando sólo se dispone de datos de funcionamiento normal, el único criterio para realizar el ajuste es medir si el sistema de detección de fallos emite alguna falsa alarma. Esto se comprueba haciendo una simulación de su comportamiento que utiliza los datos de funcionamiento normal. Si en caso de fallo los residuos aumentan claramente, no habrá problema para detectar la anomalía. Sin embargo el criterio seguido en el ajuste puede fijar unos umbrales tan altos que determinados fallos no sean detectados o que procesos de degradación no sean identificados a tiempo. El comportamiento del sistema de detección en caso de fallo es algo que sólo puede comprobarse

con datos reales de degradación, simulando el comportamiento del sistema con distintas historias de fallo.

El método de ajuste que propone esta tesis aporta tres conceptos fundamentales: la utilización de historias completas de fallo, el ajuste global de todos los parámetros que intervienen en el sistema de detección y la utilización de técnicas de optimización multi–objetivo para obtener el subconjunto óptimo de soluciones.

La **utilización de historias de fallo** permite comprobar el funcionamiento del sistema de detección tanto en situaciones normales de funcionamiento como ante distintas situaciones de fallo. Esta comprobación es fundamental para conocer la eficacia del sistema de detección y poder compararlo con otros sistemas. Los métodos de ajuste tradicionales, que no utilizan datos de situaciones reales de fallo, sólo pueden refinar los modelos para que reproduzcan fielmente el comportamiento normal, pero no pueden especializarlos para detección incipiente de fallos.

La efectividad de un sistema de detección depende de los parámetros utilizados en el modelo y de los parámetros del resto del sistema(de los módulos encargados de analizar los resultados del modelo). En lugar de ajustar los parámetros del modelo de manera independiente, se propone hacer un **ajuste simultáneo de todos los parámetros del sistema**. Este ajuste se realiza aplicando técnicas de optimización que maximizan la capacidad de detección incipiente de fallos. El método de optimización utiliza información sobre la efectividad de sistemas que utilizan parámetros diferentes para proponer un nuevo conjunto de parámetros, hasta obtener el mejor sistema de detección.

Existen diversos criterios para valorar la efectividad de un sistema de detección; generalmente estos criterios analizan cualidades diferentes. Esta pluralidad de criterios dificulta la definición de un objetivo concreto de optimización ya que puede no estar claro a qué criterio se le debe mayor importancia relativa. El problema se resuelve aplicando **técnicas de optimización multi–objetivo**, mediante las cuales se obtiene el mejor subconjunto de soluciones, dejando para la etapa final la definición de la importancia relativa que se da a cada criterio.

Para poder comparar de forma objetiva sistemas de detección de fallos que utilicen conjuntos de parámetros diferentes, es necesario calcular una

serie de características o índices que valoren el comportamiento del sistema de detección en distintos aspectos. Se define por lo tanto un nuevo conjunto de atributos a los que llamaremos **atributos de detección**.

4.5.1 Atributos de detección

Son funciones diseñadas para valorar el comportamiento de un sistema de detección de fallos. Cada atributo de detección se centra en obtener una cualidad del sistema.

El ejemplo más sencillo es el número de falsas alarmas que se producen en periodos de funcionamiento normal. Este atributo de detección puede considerarse como un contador, más o menos sofisticado, del número de instantes en que la función de detección de fallos vale 1 sin que realmente exista una situación de anomalía. Se obtiene a partir de las simulaciones realizadas con datos de funcionamiento normal. Durante el proceso de ajuste interesa minimizar este atributo de detección hasta el caso ideal en que valga cero.

El atributo de detección **número de falsas alarmas** no aporta nada nuevo al proceso de ajuste de sistemas de detección sencillos, es sólo una formalización del proceso que normalmente se aplica. En general se ajustan los sistemas de detección de forma que no se produzcan alarmas durante la situación de funcionamiento normal, lo que equivale a hacer cero este atributo de detección. Sin embargo pueden definirse otros atributos de detección que no suelen tenerse en cuenta.

Un atributo de detección más relacionado con la detección incipiente y que sólo puede obtenerse utilizando historias de fallo es el **tiempo de detección**. Este atributo representa el tiempo que transcurre desde que se inicia una degradación en el proceso monitorizado hasta que el sistema de detección incipiente emite un diagnóstico de fallo. En cierto modo este atributo mide la sensibilidad del sistema de detección en caso de fallo. Como se ha dicho anteriormente, los residuos de un mismo sistema de detección pueden evolucionar de manera diferente en caso de fallo según los parámetros que se utilicen.

Un tercer atributo de detección que puede resultar interesante es el número de situaciones de fallo en las cuales el sistema de detección no ha

activado la alarma, o **fallos de detección**. Aunque idealmente estas situaciones no deben ocurrir, es posible que determinados modos fallo de un proceso no se manifiesten en las variables observadas por el sistema de detección, de manera que pueda producirse un fallo sin que el proceso de degradación haya sido detectado. Lo más probable es que sea necesario añadir un nuevo modelo sensible al modo de fallo no detectado. También puede ocurrir que una variable sensible al fallo y que se esté utilizando en un modelo tenga asociado un parámetro de valor muy pequeño; esto hace que sus variaciones resulten despreciables a todos los efectos. El procedimiento de ajuste de parámetros, basándose en el atributo fallos de detección, puede corregir el valor de este parámetro para aumentar la sensibilidad al modo de fallo (siempre que esto no perjudique sustancialmente al resto de los atributos de detección).

Al igual que ocurre con los atributos de fallo, puede definirse una gran cantidad de atributos de detección. Es necesario seleccionar los atributos de detección más apropiados, en función de los requisitos del proceso, para dirigir adecuadamente el proceso de optimización.

4.5.2 Esquema del procedimiento de ajuste

El procedimiento de ajuste se fundamenta en la simulación del comportamiento del sistema de detección ante diferentes situaciones de degradación. Las historias de fallo recogen los datos de entrada y salida del sistema monitorizado durante estas situaciones y se utilizan para comparar la efectividad del sistema de detección con distintos parámetros. Esta efectividad se mide por medio de varios atributos de detección, que se utilizan para dirigir el proceso de optimización. El ajuste se realiza de forma global, obteniendo todos los parámetros que intervienen en el sistema de detección; es decir los parámetros del modelo (α, β, γ...), de los atributos de fallo (a, b, c...) y de la función de detección de fallos (A, B, C...).

El esquema general del procedimiento de ajuste del sistema de detección incipiente de fallos se muestra en la figura 4.8.

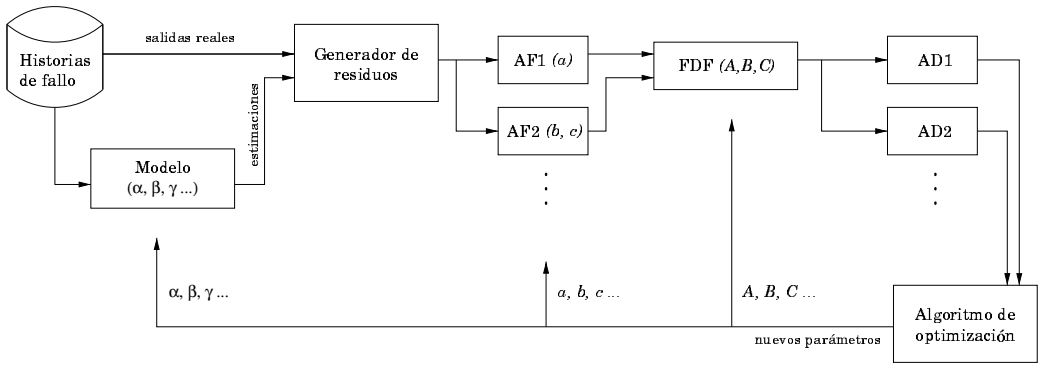


Figura 4.7: Esquema del procedimiento de ajuste

La base de datos de historias de fallo se utiliza para estudiar el comportamiento del sistema ante diferentes situaciones y poder establecer su eficacia de manera global. En la figura 4.8 se ha dibujado el sistema de detección completo, formado por el modelo, los atributos de fallo y la función de detección de fallos. A este sistema se le ha añadido el cálculo de los atributos de detección y un bloque que representa al algoritmo de ajuste de parámetros. Este algoritmo es el encargado de proponer unos nuevos parámetros del sistema en función de los valores de los atributos de detección. Los parámetros que propone el algoritmo de ajuste intervienen en todos los módulos del sistema de detección, como indica la flecha de "nuevos parámetros" que sale de este módulo en la figura 4.8.

Para mostrar el funcionamiento del procedimiento de ajuste, se describen a continuación los pasos que se siguen en el ajuste por el método tradicional y su analogía según el esquema de la figura 4.8. Se tomará como ejemplo un caso sencillo de sistema de detección de fallos, formado por un modelo lineal cuyos residuos se comparan con un umbral fijo. El procedimiento normal sería ajustar este sistema en dos fases, en la primera se ajustan los parámetros del modelo y en la segunda se fija el valor del umbral. El ajuste de los parámetros del modelo se realiza por el método de mínimos cuadrados, utilizando varios conjuntos de datos de funcionamiento normal. Este tipo de ajuste hace que los residuos sean pequeños en situación normal de funcionamiento. Seguidamente hay que establecer un límite para los residuos de forma que si son menores que el límite se considera que la situación es normal, mientras que si son mayores se

diagnostica fallo. Para que no se produzcan falsas alarmas este límite se fija a un valor ligeramente superior al valor correspondiente al máximo residuo que se produzca con los datos de funcionamiento normal. La analogía de este sistema de detección con el esquema general propuesto en esta tesis es sencilla, ya que se trata de una simplificación del mismo, pero servirá para aclarar la función de cada módulo del esquema. Se parte de un modelo lineal, que depende de varios parámetros (α, β, γ...) y cuyas predicciones se utilizan para generar residuos. Sólo existe un atributo de fallo (AF_1) que mide el nivel de sobrepaso del umbral (siendo el umbral el único parámetro del atributo de fallo). En función del valor de este atributo de fallo se decide si existe o no una anomalía. Por lo tanto la función de detección de fallos es simplemente un análisis para ver si el atributo de fallo es positivo (en este caso FDF no tiene ningún parámetro):

$$FDF = AF_1 > 0 \qquad \textbf{(4.2)}$$

De acuerdo al método tradicional, el procedimiento de ajuste sólo se basa en fijar los parámetros de tal forma que no se produzcan falsas alarmas. Por lo tanto el único atributo de detección es AD_1= "número de falsas alarmas" y el procedimiento de ajuste lo hace cero. Como no se permite modificar los parámetros del modelo, está claro que el procedimiento de ajuste consiste en modificar el umbral (único parámetro del sistema de detección aparte de los parámetros del modelo) hasta anular el número de falsas alarmas. Por otro lado las historias de fallo sólo incluyen datos sobre el funcionamiento normal del proceso y no se utilizan datos sobre degradación del mismo. Por lo tanto el algoritmo de ajuste propone un valor para el umbral, se analizan los datos para contar (mediante AD_1) el número de falsas que se producen y entonces el algoritmo de ajuste propone un umbral mayor; hasta que el número de falsas alarmas se anule.

Este pequeño ejemplo muestra que el método tradicional de ajuste de sistemas sencillos de detección que se ajustan sin considerar situaciones reales de fallo, es un caso particular de la estructura de ajuste más general que se propone en esta tesis.

El planteamiento del algoritmo de ajuste es la minimización sin restricciones de los atributos de detección:

$$\min \left\{ \begin{array}{l} AD_1(\alpha,\ \beta,\ \gamma \cdots,\ a,\ b,\ c \cdots,\ A,\ B,\ C \cdots), \\ AD_2(\alpha,\ \beta,\ \gamma \cdots,\ a,\ b,\ c \cdots,\ A,\ B,\ C \cdots),\ \cdots \end{array} \right\} \tag{4.3}$$

Esta formulación matemática quiere representar la minimización conjunta de todos los atributos de detección. Cuando se quieren minimizar varias funciones, suele tomarse como función objetivo la suma de las mismas; es decir, en este caso se plantearía el problema como:

$$\min \left\{ AD_1 + AD_2 + \cdots \right\} \tag{4.4}$$

Sin embargo, esta formulación da igual importancia relativa a todos los atributos de detección. Otras combinaciones lineales de los atributos de detección también provocan una minimización conjunta de los mismos, pero llegando a un resultado diferente. El siguiente planteamiento es un ejemplo en el cual el atributo 1 se considera 4 veces más importante que los otros:

$$\min \left\{ 4 \cdot AD_1 + AD_2 + \cdots \right\} \tag{4.5}$$

Como la importancia relativa de los distintos atributos de detección es algo difícil de decidir *a priori*, se ha optado por utilizar la notación general de la ecuación 4.3 y resolver el problema por optimización multi–objetivo (como se expone en el apartado siguiente).

En la ecuación 4.3 se han representado los atributos de detección como funciones de todos los parámetros del sistema. Ciertamente, la variación de alguno de los parámetros del modelo (α, β, γ...) produce un cambio en los residuos generados, lo cual puede cambiar los instantes en que se detectan los fallos y por lo tanto los valores de los atributos de detección. Por otro lado, un cambio en los parámetros de los atributos de fallo (a, b, c...) —por ejemplo el nivel de un umbral— también modifica los diagnósticos. Y por último los parámetros de la función de detección de fallos (A, B, C...) que definen la manera de combinar los atributos de fallo, también afectan al funcionamiento del sistema. Sin embargo, aunque está claro que existe una dependencia entre los atributos de detección y los parámetros del sistema,

la relación no puede representarse matemáticamente ya que los atributos de detección se calculan de forma empírica.

En definitiva se trata de un problema de optimización **no–lineal**, **sin restricciones** y con **gradiente desconocido**. Es no–lineal porque la función objetivo no puede expresarse como combinación lineal de las variables del problema (en este caso, los parámetros). No tiene restricciones porque los parámetros pueden tomar cualquier valor y no deben guardar ninguna proporción fija entre ellos. Y es de gradiente desconocido porque al no poder expresar matemáticamente la relación de los atributos de detección con los parámetros del sistema, tampoco se pueden obtener las derivadas de la función objetivo en las direcciones de cada uno de los parámetros. En el capítulo 5 (página 81) se exponen varios métodos generales para resolver este tipo de optimización así como métodos específicos para la solución de problemas propios del ajuste del sistema de detección incipiente de fallos.

4.6 Optimización multi-objetivo

Al ajustar un sistema de detección de fallos existe siempre una situación de compromiso entre robustez y velocidad de detección. Se entiende por robusto aquel sistema de detección que no se precipita en dar señales de alarma a menos que se haya confirmado la existencia de un fallo incipiente. El caso contrario son sistemas que producen falsas alarmas ante la mínima perturbación. Estos sistemas pueden confundir y cansar al personal encargado del mantenimiento y finalmente pueden ser ignorados.

Igual que robustez y velocidad parecen dos atributos de detección útiles para determinar la calidad de un sistema de detección de fallos, también puede haber otros criterios interesantes en cada caso particular. En general resultará difícil evaluar de forma conjunta todos los atributos de detección, ya que tienen una importancia relativa que puede depender de otros factores no considerados en el proceso de ajuste. Los atributos de detección robustez y velocidad de detección están relacionados mediante la característica de sensibilidad del sistema, sin embargo no se puede definir a priori qué proporción de los atributos es la idónea.

Este problema para fijar los criterios de la optimización no es algo exclusivo del tipo de ajuste que aquí se plantea, son muchas las situaciones en las que aparecen contradicciones al definir la función objetivo. Por ejemplo podríamos pensar en un sistema de limpieza de gases de una central térmica, donde los atributos deseables serían un bajo impacto ambiental y a la vez un bajo coste. Sin embargo estos dos atributos entran en conflicto, ya que el mejor sistema utilizará tecnología y equipos caros y seguramente resultará el más costoso, mientras que los sistemas sencillos y baratos tendrán una eficacia limitada. En un problema de optimización donde se quiere minimizar el coste existe una única función objetivo, la función de costes, y al minimizarla se obtiene una única solución del problema. Si en este ejemplo se puede valorar el impacto ambiental en forma de penalización económica se podría definir la siguiente función objetivo:

$$F.O. = Coste\_instalación + Coste\_ambiental \qquad \textbf{(4.6)}$$

Donde el *Coste_ambiental* es una penalización que depende de la eficacia del sistema que se instale. En este caso la solución sería única. Pero si el impacto ambiental no se puede expresar económicamente, nos encontramos ante un problema donde se quieren optimizar dos funciones objetivo, un problema multi-objetivo.

4.6.1 Representación multi–objetivo

En el caso de optimizaciones multi–objetivo no se puede obtener una solución única al problema, lo que interesa es obtener el conjunto de las mejores soluciones. Este conjunto de soluciones puede representarse en forma de gráfico multi–objetivo (o gráfico multi-atributo [deCuadra90]) para ayudar a realizar la selección. En los gráficos multi–objetivo cada eje representa un atributo o función objetivo que se quiere minimizar. Cada solución factible viene representada por un punto definido en función de sus atributos; todas las soluciones factibles forman una nube de puntos en el gráfico.

Siguiendo con el ejemplo anterior, tenemos dos atributos y por lo tanto el conjunto de soluciones se representa en un sistema bidimensional donde

el eje x representa el impacto ambiental y el eje y el coste. Un ejemplo del gráfico resultante sería el siguiente.

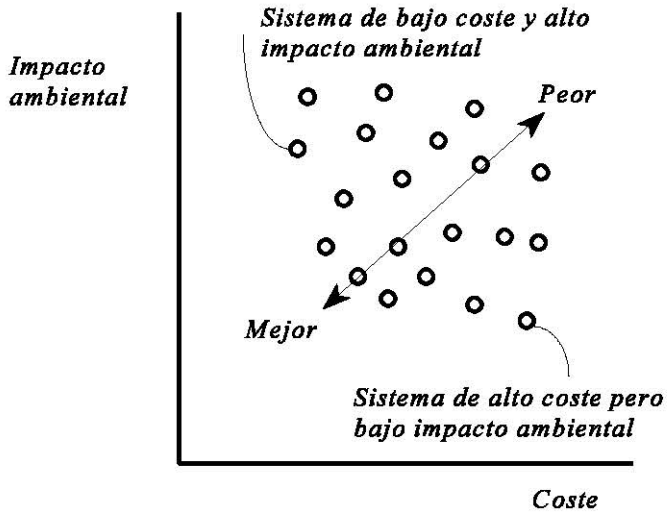


Figura 4.8: Gráfico multi–objetivo

En este tipo de representación las soluciones más próximas al origen de coordenadas son mejores porque el origen representa el punto donde todos los atributos son mínimos. Las soluciones cuya coordenada x sea pequeña tendrán en este ejemplo un coste bajo mientras que las soluciones cuya coordenada y sea pequeña tendrán poco impacto ambiental.

En el gráfico de la figura 4.8 se han dibujado algunas soluciones claramente peores que otras ya que presentan alto coste y malas prestaciones simultáneamente. Estas soluciones carecen de interés ya que nunca serán seleccionadas y por lo tanto pueden ser eliminadas.

4.6.2 Hipersuperficie óptima

Los únicos puntos de interés en una representación multi–objetivo son aquéllos que resultan mejores que el resto de los puntos en al menos una función objetivo. En el caso bidimensional la envolvente de estos puntos es una curva que tiende hacia el origen (ver ejemplos más adelante). En casos de mayor dimensión la envolvente de los puntos de interés forma una hipersuperficie a la que llamaremos **hipersuperficie óptima** (nombre tomado de [deCuadra90]).

Existen varios métodos para eliminar los puntos que carecen de interés y que permiten obtener la hipersuperficie óptima. Los dos métodos principales se basan en la definición de los criterios de **dominación estricta** y **dominación significativa** [Schweppe86]. Estos criterios está representados en la figura 4.9, donde se han rayado las zonas dominadas por una solución según cada criterio. Es decir, cualquier otro punto que se encuentre en la zona rayada carece de interés frente al punto dibujado.

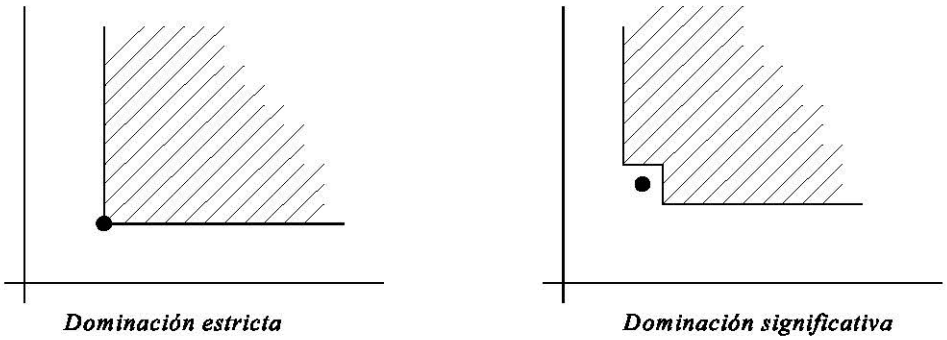


Figura 4.9: Criterios de dominación

Según el criterio de dominación estricta carecen de interés los puntos que sean peores que otra solución en todos los atributos. El criterio de dominación significativa es más refinado ya que rechaza además los puntos que producen una ligera mejoría en un atributo mientras empeoran significativamente el resto y sin embargo no rechaza puntos muy próximos aunque estén dominados estrictamente.[1]

[1]Hay que tener en cuenta que puntos muy próximos en el espacio de soluciones pueden corresponder a sistemas de detección de fallos con parámetros totalmente distintos.

En el caso de sistemas de detección de fallos se propone utilizar este tipo de representaciones para poder tomar la mejor decisión final dependiendo de cada situación. La figura 4.10 muestra tres representaciones multi–objetivo correspondientes a un sistema de detección de la descarga de una batería. En abscisas se ha representado el tiempo de detección y en

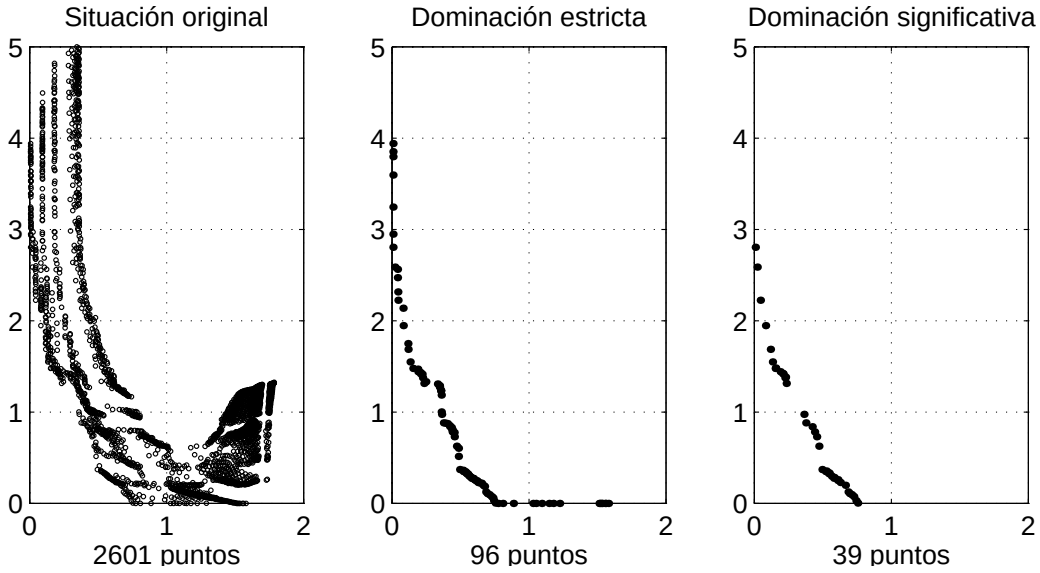


Figura 4.10: Obtención de la hipersuperficie óptima

ordenadas el número de falsas alarmas (en valores normalizados).

En la primera gráfica se han representado los puntos correspondientes a 2601 sistemas de detección que se han obtenido variando los parámetros del sistema de forma discreta. La segunda gráfica representa la hipersuperficie óptima según el criterio de dominación estricta y la tercera según el criterio de dominación significativa.

4.6.3 Obtención de la hipersuperficie óptima

El procedimiento de obtención de soluciones utilizado para generar la figura 4.10 (basado en la variación discreta de parámetros) no es aplicable para ajustar sistemas de detección, especialmente si el número de parámetros es alto. Sin embargo es posible obtener directamente la hipersuperficie óptima de una optimización multi–objetivo mediante optimizaciones mono–objetivo. Fernando de Cuadra describe en su tesis doctoral [deCuadra90] el concepto de optimalidad respecto a una familia de

funciones objetivo que permite obtener directamente la hipersuperficie óptima. El procedimiento, aplicado al ajuste del sistema de detección incipiente de fallos, consiste en asociar un parámetro λ a cada atributo de detección y definir un conjunto de funciones objetivo dando distintos valores a estos parámetros. Cada parámetro λ pondera la importancia relativa de un atributo de detección, matemáticamente las funciones objetivo de una familia pueden expresarse según la ecuación siguiente:

$$FO_i = \lambda_{i,1} \cdot AD_1 + \lambda_{i,2} \cdot AD_2 + \cdots \lambda_{i,n} \cdot AD_n \qquad \textbf{(4.7)}$$

Cada función objetivo es una combinación lineal de atributos de detección que puede resolverse mediante técnicas clásicas de optimización (ver capítulo 5 en página 81) y da lugar a un punto de la hipersuperficie óptima (siempre que no existan problemas de convergencia en la optimización). En caso de existir problemas de mínimos locales, puede aplicarse el criterio de dominación estricta para eliminar puntos poco interesantes.

4.6.4 Aplicación de técnicas de procesamiento paralelo

El número de funciones objetivo de una familia debe ser suficiente como para obtener un conjunto significativo de puntos en el gráfico multi–objetivo. Por lo tanto es necesario realizar muchas optimizaciones de la misma función objetivo en las cuales varían los valores de los parámetros λ. Sin embargo, todas las optimizaciones son independientes y los resultados que se obtienen para unos valores de λ no son necesarios para el cálculo con valores de λ diferentes. Hay que tener en cuenta que valores muy próximos en el diagrama multi–objetivo sólo se parecen en los valores de los atributos de detección, pero pueden corresponder a conjuntos de atributos totalmente diferentes.

Con objeto de reducir el tiempo de cálculo en el proceso de obtención de la hipersuperficie óptima se propone realizar los cálculos en paralelo. En un ordenador para procesamiento paralelo el tiempo de cálculo de este tipo de problemas se puede reducir por el número de procesadores. Asimismo hay que tener en cuenta que sólo es necesario acceder a datos comunes durante en inicio de cada optimización simple, después cada proceso evoluciona de manera independiente. Debido a estas condiciones lo más

apropiado en este caso es aplicar ejecución distribuida, es decir repartiendo los procesos por un conjunto de estaciones de trabajo conectadas en red.

La ejecución distribuida sólo resulta ventajosa cuando el flujo de datos entre los distintos procesos que corren en paralelo es pequeño, ya que en caso contrario se pueden producir importantes retrasos como consecuencia de la saturación de la red local. Sin embargo resulta generalmente más sencillo poder disponer de un gran conjunto de estaciones de trabajo en red que de un ordenador multiprocesador. En el caso de la obtención de la hipersuperficie óptima se dan las condiciones de poca comunicación entre procesos y por lo tanto resulta apropiado aplicar programación distribuida.

4.7 Conclusiones

El procedimiento de ajuste que se propone se basa realmente en el análisis del comportamiento de sistema de detección completo. El estudio del comportamiento del sistema se obtiene utilizando historias de fallo de diferentes situaciones, tanto de funcionamiento normal como situaciones de degradación. Mediante simulaciones con las historias de fallo se puede conocer en qué condiciones el sistema de detección indica la existencia de fallos y se puede comprobar si el diagnóstico es correcto.

Los atributos de detección se encargan de obtener medidas objetivas para poder comparar diferentes sistemas de detección en todos los posibles modos de funcionamiento. Todos los parámetros que intervienen en el sistema de detección incipiente de fallos afectan al comportamiento del mismo y por lo tanto cualquier variación altera los valores de los atributos de detección. El ajuste del sistema se plantea como una optimización global de todos los parámetros que intervienen en el sistema de detección de manera que se minimicen los atributos de detección (ver la siguiente figura).

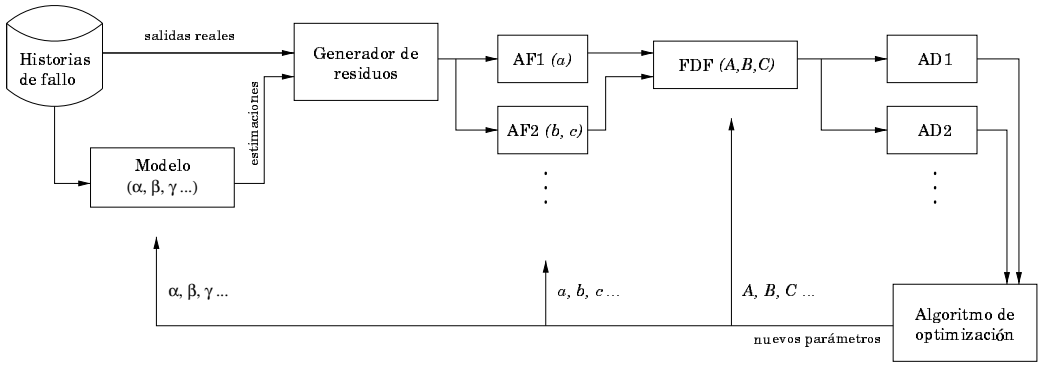


Figura 4.11: Esquema general del ajuste del sistema de detección

El ajuste simultáneo de todos los parámetros, la utilización de datos tanto de comportamiento normal como de fallo y el hecho de basar el ajuste en los atributos de detección, suponen las diferencias fundamentales frente a otros tipos de ajuste.

La utilización simultánea de varios atributos de detección lleva al planteamiento del problema del ajuste como una optimización multi–objetivo, en la cual no es necesario definir a priori la importancia relativa entre cada atributo de detección. Por lo tanto el resultado final es un diagrama que recoge el conjunto de sistemas de detección que resultan mejores en todas las cualidades. Cada punto de este diagrama tiene asociado el conjunto de todos los parámetros que definen el sistema de detección incipiente de fallos.

Capítulo 5

Optimización del sistema

5.1 Introducción

En el ajuste de parámetros de modelos siempre hay implícito un proceso de optimización. Generalmente se ajustan los parámetros con objeto de conseguir el máximo parecido posible entre las predicciones del modelo y los datos experimentales.

La manera más utilizada de valorar el parecido entre el modelo y la realidad es calcular la suma de los cuadrados de los errores de cada estimación. Cuando el modelo es lineal, y se plantea como función objetivo minimizar la suma de los cuadrados de los residuos, los parámetros del modelo se obtienen de forma inmediata resolviendo un pequeño sistema de ecuaciones (de la dimensión del número de parámetros). En otras situaciones, como es el caso del ajuste de parámetros del sistema de detección de fallos propuesto en esta tesis, hay que aplicar procedimientos iterativos de optimización para obtener el conjunto de parámetros que da

lugar al mejor modelo (según los criterios que se definan para valorar la calidad del modelo).

Este capítulo comienza con una descripción de métodos de optimización unidimensional y multidimensional que resultan de interés para el ajuste del sistema de detección de fallos. Tras esta revisión se presentan los resultados de una serie de estudios comparativos de métodos de optimización que se han realizado con objeto de estudiar tanto la velocidad de convergencia como la robustez de los métodos.

La existencia de mínimos locales supone un problema importante que dificulta los procesos de optimización. En el apartado 5.4 se analiza el origen de los mínimos locales en el problema del ajuste del sistema de detección de fallos y se aportan soluciones para el caso unidimensional y para el caso multidimensional.

Por último, en el apartado 5.5 se describen dos técnicas aplicadas específicamente al caso del sistema de detección y que mejoran el proceso de ajuste del mismo.

5.2 Revisión de métodos de optimización unidimensional

La optimización en una variable suele aplicarse en distintas direcciones para minimizar una función de n variables. En algunos casos las técnicas de optimización en una variable pueden extenderse al caso de varias dimensiones.

La optimización en una dirección, que se llama **búsqueda lineal**, se realiza mediante técnicas básicas de optimización unidimensional, ya que la optimización en una dirección fija del espacio es equivalente a la optimización de una función unidimensional que se obtiene realizando un cambio de variable. Las técnicas de optimización unidimensional pueden clasificarse en dos grupos: técnicas basadas en la comparación de los valores de la función objetivo y técnicas basadas en la aproximación de la función objetivo.

5.2.1 Optimización basada en la comparación de valores de la función objetivo

Estos métodos se basan en el tratamiento de los valores de la función objetivo evaluada en distintos puntos. Los puntos considerados definen un intervalo en el que se encuentra la solución, el intervalo se va reduciendo en cada iteración al obtener el valor de la función objetivo en un nuevo punto del intervalo y compararlo con los anteriores.

Los procedimientos más conocidos son el método de Fibonacci [Murray72] [Pike86], *bisection* [Murray72], *Golden Section* [Murray72] [Pike86], y *Lattice search* (para variables discretas) [Pike86], que se diferencian en los criterios que siguen para seleccionar el nuevo punto de evaluación y para reducir el intervalo que incluye la solución. Veamos por ejemplo el método de *Golden Section*. Este método utiliza los valores de la función en cuatro puntos de un intervalo para decidir qué nuevo subintervalo de tres puntos toma. Conocido el valor de la función objetivo en cuatro puntos consecutivos α_1, α_2, α_3 y α_4 se selecciona el intervalo (α_1, α_3) ó (α_2, α_4) comparando los valores de la función objetivo en α_2 y α_3. Los cuatro puntos se toman de acuerdo al siguiente criterio:

$$\alpha_3 - \alpha_1 = \alpha_4 - \alpha_2 = \beta(\alpha_4 - \alpha_1)$$
$$\text{donde} \quad \beta = \frac{2}{1+\sqrt{5}} = 0{,}618... \qquad \textbf{(5.1)}$$

Es decir, por condiciones de simetría el intervalo se reduce según el coeficiente β en cada iteración, y el cuarto punto del subintervalo seleccionado se define de manera que vuelva a cumplirse la condición de la ecuación anterior. Si el valor de la función en α_3 es mayor que el valor de función en α_2, se selecciona el intervalo (α_1, α_3).

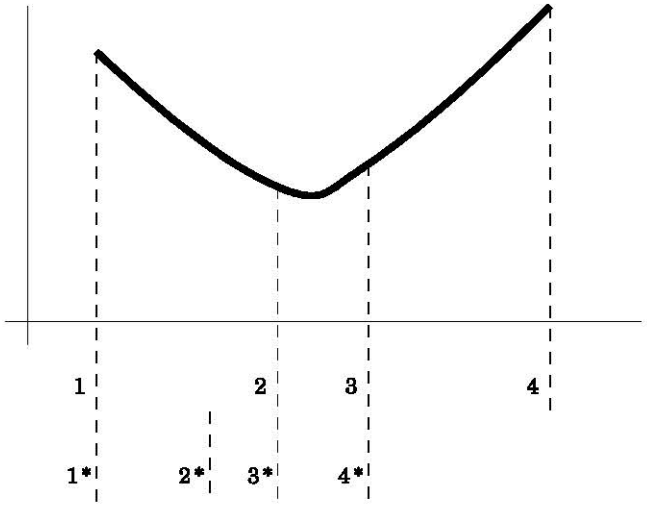


Figura 5.1: Disminución del intervalo por el método *Golden Section*

El tercer punto del nuevo intervalo (α_3^*) debe cumplir las ecuaciones 5.1 aplicadas al nuevo intervalo; pero β está definido de manera que $\alpha_3^* = \alpha_2$ ya que para cualquier valor de α_1 y de α_4 se cumple:

$$= \alpha_1^* + \beta(\alpha_4^* - \alpha_1^*) = \alpha_1 + \beta(\alpha_3 - \alpha_1) = \alpha_1 + \beta(\alpha_1 + \beta(\alpha_4 - \alpha_1) - \alpha_1)$$
$$= \alpha_1 + \beta^2(\alpha_4 - \alpha_1) \qquad \textbf{(5.2)}$$
$$= \alpha_1 - \alpha_3 + \alpha_4 = \alpha_1 - \left(\alpha_1 + \beta(\alpha_4 - \alpha_1)\right) + \alpha_4 =$$
$$= \alpha_1 + (1-\beta)\cdot(\alpha_4 - \alpha_1)$$

pero β está definido de tal forma que $1-\beta=\beta^2$, como puede demostrarse a partir de la definición de β en la ecuación 5.1, por lo tanto $\alpha_3^* = \alpha_2$ y sólo es necesario calcular α_2^*. Esta manera hábil de seleccionar el coeficiente β hace que sólo sea necesaria una evaluación de la función objetivo en cada iteración.

Todos estos métodos basados en la comparación de valores de la función objetivo utilizan exclusivamente la información de los puntos del último intervalo, ignorando el resto. Aunque esté garantizado que sólo existe un

mínimo en el intervalo, el tiempo de convergencia puede ser muy grande. Sin embargo se puede calcular el número necesario de iteraciones para conseguir una determinada precisión. En el caso del método de *Golden Section* el número de iteraciones N viene dado en función de la precisión P y de la anchura inicial del intervalo A según la ecuación siguiente:

$$N = \frac{\ln P - \ln A}{\ln \beta} \tag{5.3}$$

El método de Fibonacci es el que consigue una reducción más rápida del intervalo conociendo el número de iteraciones. Cuando se conoce inicialmente el número de iteraciones, por ejemplo para aplicaciones en tiempo real donde el número de iteraciones sea fijo, el método de Fibonacci proporciona una estrategia óptima de reducción del intervalo que viene descrita por una serie de valores (series de Fibonacci) equivalentes al coeficiente β utilizado antes. Una buena descripción sobre el método de Fibonacci se encuentra en [Pike86] [Gill81]. También hay descripciones y ejemplos del método *Golden Section* en [Murray72] [Luenberger84] [Pike86] [Gill81].

5.2.2 Optimización basada en aproximaciones de la función objetivo

En este caso la función objetivo se aproxima por una función $\hat{F}(x)$ que coincide en un determinado número de puntos con la primera. También se pueden hacer aproximaciones de forma que coincidan incluso las derivadas en determinados puntos.

La función $\hat{F}(x)$ debe ser una función fácilmente ajustable y con mínimos conocidos y fácilmente calculables, normalmente se toman polinomios de segundo o tercer grado. El procedimiento de cálculo consiste en ajustar la función $\hat{F}(x)$ a los valores y derivadas de la función objetivo en un conjunto de puntos y luego calcular el valor α para el cual es mínima. El nuevo punto α pasará a formar parte de los puntos de ajuste para la siguiente iteración.

La selección del número de puntos que forman el conjunto de ajuste no es sencilla, y la mayor parte de los problemas de estos métodos de ajuste residen en esta cuestión. Para ajustar polinomios de segundo grado hacen falta 3 puntos en los que se conozca el valor de la función y para ajustar

polinomios de tercer grado suelen utilizarse 2 puntos donde se conozca el valor de la función y la primera derivada en la dirección de optimización. Al evaluar la función en un nuevo punto será necesario eliminar uno de los puntos anteriores si se quiere realizar una interpolación con polinomios del mismo grado. El criterio de sustitución de unos puntos por otros y el comportamiento de la función puede llevar a situaciones "inestables" en las cuales la estimación del mínimo queda totalmente apartada de los puntos considerados en el ajuste.

En la siguiente figura aparecen, como ejemplo de este problema, los primeros pasos de la minimización de la función $y=x\cdot \text{sen}(2\pi x)$ (dibujada en trazo continuo) en el intervalo $x\epsilon[0.37,\ 1.05]$. En este ejemplo se ha utilizado interpolación parabólica, lo que requiere utilizar tres puntos conocidos de la función para obtener $\hat{F}(x)$. El mínimo de la función $\hat{F}(x)$ nos proporciona un cuarto punto que generalmente pertenece al intervalo de la solución. El punto que tiene máximo valor de la función objetivo, que suele ser uno de los extremos del intervalo, se elimina para hacer la siguiente interpolación utilizando los tres mejores. En la figura se parte de los puntos $x1=0.37$, $x4=1.05$ y se calculan dos puntos del interior del intervalo según las proporciones propuestas en el método *Golden Section*. Los cuatro puntos aparecen marcados con círculos negros (●), entonces se elimina el punto más desfavorable ($x=1.05$) y se interpola la parábola que aparece en trazo de puntos. El mínimo de la parábola está indicado con una circunferencia (○) y al caer fuera del intervalo y con un valor superior al de $x1$, el procedimiento queda bloqueado.

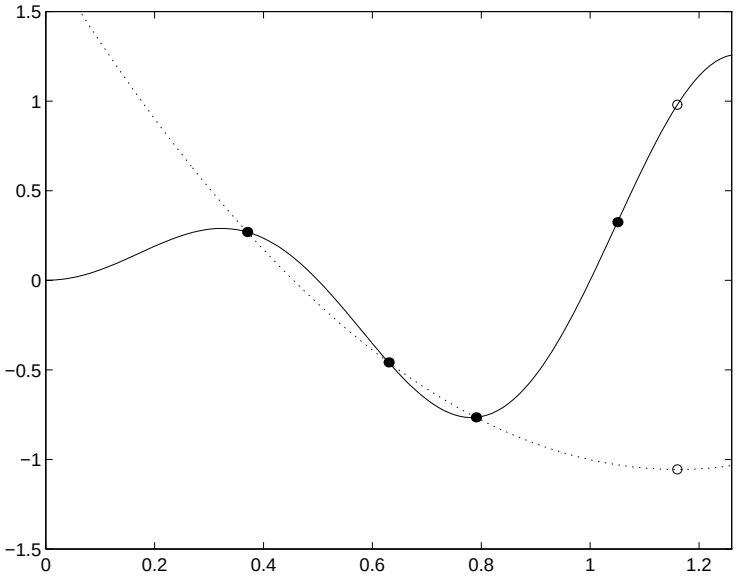


Figura 5.2: Minimización de una función por ajuste parabólico

Aunque la convergencia de los métodos de aproximación funcional suele ser más rápida que la de métodos basados en comparación de valores, en algunos casos pueden presentarse los problemas comentados, por lo tanto son menos robustos. La solución es utilizar una estrategia mixta entre comparación de valores y aproximación funcional. Un **método combinado** de optimización [Murray72] consiste en acotar el intervalo en el que se encuentra la solución y encontrar el mínimo de la función aproximada, este valor se utilizará sólo si se encuentra dentro del intervalo. El mínimo estimado se ignora cuando queda fuera del intervalo y en ese caso se avanza en la optimización reduciendo el intervalo por un método de comparación funcional.

5.3 Revisión de métodos de búsqueda directa

Los métodos numéricos para determinar el óptimo de una función objetivo no lineal de varias variables son procedimientos iterativos que parten de una solución inicial y van obteniendo una secuencia de puntos que mejoran la solución progresivamente. Los métodos se diferencian en la estrategia

que utilizan para encontrar el siguiente punto a partir de una solución factible no óptima.

Los métodos de optimización no lineal pueden clasificarse en tres grupos: métodos Quasi-Newton, direcciones conjugadas y búsqueda directa.

Los métodos **Quasi-Newton** son simplificaciones del método de Newton, que basan la búsqueda de soluciones en el valor de la función objetivo y en el valor de las primeras derivadas. Partiendo de una solución inicial calculan una dirección en la cual la solución mejora y avanzan en dicha dirección.

Matemáticamente los métodos Quasi-Newton se fundamentan en la aproximación de la función objetivo alrededor de un punto, en cada iteración sólo se considera el valor de la función objetivo y de sus derivadas en el punto actual para avanzar al nuevo punto. La información relativa a los puntos de iteraciones anteriores no se considera para el cálculo de la nueva solución.

Existen muchas variaciones de estos métodos que tienen distinto rendimiento dependiendo de la forma de la función objetivo y de la velocidad de evaluación de la función y de sus derivadas. Cuando el gradiente de la función objetivo no es conocido, puede realizarse un cálculo aproximado del mismo mediante evaluaciones de la función alrededor del punto de interés. Sin embargo, este tipo de cálculo ralentiza mucho el proceso, especialmente si el número de variables de optimización es grande. En estos casos resulta más eficiente utilizar métodos de búsqueda directa (descritos más adelante) que métodos basados en la información del gradiente.

Los métodos de **direcciones conjugadas** no están basados en el valor del gradiente para obtener la dirección de avance a partir de una solución. Realizan avances según direcciones perpendiculares entre sí y luego calculan la dirección de "aceleración" a partir del punto de solución anterior y del punto que resulta de los n movimientos perpendiculares. Estos métodos son muy útiles cuando no es fácil o no se puede calcular la derivada de la función objetivo y el comportamiento es bueno cuando la función es suave, tipo cuadrático. La velocidad del método depende fundamentalmente del algoritmo de búsqueda lineal utilizado.

El método de direcciones conjugadas más utilizado, por su sencillez, es el método de Powell [Powell64], que utiliza las direcciones de los ejes como

direcciones de avance. Otra alternativa, relacionada con los métodos quasi-newton, es utilizar como dirección de avance la dirección del gradiente de la función, ya que al realizar una optimización unidireccional en la dirección del gradiente se llega a un punto en el cual el gradiente es perpendicular a la dirección de avance (salvo errores de redondeo). Este último método, que se conoce como **gradiente conjugado** sólo se diferencia del método del gradiente puro (o *steepest descent*) en el paso de aceleración que tiene lugar cada n iteraciones del método del gradiente puro.

En algunos problemas de optimización, como es el caso del ajuste del sistema de detección de fallos, las funciones objetivo vienen dadas en forma de algoritmos. Los algoritmos pueden ser complicados y resultar lentos de calcular. El cálculo del gradiente tampoco es posible de forma analítica, por lo tanto la única información disponible es el valor de la función. Además la función objetivo puede ser muy irregular y presentar cambios bruscos o puede tener ruido asociado. En estos casos el comportamiento de los métodos de direcciones conjugadas no es tan bueno y los métodos de fundamento teórico como aquellos basados en gradientes no son aplicables, pero se puede recurrir a los métodos de **búsqueda directa**, también conocidos como métodos lógicos o métodos de orden cero.

Los métodos de búsqueda directa se caracterizan por no utilizar información del gradiente de la función objetivo y son idóneos en los casos en que no se conocen o no son continuas las derivadas parciales de primer orden de la función. La única información que utilizan es el valor de la función en diferentes puntos, y el único requisito es la continuidad de la función.

Aunque los métodos directos han sido desarrollados heurísticamente y su convergencia no ha sido probada matemáticamente, presentan el mejor comportamiento para el ajuste del sistema de detección de fallos que se propone en esta tesis, como se demuestra más adelante en este capítulo.

A continuación se explican los métodos de búsqueda directa utilizados, en concreto los métodos *Pattern Search*, *Razor Search* y el método *Simplex* con algunas de sus variaciones más utilizadas.

5.3.1 Método de optimización *Pattern Search*

Es un método de optimización propuesto por Hooke y Jeeves [Hooke & Jeeves 61] que realiza movimientos de exploración y movimientos de avance para generar una secuencia de "puntos básicos" que mejoran la solución progresivamente.

El movimiento de exploración se realiza alrededor de un punto $\boldsymbol{x}$ mediante movimientos sucesivos de paso constante en la dirección de cada variable, siempre que mejoren el valor de la función objetivo de punto $\boldsymbol{x}$, y da lugar a un nuevo punto básico $\boldsymbol{x}^{(k)}$ (que puede ser el propio punto $\boldsymbol{x}$ si todos los movimientos fracasan en encontrar un punto mejor). El método original contempla la posibilidad de realizar movimientos en las direcciones de todas las variables, sin embargo el punto básico que se obtiene depende del orden en que se realiza la exploración; existen distintas maneras de realizar este movimiento.

El movimiento de avance queda definido en magnitud y sentido por el vector trazado desde el punto básico anterior $\boldsymbol{x}^{(k-1)}$ hasta el punto básico actual $\boldsymbol{x}^{(k)}$ según la ecuación:

$$\boldsymbol{x} = \boldsymbol{x}^{(k)} + \left(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)} \right) \qquad \textbf{(5.4)}$$

El movimiento de avance da lugar a un nuevo punto $\boldsymbol{x}$ alrededor del cual se realiza un movimiento de exploración para obtener el punto básico $\boldsymbol{x}^{(k+1)}$. Para que el movimiento de avance sea aceptado el valor de la función objetivo debe disminuir; es decir tiene que cumplirse que $\mathrm{F}\left(\boldsymbol{x}^{(k+1)}\right) \leq \mathrm{F}\left(\boldsymbol{x}^{(k)}\right)$.

El proceso de optimización parte de un punto inicial $\boldsymbol{x}^{(1)}$ mediante un movimiento de exploración que da lugar al segundo punto básico $\boldsymbol{x}^{(2)}$. Luego realiza movimientos de avance (que conllevan un movimiento de exploración) hasta que el movimiento de avance sea desestimado. Puede comprobarse que cuando la dirección es favorable, los movimientos de avance son cada vez mayores, y las exploraciones que tienen lugar van ajustando la dirección de avance.

La figura 5.3 muestra los primeros pasos de la minimización de la función $z=(y-x^2)^2+(x-1)^2$, partiendo del punto $\boldsymbol{x}^{(1)}=[-0.3,-0.8]$ y con un paso inicial de 0.1 para el movimiento de exploración.

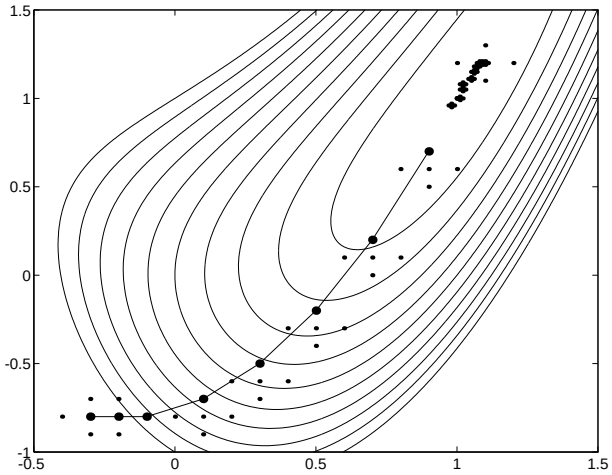


Figura 5.3: Optimización por el método de Hooke y Jeeves

En este ejemplo el movimiento de exploración se ha programado de forma que sólo hay desplazamiento según uno de los ejes, pero se comprueban las cuatro posiciones finales posibles para elegir la mejor. Se han marcado con círculos pequeños los puntos en los que se ha evaluado la función objetivo en los movimientos de exploración, y con círculos grandes los puntos básicos seleccionados.

Cuando el movimiento de avance es desestimado, se realiza un movimiento de exploración alrededor del punto básico actual para obtener un nuevo punto básico (como al iniciar el procedimiento). Esto frena los movimientos de avance.

En las proximidades de la solución el movimiento de exploración alrededor del punto básico tampoco tendrá éxito, lo que obliga a disminuir el incremento que se aplica a las variables en la exploración. Cuando los incrementos de las variables sean menores que la precisión necesaria sin que estas variaciones consigan mejorar la solución, se dará por concluida la búsqueda.

El método es bastante rápido y fácil de programar. Requiere muy poca memoria, a menos que en un problema de dimensión grande se quiera evitar la re-evaluación de la función objetivo cuando fallan los movimientos de avance.

El número de evaluaciones de la función objetivo depende mucho del procedimiento que se utilice en el movimiento de exploración. Realizando movimientos sucesivos según todas las direcciones pero sin que el resultado sea óptimo requiere $2n+1$ evaluaciones de la F.O. en el caso más desfavorable (por ejemplo cuando no haya movimiento). Sin embargo, para localizar la mejor posición admitiendo desplazamientos según las n direcciones simultáneamente, es necesario evaluar la F.O. en 3^n puntos. Otra posibilidad, que se ha utilizado para generar la figura 5.3, es localizar la mejor posición admitiendo un único desplazamiento según la variable más sensible. Este último criterio requiere $2n+1$ evaluaciones de la F.O. en cualquier caso.

El mayor inconveniente del método de optimización *Pattern search* es que puede quedar atrapado en un valle escarpado que no siga las direcciones de los ejes, como consecuencia de una disminución acelerada del paso utilizado en el movimiento de exploración. Para evitar estos problemas existen pequeñas modificaciones como el método de Rosenbrock propuesto en 1960 [Murray72] [Rosenbrock60] y el método D.S.C. propuesto por Davies, Swann y Campey [Murray72] [Swann]. Estos métodos se basan fundamentalmente en la variación de las direcciones de los movimientos de exploración en la búsqueda local. Con ello se consigue una mejor adaptación al problema evitando tener que reducir el paso de exploración. Otra variante muy interesante que evita los problemas de valles escarpados y algunos problemas de mínimos locales es el método *Razor Search* que se expone a continuación.

5.3.2 Método *Razor Search*

Este método fue desarrollado por Bandler y McDonald en 1969 [Bandler69] [Murray72] para la minimización de la desviación máxima entre la respuesta de un sistema de redes de micro-ondas y la respuesta ideal deseada.

Es una variante del método de optimización *Pattern search* enfocada a la optimización de funciones objetivo mal condicionadas que presenten valles escarpados. En este tipo de problemas el método *Pattern search* no tiene buen comportamiento porque si la dirección de valle no coincide con la dirección de uno de los ejes, los movimientos de avance no tienen éxito

y por lo tanto los puntos básicos sólo cambian por medio de movimientos de exploración. También es frecuente que la optimización termine antes de haber alcanzado el óptimo.

El método *Razor search* intenta mantener una dirección de avance para evitar que las reducciones en el paso del movimiento de exploración frenen la evolución del método. Para ello actúa de dos maneras: haciendo variar el paso del movimiento de exploración en función de la velocidad de avance, y realizando movimientos aleatorios cuando el método se ralentiza para buscar una dirección de avance en la dirección de valle.

En el artículo publicado por Bandler y MacDonald [Bandler69] están descritos el método, todos los parámetros que intervienen, los diagramas de flujo de los algoritmos y una comparación de rendimientos.

En concreto el método *Razor search* propone aplicar la siguiente ecuación para el movimiento de avance:

$$\boldsymbol{x} = \boldsymbol{x}^{(k)} + \alpha \left(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)} \right) \tag{5.5}$$

donde α normalmente vale 1, siendo el método original, pero si el movimiento no tiene éxito se intenta un avance con $\alpha=0.5$ y si es necesario otro con $\alpha=-0.5$ antes de descartarlo y pasar a una búsqueda local. Esto permite una mejor adaptación de la dirección de avance a la función objetivo que en el ejemplo de la figura 5.3 cuando el valle es más escarpado.

Sólo en caso de fracasar en todos los intentos de avance se pasará a intentar un movimiento de exploración. Sin embargo, esto no ralentiza tanto el procedimiento porque todas las búsquedas locales se realizan con un paso de valor:

$$\frac{\boldsymbol{x} - \boldsymbol{x}^{(k)}}{\sqrt{n}} \tag{5.6}$$

donde n es la dimensión del espacio. En caso de desestimar el movimiento de exploración, no queda más remedio que reducir el paso en un factor constante. La búsqueda termina cuando el paso del movimiento de exploración es menor que un valor ϵ.

Para evitar que el método evolucione sólo mediante pequeños movimientos de exploración, seguramente porque se ha llegado a un valle escarpado cuyo trayecto se sigue lentamente, se intenta buscar una nueva dirección de avance mediante un movimiento aleatorio. El movimiento aleatorio y la posterior optimización permiten llegar a un nuevo punto básico en el valle que se utilizará para definir la nueva dirección de avance. Parece lógico pensar que en las primeras optimizaciones el valor de ϵ debe ser mayor que la precisión deseada en la solución. Los autores del método proponen definir el número m de movimientos aleatorios (o reinicios de optimización) y entonces partir del valor dado por la siguiente fórmula:

$$\epsilon = \epsilon_{min} \cdot \eta^m \qquad \textbf{(5.7)}$$

donde ϵ_{min} es la precisión final y η es un factor de escala. Después de cada movimiento aleatorio se recalcula ϵ dividiendo el valor anterior por η. Se comprueba fácilmente que el valor final de ϵ en la última optimización parcial es igual a ϵ_{min}.

El movimiento aleatorio se obtiene sumando a cada coordenada del punto básico bloqueado una cantidad según la siguiente ecuación:

$$x_i^{(k)} = x_i^{(k-1)} + \rho \cdot R(1) \cdot \epsilon \qquad \text{para} \quad i = 1..n \qquad \textbf{(5.8)}$$

donde R(1) es un valor aleatorio de distribución uniforme entre −1 y 1, y ρ es un factor de escala que puede hacerse dependiente de los límites de las variables o de la distancia recorrida en la última optimización parcial.

Este método fue desarrollado por Bandler y McDonald en 1969 [Bandler69] [Murray72] para la minimización de la desviación máxima entre la respuesta de un sistema de redes de micro-ondas y la respuesta ideal deseada.

Es una variante del método de optimización *Pattern search* enfocada a la optimización de funciones objetivo mal condicionadas que presenten valles escarpados. En este tipo de problemas el método *Pattern search* no tiene buen comportamiento porque si la dirección de valle no coincide con la dirección de uno de los ejes, los movimientos de avance no tienen éxito y por lo tanto los puntos básicos sólo cambian por medio de movimientos

de exploración. También es frecuente que la optimización termine antes de haber alcanzado el óptimo.

El método *Razor search* intenta mantener una dirección de avance para evitar que las reducciones en el paso del movimiento de exploración frenen la evolución del método. Para ello actúa de dos maneras: haciendo variar el paso del movimiento de exploración en función de la velocidad de avance, y realizando movimientos aleatorios cuando el método se ralentiza para buscar una dirección de avance en la dirección de valle.

En el artículo publicado por Bandler y MacDonald [Bandler69] están descritos el método, todos los parámetros que intervienen, los diagramas de flujo de los algoritmos y una comparación de rendimientos.

En concreto el método *Razor search* propone aplicar la siguiente ecuación para el movimiento de avance:

$$\boldsymbol{x} = \boldsymbol{x}^{(k)} + \alpha \left(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)} \right) \tag{5.9}$$

donde α normalmente vale 1, siendo el método original, pero si el movimiento no tiene éxito se intenta un avance con $\alpha=0.5$ y si es necesario otro con $\alpha=-0.5$ antes de descartarlo y pasar a una búsqueda local. Esto permite una mejor adaptación de la dirección de avance a la función objetivo que en el ejemplo de la figura 5.3 cuando el valle es más escarpado.

Sólo en caso de fracasar en todos los intentos de avance se pasará a intentar un movimiento de exploración. Sin embargo, esto no ralentiza tanto el procedimiento porque todas las búsquedas locales se realizan con un paso de valor:

$$\frac{\boldsymbol{x} - \boldsymbol{x}^{(k)}}{\sqrt{n}} \tag{5.10}$$

donde n es la dimensión del espacio. En caso de desestimar el movimiento de exploración, no queda más remedio que reducir el paso en un factor constante. La búsqueda termina cuando el paso del movimiento de exploración es menor que un valor ϵ.

Para evitar que el método evolucione sólo mediante pequeños movimientos de exploración, seguramente porque se ha llegado a un valle escarpado cuyo trayecto se sigue lentamente, se intenta buscar una nueva dirección de avance mediante un movimiento aleatorio. El movimiento aleatorio y la posterior optimización permiten llegar a un nuevo punto básico en el valle que se utilizará para definir la nueva dirección de avance. Parece lógico pensar que en las primeras optimizaciones el valor de ϵ debe ser mayor que la precisión deseada en la solución. Los autores del método proponen definir el número m de movimientos aleatorios (o reinicios de optimización) y entonces partir del valor dado por la siguiente fórmula:

$$\epsilon = \epsilon_{min} \cdot \eta^m \qquad \textbf{(5.11)}$$

donde ϵ_{min} es la precisión final y η es un factor de escala. Después de cada movimiento aleatorio se recalcula ϵ dividiendo el valor anterior por η. Se comprueba fácilmente que el valor final de ϵ en la última optimización parcial es igual a ϵ_{min}.

El movimiento aleatorio se obtiene sumando a cada coordenada del punto básico bloqueado una cantidad según la siguiente ecuación:

$$x_i^{(k)} = x_i^{(k-1)} + \rho \cdot R(1) \cdot \epsilon \qquad \text{para } i = 1..n \qquad \textbf{(5.12)}$$

donde R(1) es un valor aleatorio de distribución uniforme entre −1 y 1, y ρ es un factor de escala que puede hacerse dependiente de los límites de las variables o de la distancia recorrida en la última optimización parcial.

5.3.3 Método Simplex

El método Simplex para optimización no lineal sin restricciones propuesto por Spendley, Hext y Himsworth en 1962 [Murray72] se basa en la evaluación de la F.O. en un conjunto de puntos mutuamente equidistantes. Para un problema de dimensión n este conjunto está formado por $n+1$ puntos, en el caso de dos dimensiones el conjunto forma un triángulo equilátero, en el caso de tres dimensiones un tetraedro regular, etc. Estos elementos tienen la propiedad de permitir la generación de un nuevo

elemento que tiene en común un lado (o cara) con el primero al hallar la simetría del punto libre respecto al lado común.

La simetría del punto se calcula con respecto al centroide $\boldsymbol{x}_c^{(k)}$ de la cara opuesta; es decir, si llamamos $\boldsymbol{x}_0^{(k)}$, $\boldsymbol{x}_1^{(k)}$, ... $\boldsymbol{x}_n^{(k)}$ a los $n+1$ puntos del elemento en la iteración k, la simetría del punto j se calcula mediante la siguiente ecuación vectorial:

$$\boldsymbol{x}_j^{(k+1)} = \boldsymbol{x}_c^{(k)} + \left(\boldsymbol{x}_c^{(k)} - \boldsymbol{x}_j^{(k)} \right) \qquad \textbf{(5.13)}$$

o bien:

$$\boldsymbol{x}_j^{(k+1)} = \frac{2}{n} \left(\boldsymbol{x}_0^{(k)} + \boldsymbol{x}_1^{(k)} + \ \ldots \ + \boldsymbol{x}_{j-1}^{(k)} + \boldsymbol{x}_{j+1}^{(k)} + \ \ldots \ + \boldsymbol{x}_n^{(k)} \right) - \boldsymbol{x}_j^{(k)} \qquad \textbf{(5.14)}$$

Por lo tanto cada iteración sólo requiere evaluar la F.O. en un punto.

El algoritmo parte del punto $\boldsymbol{x}_0^{(1)}$ a partir del cual se calculan los otros n puntos que forman el elemento básico inicial. El punto i del elemento se obtiene sumando la cantidad δ_2 a la componente i de $\boldsymbol{x}_0^{(1)}$ y sumando δ_1 al resto de las componentes [Murray72]. La definición de δ_1 y δ_2 es la siguiente:

$$\delta_1 = \frac{\sqrt{n+1} + n - 1}{n\sqrt{2}} \ ; \qquad \delta_2 = \frac{\sqrt{n+1} - 1}{n\sqrt{2}} \qquad \textbf{(5.15)}$$

Una vez definido el poliedro inicial, el método evoluciona calculando la simetría del punto cuya F.O. es más desfavorable. Si en dos iteraciones consecutivas el punto más desfavorable es el mismo, se calcula la simetría del segundo punto más desfavorable; esto evita que el algoritmo se quede oscilando. Si a pesar de todo existe al menos un vértice que no ha cambiado en las últimas α iteraciones, será necesario reducir el tamaño del poliedro hasta un tamaño correspondiente a la precisión deseada. El número necesario de iteraciones (α) para decidir la reducción del elemento es experimental, el valor propuesto por Spendley, Hext y Himsworth es función del número de variables del problema: $\alpha = 1.65n + 0.05n^2$

5.3.4 Método de Nelder-Mead

Probablemente la versión más eficiente del método Simplex es la que incluye las modificaciones propuestas por Nelder y Mead en 1965 [Nelder & Mead 64] [Murray72].

En este método el elemento poliedro se autoescala dinámicamente de acuerdo con el comportamiento local de la función objetivo. En este caso la ecuación 5.13 queda modificada de la siguiente manera:

$$\boldsymbol{x}_j^{(k+1)} = \boldsymbol{x}_c^{(k)} + \gamma \left(\boldsymbol{x}_c^{(k)} - \boldsymbol{x}_j^{(k)} \right) \qquad \textbf{(5.16)}$$

donde γ es el coeficiente de reflexión. Si $\gamma=1$ es el caso de reflexión sin deformación del método Simplex original y da lugar al punto reflejado x_r.

Cuando el valor de la F.O. en este punto, F_r, es mejor que en todos los demás puntos del poliedro, se prueba con un valor $\gamma>1$ (por ejemplo $\gamma=2$) para llegar al punto x_e que expande el poliedro. Si F_e sigue siendo el mejor punto se acepta la expansión y por tanto $\boldsymbol{x}_j^{(k+1)}=\boldsymbol{x}_e$ (Figura 5.4 caso 1).

Si el punto reflejado no mejora la situación, es decir $\mathrm{F}\left(\boldsymbol{x}_j^{(k)}\right)<F_r$, entonces se realiza una contracción sin reflexión (Figura 5.4 caso 2), tomando γ entre -1 y 0 (por ejemplo $\gamma=-0.5$).

Si el punto reflejado mejora la situación aunque sigue siendo el peor del nuevo poliedro (Figura 5.4 caso 3), entonces se realiza una contracción tomando un valor de γ entre 0 y 1 (por ejemplo $\gamma=0.5$).

El coeficiente de expansión $\gamma=2$ y los de contracción $\gamma=\pm0.5$ se consideran los mejores tras diversos ensayos [Nelder & Mead 64].

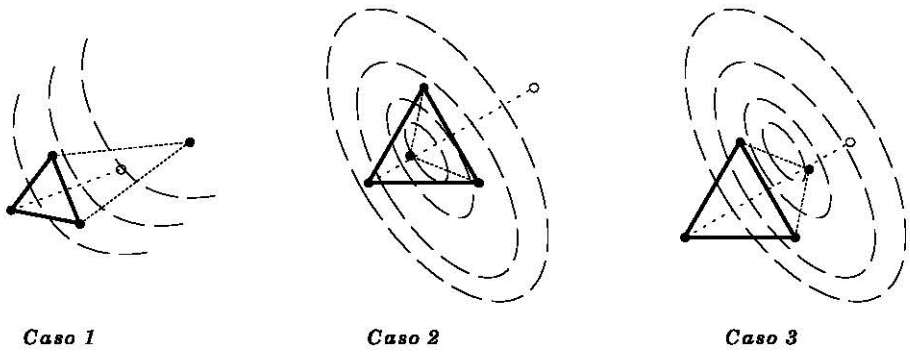


Figura 5.4: Modificaciones al método simplex propuestas por Nelder y Mead

En cualquier caso el nuevo punto no puede ser el peor del poliedro, para evitar que el método se quede iterando inútilmente. Si esto no se ha conseguido con las expansiones y contracciones comentadas, será necesario reducir el tamaño del poliedro. Llamando $\boldsymbol{x}_m$ al punto cuyo valor de la función objetivo es mínimo, la reducción del poliedro se realiza moviendo el resto de los puntos hacia $\boldsymbol{x}_m$ la mitad de la distancia que los separa, es decir

$$\boldsymbol{x}_i^{(k+1)} = \frac{\boldsymbol{x}_i^{(k)} + \boldsymbol{x}_m^{(k)}}{2} \qquad \text{para } i=0..n,\ i\neq m \qquad \textbf{(5.17)}$$

Las condiciones que dan lugar a un movimiento de contracción evitan que se realice una reducción precipitada del poliedro. Pensando en un problema de dimensión 2 en las proximidades de un valle escarpado, esta condición deforma el triángulo en la dirección del valle, permitiendo una evolución rápida a lo largo del mismo. Si se reduce el tamaño del triángulo los avances se ralentizan.

5.3.5 Selección del método de optimización

En este apartado se han propuesto algunos de los métodos de optimización por búsqueda directa más utilizados. Todos ellos se consideran adecuados para realizar el ajuste del sistema de detección de fallos. Son algoritmos

sencillos, fáciles de entender y de programar, han sido aplicados con éxito en distintos tipos de aplicaciones y sin embargo resulta difícil establecer cuál es el mejor.

En la mayoría de los artículos donde se describen nuevos algoritmos de optimización o donde se proponen pequeñas modificaciones, suelen hacerse estudios comparativos con otros métodos sobre un problema concreto. Sin embargo hay que tener en cuenta que cada función objetivo tiene sus particularidades y un algoritmo puede tener un comportamiento espectacular en unos casos y nefasto en otros. Incluso ocurre, como se demuestra más adelante, que el mismo algoritmo sobre el mismo tipo de problema puede tener un comportamiento muy diferente sólo con variar ligeramente las condiciones iniciales (coordenadas del punto de partida). Algunos algoritmos de optimización tienen muchos parámetros internos que marcan decisivamente la eficacia del método; una pequeña variación de estos parámetros puede cambiar el comportamiento del método. Lo mismo ocurre con las condiciones de finalización, el número de iteraciones puede aumentar o incluso el método pude no converger por culpa de unas condiciones de finalización mal ajustadas. Este problema es especialmente importante en los métodos que aplican varias condiciones en serie, por ejemplo una búsqueda lineal dentro de un método multidimensional.

Cuando se van a resolver muchos problemas del mismo tipo puede compensar la selección de un método rápido y bien adaptado al problema. Esto supone realizar muchos ensayos con distintos algoritmos y variando los parámetros internos de cada uno. Pero en general es más importante la robustez general que la velocidad en un caso concreto. Para ilustrar esta idea se muestran a continuación los resultados de eficacia y robustez de diversos métodos de optimización aplicados al caso de minimizar la función de Rosenbrock [Gill81]:

$$FO(\boldsymbol{x}) = 100\left(x_2 - x_1^2\right)^2 + \left(1 - x_1\right)^2 \qquad \textbf{(5.18)}$$

Esta función bidimensional tiene el mínimo en el punto (1, 1), en un valle escarpado en forma curva, y suele utilizarse para comparar la velocidad de diversos métodos. La función es continua y derivable, lo que permite aplicar algoritmos de optimización que se basan en la información del gradiente.

La mayoría de los métodos de este tipo utilizados en la prueba han sido recogidos de [Muñoz96, sección 2.4]. La tabla 5.1 muestra los nombres y una breve descripción de los métodos, en las referencias indicadas pueden encontrarse explicaciones más detalladas.

Tabla 5.1: Métodos de optimización utilizados

`pol_rib`	Polak Ribiere [Bertsekas79]
`flet_ree`	Fletcher Reeves [Bertsekas79]
`lmqn0`	Quasi Newton de baja memoria [Gill81]
`lmqn1`	Quasi Newton de baja memoria [Luenberger84]
`scqn`	Quasi Newton autoescalado [Luenberger84]
`scqnr`	Quasi Newton autoescalado con restart [Luenberger84]
`dirconj`	Direcciones conjugadas
`newtsimp`	Newton simplificado
`stepdesc`	Descenso del gradiente
`hj1`	Hooke and Jeeves
`hj2`	Hooke and Jeeves
`razor`	Variación de HJ propuesta por Bandler y McDonald
`simplex`	Simplex con las variaciones de Nelder y Mead

En el caso del ajuste del sistema de detección de fallos, el tiempo empleado en la evaluación de la función objetivo es mucho mayor que el tiempo empleado por el algoritmo de optimización entre dos iteraciones. Por lo tanto la velocidad de cada método viene marcada por el número de evaluaciones de la función objetivo, que es fácil de medir. En general la función no es derivable y por lo tanto los métodos basados en la información del gradiente sólo pueden aplicarse calculando el gradiente aproximado. Como se ve más adelante, esto incrementa el número de evaluaciones de la función objetivo y pone estos métodos en desventaja frente a los métodos de búsqueda directa descritos en este capítulo.

Un buen punto inicial para probar la eficacia de los métodos sobre la función de Rosenbrock es el punto (-1,1); sin embargo no se ha realizado una única optimización por cada método sino que se han tomado como puntos de partida 441 puntos repartidos uniformemente en el intervalo $x \in [-1.2,-0.8]$, $y \in [0.8,1.2]$. A pesar de lo próximos que son los puntos iniciales, los resultados de la optimización son muy diferentes en algunos casos (por ejemplo el método `scqn` no converge en todos los casos).

En las siguientes tablas de resultados se recogen los valores mínimo, máximo y medio del número de iteraciones, y un factor de dispersión definido como el cociente de la varianza y de la media. Este factor de dispersión es un buen indicativo de la robustez del método, ya que un algoritmo que sea rápido generalmente pero muy lento en algunos casos tiene un factor de dispersión mayor que otro que tarde siempre lo mismo.

Tabla 5.2: Optimizaciones utilizando el valor del gradiente

Método	mínimo	máximo	media	dispersión
`pol_rib`	53	94	67.74	0.08
`lmqn1`	47	130	70.73	0.16
`lmqn0`	46	103	70.95	0.13
`scqnr`	98	119	109.79	0.04
`flet_ree`	117	391	203.90	0.27
`scqn`	60	no converge	409.44	3.66
`dirconj`	959	2003	1657.00	0.11
`stepdesc`	5362	7420	6850.69	0.05
`newtsimp`	8129	10368	9568.93	0.03

El método `scqn` aunque parecía muy bueno inicialmente, por ser el cuarto más rápido en el mejor de los casos, no converge en algunas situaciones. Otra observación interesante de la primera tabla de resultados es la comparación entre los métodos `flet_ree` y `scqnr`. En el mejor de los casos tienen un comportamiento parecido sin embargo el método `flet_ree` es

menos robusto, con un factor de dispersión casi 7 veces mayor, esto hace que sea mucho peor en algunos casos.

El problema más interesante en el ámbito de esta tesis es la optimización de funciones con derivada no conocida, ya que este es el caso del ajuste del sistema de detección de fallos (ver apartado 4.5.2 en página 69). En la tabla 5.3 se recogen los resultados de la optimización del mismo problema pero sin utilizar las fórmulas de las derivadas parciales de la función objetivo para obtener el valor del gradiente. En este caso los métodos de búsqueda directa parecen mejores, sin embargo los métodos basados en Hooke and Jeeves tiene un factor de dispersión alto lo que parece favorecer a los métodos robustos basados en el gradiente a pesar de estar utilizando un método aproximado para calcular las derivadas. Por ejemplo el método `hj1`, que resulta el tercero más rápido en el mejor de los casos pasa a ser octavo en tiempo medio. Hay que tener en cuenta que el cálculo aproximado del gradiente requiere muchas evaluaciones de la F.O. al aumentar la dimensión del problema, lo que seguramente deja en desventaja a los métodos basados en el mismo.

Tabla 5.3: Optimizaciones sin utilizar la fórmula del gradiente

Método	mínimo	máximo	media	dispersión
simplex	104	222	159.05	0.12
pol_rib	265	470	338.78	0.08
hj2	84	777	342.65	0.52
lmqn1	235	650	353.64	0.16
lmqn0	230	515	354.54	0.13
razor	170	1732	488.84	0.49
scqnr	490	595	548.97	0.04
hj1	169	1240	602.93	0.34
flet_ree	585	2125	1018.73	0.27
scqn	300	no converge	2019.21	3.70
dirconj	4795	10015	8285.22	0.11
newtsimp	17557	22440	20709.54	0.03
stepdesc	26810	37100	34253.31	0.05

El método simplex, descrito en el apartado 5.3.4, resulta en promedio el más rápido siendo además uno de los menos dispersos. Aparentemente es el mejor método, pero no hay que olvidar que las 441 optimizaciones de cada método se han realizado sobre una misma F.O. y por lo tanto no se pueden generalizar los resultados a otros casos. El objetivo fundamental de estos cálculos es precisamente demostrar que los resultados de ensayos concretos deben admitirse con ciertas reservas, y lo mismo ocurre con estos.

5.4 Descripción del problema de los mínimos locales y aportación de soluciones

Los métodos de optimización están orientados a la localización del mínimo de una función objetivo a partir de un punto inicial.

Las funciones **unimodales** son aquellas en las que el único punto que cumple las condiciones de mínimo es la solución del problema. En este tipo de funciones se puede demostrar la convergencia de los métodos de optimización basados en principios matemáticos desde cualquier punto del espacio de soluciones.

Las funciones **multimodales** son aquellas en las que existen varios puntos que cumplen las condiciones de mínimo, estos puntos se llaman mínimos locales. En el caso de funciones multimodales, sólo puede garantizarse la convergencia a uno de los mínimos locales y siempre que el punto de partida se encuentre dentro de la "zona de influencia" de dicho mínimo local.

Los métodos de búsqueda directa están basados en criterios empíricos y su convergencia no puede demostrarse matemáticamente, sin embargo son métodos que experimentalmente han presentado un buen comportamiento en regiones convexas del espacio.

Cuando las funciones son multimodales no existe ninguna manera de demostrar la convergencia a la solución del problema por ningún método de optimización, al menos utilizando aritmética convencional. Matemáticamente es posible demostrar la localización del mínimo global de una función, que cumpla determinadas condiciones de derivabilidad, mediante las técnicas de análisis por intervalos. El número 165 de la serie "*Pure and applied mathematics*" [Hansen92] trata el tema de la optimización desde el punto de vista del análisis por intervalos.

En la mayoría de los problemas de ingeniería los mínimos locales no suponen un problema importante ya que se trabaja con funciones objetivo lineales o se busca una solución en una región acotada del espacio donde se cumplen las condiciones de convergencia (soluciones en otras regiones del espacio pueden no tener sentido físico). Por esta razón y por la falta de procedimientos eficientes capaces de resolver el problema de optimización

multi-modal de forma sistemática, es un tema poco tratado en los libros de ingeniería o de optimización.

Cuando la función objetivo es lineal, no existe problema de mínimos locales, ya que estas funciones tienen forma de hiperplano. En algunos casos es posible recurrir a aproximaciones lineales de problemas reales para obtener una solución inicial factible y luego mejorarla aplicando técnicas de optimización no lineal sobre el problema original. Puesto que los métodos de optimización lineal suelen ser más rápidos que los de optimización no lineal, esta manera de proceder puede ahorrar tiempo de cálculo y evitar problemas de mínimos locales en muchos casos prácticos.

Hay que tener en cuenta que aunque la función objetivo sea lineal, si existen restricciones no lineales el problema puede presentar mínimos locales, como en el ejemplo que se muestra a continuación.

$$\begin{aligned} &\text{minimizar} && z = x + 5\,y \\ &\text{sujeto a} && \frac{1}{4}x^4 - 2\,x^3 + \frac{11}{2}x^2 - 6\,x + \frac{13}{4} - y \geq 0 \end{aligned} \qquad \textbf{(5.19)}$$

La figura 5.5 muestra un gráfico representativo del problema. Se han marcado con puntos negros (●) los dos mínimos locales, siendo el óptimo el correspondiente a $x=1$, $y=1$. En función del método de optimización utilizado y de las condiciones iniciales del cálculo (punto de partida, parámetros iniciales del algoritmo de optimización...) se puede obtener como resultado cualquiera de los dos puntos.

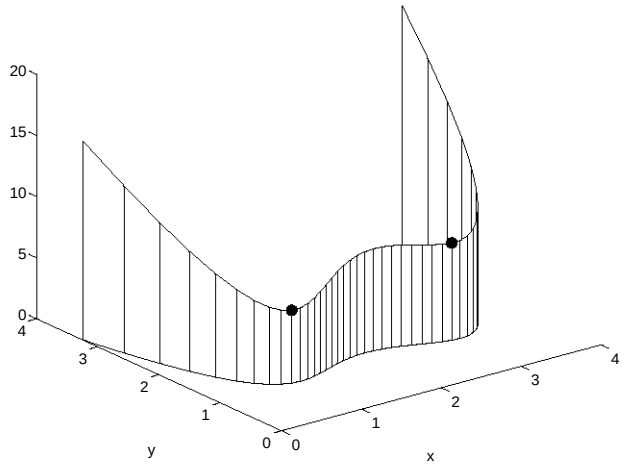


Figura 5.5: Optimización lineal con restricciones no lineales

En optimizaciones con función objetivo no lineal, tanto con restricciones como sin ellas, puede haber problemas de mínimos locales porque las funciones no lineales pueden ser multimodales. Sólo en el caso de funciones estrictamente convexas puede garantizarse la existencia de un solo punto que cumple las condiciones de mínimo [Polyak87, sección 1.3, teorema 3]. Por ejemplo el problema no lineal sin restricciones de minimizar $y=\text{sen}(x)$ tiene infinitas soluciones que son $x=\frac{3\pi}{2}+2k\pi$ con k entero. Cualquier método de búsqueda lineal de los descritos anteriormente localizará sólo una de estas soluciones o dará error de convergencia dependiendo del intervalo inicial que se defina.

Algunos problemas de optimización pueden tener funciones objetivo no derivables, con discontinuidades, con ruido, etc. Esto dificulta aún más el proceso de cálculo y la localización del óptimo.

5.4.1 Mínimos locales en el ajuste del sistema de detección de fallos

El problema de optimización que se plantea en esta tesis para ajustar el sistema de detección de fallos presenta problemas de mínimos locales. El problema aparece en la optimización multi–objetivo al intentar minimizar

una función objetivo global que tiene sumandos "contradictorios" como son el tiempo de detección y el número de falsas alarmas.

En la optimización multi–objetivo estamos minimizando la suma ponderada de dos funciones. Si estas funciones varían de forma suave no hay problemas de convergencia, pero basta con que tengan cambios bruscos de pendiente, o bien ondulaciones, para que la suma presente mínimos locales. La figura 5.6 muestra un ejemplo de optimización multi–objetivo de una variable donde una de las funciones objetivo disminuye al aumentar la variable y contrariamente la otra función objetivo aumenta. En línea de puntos se ha dibujado la suma de las dos funciones, que es lo que realmente se intenta minimizar. Se observa que debido a los cambios de pendiente de las dos funciones objetivo, la suma presenta mínimos locales (indicados mediante puntos ●).

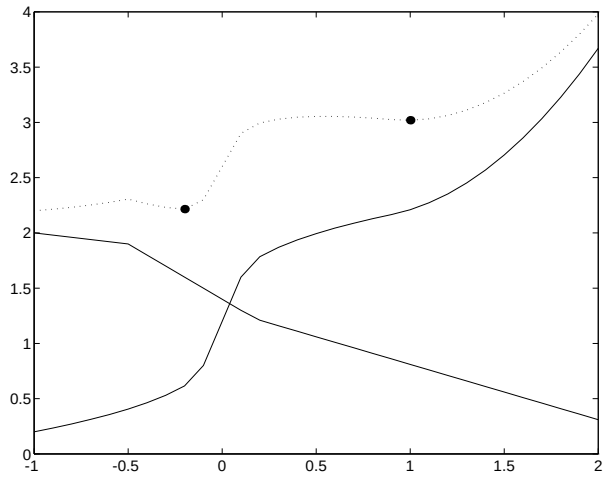


Figura 5.6: La suma de funciones objetivo puede dar lugar a mínimos locales

Durante el ajuste del sistema de detección de fallos se producen cambios bruscos en los atributos de detección por la propia estructura del problema. Estos cambios bruscos en los atributos de detección son la causa de la aparición de mínimos locales al plantear su minimización de forma conjunta.

Suponiendo el caso de un proceso industrial para el cual se desea adaptar un sencillo sistema de detección de fallos, basado en la comparación del nivel de los residuos con un umbral fijo. El sistema de detección avisará del fallo incipiente si el valor de los residuos supera el umbral, por lo tanto

el tiempo de detección viene dado por el tiempo que tardan los residuos en alcanzar el umbral después de comenzar la degradación del sistema. La figura 5.7 muestra dos gráficas, la primera es la evolución temporal de los residuos del sistema al final de su vida útil (datos simulados) y la segunda es la evolución del tiempo de detección en función del parámetro umbral que se fije en el sistema de detección de fallos. La detección tiene lugar en un instante que depende del valor del umbral. Si el umbral es pequeño puede observarse en la primera gráfica que la detección tendrá lugar muy pronto, por lo tanto en la segunda gráfica se obtienen tiempos de detección pequeños para umbrales pequeños y tiempos de detección grandes para umbrales grandes. Sin embargo, en determinados puntos se producen grandes cambios en el atributo "tiempo de detección" para pequeñas variaciones en el parámetro "umbral", lo que da lugar a una función de tipo escalonado.

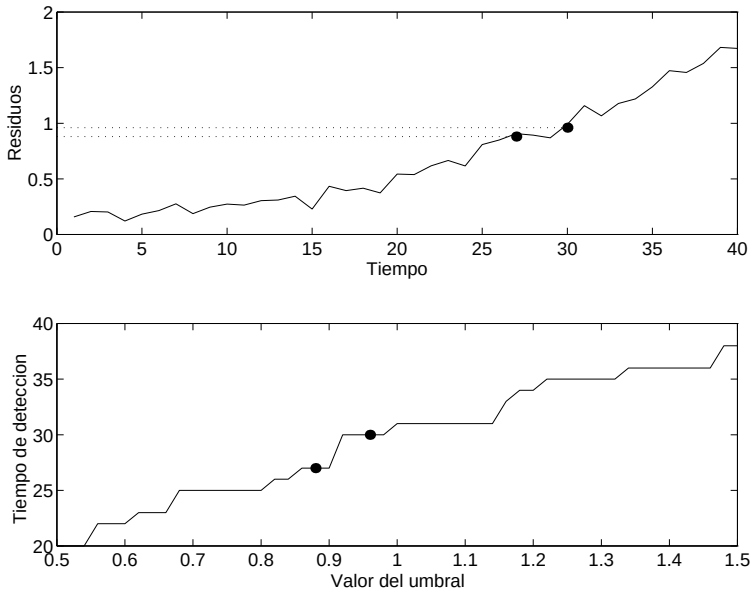


Figura 5.7: Evolución del tiempo de detección en función del umbral

Aunque la figura 5.7 muestra la sensibilidad de un atributo de detección frente a un solo parámetro y basándose en datos simulados, este comportamiento se ha observado en la práctica en distintos atributos de detección frente a la variación de cualquier parámetro. La forma escalonada de los atributos de detección y la necesidad de realizar una optimización

multi–objetivo en la que intervienen varios atributos de detección indican que el proceso de ajuste se verá afectado por problemas de mínimos locales.

Esta tesis no tiene por objeto el desarrollo de nuevos algorítmos de optimización, lo que requiere un trabajo largo y riguroso de comparación con otros métodos en robustez y velocidad de convergencia con distintos tipos de funciones y en distintas condiciones iniciales, etc. Sin embargo no se puede dejar sin proponer una manera práctica de resolver el ajuste de parámetros para llegar a obtener el sistema de detección de fallos óptimo. A continuación se dan unas orientaciones sobre la manera de plantear una optimización cuando se sospecha que el problema tiene mínimos locales en contexto semejante al de esta tesis.

5.4.2 Optimización multi-modal unidimensional

En los métodos de optimización unidimensional descritos anteriormente se suponían unas condiciones de convexidad de las funciones que no se dan en el caso de funciones multimodales en todo el espacio de solución.

El método más intuitivo para localizar el mínimo global, también aplicable en el caso de funciones multimodales, es realizar muchas evaluaciones de la función objetivo cambiando las variables mediante pequeños pasos; el menor valor que se obtenga corresponde al mínimo global[1]. Pero en la práctica este método no es aplicable ya que la precisión deseada en las variables da lugar a unos pasos muy pequeños y a un número de evaluaciones de la función objetivo muy grande. Además, incluso en el caso de funciones continuas sólo se garantiza la localización del mínimo global si el paso en las variables es menor que la zona de influencia del óptimo (figura 5.8, caso 1). Por otro lado, aunque el punto localizado tenga precisión suficiente en la variable, puede ocurrir que el valor obtenido de la función tenga mucho error (figura 5.8, caso 2).

[1] Se puede demostrar que el valor mínimo de la función objetivo de un conjunto de puntos distribuidos uniformemente en el espacio de soluciones tiende en términos de probabilidad al mínimo global [Polyak87, sección 6.2, teorema 2].

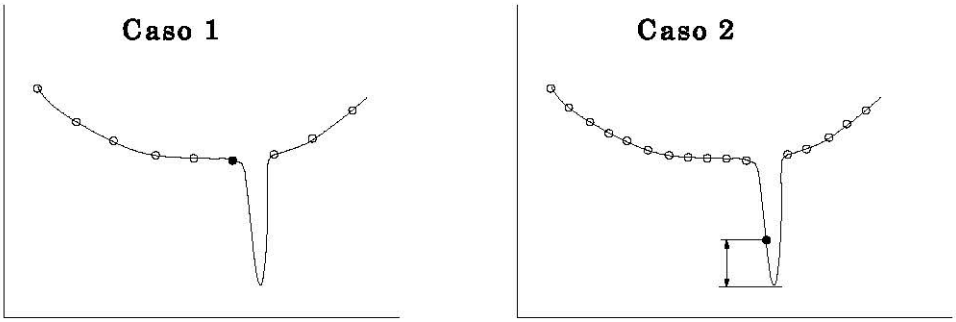


Figura 5.8: Problemas en la localización del mínimo

Por estas razones resulta más práctico en el caso de funciones continuas utilizar algoritmos combinados entre muestreo y optimización. En este tipo de algoritmos se realiza un muestreo para localizar las zonas en las cuales se encuentran los mínimos locales y posteriormente se lanza un método de optimización bajo la suposición de continuidad y convexidad de la función en las proximidades del mínimo local. Un ejemplo de este tipo de algoritmos para el caso unidimensional es el método *Ripple Search* descrito en [Bandler69] que utiliza una búsqueda de Fibonacci para obtener todos los mínimos y máximos de una función en un intervalo dado.

Para el ajuste del sistema de detección de fallos y en otros muchos problemas de optimización no interesa tanto la localización de todos los valores extremos sino la localización del mínimo global. Con esta idea se propone a continuación un método para obtener el mínimo global que pretende reducir el número de evaluaciones de la función objetivo.

En este método en lugar de realizar un muestreo de la función e inmediatamente lanzar optimizaciones en todos los puntos sospechosos de ser mínimos, se intenta obtener una confirmación sobre la posible existencia del mínimo mediante dos evaluaciones (como máximo) de la función. Después de realizar el muestreo, puede considerarse que existe un mínimo cuando un punto está rodeado por otros dos con valores superiores en la función objetivo (llamaremos a estos puntos mínimos de orden 1). Sin embargo este criterio no es suficiente para poder lanzar con seguridad un algoritmo de optimización porque no se garantiza la convexidad de la función. Cuando se quiere mantener un número bajo de evaluaciones se propone exigir que los dos puntos que rodean al máximo sean a su vez menores que los dos puntos siguientes (mínimos de orden 2). Las posibilidades de éxito en la optimización con esta condición son mayores,

aunque tampoco se garantiza la unicidad de solución en el intervalo ni la convergencia. De hecho pueden producirse fenómenos parecidos al *aliasing* en muestreo de señales como en el ejemplo de la figura 5.9.

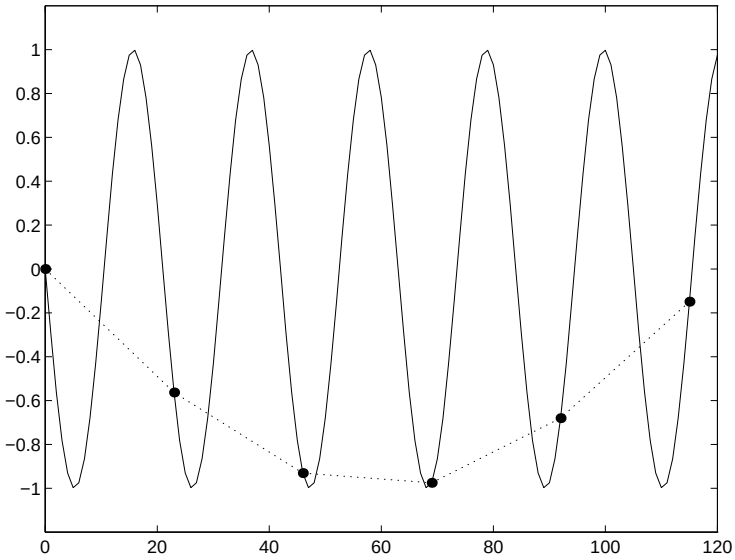


Figura 5.9: Efecto de *aliasing* al explorar la función

El método propuesto consiste en realizar un muestreo no muy fino de la función. Los mínimos de orden 2 quedan almacenados para su procesamiento posterior sin embargo en el caso de mínimos de orden 1 que no cumplan la segunda condición se realizarán evaluaciones por la derecha y por la izquierda. Esta nuevas evaluaciones normalmente confirmarán la localización en el mismo punto o en uno de los recién evaluados.

En la figura 5.10 se muestra un ejemplo del método de localización de mínimos. Inicialmente se realiza una exploración en 8 puntos de la función (primera gráfica) lo que da lugar a dos mínimos: los puntos 2 y 6 indicados mediante círculos ●. Los puntos 5 y 7 que rodean al segundo mínimo son a su vez menores que los puntos 4 y 8 respectivamente, lo que confirma la localización de un mínimo en el punto 6 (mínimo de orden 2). Sin embargo los puntos 1 y 3 no cumplen esta segunda condición, ya que el intervalo formado por los puntos 1, 2 y 3 no es convexo, esto puede hacer fallar el algoritmo de optimización. Por lo tanto se procede a subdividir los segmentos que rodean el primer mínimo local hasta conseguir un mínimo de orden 2 (tercera gráfica). En este caso son necesarias 4 nuevas

evaluaciones de la función, puede observarse que sólo en la zona del primer mínimo se realizan nuevas evaluaciones ya que el segundo cumple la condición de orden 2 desde el principio.

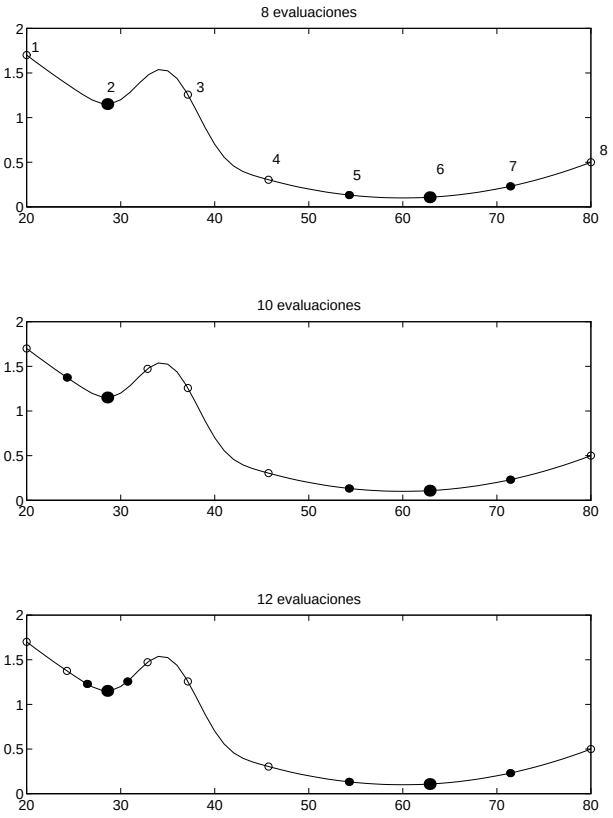


Figura 5.10: Método de localización de mínimos locales propuesto para funciones unidimensionales

Las ventajas fundamentales del método son las siguientes:

- Se obtiene mayor seguridad en la localización del mínimo con unas pocas evaluaciones adicionales (que en cualquier caso se iban a realizar dentro del algoritmo de optimización).
- Cuando sólo interesa localizar el mínimo global se pueden despreciar unas zonas frente a otras, que presenten valores mucho menores, en lugar de lanzar el algoritmo de optimización en cada una de ellas. En la suposición de que los 3 puntos que definen el

mínimo local se encuentran en una zona unimodal convexa, una buena estimación del valor del mínimo se puede obtener interpolando una parábola.

- Se ha comprobado experimentalmente que en caso de funciones oscilantes donde se ha tomado inicialmente un paso demasiado grande, no se localizan todos los mínimos locales. Sin embargo al aplicar sucesivamente la subdivisión alrededor de los primeros puntos, van apareciendo otros mínimos locales que habían pasado desapercibidos. Esto permite fijar un paso de muestreo bastante grande, lo que reduce el número de evaluaciones, con la tranquilidad de saber que el método realizará cierta adaptación automática al problema en el caso en que aparezca una función más oscilante. La figura 5.11 muestra la secuencia de subdivisiones que tienen lugar al muestrear la función $y = \frac{\text{sen}\left(\frac{2\pi}{19}\, x\right)}{\sqrt{x}}$ en el intervalo $x \in [20,60]$. Puede observarse que inicialmente sólo se detectan dos mínimos, pero las sucesivas subdivisiones que van teniendo lugar terminan localizando los tres mínimos locales que existen en este ejemplo.

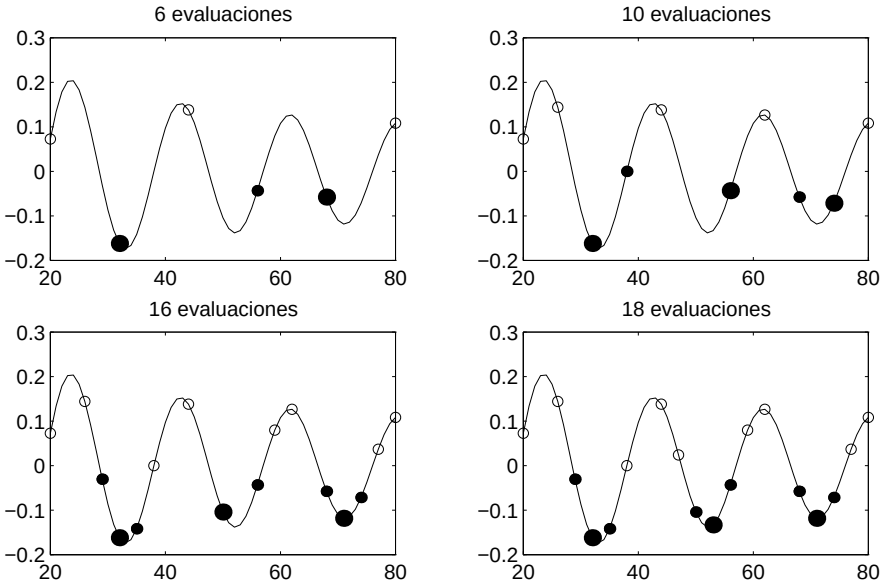


Figura 5.11: Secuencia de evolución del método de localización de mínimos locales propuesto para el caso unidimensional

En resumen, para problemas sencillos el paso inicial determina el número de evaluaciones de la función objetivo, pero cuando aparece una función más complicada donde el paso anterior resulta insuficiente, el método realiza automáticamente más evaluaciones.

5.4.3 Optimización multi-modal multidimensional

En el caso multidimensional se puede proceder de manera equivalente para localizar los mínimos locales: haciendo un muestreo de la función, obteniendo los puntos que indiquen la existencia de mínimos y lanzando algoritmos de optimización multidimensional.

El problema fundamental es que el número necesario de evaluaciones de la función objetivo es muchísimo mayor, especialmente si la dimensión del problema es grande. Pensando por ejemplo en 10 divisiones por variable, un problema bidimensional requiere 100 evaluaciones, uno tridimensional 1000, etc.

La localización de mínimos es sencilla de calcular, requiere comparar cada punto con los $2N$ puntos que lo rodean (siendo N la dimensión del espacio), pero este cálculo es rápido comparado con el tiempo de evaluación de la función en cada punto.

Como ejemplo se muestra a continuación una aplicación sobre la función `peaks` del programa Matlab [Mat1]. Esta función tiene la siguiente expresión matemática:

$$z = 3(1-x)^2 e^{-x^2-(y+1)^2} - 10\left(\frac{x}{5}-x^3-y^5\right) e^{-x^2-y^2} - \frac{1}{3} e^{-(x+1)^2-y^2} \quad \textbf{(5.20)}$$

Presenta tres máximos y tres mínimos en el intervalo [-3..3, -3..3] como puede verse en la gráfica de la figura 5.12.

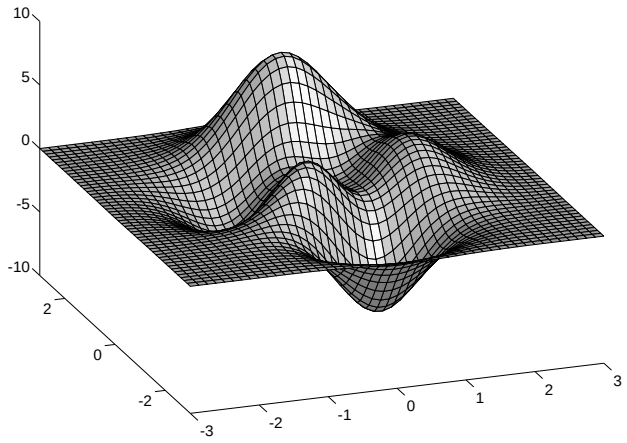


Figura 5.12: Función peaks

Para localizar todos los mínimos nos encontramos de nuevo con el problema de seleccionar un número de divisiones por variable suficientemente pequeño. En la figura 5.13 se ha dibujado una malla tridimensional de los puntos de la función `peaks` muestreados, destacando aquellos que cumplen la condición de mínimo local aplicada al caso bidimensional y su proyección sobre las curvas de nivel de la función original. Puede observarse que en el primer caso la aproximación es muy buena (comparando con la figura 5.12) y la localización de los tres mínimos locales no presenta ningún problema, pero el número de evaluaciones necesario es de 150. En el segundo caso se ha reducido el número de evaluaciones a 20 y como consecuencia sólo se localizan 2 de los 3 mínimos locales.

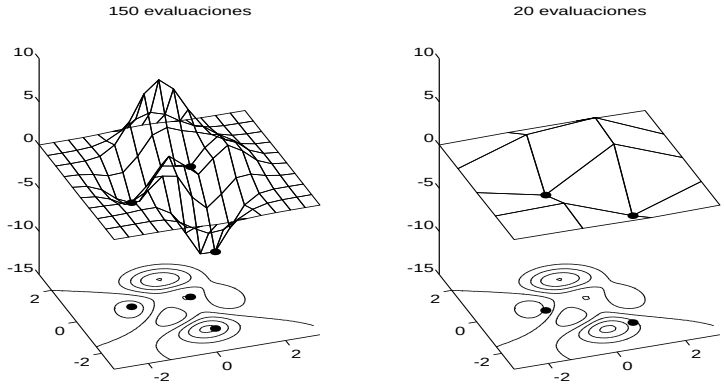


Figura 5.13: Localización de mínimos locales en la función peaks

En el ejemplo se han mostrado los mínimos locales que se han obtenido buscando aquellos puntos cuyos valores de la función objetivo son menores que los valores de los cuatro puntos que los rodean. En realidad, en el caso general multidimensional la comparación se realiza con los puntos anterior y posterior en la dirección de cada variable.

De manera análoga también es posible comprobar la segunda condición de mínimo descrita en el apartado anterior tomando dos puntos más en la dirección de cada variable. Sin embargo no es fácil definir un método automático de subdivisión de intervalos para el caso de N dimensiones. Un mecanismo automático de subdivisión tendría la ventaja de permitir seleccionar un paso inicial de exploración mayor, lo que da lugar a un número de evaluaciones de la F.O. menor. Esta sería la ventaja fundamental, ya que la ventaja de garantizar la convergencia del algoritmo de optimización, que también se apuntaba antes, ahora no tiene tanta importancia. Los métodos de optimización multidimensional de búsqueda directa pueden partir directamente de un punto del espacio, a diferencia de los métodos de búsqueda lineal que deben partir generalmente de un intervalo convexo.

Por último mencionar que algunos métodos de localización de mínimos locales se basan en únicamente en evaluaciones de la función sin necesidad de aplicar algoritmos de optimización. El método *Stuckman* por ejemplo se basa en información estadística mediante aproximaciones a funciones gaussianas, en el artículo [deCuadra90] se realizan comparaciones con otros

métodos. El número de evaluaciones de estos métodos es bastante grande o al menos comparable con un muestreo seguido de una optimización. La función `rcos`, que se muestra en la figura 5.14, tiene la siguiente expresión matemática:

$$z = a \cdot \left(y - b \cdot x^2 + c \cdot x - d\right)^2 - e \cdot (1 - f) \cdot \cos(x) - e; \qquad \textbf{(5.21)}$$

donde $a=1;\ b=\frac{5.1}{4\pi^2};\ c=\frac{5}{\pi};\ d=6;\ e=10;\ f=\frac{1}{8\pi};$

Esta función presenta 3 mínimos de valor z=−19.6 en el intervalo $x=[-1,\ 13]$, $y=[-2,\ 15]$, concretamente en los puntos: x1=(0.00, 6.00), x2=(6.28, 1.10) y x3=(12.57, 6.40).

La figura 5.14 muestra una imagen de la función `rcos` donde se han marcado los puntos que corresponden a los mínimos.

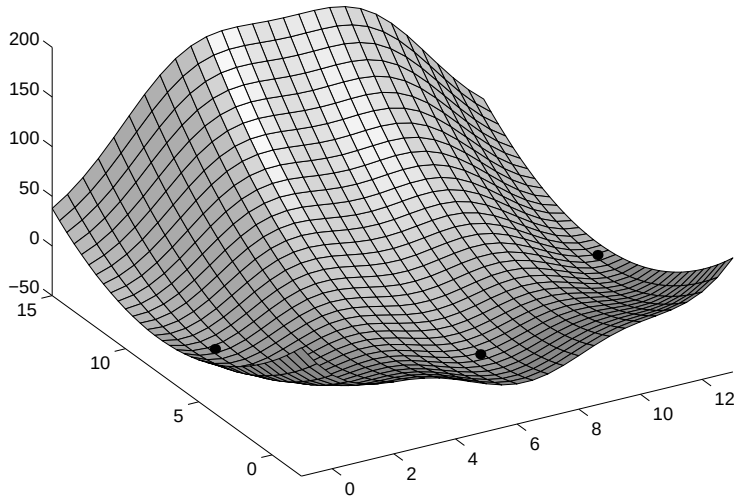


Figura 5.14: Función rcos

El número necesario de evaluaciones por el método de *Stuckman* es del orden de 500. Sin embargo realizando 6 divisiones en el eje x y 5 en el eje y pueden localizarse sin problemas los tres mínimos, tal y como se muestra en los gráficos de la figura 5.15. Una optimización posterior por el método simplex requiere menos de 60 evaluaciones en cualquiera de los casos y

finalmente se obtienen los valores de los tres mínimos con un error máximo de 0,0001 en x o y.

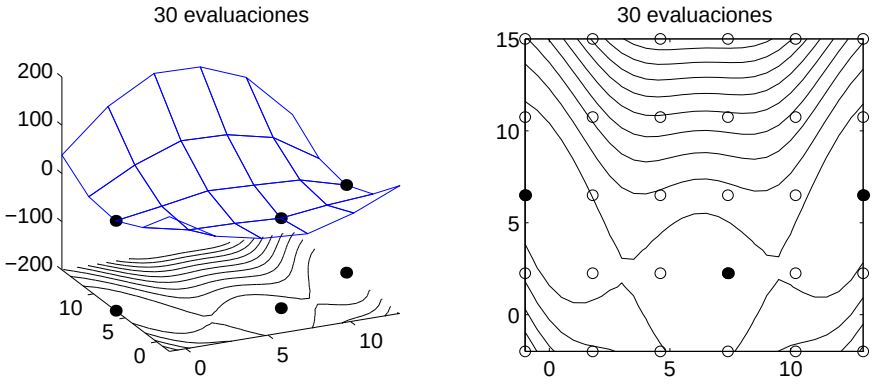


Figura 5.15: Proceso de localización de mínimos locales en la función rcos

En conclusión, el procedimiento de realizar evaluaciones en la función objetivo, comprobar las condiciones de mínimo local y lanzar optimizaciones en los puntos obtenidos, tiene una eficacia al menos comparable con otros métodos especialmente diseñados para obtener mínimos locales. En el caso del ajuste del sistema de detección de fallos, o en situaciones en las que sólo interese el valor del mínimo global, tiene también la ventaja de poder ignorar desde el primer momento aquellas zonas que tienen mínimos locales de valor muy superior al mínimo global.

5.5 Desarrollo de técnicas específicas para el ajuste del sistema de detección de fallos

En algunos problemas de optimización, como en el ajuste global del sistema de detección de fallos, ocurre que el tiempo de evaluación de la función objetivo es variable. En estos casos es conveniente tener en cuenta la estructura del problema concreto para organizar el algoritmo de optimización en forma de optimizaciones parciales [deCuadra90].

Un ejemplo de búsqueda por optimizaciones parciales es la extensión multivariable de métodos de búsqueda unidimensional, como el método *Generalized Fibonacci Search* [Murray72]. En el caso de dos variables el método consiste en realizar una optimización unidimensional en x_1, pero para cada valor de x_1 se calcula el valor de x_2 que minimiza la función. Es

decir, para cada valor x_1 se realiza una optimización unidimensional en x_2 (con x_1 fijo) y el resultado se considera el valor de la F.O. en x_1.

En la figura 5.16 se muestra el estado de la optimización en la primera iteración de x_1 de una función en el intervalo $x_1 \in [0,1.5]$, $x_2 \in [-0.5,2.5]$. En este ejemplo se ha utilizado el método *Golden section* y por lo tanto se parte de 4 puntos x_1 distribuidos en el intervalo de partida según la ecuación 5.1 (ver apartado 5.2.1 en página 83). En el gráfico se han dibujado todos los puntos en los que se ha evaluado la función objetivo, remarcando aquellos que son mínimos tras la optimización unidimensional en x_2 con x_1 fijo. Teniendo en cuenta que a cada punto x_1 se le asigna el valor remarcado, el paso siguiente en la optimización sería seleccionar el intervalo $x_1 \in [0.57, 1.5]$ ya que el mínimo valor de la función objetivo corresponde a $x_1=0.93$. Continuando con el proceso habría que realizar una optimización unidimensional para $x_2 \in [-0.5,2.5]$ con $x_1=1.15$ (que se calcula aplicando la ecuación 5.1 al nuevo intervalo seleccionado).

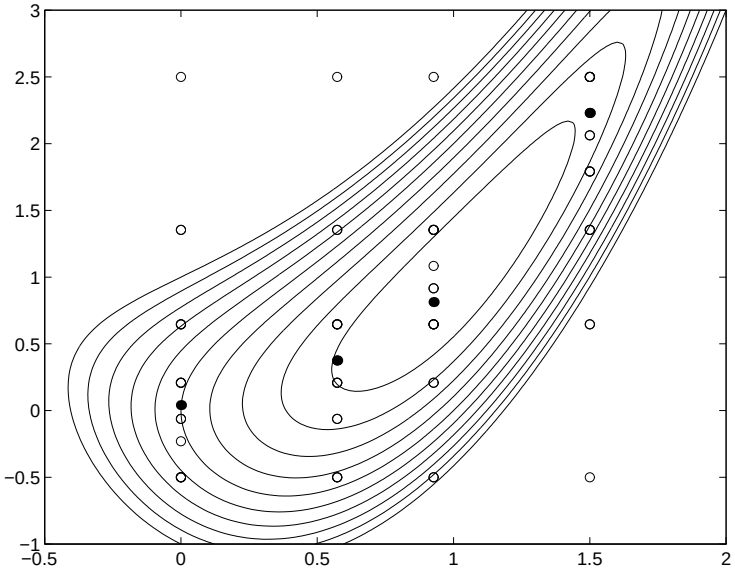


Figura 5.16: Ejemplo de optimización por el método de Fibonacci generalizado

Si las evaluaciones de la F.O. para distintos valores de x_2 son rápidas suponiendo que no cambia x_1, entonces este método puede resultar más eficaz que realizar una optimización multidimensional en la que tanto x_1 como x_2 cambian en cada iteración. Aunque el número de evaluaciones de la F.O. en el caso de optimizaciones parciales sea generalmente mayor, si

estas evaluaciones son en promedio más rápidas el tiempo total de la optimización puede ser menor.

5.5.1 Ajuste por optimizaciones parciales

Veamos el caso de ajuste global del sistema de detección de fallos. En este sistema, cuyo diagrama de bloques aparece en la figura 5.17, existen tres grupos de parámetros: los parámetros del modelo (denominados α, β, γ ...), los parámetros de los atributos de fallo (denominados a, b, c...) y los parámetros de la función de detección de fallos (denominados A, B, C...).

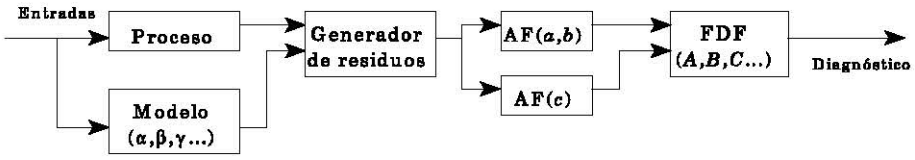


Figura 5.17: Esquema del sistema de detección de fallos

Conocidos los parámetros de modelo es posible calcular la evolución de los residuos para todos los conjuntos de entrenamiento. Este cálculo supone evaluar el modelo tantas veces como el número de muestras de todas las historias de fallo. Lo mismo ocurre con los parámetros a, b, c... que permiten calcular los atributos de fallo para cada curva de evolución de los residuos. Finalmente la función de detección de fallos (FDF) es una combinación lógica de los valores instantáneos de los atributos de fallo (AF) que se obtiene rápidamente.

Durante el proceso de optimización, la variación de un parámetro del segundo grupo sólo requiere recalcular el atributo de fallo al que afecta y la FDF para obtener los nuevos diagnósticos. Mientras que la variación de uno de los parámetros del modelo obliga a recalcular todos los residuos y todos los AF, lo que requiere un tiempo de cálculo mucho mayor. En consecuencia, un ajuste mediante optimizaciones parciales es indicado para el caso que estamos tratando.

Como la variación en cualquiera de los parámetros del modelo obliga a recalcular todo el sistema, estos parámetros deben resolverse mediante algoritmos multidimensionales. Sin embargo los parámetros de los AF son

bastante independientes, ya que modificar uno de ellos no requiere recalcular todos los AF, esto permite realizar optimizaciones unidimensionales para cada parámetro.

5.5.2 Eliminación de mínimos locales mediante filtrado de los residuos

Algunos problemas de mínimos locales se producen por ruido en la curva de residuos, como se mostraba en la figura 5.7. Este ruido da lugar a pequeños mínimos locales que tienen poca influencia sobre la forma de la F.O. a gran escala pero dificultan el proceso de optimización.

Para aliviar el problema se propone suavizar la curva de residuos mediante filtrado. Este proceso de filtrado tiene la ventaja de eliminar muchos de los problemas de mínimos locales que pueden "atascar" los algoritmos de optimización como puede comprobarse más adelante. El esquema que se sigue para ajustar el sistema de detección de fallos es el que se muestra en la figura siguiente.

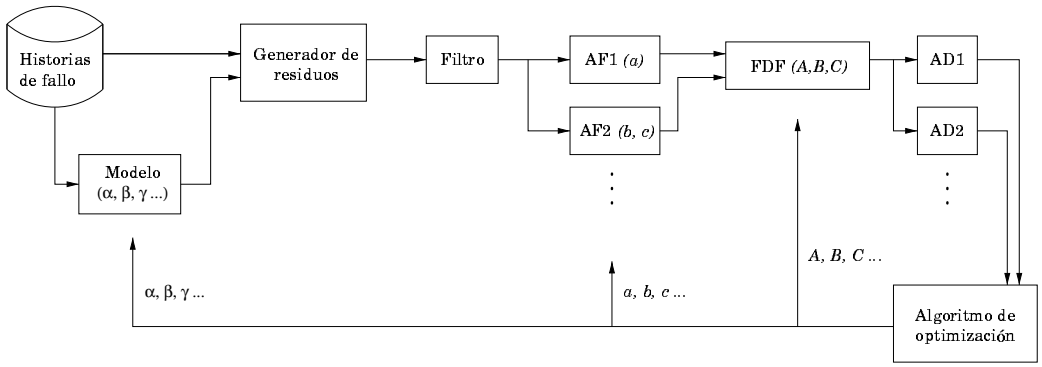


Figura 5.18: Esquema general del procedimiento de ajuste con filtrado

En este esquema se utiliza el mismo sistema de detección que se utilizará en tiempo real, pero con la incorporación de un filtro que afecta a los residuos. Cada historia de fallo contiene los datos de entrada y de salida del sistema durante un ciclo de funcionamiento en el que el sistema empieza funcionando correctamente y se degrada hasta una situación de fallo.

El procedimiento de ajuste con filtrado calcula los residuos de todos los puntos de cada historia de fallo y luego aplica un filtro. Este filtro suaviza

la evolución dinámica de los residuos, atenuando el problema de los mínimos locales descrito en el apartado 5.4.1 (página 107). A partir de las curvas de residuos suavizadas, se calculan los atributos de detección y la función de detección de fallos de la misma manera que en el ajuste normal (figura 5.18). Una vez conocidos los instantes en que el sistema de detección avisa de la existencia de fallos, pueden calcularse los atributos de detección, que combinados forman la función objetivo que se quiere minimizar. Se ha comprobado que esta función objetivo presenta menos problemas de mínimos locales que la función objetivo que se obtiene sin aplicar el filtrado.

La figura 5.19 muestra una comparación entre la F.O. sin suavizar y las F.O. que se obtienen aplicando filtros de distinto tamaño. Estos resultados se han obtenido para el caso del sistema de detección de descarga de baterías, calculando el valor de la función objetivo (FO= tiempo de detección + nº de falsas alarmas) para distintos valores de los parámetros del modelo. La primera columna muestra la forma de la F.O. para los parámetros $V \in [1, 1.5]$ y $R \in [-0.5, 0]$; estas gráficas parecen suaves y sin mínimos locales, con el óptimo situado en un valle. Sin embargo diferentes optimizaciones lanzadas desde varios puntos no convergen hacia el valle sino que quedan atrapadas en pequeños mínimos locales que se forman a un nivel más fino. La segunda columna de gráficas es una ampliación de la primera correspondiente a los parámetros $V \in [1.25, 1.35]$ y $R \in [-0.35, -0.25]$ donde se han dibujado los puntos finales de la optimización obtenidos por distintos métodos y partiendo de distintos puntos. Puede observarse que no siempre se alcanza el valle solución, sino que a veces la optimización parece terminar en zonas planas. La tercera columna de gráficas desvela mediante una ampliación mayor la existencia de una estructura irregular en estas zonas.

Las gráficas han sido obtenidas aplicando distintos grados de filtrado. La primera fila corresponde al ajuste sin filtrado, en este caso se aprecia la existencia de irregularidades superficiales que producen muchos fallos de convergencia. Las otras tres filas de gráficas corresponden a filtrados de media movil utilizando ventanas de 10, 20 y 30 muestras respectivamente. Puede observarse como el filtrado de los residuos suaviza la forma de la función objetivo y por lo tanto el éxito de los algoritmos de optimización es mayor.

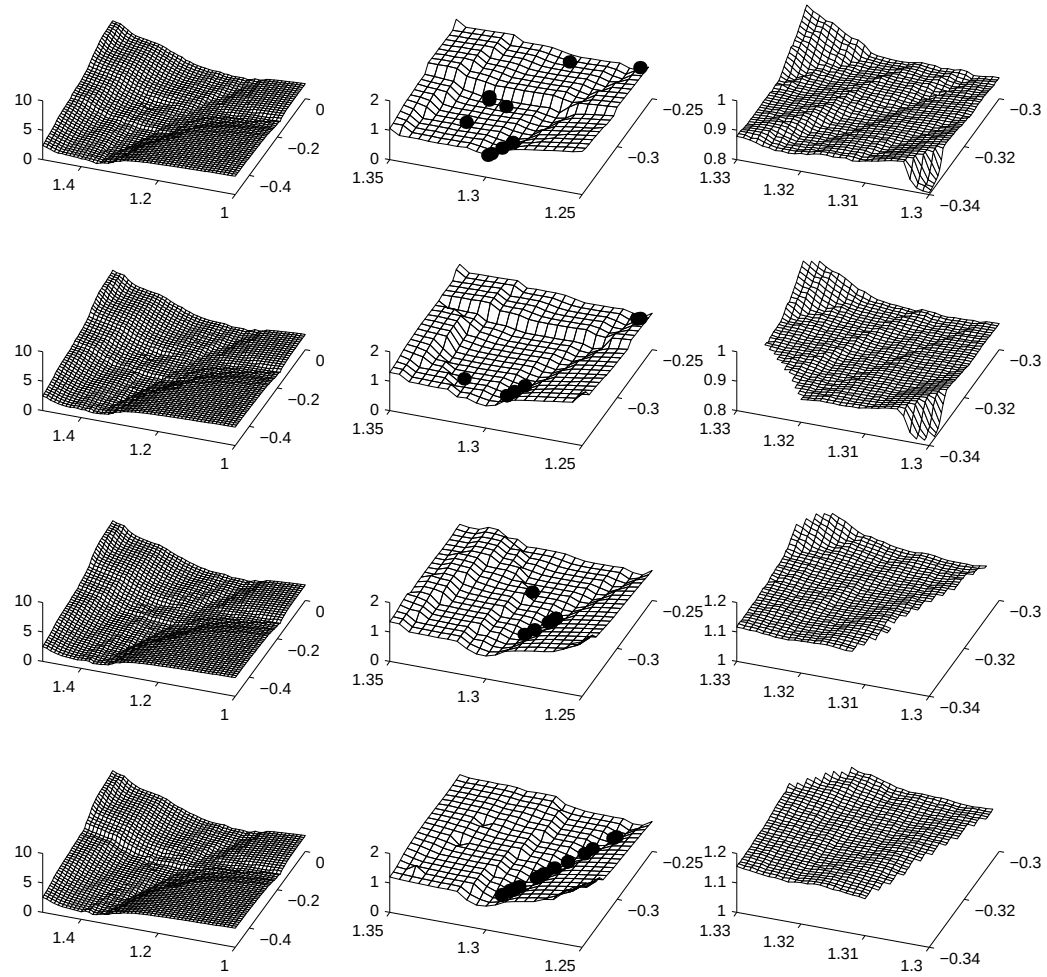


Figura 5.19: Efecto de filtrado de los residuos durante el proceso de ajuste

Sin embargo hay que tener en cuenta que el proceso de filtrado también puede enmascarar problemas reales del ajuste del modelo que en situaciones normales darían lugar a falsas alarmas o podrían afectar al tiempo de detección, retrasando o adelantando los diagnósticos. De hecho puede observarse en las gráficas que la forma de la función objetivo se ve alterada, sobre todo en valor absoluto. Es decir, el filtrado de los residuos no es equivalente a un filtrado de la función objetivo, que sería lo deseable, y por lo tanto una vez localizado el óptimo por esta técnica, será necesario volver a lanzar el proceso de ajuste original; aunque en una región del espacio más reducida.

Capítulo 6

Detección de fallos mediante estimación de parámetros

6.1 Introducción

El objetivo de este capítulo es analizar en profundidad los sistemas de detección de fallos basados en el seguimiento de los parámetros y comparar su funcionamiento con los sistemas de detección basados en el análisis de residuos.

En el capítulo 2 se han explicado en rasgos generales los fundamentos de los dos métodos de detección. Posteriormente, en el apartado ? se ha explicado cómo se aplica la estructura del sistema de detección incipiente de fallos en cada caso. Este capítulo proporciona mayor detalle sobre la manera de implantar un sistema de detección basado en el seguimiento de parámetros. Concretamente se describe la manera de implantar algoritmos

de estimación recursiva de parámetros para su utilización en detección incipiente de fallos.

Otro de los objetivos fundamentales de este capítulo es mostrar las limitaciones que pueden sufrir los algoritmos de estimación recursiva de parámetros cuando se aplican a la detección incipiente de fallos. Es importante analizar si en el proceso que se quiere monitorizar se producen estas situaciones, con el fin de decidir con mayor criterio qué método de detección se debe aplicar.

También se compara el funcionamiento de estos sistemas de detección con aquéllos basados en el análisis de los residuos. Utilizando un ejemplo sencillo se muestra que ambas técnicas son válidas para detección incipiente de fallos, aunque se hace especial énfasis en mostrar con claridad las limitaciones del análisis basado en la evolución de los parámetros.

6.2 Detección de fallos mediante estimación de parámetros

Las técnicas de estimación de parámetros se pueden comportar como un mecanismo indirecto de medida de las características físicas de un equipo industrial. Cuando se conocen las ecuaciones matemáticas que describen el comportamiento del equipo como función de una serie de características, se tiene en realidad un modelo matemático (basado en fundamentos físicos) que depende de unos parámetros desconocidos. Haciendo trabajar el equipo en diferentes condiciones y midiendo las variables que aparecen en el modelo, se pueden obtener los valores de los parámetros aplicando técnicas de estimación de parámetros. Estos parámetros corresponden a características de los componentes y por lo tanto el proceso de identificación de estos parámetros equivale en la práctica a medir las características que representan.

Una ventaja adicional es que se obtienen los valores de las características de los componentes montados y funcionando en condiciones reales. En algunos casos estos valores pueden diferir de aquellos obtenidos al ensayar los componentes de manera independiente. La razón fundamental que justifica esta diferencia es la dificultad de reproducir en

el banco de ensayos las mismas condiciones de funcionamiento y de montaje que en el sistema real.

Los procedimientos de estimación de parámetros, entendidos como métodos indirectos de medir las características físicas de los componentes, son aplicables para detectar fallos en equipos industriales. La estimación de parámetros se ha aplicado con éxito como técnica de control de calidad, por ejemplo para detectar fallos de fabricación en motores eléctricos [Filbert92]. En este caso el método consiste en obtener las características de los componentes y compararlas con los valores de diseño.

Como técnica de detección incipiente de fallos puede aplicarse el procedimiento cada cierto tiempo y comprobar si todos los parámetros siguen dentro de los márgenes establecidos. El mayor inconveniente es que las pruebas necesarias para encontrar los valores de los parámetros pueden interferir con el funcionamiento programado para el equipo, disminuyendo su capacidad de producción. La solución es utilizar técnicas de estimación de parámetros en continuo, también conocidas como técnicas de estimación recursiva de parámetros [Wellstead91]. Mediante estas técnicas se puede diseñar un sistema de monitorización y detección de fallos que trabaje en continuo y sin interferir con el funcionamiento del equipo.

En realidad la estimación recursiva de parámetros no requiere que el modelo esté basado en fundamentos físicos y su aplicación como herramienta de detección de fallos puede extenderse a otros tipos de modelo. Cuando se utilizan modelos físicos, los parámetros están asociados a características de los componentes y es fácil saber si el componente se ha deteriorado a partir del valor del parámetro. Por el contrario, en los modelos de caja negra, los parámetros no tienen interpretación física y por lo tanto un valor aislado de los mismos no aporta la información necesaria para establecer un diagnóstico definitivo. Sin embargo los algoritmos de estimación recursiva proporcionan los valores de los parámetros en cada instante de muestreo, lo que permite realizar una análisis de la evolución de los mismos como si se tratara de los residuos o de cualquier otra señal del sistema. Las variaciones en la evolución de los parámetros, tanto si tienen interpretación física como si no, son un indicativo de fallo o de degradación del equipo monitorizado.

Las principales desventajas de utilizar modelos que no estén basados en fundamentos físicos son una mayor dificultad en el módulo de detección y poca información útil para el diagnóstico. El módulo de detección es más complicado porque los parámetros no tienen interpretación física por sí mismos y por lo tanto es necesario comparar con los valores del pasado y decidir si cierto nivel de variación (de unidades desconocidas) debe considerarse peligroso. Sin embargo este problema se soluciona aplicando la estructura general del sistema de detección incipiente de fallos y la metodología de ajuste que se propone en esta tesis, ya que los umbrales de cada parámetro quedarán definidos de forma automática teniendo en cuenta todas las historias de fallo disponibles.

El diagnóstico con modelos de caja negra sólo es posible a nivel de equipo completo, ya que el fallo de componentes diferentes puede manifestarse en variaciones de un mismo subconjunto de parámetros. Por el contrario, en modelos basados en fundamentos físicos, cada parámetro representa una característica de un componente y la identificación del fallo es inmediata. Por ejemplo, en el modelo de un sistema de suspensión, donde se observa una degradación del parámetro que representa el amortiguamiento, está claro que la causa del fallo es un envejecimiento del amortiguador.

La estimación de parámetros en continuo puede tener mayor o menor complejidad en función del tipo de modelo que se utilice (lineal o no-lineal) y en función del criterio de ajuste que se aplique (mínima suma de errores cuadráticos, mínima suma de errores absolutos, mínimo error máximo, etc). En el caso de modelos lineales y de ajuste basado en mínimos cuadrados, el cálculo se puede hacer de una manera sencilla. En los dos apartados siguientes se explican el algoritmo de ajuste por mínimos cuadrados en su forma tradicional y en forma recursiva [Wellstead91].

6.2.1 Estimación de parámetros en continuo

Para poder ajustar los parámetros de un modelo hace falta como mínimo tantos datos como número de parámetros. Sin embargo los datos normalmente se obtienen de manera experimental y están afectados de un cierto nivel de ruido. Con objeto de evitar los efectos del ruido sobre el valor de los parámetros interesa utilizar un volumen de datos mayor que el número de parámetros a estimar. La estimación de parámetros en continuo

permite ir reduciendo los efectos del ruido a medida que se dispone de más información sobre el comportamiento de un proceso.

Suponiendo que un proceso tiene un comportamiento lineal, este proceso se puede expresar matemáticamente de la siguiente manera:

$$y(t) = \boldsymbol{x}^{\mathrm{T}}(t) \cdot \theta \tag{6.1}$$

donde $y(t)$ es el valor de la salida en el instante t, $\boldsymbol{x}(t)$ es un vector de dimensión np (número de parámetros) que contiene todas las entradas y θ es un vector con dimensión np que contiene las características del proceso.

El siguiente modelo obtiene una estimación de la variable de salida $y(t)$, a la cual se denominará $\bar{y}(t)$, a partir de los valores reales $\boldsymbol{x}(t)$ y de una estimación de los parámetros, que se denominará $\bar{\theta}$.

$$\bar{y}(t) = \boldsymbol{x}^{\mathrm{T}}(t) \cdot \bar{\theta} \tag{6.2}$$

El vector $\boldsymbol{x}(t)$ no se ha considerado un entrada única porque de esta manera no se pierde generalidad en la explicación, ya que el vector puede estar formado por los valores de múltiples señales y variaciones de éstas, por ejemplo:

$$\boldsymbol{x}^{\mathrm{T}}(t) = [\ 1 \quad x(t) \quad x^2(t) \quad x(t-1) \quad z(t) \quad y(t-1) \quad \cdots\ ] \tag{6.3}$$

Cada uno de los elementos de este vector se multiplica por uno de los parámetros del modelo. El comportamiento del modelo, en paralelo con el comportamiento real del proceso, puede representarse gráficamente de la siguiente manera:

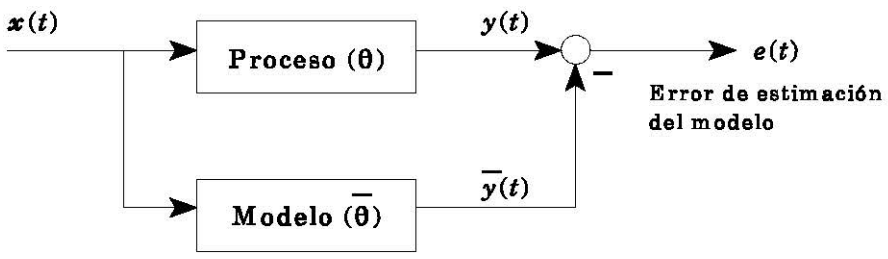


Figura 6.1: Esquema del ajuste del modelo

En cada instante de muestreo se cumple la siguiente relación, que tiene en cuenta el error de estimación del modelo:

$$y(t) = \boldsymbol{x}^{\mathrm{T}}(t) \cdot \bar{\theta} + e(t) \tag{6.4}$$

Considerando un conjunto de datos de funcionamiento, para cada instante se tiene un valor real de la salida, un vector de valores de entrada y un error de estimación. La ecuación 6.4 puede expresarse de forma matricial como:

$$\boldsymbol{y}(n) = X(n) \cdot \bar{\theta} + \boldsymbol{e}(n) \tag{6.5}$$

donde $\boldsymbol{y}(n)$ es un vector de dimensión n (número de muestras) que contiene los valores de la salida, $X(n)$ es una matriz de dimensión $n \times np$ formada por la acumulación de n vectores $\boldsymbol{x}^{\mathrm{T}}(t)$, $\bar{\theta}$ es el vector de parámetros del modelo y $\boldsymbol{e}(t)$ es un vector de dimensión n con los valores instantáneos del error de estimación. Hay que destacar la diferencia entre la notación de la ecuaciones 6.4 y 6.5. El valor de la salida en el instante t viene dado por $y(t)$ en la ecuación 6.4, mientras que el vector $\boldsymbol{y}(n)$ de la ecuación 6.5 contiene n valores de la señal de salida.

$$\boldsymbol{y}(n) = \begin{Bmatrix} y(1) \\ y(2) \\ \vdots \\ y(n) \end{Bmatrix}; \quad X(n) = \begin{bmatrix} \boldsymbol{x}^{\mathrm{T}}(1) \\ \boldsymbol{x}^{\mathrm{T}}(2) \\ \vdots \\ \boldsymbol{x}^{\mathrm{T}}(n) \end{bmatrix}; \quad \bar{\theta} = \begin{Bmatrix} \bar{\theta}_1 \\ \bar{\theta}_2 \\ \vdots \\ \bar{\theta}_{np} \end{Bmatrix}; \quad \boldsymbol{e}(n) = \begin{Bmatrix} e(1) \\ e(2) \\ \vdots \\ e(n) \end{Bmatrix} \tag{6.6}$$

El método de ajuste por mínimos cuadrados obtiene los parámetros que hacen que la suma de los cuadrados de los elementos del vector $\boldsymbol{e}(n)$ sea mínima. Estos parámetros se obtienen directamente mediante la siguiente ecuación [Golub89] [Press92]:

$$\bar{\theta} = \left[X^{\mathrm{T}}(n) \cdot X(n)\right]^{-1} \cdot \left\{X^{\mathrm{T}}(n) \cdot \boldsymbol{y}(n)\right\} \tag{6.7}$$

Para resolver esta ecuación es necesario calcular la inversa de una matriz de dimensión $np \times np$. Sin embargo esta matriz se obtiene a partir del producto de dos matrices cuyo tamaño depende del número de datos considerado. En un sistema de muestreo continuo es interesante poder ir

actualizando el valor de los parámetros a medida que se dispone de más datos del proceso. De esta manera van disminuyendo los efectos del ruido y el valor de los parámetros mejora progresivamente. La figura 6.2 muestra las curvas de evolución de los parámetros con el número de muestras en un proceso en el existe ruido asociado a las medidas.

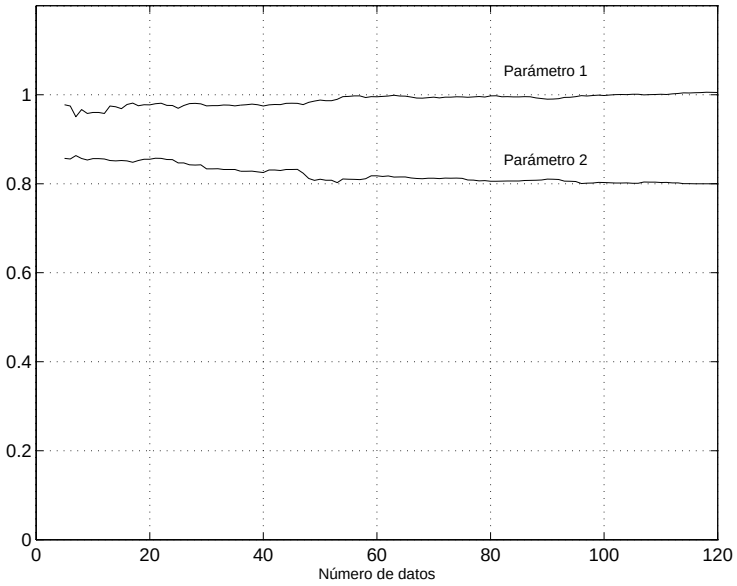


Figura 6.2: Ejemplo de estimación de parámetros en continuo

Esta gráfica ha sido generada suponiendo un proceso lineal en el cual la salida y puede calcularse a partir de la entrada x mediante la ecuación $y=1+0.8x$. Se trata por lo tanto de un proceso con dos parámetros (1 y 0.8), los vectores $\boldsymbol{x}^{\mathrm{T}}$ y θ según la ecuación 6.1 son los siguientes:

$$\boldsymbol{x}^{\mathrm{T}} = \{\, 1 \quad x \,\}\,; \qquad \theta = \begin{bmatrix} 1 \\ 0.8 \end{bmatrix} \tag{6.8}$$

Se ha construido la matriz X a partir de una serie de números aleatorios y se han calculado los valores de la salida (que van formando el vector $\boldsymbol{y}(n)$); finalmente se ha sumado un ruido aleatorio a cada medida. Aplicando sucesivamente la ecuación 6.7 se va obteniendo la estimación del valor de los parámetros en cada instante. Inicialmente esta estimación no se aproxima mucho a los valores reales, pero al aumentar el número de datos la estimación mejora progresivamente.

En cada iteración se considera una situación más que en la iteración anterior, por lo tanto la matriz X crece en una fila y el vector $\boldsymbol{y}$ crece en un elemento. El cálculo de los parámetros a partir de la ecuación 6.7 no resulta muy apropiado por las razones siguientes:

- Requiere el cálculo de la inversa de una matriz de dimensión igual al número de parámetros en cada iteración. El resultado obtenido en la iteración anterior no aporta ninguna información útil para el cálculo, que debe repetirse desde el principio.
- Al crecer el número de muestras, los valores de la matriz $X^{T}X$ van creciendo hacia infinito, lo que conlleva problemas computacionales en el cálculo de la inversa.
- Para aplicar la ecuación directamente es necesario almacenar todos los datos de entrada y salida que se han obtenido[1].

A continuación se muestra un algoritmo que permite obtener el mismo resultado pero realizando los cálculos de una manera más efectiva.

6.2.2 Algoritmo de estimación recursiva por mínimos cuadrados

La estimación recursiva permite obtener el valor de los parámetros de la iteración n mediante una actualización de los parámetros de la iteración $n-1$. En lugar de realizar el cálculo con los datos de las n iteraciones, el método opera con las medidas más recientes para corregir el valor actual de los parámetros [Wellstead91] [Hsia77].

Para poder expresar el cálculo de $\hat{\theta}(n)$ en función de los datos del instante de muestreo anterior hace falta expresar $\boldsymbol{y}(n)$ y $X(n)$ en función de $\boldsymbol{y}(n-1)$ y de $X(n-1)$ respectivamente:

[1] En realidad esto no es imprescindible ya que se pueden actualizar la matriz $[X^{T}(n)X(n)]$ y el vector $\{\boldsymbol{X}^{T}(n)\boldsymbol{y}(n)\}$ para obtener $[X^{T}(n+1)X(n+1)]$ y $\{\boldsymbol{X}^{T}(n+1)\boldsymbol{y}(n+1)\}$.

$$X(n) = \begin{bmatrix} \boldsymbol{x}^{\mathrm{T}}(1) \\ \boldsymbol{x}^{\mathrm{T}}(2) \\ \vdots \\ \boldsymbol{x}^{\mathrm{T}}(n-1) \\ \boldsymbol{x}^{\mathrm{T}}(n) \end{bmatrix} = \begin{bmatrix} [X(n-1)] \\ \cdots\cdots \\ \{\boldsymbol{x}^{\mathrm{T}}(n)\} \end{bmatrix}; \quad \boldsymbol{y}(n) = \begin{Bmatrix} y(1) \\ y(2) \\ \vdots \\ y(n-1) \\ y(n) \end{Bmatrix} = \begin{Bmatrix} \{\boldsymbol{y}(n-1)\} \\ \cdots\cdots \\ y(n) \end{Bmatrix} \quad \textbf{(6.9)}$$

Los dos términos de la ecuación 6.7 (la matriz y el vector de la parte derecha) pueden escribirse en función de $n-1$ de la siguiente manera:

$$X^{\mathrm{T}}(n)\cdot X(n) = \left[\, [X^{\mathrm{T}}(n-1)] \;\vdots\; \{\boldsymbol{x}(n)\} \,\right] \cdot \begin{bmatrix} [X(n-1)] \\ \cdots\cdots \\ \boldsymbol{x}^{\mathrm{T}}(n) \end{bmatrix} \quad \textbf{(6.10)}$$

$$= [X^{\mathrm{T}}(n-1)\cdot X(n-1)] + [\boldsymbol{x}(n)\cdot \boldsymbol{x}^{\mathrm{T}}(n)]$$

$$X^{\mathrm{T}}(n)\cdot \boldsymbol{y}(n) = \left[\, [X^{\mathrm{T}}(n-1)] \;\vdots\; \{\boldsymbol{x}(n)\} \,\right] \cdot \begin{Bmatrix} \{\boldsymbol{y}(n-1)\} \\ \cdots\cdots \\ y(n) \end{Bmatrix} \quad \textbf{(6.11)}$$

$$= \{X^{\mathrm{T}}(n-1)\cdot \boldsymbol{y}(n-1)\} + \{\boldsymbol{x}(n)\cdot y(n)\}$$

Teniendo en cuenta que la ecuación 6.7 adquiere la expresión 6.12 para el instante $n-1$, se puede sustituir el término $\{X^{\mathrm{T}}(n-1)\cdot \boldsymbol{y}(n-1)\}$ en la ecuación 6.11 y se obtiene otra expresión para $X^{\mathrm{T}}(n)\cdot \boldsymbol{y}(n)$ que sólo depende de las medidas actuales y de matrices de la iteración anterior (ecuación 6.13).

$$\hat{\theta}(n-1) = [X^{\mathrm{T}}(n-1)\cdot X(n-1)]^{-1} \cdot \{X^{\mathrm{T}}(n-1)\cdot \boldsymbol{y}(n-1)\} \quad \textbf{(6.12)}$$

$$X^{\mathrm{T}}(n)\cdot \boldsymbol{y}(n) = [X^{\mathrm{T}}(n-1)\cdot X(n-1)]\cdot \hat{\theta}(n-1) + \{\boldsymbol{x}(n)\cdot y(n)\} \quad \textbf{(6.13)}$$

La ecuación 6.13 se simplifica al sustituir el valor actual de la señal de salida $y(n)$ por una expresión en función del error de ajuste. Este error es la diferencia entre el valor real de la salida y la estimación que hace el modelo utilizando los parámetros que se quieren corregir:

$$y(n) = \boldsymbol{x}^{\mathrm{T}}(n)\cdot\bar{\theta}(n-1) + e(n) \qquad \textbf{(6.14)}$$

sustituyendo esta ecuación en 6.13 se obtiene:

$$X^{\mathrm{T}}(n)\cdot\boldsymbol{y}(n) = \left[X^{\mathrm{T}}(n-1)\cdot X(n-1)\right]\cdot\bar{\theta}(n-1) + \boldsymbol{x}(n)\cdot e(n) + \left[\boldsymbol{x}(n)\cdot\boldsymbol{x}^{\mathrm{T}}(n)\right]\cdot\bar{\theta}(n-1) \qquad \textbf{(6.15)}$$

Teniendo en cuenta que $\left[X^{\mathrm{T}}(n-1)\cdot X(n-1)\right]+\left[\boldsymbol{x}(n)\cdot\boldsymbol{x}^{\mathrm{T}}(n)\right]$ es igual a $X^{\mathrm{T}}(n)\cdot X(n)$, según la ecuación 6.11, se puede sacar $\bar{\theta}(n-1)$ factor común y se llega a una expresión que sólo depende de $\bar{\theta}(n-1)$ y de valores en el instante n:

$$X^{\mathrm{T}}(n)\cdot\boldsymbol{y}(n) = \left[X^{\mathrm{T}}(n)\cdot X(n)\right]\cdot\bar{\theta}(n-1) + \boldsymbol{x}(n)\cdot e(n) \qquad \textbf{(6.16)}$$

Sustituyendo ahora en la ecuación original de cálculo de parámetros por el método de mínimos cuadrados (ecuación 6.7) se obtiene la fórmula para actualizar $\bar{\theta}(n)$ en función de $\bar{\theta}(n-1)$ y del error instantáneo de ajuste $e(n)$.

$$\bar{\theta}(n) = \bar{\theta}(n-1) + \left[X^{\mathrm{T}}(n)\cdot X(n)\right]^{-1}\cdot\boldsymbol{x}(n)\cdot e(n) \qquad \textbf{(6.17)}$$

Si se llama $P(n)$ a la inversa de la matriz $X^{\mathrm{T}}(n)\cdot X(n)$ se llega finalmente a la ecuación:

$$\bar{\theta}(n) = \bar{\theta}(n-1) + P(n)\cdot\boldsymbol{x}(n)\cdot e(n) \qquad \textbf{(6.18)}$$

Esta ecuación calcula el vector de parámetros como una corrección del anterior vector de parámetros. Esta corrección sólo depende de la matriz $P(n)$, del vector de entradas al proceso $\boldsymbol{x}(n)$ y del valor escalar $e(n)$ que representa el error de ajuste del modelo. Por lo tanto el mayor problema sigue siendo el cálculo de la matriz P. Sin embargo se ve a continuación la manera de obtener $P(n)$ a partir de $P(n-1)$.

El lema de inversión de matrices establece que:

$$(A+BCD)^{-1} = A^{-1} - A^{-1}B\cdot\left(C^{-1}+DA^{-1}B\right)^{-1}\cdot DA^{-1} \quad \textbf{(6.19)}$$

donde se pueden hacer las siguientes asignaciones:

$$\begin{aligned} \mathrm{A} &= X^{\mathrm{T}}(n-1)\cdot X(n-1) = P^{-1}(n-1) \\ \mathrm{B} &= \boldsymbol{x}(n) \\ \mathrm{C} &= 1 \\ \mathrm{D} &= \boldsymbol{x}^{\mathrm{T}}(n) \end{aligned} \quad \textbf{(6.20)}$$

Haciendo estas sustituciones se obtiene en la parte izquierda la inversa del término $X^{\mathrm{T}}(n-1)\cdot X(n-1) + \boldsymbol{x}(n)\boldsymbol{x}^{\mathrm{T}}(n)$, que según la ecuación 6.11 es igual a $P(n)$, y en la parte derecha una expresión que sólo depende de $P(n-1)$ y del vector de entradas en el instante n. Es decir, se llega a la siguiente fórmula recursiva para obtener la matriz P:

$$P(n) = P(n-1)\cdot\left[\mathrm{I}_{np\times np} - \boldsymbol{x}(n)\cdot\left(1+\boldsymbol{x}^{\mathrm{T}}(n)\cdot P(n-1)\cdot\boldsymbol{x}(n)\right)^{-1}\cdot\boldsymbol{x}^{\mathrm{T}}(n)\cdot P(n-1)\right] \quad \textbf{(6.21)}$$

Esta ecuación no precisa el cálculo de ninguna inversa, ya que el término que figura entre paréntesis se reduce a un escalar.

En conclusión, se ha llegado a una expresión (ecuación 6.18) que permite calcular los parámetros del modelo de forma recursiva; es decir, en función de los valores obtenidos en el instante anterior. Este cálculo puede entenderse como una corrección de los valores anteriores $\hat{\theta}(n-1)$ mediante un sumando que es función del error de ajuste, del vector de entradas y de la matriz P. En cierto modo la matriz P tiene en cuenta la historia del pasado y pondera el valor del error al calcular el factor de corrección. La matriz $P(n)$ se obtiene también de manera recursiva a partir de $P(n-1)$ y de la medida más reciente de las entradas mediante la ecuación 6.21.

6.3 Aplicación a la monitorización de los parámetros

La monitorización de parámetros tiene por objeto mantener actualizados los valores de los parámetros del modelo con el fin de poder detectar cambios significativos en los mismos. Las técnicas de estimación de parámetros en continuo descritas en los apartados anteriores son útiles para

realizar esta tarea aunque requieren una pequeña variación. En principio estas técnicas están pensadas para ir disminuyendo los errores de ajuste y permiten, por ejemplo instalar un sistema de control sin necesidad de haberlo ajustado previamente con precisión (los sistemas de control adaptativo [Wellstead91] [Sastry89]). Además establecen pequeñas correcciones en los parámetros si éstos cambian por razones exógenas, como variaciones climáticas, lo que supone una mejora frente a la opción tradicional de mantenerlos fijos.

Para poder aplicar las técnicas de estimación de parámetros en sistemas de detección incipiente de fallos es importante poder detectar con prontitud los cambios que se produzcan en las características del proceso. La capacidad de detección de estos cambios depende de la velocidad de adaptación del sistema de estimación de parámetros en continuo.

Los algoritmos de estimación recursiva descritos en los apartados anteriores consideran con igual importancia todos los datos disponibles del proceso y por lo tanto no son muy sensibles a variaciones que tengan lugar después de largos períodos de funcionamiento estable. La respuesta es análoga a una estimación de la media de una variable que ha tomado valores estables durante mucho tiempo y que cambia bruscamente. La figura 6.3 muestra en trazo de puntos los valores reales de la variable y en trazo continuo la estimación de la media.

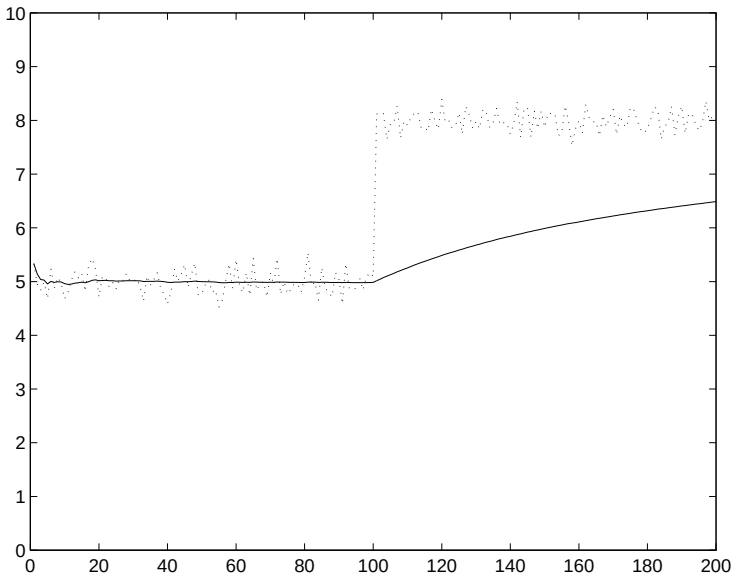


Figura 6.3: Estimación al considerar todos los datos con igual peso

En cada instante el cálculo de la media se ha realizado utilizando todos los valores del pasado. Inicialmente el valor real de la media vale 5, pero cambia bruscamente en la muestra 100 y pasa a valer 8; sin embargo, el valor de la estimación responde lentamente a este cambio. Al final de la gráfica existen tantos puntos con media 5 como puntos con media 8 y por lo tanto la estimación, que considera todos los datos disponibles con igual importancia, alcanza una situación intermedia.

Es evidente que para aplicar estas técnicas en un sistema de detección incipiente de fallos es necesario mejorar esta respuesta. Para ello se debe dar mayor relevancia a los datos más recientes, bien considerando únicamente los últimos datos o bien aplicando factores de olvido.

6.3.1 Estimación mediante ventanas rectangulares

El primer método que se propone para mejorar la dinámica de las estimaciones ante cambios en las características del proceso es utilizar sólo los últimos datos para realizar el cálculo. Todos los datos se consideran con igual importancia relativa; es decir, se le presta la misma atención al último dato que al de hace 10 muestras, por poner un ejemplo.

El método es equivalente a desplazar una ventana rectangular de anchura constante que al multiplicarla por los datos elimina aquéllos que se consideran anticuados. La figura 6.4 muestra el mismo ejemplo de la figura 6.3 pero ajustado mediante ventanas rectangulares de 20 muestras de longitud.

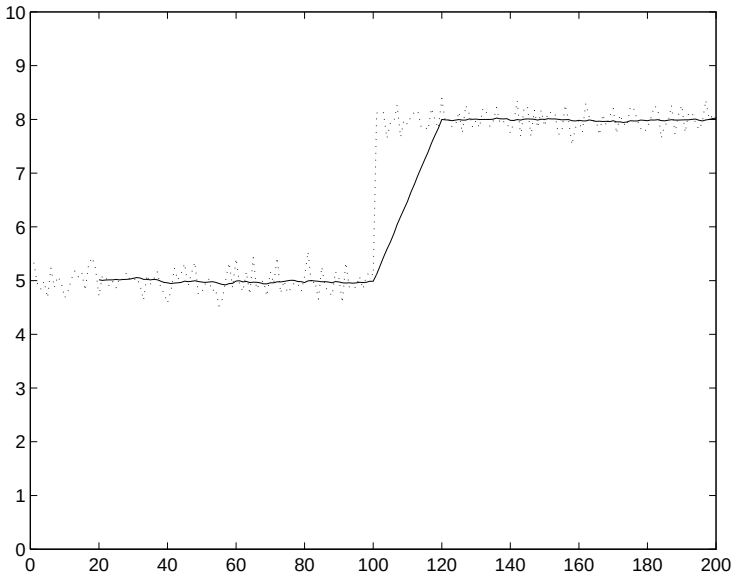


Figura 6.4: Estimación que considera una ventana de 20 muestras

Puede observarse que la respuesta es mucho más rápida que en el caso de considerar en cada instante todos los puntos del pasado. Dado que la anchura de la ventana que se ha utilizado en este ejemplo es de 20 muestras, a partir del instante 120 ya no interviene en el cálculo ningún dato de media 5, por lo que el resultado ya es correcto. Cuanto más estrecha sea la ventana de cálculo, más rápida es la respuesta; sin embargo la estimación es más sensible al ruido de las medidas. Además, considerar pocos datos puede tener problemas en la estimación si los datos no representan las diferentes condiciones de funcionamiento del proceso, como se verá más adelante en este capítulo.

El tamaño de la ventana es por lo tanto un parámetro importante en los sistemas de detección incipiente de fallos que utilicen este método de cálculo. El procedimiento de ajuste que se propone en esta tesis se encarga de encontrar este valor de forma automática.

El cálculo de los parámetros debe realizarse en cada momento utilizando sólo los últimos datos. Cualquier método de ajuste es válido en este caso. Si se adopta el método de los mínimos cuadrados debe aplicarse la fórmula siguiente:

$$\hat{\theta} = \left[X^{T}(n) \cdot X(n)\right]^{-1} \cdot \left\{X^{T}(n) \cdot \boldsymbol{y}(n)\right\} \qquad \textbf{(6.22)}$$

Esta es la misma fórmula que aparece en la ecuación 6.7 de la página 130 salvo que ahora la dimensión de las matrices es fija. En cada iteración se elimina la primera fila de la matriz X y el primer elemento del vector $\boldsymbol{y}$, y se añade una nueva fila y un nuevo elemento (los datos que se acaban de obtener). Es necesario resolver la inversa de la matriz $X^{T}X$ sin que el resultado obtenido en la iteración anterior sea de ninguna utilidad, a pesar de que casi todos los datos de la matriz X se repiten.

6.3.2 Estimación recursiva con factores de olvido

Este método consiste en aplicar una ponderación de tipo exponencial a los datos que intervienen en el cálculo. La ventana exponencial reduce progresivamente la atención que se presta a información del pasado hasta que se hace despreciable.

La figura 6.5 muestra el mismo ejemplo de las figuras 6.3 y 6.4 pero realizando el cálculo de la media en el instante n de acuerdo a la expresión:

$$\bar{x}(n) = \frac{\sum_{i=1}^{n} \gamma^{n-i} \cdot x_i}{\sum_{i=1}^{n} \gamma^{n-i}} \qquad \textbf{(6.23)}$$

En esta expresión γ tiene un valor entre 0 y 1 y reduce la importancia relativa de los primeros datos. La constante γ se conoce como **factor de olvido** [Hsia77] [Söderström89]. Cuanto mayor sea este factor más lenta es la respuesta del sistema, de hecho si $\gamma=1$ equivale al método de estimación de parámetros que considera todos los datos con igual peso. Si el factor de olvido tiene un valor pequeño, sólo los datos más recientes tienen influencia en el cálculo, por lo que se obtiene una respuesta rápida aunque generalmente ruidosa. En el ejemplo de la figura 6.5 se ha tomado el valor $\gamma=0{,}8$.

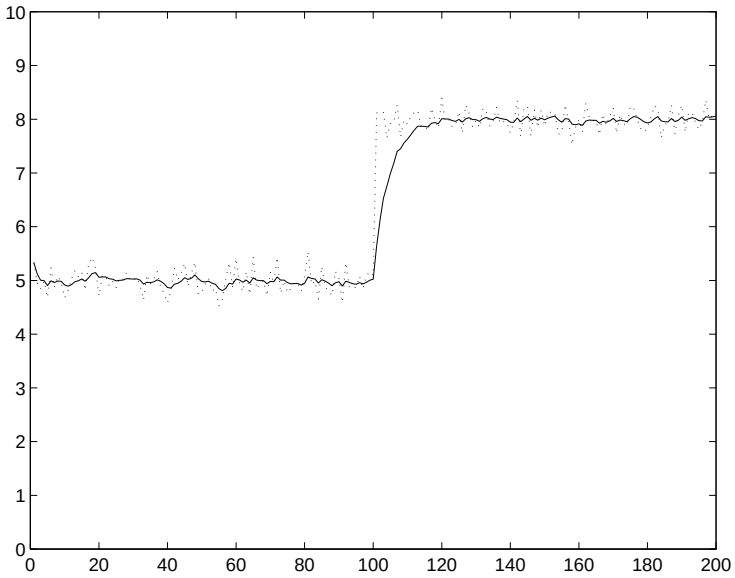


Figura 6.5: Ajuste utilizando factor de olvido

La ventaja fundamental que ha hecho tan popular la utilización de factores de olvido en la estimación de parámetros en continuo es la facilidad de su implantación. Aunque en el cálculo participan todos los datos obtenidos durante el funcionamiento del proceso, es posible hacer un planteamiento recursivo similar al que se realizó en el apartado 6.2.2. En lugar de plantear el ajuste como la minimización:

$$\min\left\{\sum_{i=1}^{n} e^2(i)\right\} \tag{6.24}$$

lo que da lugar a la ecuación 6.7 y en forma recursiva a las ecuaciones 6.18 y 6.21, se plantea ahora como:

$$\min\left\{\sum_{i=1}^{n} \gamma^{n-i} \cdot e^2(i)\right\} \tag{6.25}$$

Este planteamiento da lugar a las siguientes ecuaciones de estimación en forma recursiva [Wellstead91] [Söderström89]:

$$\bar{\theta}(n) = \bar{\theta}(n-1) + P(n) \cdot \boldsymbol{x}(n) \cdot e(n) \tag{6.26}$$

$$P(n) = \frac{1}{\gamma} P(n-1) \cdot \left[I_{np \times np} - \frac{\boldsymbol{x}(n) \cdot \boldsymbol{x}^{T}(n) \cdot P(n-1)}{\gamma + \boldsymbol{x}^{T}(n) \cdot P(n-1) \cdot \boldsymbol{x}(n)} \right] \quad \textbf{(6.27)}$$

En esta formulación aparece el parámetro γ que es el factor de olvido. El denominador de la fracción que aparece entre corchetes se reduce de nuevo a un escalar, por lo tanto según esta formulación no es necesario calcular ninguna inversa.

Hay que destacar que el factor de olvido se aplica de manera global a toda la matriz *P,* por lo tanto afecta de igual manera al cálculo de todos los parámetros del modelo. Esta situación puede no ser la ideal si la dinámica de las características del proceso no es homogénea. En función de la velocidad con que cambia cada parámetro y de su sensibilidad al ruido, puede ser interesante utilizar factores de olvido diferentes. Este es el caso, por ejemplo, de la estimación de parámetros en un motor eléctrico de inducción donde los parámetros mecánicos tienen constantes de tiempo lentas mientras que los parámetros eléctricos pueden cambiar rápidamente [Vélez-Reyes88]. Cuando interese utilizar un factor de olvido específico para el cálculo de un parámetro pueden reformularse las ecuaciones del modelo desacoplando el cálculo de dicho parámetro [Vélez-Reyes92].

6.4 Problemas en la estimación recursiva de parámetros

La estimación de parámetros de procesos puede presentar dos problemas fundamentales, ambos relacionados con la calidad de los datos utilizados para el ajuste. El primer problema es la falta de riqueza de los datos y el segundo problema es la falta de consistencia en los parámetros que se quieren estimar. La estimación recursiva equivale a una serie de ajustes de parámetros en los cuales van cambiando los conjuntos de datos de entrada y salida. Cada ajuste de parámetros requiere que los conjuntos de datos aporten suficiente información y que las características del proceso sean constantes (ya que los parámetros que se obtienen representan a estas características para todos los datos utilizados en el ajuste).

6.4.1 Falta de riqueza en los datos

Para realizar el ajuste de un modelo hacen falta como mínimo tantos datos como número de parámetros (igual número de ecuaciones que de incógnitas), pero además estos datos deben ser independientes. El volumen de datos utilizados en un ajuste no es un buen indicativo de la calidad de los mismos, es importante que los datos correspondan a situaciones de trabajo diferentes. Esto se ve claramente en el caso de modelos de regresión lineal que se ajusten por el método de mínimos cuadrados ya que los parámetros se obtienen directamente a partir de un cálculo matricial:

$$\tilde{\theta} = \left[X^{T}(n) \cdot X(n)\right]^{-1} \cdot \left\{X^{T}(n) \cdot \boldsymbol{y}(n)\right\} \qquad \textbf{(6.28)}$$

Cada situación de trabajo viene caracterizada por un vector de entradas $\boldsymbol{x}$ que se almacena, para el cálculo, en la matriz X. Sólo en el caso de que existan tantas filas de X linealmente independientes como número de parámetros, la matriz $X^{T}X$ no será singular y la ecuación 6.28 tendrá solución. Para conseguir esta independencia lineal, las condiciones de trabajo durante el registro de datos deben ser diferentes.

También es importante cubrir todo el rango de funcionamiento del proceso para obtener un conjunto de parámetros fiables. Nunca se conocen con exactitud las ecuaciones que rigen el comportamiento de un proceso real, pero se utilizan modelos matemáticos aproximados que estiman el comportamiento del proceso. El modelo será válido para un conjunto restringido de situaciones de trabajo. Es importante disponer de información sobre todas las condiciones normales de trabajo para obtener un modelo que sea válido para todas ellas. En caso contrario los parámetros que se obtienen están demasiado especializados en una condición de funcionamiento y los errores de predicción son globalmente mayores.

La figura 6.6 muestra un ejemplo del ajuste de un modelo lineal (línea fina) que se utiliza para predecir la salida de un proceso cuyo comportamiento real (línea gruesa) es algo parabólico. El proceso produce una salida y que depende del valor de la entrada x, el cual esta comprendido entre 1 y 2 durante el funcionamiento normal. El primer caso corresponde a un ajuste que se realizó utilizando valores de x comprendidos entre 1,5 y 2. En el segundo caso se ha realizado un ajuste considerando datos de entrada de todo el rango de funcionamiento normal. Puede observarse que

los errores de ajuste para $x \epsilon [1.5, 2]$ son menores en el caso 1 ya que el modelo se ha especializado en estas condiciones de trabajo. Sin embargo, el comportamiento del segundo modelo es globalmente mejor.

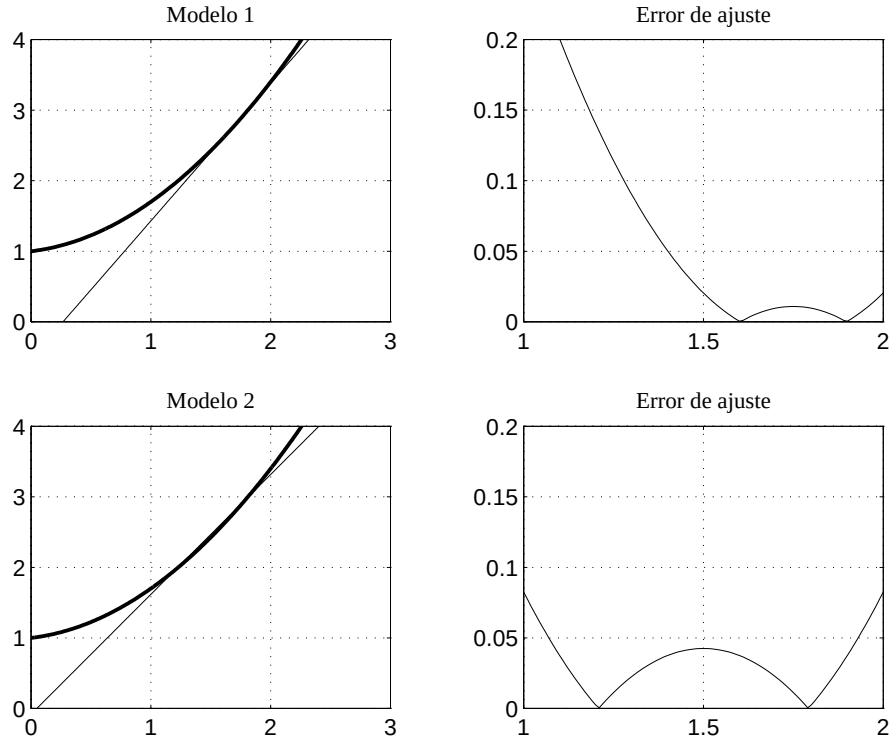


Figura 6.6: Diferencia de resultados utilizando el mismo tipo de modelo pero realizando el ajuste con distintos conjuntos de datos.

En cuanto al volumen de los datos utilizados para el ajuste, siempre interesa que sea grande con objeto de disminuir los efectos del ruido de medida. La única condición necesaria para que pueda realizarse el ajuste es que el número de puntos de operación diferentes sea mayor o igual al número de parámetros. Sin embargo no importa utilizar varios datos obtenidos en el mismo punto de operación, de hecho es beneficioso si existe ruido en las medidas. Se ve a continuación un ejemplo del ajuste de la función $y=1+0{,}8x$. Suponiendo que se dispone de cien datos para $x_1=1$ y sólo de uno para $x_2=2$, entonces el vector de salidas y debería contener cien valores 1,8 y un valor 2,6 (siempre que no exista ruido). Sustituyendo en la ecuación 6.28, se obtienen los valores correctos de los parámetros:

$$\bar{\theta} = \left[\begin{bmatrix} 1 & \cdots & 1 & 2 \\ 1 & \cdots & \cdots & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ \vdots & 1 \\ 1 & 2 \end{bmatrix} \right]^{-1} \cdot \left\{ \begin{bmatrix} 1 & \cdots & 1 & 2 \\ 1 & \cdots & \cdots & 1 \end{bmatrix} \cdot \begin{bmatrix} 1{,}8 \\ \vdots \\ 1{,}8 \\ 2{,}6 \end{bmatrix} \right\} = \begin{Bmatrix} 1{,}0 \\ 0{,}8 \end{Bmatrix} \qquad \textbf{(6.29)}$$

Sin embargo, repitiendo el mismo ejemplo pero con un ruido aleatorio (de distribución normal con media cero y varianza 0.01) que afecta al vector y se han obtenido unos parámetros ligeramente diferentes: $\bar{\theta}=[0{,}9939\ \ 0{,}8066]^T$. Con este vector de parámetros la estimación de la salida para $x_1=1$ vale $\bar{y}_1=1{,}8005$ mientras que la estimación de la salida para $x_2=2$ vale $\bar{y}_2=2{,}6070$. Puede observarse que la estimación en el punto x_2, donde sólo hay un dato, es peor que en el punto x_1 donde hay cien datos. Esto es debido a que el método de ajuste realiza el mejor trabajo posible y obtiene los parámetros que producen el valor medio de las salidas para $x=1$ (eliminando la componente de ruido casi por completo) y el único valor conocido en la situación $x=2$ (que es este caso fue 2,6070 en lugar de 2,6 como consecuencia del ruido). Cuando x toma valores distintos de 1 y 2 la estimación de la salida da lugar a errores de diferentes magnitudes. Se puede decir que el modelo se ha especializado en reproducir la situación $x=1$ de la cual existen muchos datos, y estima otras situaciones, no consideradas durante el proceso de ajuste, con peor precisión.

En conclusión, para realizar un buen ajuste por el método de mínimos cuadrados es importante disponer de datos con suficiente calidad, esto implica:

- Que existan datos de todas las condiciones de funcionamiento del proceso.
- Que existan suficientes datos de cada situación de trabajo como para disminuir los efectos del ruido.

Aparentemente la solución al problema de la calidad de los datos está en esperar a tener suficientes datos del funcionamiento del proceso antes de realizar el ajuste del modelo. Sin embargo, esto se traduce en utilizar ventanas y factores de olvido grandes, lo que perjudica la respuesta del sistema y provoca problemas por falta de consistencia en los parámetros.

6.4.2 Falta de consistencia en los parámetros

Es frecuente que los procesos industriales sufran variaciones lentas en los parámetros, debidas al envejecimiento de los componentes o a cambios en las condiciones de trabajo no considerados en los modelos (como cambio climático). Las técnicas de estimación recursiva de parámetros han experimentado un importante desarrollo para poder corregir los modelos adaptándolos a estos cambios.

Mientras las variaciones de las características del proceso sean lentas, la estimación de los parámetros del modelo tendrá lugar con un conjunto de datos homogéneos. Sin embargo, si las características del proceso cambian con rapidez o se toman conjuntos de datos demasiado largos, el ajuste de parámetros se realizará utilizando datos que no corresponden al mismo proceso.

Las técnicas de ajuste de modelos obtienen el conjunto de parámetros que da lugar al menor error de estimación a lo largo de todos los datos utilizado en el ajuste. Si las características del proceso se han mantenido constantes durante todo el tiempo de muestreo, los datos son homogéneos y el ajuste del modelo obtendrá como valor de los parámetros el valor de las características. Por el contrario, si algunos datos corresponden al funcionamiento del proceso con unas características y otros corresponden al mismo proceso pero con características diferentes, se produce una inconsistencia que afecta al ajuste.

En la figura 6.3 (página 136) se ha visto que la estimación de la media de una variable da lugar a un valor intermedio cuando el conjunto de datos no es homogéneo. Pero otros algoritmos tienen un comportamiento menos intuitivo. Se muestra a continuación el caso de un estimador de la frecuencia fundamental de una señal. Este estimador se basa en el cálculo de la transformada de Fourier de la señal [Ludeman87] [Oppenheim83]. En la figura 6.7 se han dibujado dos señales básicas, una de amplitud 3 y frecuencia 0,5 Hz, y la otra de amplitud 2 y frecuencia 0,3 Hz. Como es sabido, el cálculo de la transformada de Fourier de la suma de estas señales identifica correctamente sus amplitudes y sus frecuencias. Sin embargo, si se genera un conjunto de datos en el que la primera parte corresponde a la señal de 0,5 Hz y la segunda mitad a la señal de 0,3 Hz, la transformada

de Fourier que se obtiene refleja la existencia de múltiples frecuencias diferentes. En este caso no aparecen las frecuencias originales en su magnitud real, ni una única componente de frecuencia intermedia, como se podría esperar.

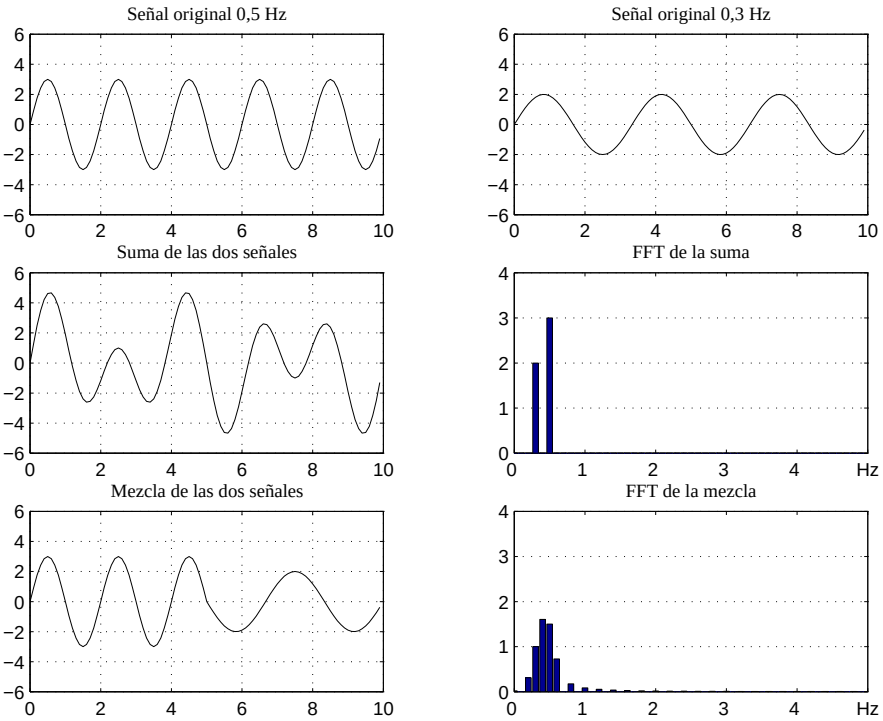


Figura 6.7: Estimación de la frecuencia de una señal

En consecuencia, el ajuste de modelos con datos correspondientes a situaciones en las cuales las características físicas de los componentes del proceso han cambiado, puede dar lugar a parámetros ficticios. Por lo tanto no es aconsejable utilizar un volumen de datos muy grande porque probablemente se estén mezclando datos antiguos con datos nuevos.

En caso de aparición de un fallo incipiente, y dificultan la detección del fallo.

Lógicamente siempre que aparezca un fallo se iniciará una variación en las características físicas del sistema. Entonces los datos nuevos corresponden a la situación de fallo, mientras que los antiguos corresponden a la situación de funcionamiento normal. Por lo tanto el algoritmo de estimación tendrá que trabajar con un conjunto de datos no homogéneo. Como consecuencia, durante un periodo transitorio, los parámetros que se obtienen no son los

que corresponden a las características reales y esto dificulta y retrasa la detección.

6.5 Comparación entre detección basada en residuos y detección basada en parámetros

Como resumen de este capítulo, se muestra a continuación un ejemplo sobre el que se han aplicado las técnicas de estimación recursiva de parámetros. La evolución de los parámetros, en el caso de aplicar estimación recursiva, se compara con la evolución de los residuos, en el caso de utilizar modelos no adaptativos.

Se ha tomado como ejemplo el caso de un circuito eléctrico formado por una batería y una resistencia externa. Se trata del mismo ejemplo utilizado en el apartado 2.2 y en [Palacios97] salvo que ahora se han utilizado datos simulados en lugar de datos experimentales. La utilización de datos simulados permite mostrar con mayor claridad el funcionamiento de las técnicas de estimación de parámetros.

Los datos han sido generados suponiendo que la batería tiene un comportamiento equivalente a una fuente de tensión ideal v_i y una resistencia interna r_i (ver figura 6.8).

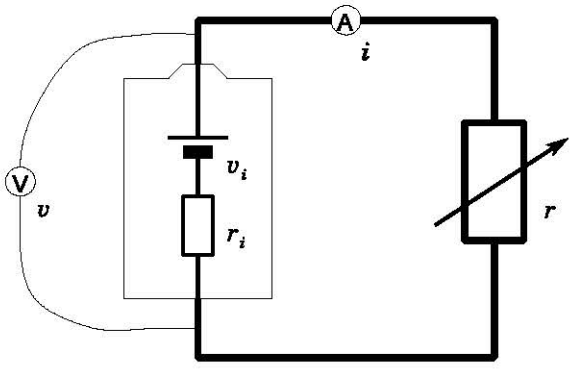


Figura 6.8: Esquema del circuito eléctrico

Las dos variables medibles de este circuito son la tensión total de la batería (v) y la corriente del circuito (i) que dependen del valor que se ajuste en la

resistencia variable r. La manera de simular estas señales se basa en las siguientes ecuaciones, que dependen de los valores de la resistencia externa r y de las características físicas v_i y r_i de la batería.

$$i = \frac{v_i}{(r_i + r)} \qquad v = r \cdot i \tag{6.30}$$

Posteriormente se ha añadido una componente de ruido blanco a ambas señales para simular los efectos del ruido en los equipos de medida.

6.5.1 Condición normal de funcionamiento

Inicialmente se han tomado unos valores constantes para representar las características físicas v_i y r_i. Los datos representados en la figura 6.9 se han obtenido para unos valores de las características físicas v_i=1,3 V y r_i=0,35 Ω y para un valor de la resistencia r que varía de manera aleatoria. La primera gráfica representa la evolución de la tensión (línea continua) y de la intensidad (línea discontinua) del circuito. Es importante destacar que las oscilaciones que presentan estas señales son consecuencia, principalmente, de las variaciones de la carga del circuito y no del ruido introducido. La curva de carga del circuito está representada en la segunda gráfica de la figura 6.9. La variación de la carga, siendo aleatoria, tiene una distribución tal que los valores que toman la tensión y la intensidad del circuito tiene una distribución uniforme.

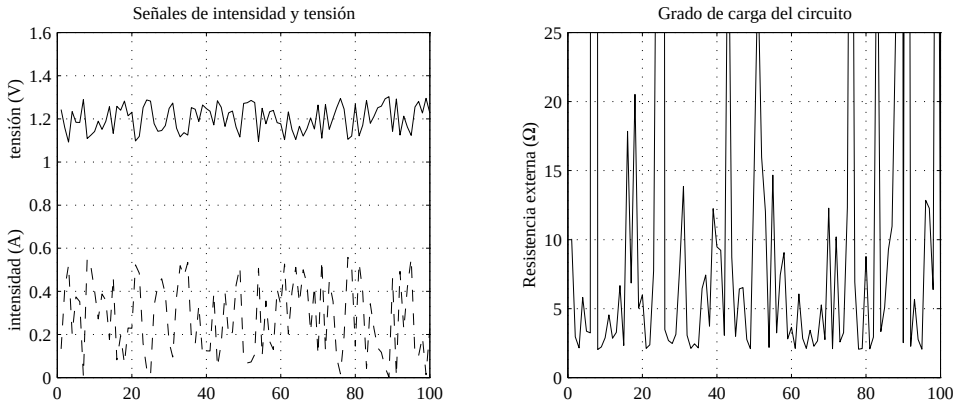


Figura 6.9: Simulación de una condición normal de funcionamiento

En este caso las características físicas de la batería se han mantenido constantes y el grado de carga varía en todo el rango de funcionamiento. En estas condiciones la estimación recursiva de parámetros no presenta ningún inconveniente.

La figura 6.10 muestra el comportamiento de las dos técnicas de detección. La primera gráfica representa la evolución de los residuos en el caso de utilizar un modelo de parámetros fijos. Puede observarse que los residuos se mantienen en un nivel mínimo a pesar de la pequeña componente de ruido que sufren las señales. Las pequeñas oscilaciones en el nivel del error son despreciables con la escala utilizada en la gráfica, que es la misma escala que se utiliza más adelante en caso de fallo. La segunda gráfica representa los valores obtenidos por el algoritmo de estimación recursiva de parámetros con un factor de olvido γ=0,8. Los valores de los parámetros se aproximan con mucha precisión a los valores de las características físicas utilizados en la simulación, salvo unas oscilaciones como consecuencia del ruido de las señales.

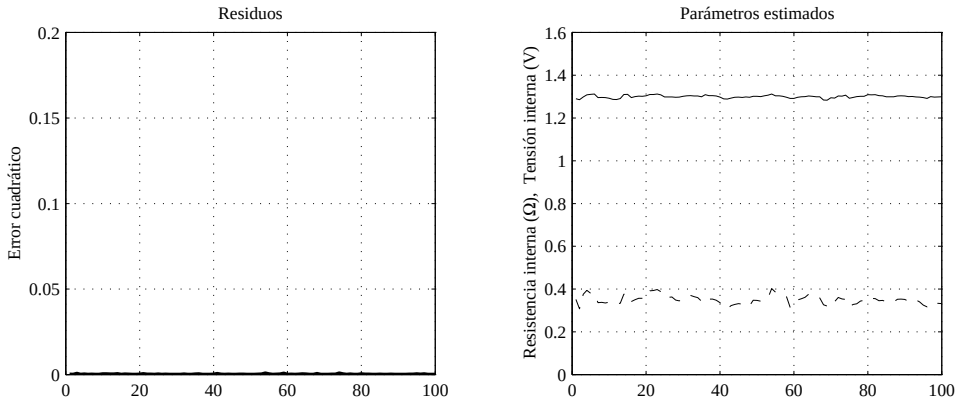


Figura 6.10: Comparación de las técnicas de detección

6.5.2 Degradación del proceso

Como segundo ejemplo se ha simulado una degradación de la batería hasta una situación que se puede considerar de fallo del circuito. Estos datos se han obtenido simulando una variación de las características físicas de la batería. Se han introducido simultáneamente una disminución de la tensión interna v_i y un aumento progresivo de la resistencia interna r_i. El resto de

las condiciones se ha mantenido igual que en el caso anterior; es decir, la variación de carga de forma aleatoria y el ruido de las señales. La figura 6.11 representa este segundo caso.

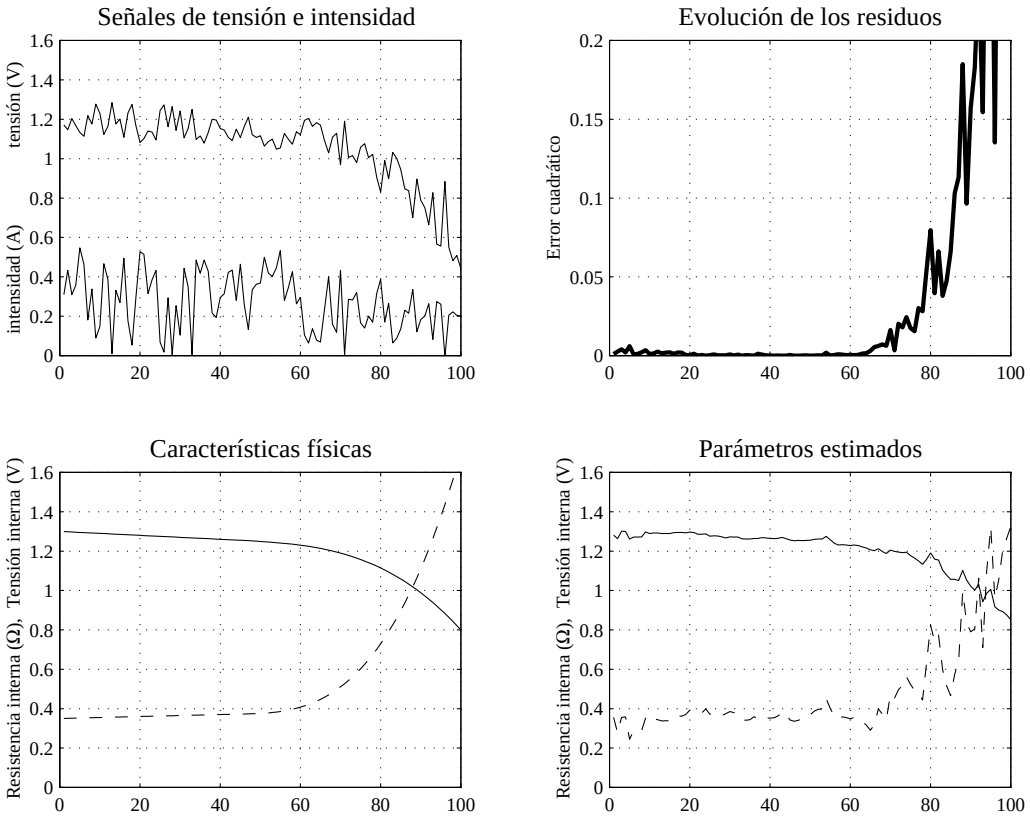


Figura 6.11: Simulación de un caso de degradación

La primera gráfica representa las señales de tensión e intensidad del circuito, que en la parte final experimentan una fuerte disminución como consecuencia de la degradación de la batería. Ante esta situación un **modelo de parámetros fijos** da lugar a una curva de residuos que se ha representado en la segunda gráfica. Inicialmente los residuos son despreciables, como ocurría en la figura 6.10, pero en la parte final experimentan un aumento muy importante. Un sistema de detección basado en el análisis de residuos debería detectar el fallo incipiente del circuito al observar este aumento tan claro en los residuos.

La tercera gráfica de la figura 6.11 muestra la **evolución de las características físicas** utilizadas durante la simulación. La degradación de ambas características comienza en el instante 50, aunque de manera muy suave. La última gráfica de la figura representa los valores obtenidos por

estimación recursiva cuando se utiliza un **modelo adaptativo**. Comparando con la gráfica de características físicas se observa que el seguimiento es bueno aunque con un ligero retraso. Como se ha explicado anteriormente en este capítulo, tomando factores de olvido más pequeños se obtiene una respuesta más rápida pero los resultados se ven más afectados por el ruido. Un sistema de detección de fallos basado en el análisis de la evolución de los parámetros debería detectar el fallo incipiente del circuito al observar la disminución que se produce en el parámetro de la tensión interna o el aumento que experimenta el parámetro de la resistencia interna. En función del factor de olvido que se utilice puede ocurrir que la detección se realice con cierto retraso o que las oscilaciones debidas al ruido dificulten el proceso de detección.

6.5.3 Condición estable de funcionamiento

En los dos casos anteriores se ha utilizado una carga del circuito que varía de forma aleatoria, lo que proporciona abundante información sobre las características del proceso. Esto es una situación excesivamente favorable, ya que en la práctica es habitual que los procesos experimenten períodos de funcionamiento en los que las condiciones de trabajo no cambian. En el tercer caso ejemplo (figura 6.12) se ha realizado una simulación en la cual la carga del circuito varía aleatoriamente durante la primera mitad pero permanece constante al final. Al mantenerse constante la carga, también se mantienen constantes la tensión y la intensidad del circuito, salvo un pequeño rizado del ruido de medida.

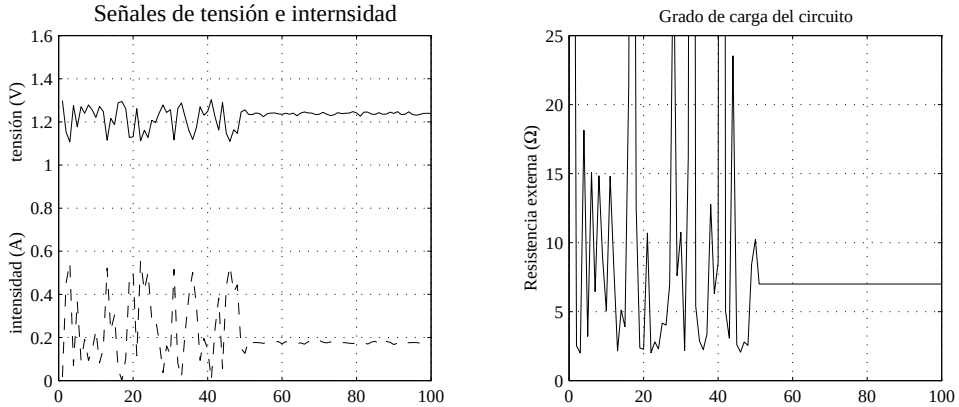


Figura 6.12: Caso en el que la carga deja de variar

En este caso el modelo de parámetros constantes no experimenta ningún cambio y los residuos no se alteran como consecuencia de la estabilidad en las condiciones de trabajo (primera gráfica en la figura 6.13). Sin embargo, la estimación de parámetros presenta problemas cuando no existe suficiente riqueza en los datos. La primera parte de la gráfica de parámetros estimados es igual que en la figura 6.10, pero en la parte final se produce una inestabilidad como consecuencia de la falta de información en los datos utilizados para el ajuste. Como el ajuste de parámetros siempre se realiza utilizando datos del pasado, los problemas de estimación no aparecen inmediatamente al estabilizarse la carga. Un cierto tiempo después, que depende del valor del factor de olvido, es cuando los datos con riqueza de información se hacen despreciables y los datos más recientes no permiten hacer la estimación.

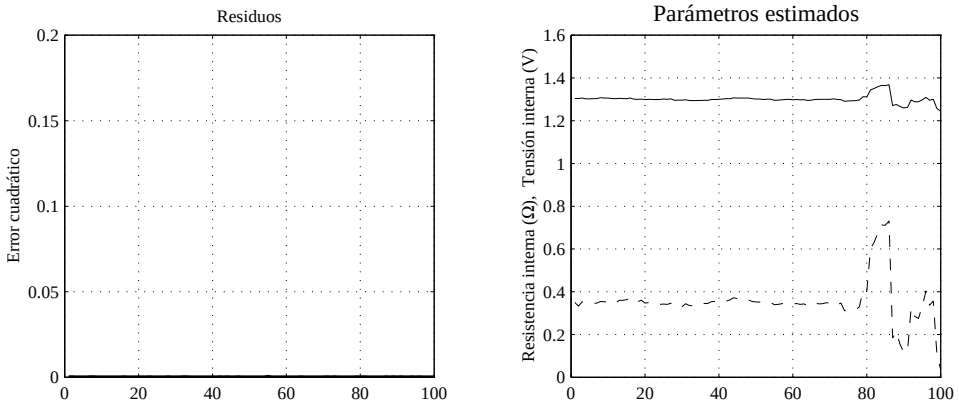


Figura 6.13: Inestabilidad en los parámetros por la falta de información

Esta situación de inestabilidad de cálculo puede detectarse por varios métodos analizando las matrices que intervienen en el ajuste [Wellstead91] [Basseville93]. Sin embargo, el único interés que tiene la identificación de estas situaciones es evitar la obtención de parámetros ficticios o que el programa de cálculo produzca un error. La solución que puede adoptarse es mantener los parámetros anteriores en lugar de intentar estimar unos nuevos. Esta solución deja anulada la posibilidad de detección, pero evita que se produzcan falsas alarmas debidas a las oscilaciones ficticias que aparecen en los parámetros como consecuencia de la falta de información sobre el funcionamiento del proceso.

6.6 Conclusiones

En este capítulo se han explicado detalles sobre la manera de realizar el cálculo de parámetros en continuo; tanto en formulación clásica como en la formulación recursiva. Se ha destacado la necesidad de utilizar factores de olvido u otros mecanismos que ayuden a aumentar la importancia de los datos más recientes, para que estas técnicas sean aplicables a la detección incipiente de fallos.

También se han presentado los dos problemas fundamentales que pueden presentarse en los sistemas de detección basados en el seguimiento de los parámetros: la variación rápida de las características físicas y la falta de información sobre distintas condiciones de funcionamiento. Estos dos problemas dependen del factor de olvido utilizado, pero en algunos procesos pueden ser tan importantes que hagan inefectiva la detección basada en el seguimiento de parámetros (fundamentalmente procesos en los cuales las condiciones de trabajo son muy estables).

Una solución combinada puede resolver problemas de falsas alarmas en procesos donde las características cambian muy lentamente (cambios de invierno a verano o envejecimiento lento). En estos casos se puede plantear un sistema de detección basado en residuos, en el cual se reajustan los parámetros del modelo de forma recursiva para corregir las pequeñas variaciones de funcionamiento de evolución lenta. La estimación de parámetros sólo tendría por objeto mejorar la precisión del modelo, lo que permite fijar unos umbrales de detección más estrechos sin que se produzcan falsas alarmas; esto mejora la sensibilidad del sistema de detección. Queda pendiente, como futuro trabajo, comprobar si en algún proceso real es mejor este planteamiento combinado que utilizar un modelo de parámetros fijos pero dejando que el procedimiento de ajuste obtenga los parámetros. Hay que tener en cuenta que el procedimiento de ajuste que se plantea en esta tesis es capaz de encontrar conjuntos de parámetros que mejoran la capacidad de detección sin que estos parámetros sean los que producen una máxima precisión del modelo [Palacios97].

Por último aclarar que en el caso de utilizar modelos adaptativos, el procedimiento de ajuste del sistema de detección no se encarga de obtener los parámetros del modelo, en todo caso ajustaría valores del estimador (como el factor de olvido). Este ajuste se realiza conjuntamente con el resto

de los parámetros del sistema de detección; es decir, los parámetros de los atributos de fallo y de la función de detección de fallos.

Capítulo 7

Ejemplo práctico

7.1 Descripción general del ejemplo

Como ejemplo de aplicación de la metodología propuesta en esta tesis, se ha seleccionado el caso de la detección incipiente de fallos en trenes. Se ha elegido este ejemplo porque se trata a la vez de un caso real y de un caso general, para mostrar la aplicabilidad de la tesis a cualquier tipo de sistema.

En concreto se ha analizado el sistema de suspensión de los trenes denominados UT-450 y UT-451, fabricados por la empresa GEC-Alsthom. Se trata de trenes eléctricos utilizados para el transporte rápido y masivo de viajeros en líneas de cercanías urbanas y suburbanas. Son trenes de 2 pisos con una capacidad de casi 200 personas sentadas por cada coche (algunos de ellos son los utilizados para transporte de viajeros en la Comunidad de Madrid). Se ha incluido una breve descripción de estos trenes más adelante.

Para poder mostrar el funcionamiento del sistema de detección de fallos, se ha desarrollado un modelo de simulación, que permite generar datos sobre el comportamiento del tren en distintas situaciones. Con este sistema de simulación pueden obtenerse datos suficientes sobre comportamiento normal y sobre comportamiento en caso de fallo. Estos datos se utilizan para ajustar y comprobar el sistema de detección que se presenta como ejemplo.

El sistema de detección incipiente de fallos utiliza un modelo autoregresivo que sólo se basa en variables medibles desde el interior del tren con sensores convencionales. Este sistema es fácilmente integrable en los coches y permite disponer en tiempo real de un indicador del estado de salud de los componentes del sistema de suspensión. Si este sistema se instalase en los trenes de cercanías, el gran volumen de datos obtenido debería utilizarse para reajustar el sistema de detección, sustituyendo los datos obtenidos por simulación. De esta manera el sistema de detección podría tener en cuenta efectos no considerados durante las simulaciones.

7.1.1 Tipos de coches que componen un tren

Las unidades UT-450 y UT-451 están compuestas por combinaciones de tres tipos diferentes de coches: la motriz, los remolques y el remolque con cabina.

La **unidad motriz** se muestra en la figura 7.1, está compuesta por dos bogies motores y una caja que incluye cabina, equipos de control y departamento para 128 pasajeros sentados.

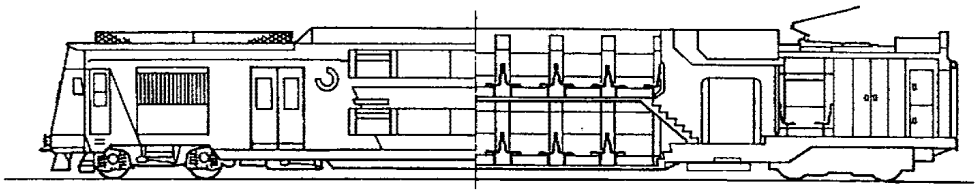


Figura 7.1: Esquema de la unidad motriz

Esta unidad está alimentada a tensión continua de 3 kV y puede desarrollar una potencia de 1480 kW. Las características de dimensiones y masas se presentan más adelante en la tabla 7.1 comparadas con los remolques.

El **remolque** está formado por dos bogies no motorizados y una caja con capacidad para 188 pasajeros sentados. Un esquema del remolque de las unidades UT-450 y UT-451 se muestra a continuación en la figura 7.2.

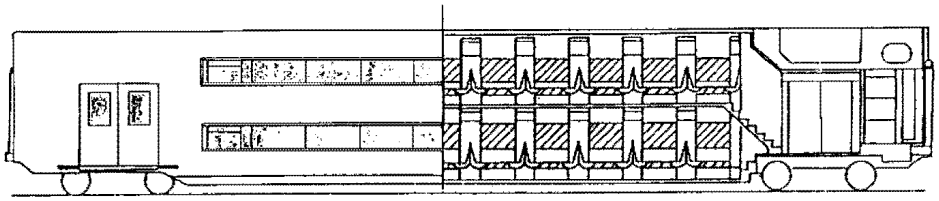


Figura 7.2: Esquema del remolque

Existen dos tipos de coches remolques, denominados R y Re (remolque normal y remolque especial). Los coches son idénticos exteriormente y sólo se diferencian en algunos detalles de las instalaciones eléctricas y neumáticas que no influyen en los cálculos que se muestran en este capítulo. Las dimensiones y masas son iguales en los dos tipos de remolques.

Una variante importante es el **remolque con cabina** (figura 7.3). La caja es similar al remolque pero incluye una cabina que permite conducir el tren, sin embargo el remolque con cabina no tiene capacidad de tracción. Se utilizan dos bogies, de dimensiones similares al bogie del remolque; ninguno está motorizado. El compartimento de viajeros queda reducido a 182 pasajeros sentados.

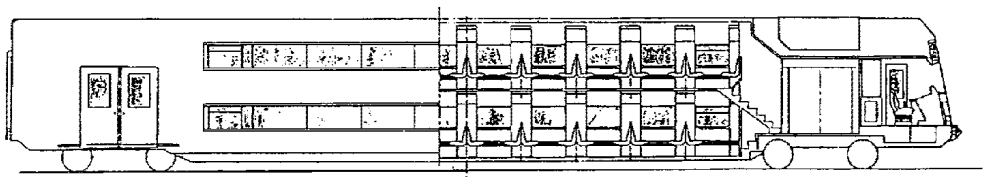


Figura 7.3: Esquema del remolque con cabina

La tabla 7.1 recoge los valores de dimensiones en milímetros y masas en toneladas de los diferentes tipos de coches.

Tabla 7.1: Dimensiones (mm) y masas (t) de los diferentes tipos de coches

	Unidad motriz	Remolque	Remolque con cabina	
Longitud total (entre topes)	26900	26400	27280	
Distancia entre centros de bogies	20000	20000	20000	
Distancia entre ejes del bogie	2650	2400	2400	
Diámetro de la rueda	1020	840	840	
Masa de la caja	45,2	38,8	40,8	
Masa del bogie	12,5	6,9	7,1	6,9
Masa total del coche en vacío	70,2	52,6	54,8	
Masa total con carga normal	86,7	76,7	77,9	

*En la casilla de la columna del remolque con cabina en la que aparecen dos valores, el primer número corresponde al bogie delantero y el segundo al bogie trasero.

7.1.2 Composición de las unidades UT-450 y UT-451

La unidad UT-450 está formada por 2 unidades motrices y 4 remolques. Las unidades motrices van acopladas en ambos extremos del tren y proporcionan una potencia total de 2960 kW. La capacidad total de viajeros es de 1008 plazas sentadas más 836 plazas de pie con carga normal. La longitud total del tren es de 159,4 m y la masa total 350,8 t en vacío o 481 t en carga normal.

La unidad UT-451 es aproximadamente la mitad que la 450, está formada por una unidad motriz, un remolque intermedio y un remolque con cabina. Esta composición básica tiene una capacidad de viajeros de 498 plazas sentadas más 410 plazas de pie con carga normal. La longitud es de 80,6 m y la masa total de 177,6 t en vacío o 241,4 t en carga normal. También pueden unirse dos unidades básicas 451 para formar composiciones con dos motrices, dos remolques y dos remolques con cabina, en distintas combinaciones.

La tabla 7.2 muestra los datos más significativos de las prestaciones de las unidades UT-450 y UT-451.

Tabla 7.2: Prestaciones de las unidades UT-450 y UT-451

Velocidad máxima	140 km/h
Aceleración inicial	0,62 m/s^2
Aceleración normal entre 0 y 60 km/h	0,50 m/s^2
Deceleración con freno eléctrico	0,33 m/s^2
Deceleración con freno eléctrico y neumático	0,90 m/s^2
Deceleración máxima de urgencia	1,20 m/s^2

7.2 Descripción del sistema de suspensión

Los tres movimientos básicos de un vehículo como sólido rígido son el movimiento vertical, el balanceo y el cabeceo (ver figura 7.4). El movimiento vertical se caracteriza porque todos los sistemas de suspensión del vehículo se mueven en fase, mientras que en los otros dos casos hay sistemas en fase y en contrafase. En el movimiento de balanceo los sistemas de suspensión vertical de la parte derecha trabajan en contrafase con los sistemas de suspensión vertical de la parte izquierda. En el movimiento de cabeceo los sistemas de suspensión de la parte delantera actúan en contrafase con la parte trasera. Estos tres movimientos básicos se ven excitados por las irregularidades del trazado: baches puntuales, ondulaciones del suelo y curvas. El movimiento de cabeceo también se provoca con las aceleraciones y frenadas del vehículo.

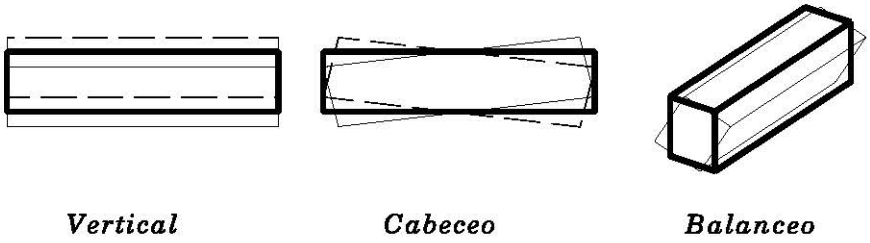


Figura 7.4: Principales movimientos de un vehículo

Los efectos de los movimientos vertical y de cabeceo se atenúan mediante sistemas de suspensión formados por amortiguadores y conjuntos de muelles o de ballestas. Para limitar el balanceo se montan sistemas que incluyen generalmente barras que trabajan a torsión, pero este movimiento es absorbido por los amortiguadores del sistema de suspensión vertical.

En el caso de los trenes existe otro movimiento muy característico, el movimiento de lazo, que se induce por el contacto rueda–carril. El movimiento de lazo se traduce en oscilaciones alrededor del eje vertical del tren y puede producirse en tramos rectos. Este movimiento se atenúa mediante amortiguadores que se oponen a la rotación del bogie respecto a la caja.

En menor medida se producen los otros dos movimientos de translación, que son las oscilaciones en dirección longitudinal y transversal. No existen sistemas de suspensión que permitan grandes desplazamientos en estas direcciones por lo que las amplitudes de estos movimientos son pequeñas comparadas con el movimiento vertical. Los sistemas de suspensión están pensados principalmente para permitir un movimiento vertical que haga más confortable la marcha de los viajeros, las oscilaciones en las otras dos direcciones del espacio están limitados por sistemas mucho más rígidos.

El sistema de suspensión de las unidades UT-450 y UT-451 (diseñado por GEC-Alsthom) es de los más completos y está formado por una doble suspensión vertical, un sistema anti-balanceo, amortiguadores anti-lazo y amortiguadores transversales. El esquema principal del bogie de la unidad

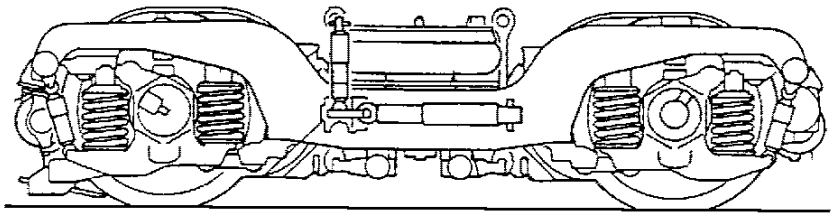


Figura 7.5: Esquema del bogie en alzado

motriz se presenta en las figuras 7.5 y 7.6. En el alzado (figura 7.5) se observan los sistemas de suspensión vertical formados por muelles, amortiguadores y balonas, que se describen más adelante. También se observa el amortiguador anti-lazo, montado en posición horizontal, y las bielas y manivelas del sistema anti-balanceo (en la parte inferior). En la planta (figura 7.6) se identifican los dos ejes de las ruedas, dispuestos de manera anti-simétrica. Sobre cada eje están montadas las ruedas, un freno de disco y una reductora, que conecta con el motor eléctrico. Con mayor dificultad, pueden verse en el centro del bogie los dos amortiguadores que absorben los movimientos laterales; estos amortiguadores van montados horizontalmente y en dirección paralela a los ejes.

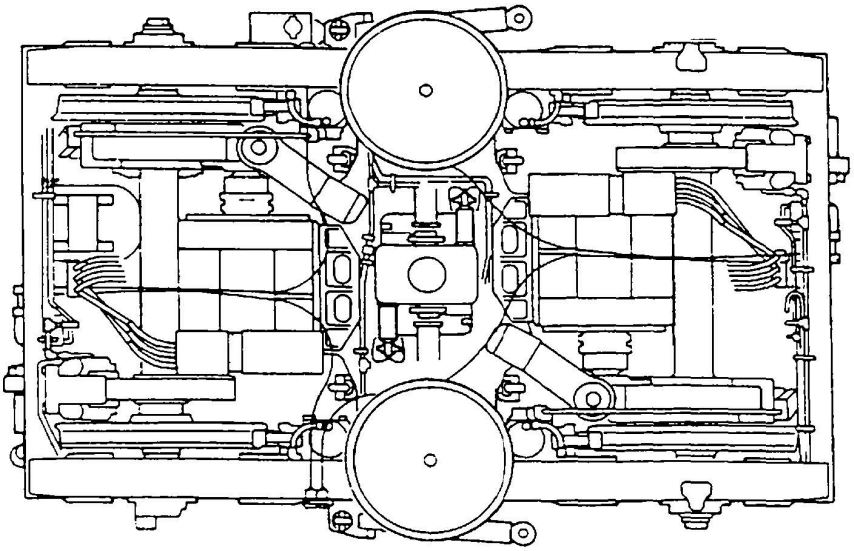


Figura 7.6: Esquema del bogie en planta

A continuación se describen en más detalle los sistemas de suspensión vertical (primaria y secundaria) que serán analizados en profundidad. El sistema de detección de fallos que se desarrolla más adelante se basa únicamente en señales que miden el movimiento vertical.

7.2.1 Suspensión primaria

La suspensión primaria actúa entre el eje y la estructura del bogie, permitiendo un movimiento relativo que aísla al resto del coche de oscilaciones, especialmente a frecuencias altas o baches puntuales.

El sistema de suspensión primaria está colocado sobre la pieza que incluye los rodamientos y las cajas de grasa donde se monta el eje (pieza gris en la figura 7.7).

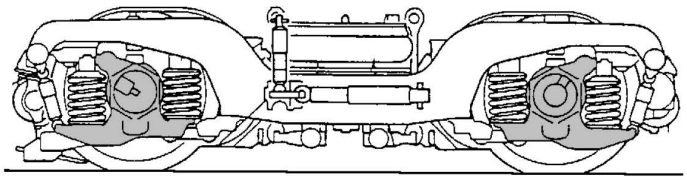


Figura 7.7: Pieza fundamental de la suspensión primaria

El movimiento de esta pieza es solidario al eje y por lo tanto sigue las irregularidades de la vía. El bogie se une a la pieza mediante dos bielas horizontales, que transmiten los esfuerzos de aceleración y frenada, y verticalmente apoya sobre ella mediante dos muelles y un amortiguador.

Los muelles utilizados son dobles, para obtener una alta rigidez en un sistema poco voluminoso. Los dos muelles, interior y exterior, tienen la misma longitud natural y por lo tanto actúan siempre de manera simultánea. La constante de rigidez conjunta del muelle doble es de 475.000 N/m y por lo tanto los dos muelles dobles que actúan sobre cada rueda suponen una rigidez de 950.000 N/m. El comportamiento global de los muelles de la suspensión primaria puede modelarse según la ecuación siguiente:

$$f_m = 0{,}95 \cdot x + 77{,}26 \qquad \textbf{(7.1)}$$

donde x es el desplazamiento (en mm) con respecto a la posición de equilibrio y f_m es la fuerza (en kN) de los muelles. La fuerza se considera positiva si es de sustentación, es decir si los muelles están comprimidos.

Las dos bielas horizontales pueden localizarse fácilmente en la figura 7.8, donde se han remarcado en color gris. Están diseñadas para permitir el movimiento relativo entre el eje y el bogie en dirección vertical e impedir movimientos relativos en dirección horizontal. Las bielas están montadas mediante *silentblocs*, que producen pares en función del ángulo de giro. Estos pares se traducen en fuerzas de sustentación que se suman a las fuerzas que producen los muelles.

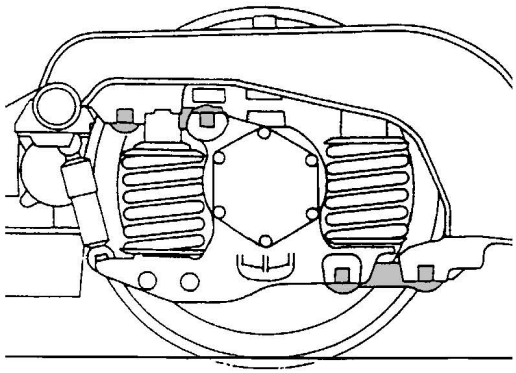


Figura 7.8: Bielas de la suspensión primaria

El efecto de las bielas en la suspensión primaria es mucho menor que el de los dos muelles dobles, por lo que puede considerarse lineal y modelarse sin mucho error de la siguiente manera:

$$f_b = 0{,}410 \cdot x - 3{,}81 \qquad \textbf{(7.2)}$$

donde x es el desplazamiento vertical (en mm) con respecto a la posición de equilibrio y f_b es la fuerza de sustentación (en kN) que se produce como consecuencia de la torsión de los *silentblocs* de las bielas.

El sistema de suspensión primaria cuenta con un tope de goma que actúa en caso de producirse una importante deformación de los muelles. La posición del tope de goma se ha remarcado en la figura 7.9.

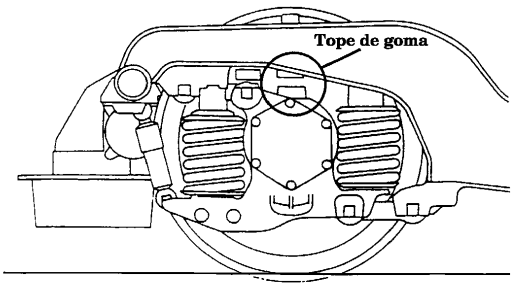


Figura 7.9: Tope de la suspensión primaria

Puede observarse que el tope no actúa en condiciones normales de marcha sino que está pensado para situaciones extremas. Debe producirse un acortamiento de los muelles de la suspensión de 30 mm (con respecto a la posición de equilibrio en vacío) para que el tope empiece a actuar. Por lo tanto, a partir de esta deformación el comportamiento de la suspensión cambia drásticamente, lo que supone una no–linealidad importante.

Por otro lado, la característica desplazamiento–fuerza del tope de goma es altamente no lineal, como se muestra en las curvas del fabricante (figura 7.10). Las fuerzas necesarias para conseguir grandes desplazamientos crecen rápidamente y la deformación máxima se estima en tan sólo 12 mm.

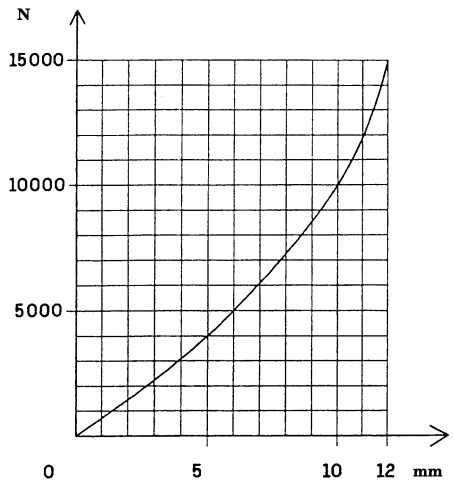


Figura 7.10: Curva característica del tope de goma de la suspensión primaria

Esta curva puede aproximarse matemáticamente mediante la ecuación 7.3. Esta ecuación ha sido obtenida por interpolación de 7 puntos leídos de la gráfica de la figura 7.10 y midiendo el desplazamiento con respecto a la posición de equilibrio en vacío.

$$ft = \frac{1{,}4128 - \sqrt{6{,}9479 - 0{,}1651x}}{0{,}0825} \qquad \text{para } x \geq 30 \qquad \textbf{(7.3)}$$

La figura 7.11 muestra la curva característica del tope de acuerdo a la ecuación 7.3. En esta figura se han remarcado los puntos utilizados para la interpolación.

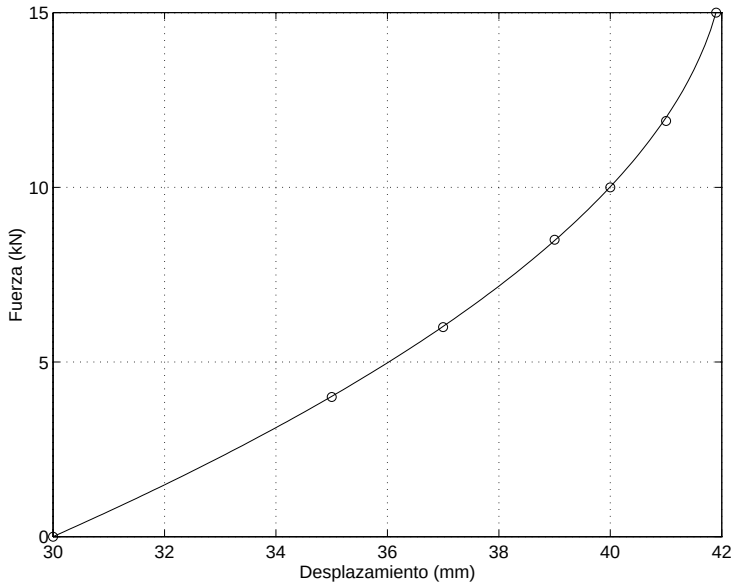


Figura 7.11: Estimación de la característica del tope de goma

Para concluir, el comportamiento elástico de la suspensión primaria se resume en la ecuación 7.4. En esta ecuación se ha considerado la rigidez conjunta de los dos muelles dobles que existen por cada rueda, el efecto de los *silentblocs* utilizados para el montaje de las bielas y el efecto del tope de goma que actúa para grandes desplazamientos de compresión.

$$\begin{aligned} f_p &= 1{,}36 \cdot x + 73{,}5 && \text{para } x \leq 30 \\ f_p &= 1{,}36 \cdot x + 73{,}5 + \frac{1{,}4128 - \sqrt{6{,}9479 - 0{,}1651x}}{0{,}0825} && \text{para } x \geq 30 \end{aligned} \quad \textbf{(7.4)}$$

En esta ecuación f_p representa la fuerza que ejerce la suspensión primaria (medido en kN), siendo x el desplazamiento en milímetros con respecto a la posición de equilibrio en vacío o VOM (vehículo en orden de marcha). La ecuación 7.4 resume el comportamiento elástico de la suspensión primaria y tiene la representación gráfica que se muestra en la figura 7.12. Es importante remarcar que las fuerzas están referidas a una sola rueda, es decir que considerando las 8 ruedas de una unidad motriz la fuerza total en vacío es de 587,6 kN. Esta fuerza equivale a un peso de 60 t, que

corresponde a la masa de la unidad motriz en vacío descontando la masa de los cuatro ejes.

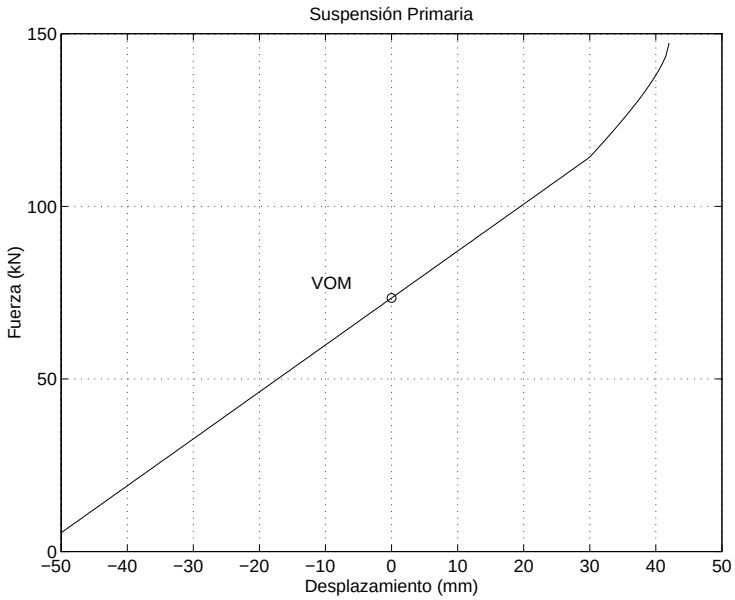


Figura 7.12: Curva característica de la suspensión primaria

Existe un amortiguador para cada rueda que absorbe la energía de los movimientos relativos entre el eje y el bogie (es el único amortiguador que aparece en la figura 7.9). Estos amortiguadores tiene el nombre de **amortiguadores anti–galope** porque están colocados de tal forma que absorben con gran eficacia los movimientos de cabeceo del bogie con respecto a la vía (en la figura 7.5 de la página 161 puede verse la disposición de los amortiguadores en cada extremo del bogie).

El amortiguador produce una fuerza positiva o negativa proporcional a la velocidad relativa entre el eje y la estructura del bogie. La característica nominal de este amortiguador es bastante lineal para velocidades bajas aunque las fuerzas son proporcionalmente menores si la velocidad aumenta, tal y como se muestra en la figura 7.13.

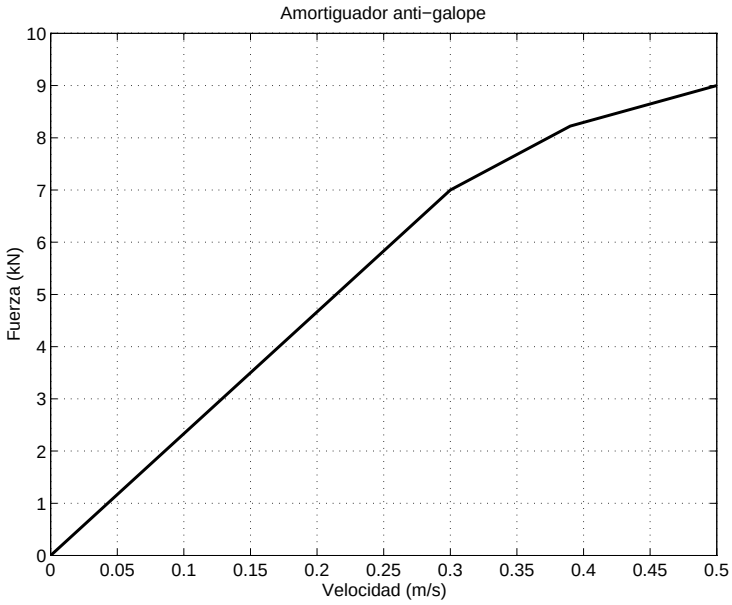


Figura 7.13: Característica del amortiguador anti–galope

Matemáticamente el comportamiento puede expresarse de la siguiente manera:

$$\begin{aligned} f_{ag} &= 23{,}333\ \dot{x} && |\dot{x}| \le 0{,}30 \\ f_{ag} &= 13{,}611\ \dot{x} + 2{,}917 && 0{,}30 \le |\dot{x}| \le 0{,}39 \\ f_{ag} &= 7{,}045\ \dot{x} + 5{,}477 && 0{,}39 \le |\dot{x}| \le 0{,}50 \end{aligned} \qquad \textbf{(7.5)}$$

donde f_{ag} es la fuerza (en kN) que ejerce el amortiguador para una velocidad relativa $\dot{x}$ (en ms^{-1}).

7.2.2 Suspensión secundaria

La suspensión secundaria actúa entre la caja y el bogie, permitiendo movimientos relativos de translación y rotación en todas las direcciones.

El sistema de suspensión secundaria de cada bogie está formado principalmente (ver figura 7.14, de izquierda a derecha) por dos balonas, dos amortiguadores verticales, dos amortiguadores anti–lazo, el sistema anti–balanceo y dos amortiguadores transversales.

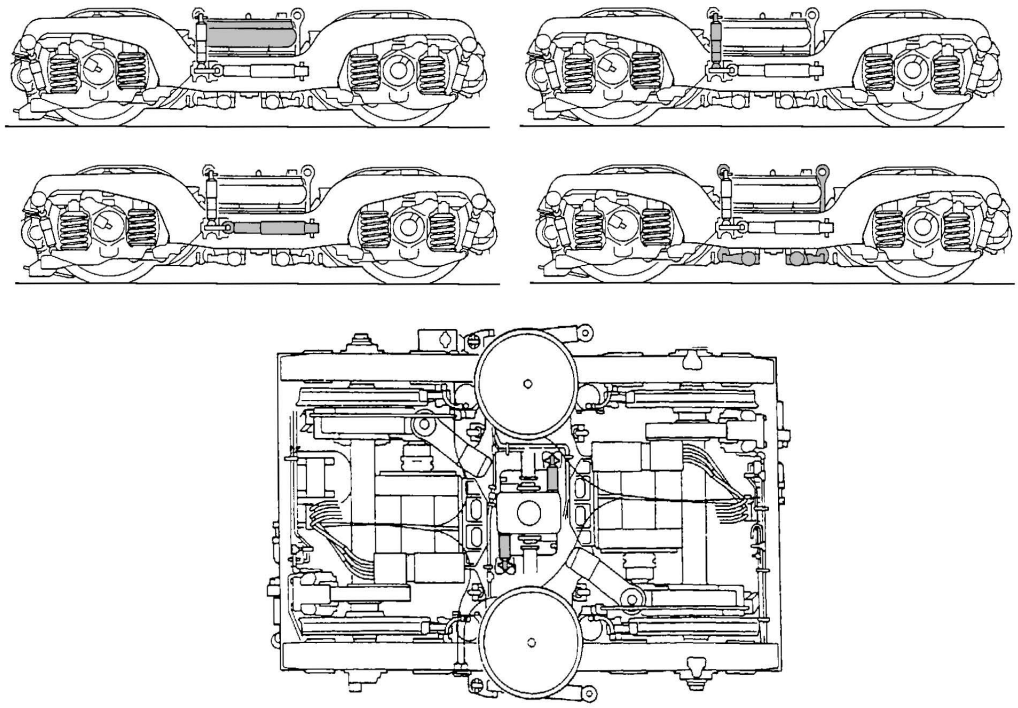


Figura 7.14: Elementos principales de la suspensión secundaria

Analizando el movimiento vertical, el componente fundamental del sistema de suspensión secundaria es la balona. Se trata de un sistema hidráulico de suspensión formado por una pastilla de caucho llena de aire a una presión aproximada de 4 bar. Los desplazamientos verticales y laterales modifican la presión interior y la forma de la balona; el comportamiento elástico no es lineal.

Para obtener una expresión matemática que represente el comportamiento de la balona se ha realizado una interpolación parabólica de datos reales leídos de las curvas experimentales proporcionadas por el fabricante. La ecuación obtenida es la siguiente:

$$f_s = 0{,}0021x^2 + 0{,}7033x + 110{,}71 \qquad \textbf{(7.6)}$$

donde f_s es la fuerza de sustentación (en kN) y x es el desplazamiento vertical (en mm) respecto a la posición de equilibrio en vacío y considerando desplazamientos positivos aquéllos que producen compresión.

La figura 7.15 es la representación gráfica de la ecuación 7.6, donde también se han dibujado los puntos utilizados para la interpolación.

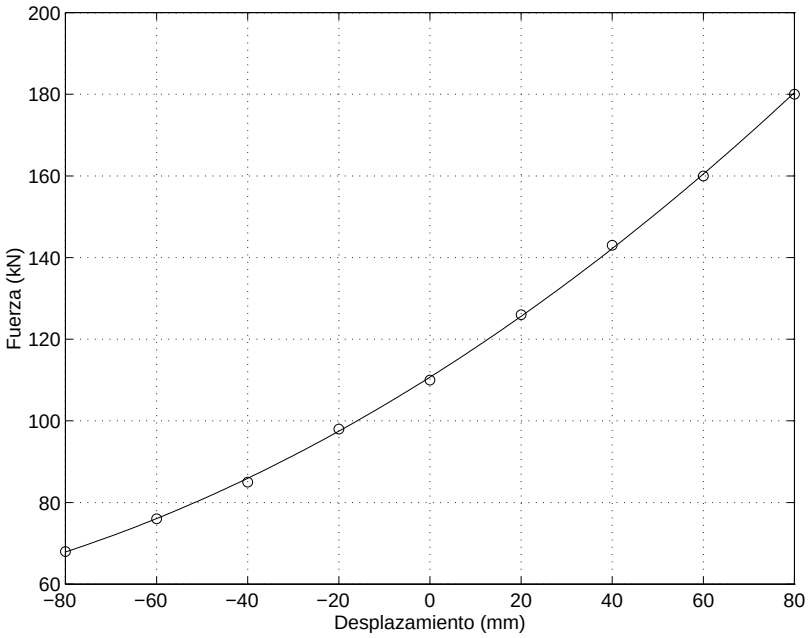


Figura 7.15: Característica elástica de la balona

Para este ejemplo sólo se han considerado los efectos elásticos y se han despreciado los efectos amortiguadores de la balona. Aunque el efecto amortiguador en sistemas de muelles siempre es despreciable, en sistemas de suspensión hidráulica el efecto puede ser importante. En este caso, sin embargo, se ha considerado que la energía absorbida por la balona es despreciable frente al efecto de los amortiguadores verticales.

Internamente la balona cuenta con un sistema de muelles, como puede apreciarse en el gráfico de sección de la figura 7.16. Este conjunto de muelles sólo actúa cuando no hay presión en la balona o en caso de producirse grandes desplazamientos, se denomina suspensión de socorro. Los muelles se montan comprimidos y no actúan durante el movimiento vertical normal ni interfieren con los movimientos transversales, en condiciones normales todos los movimientos de la suspensión secundaria dependen exclusivamente de las características de la balona. En caso de producirse grandes desplazamientos verticales, la chapa superior choca contra el conjunto de muelles, que actúa como un tope de seguridad.

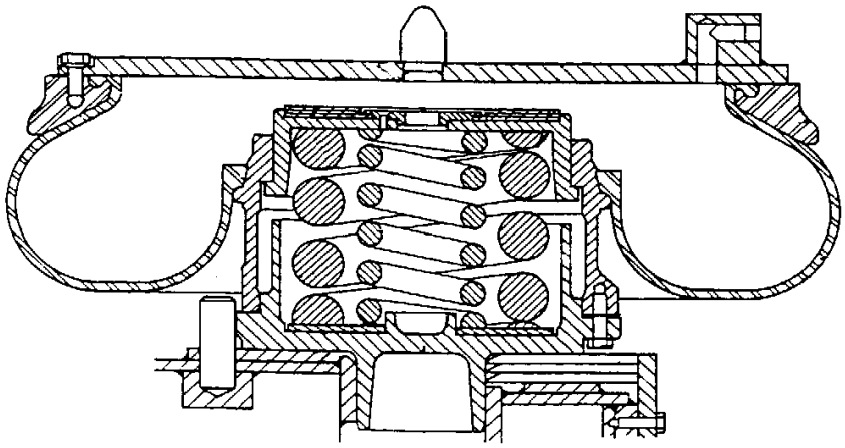


Figura 7.16: Sección de la balona

El desplazamiento que debe producirse para que el muelle doble empiece a actuar es de 57 mm, esto supone una fuerza adicional de 50 kN que equivale a un peso extra de la caja de casi 20 t (incremento del 45%). Alcanzada esta situación, los muelles empiezan a actuar con una rigidez conjunta de 4,33 kN/mm, que es muy superior a la rigidez media de la balona en ese momento (aproximadamente de 1 kN/mm). Pero el efecto más importante es la necesidad de aplicar una fuerza adicional de 117 kN (precarga del conjunto de muelles) para iniciar la deformación de la suspensión de socorro. Esto supone una alta no–linealidad (figura 7.17) que sólo se ve amortiguada por una goma que evita el choque directo de las dos chapas metálicas.

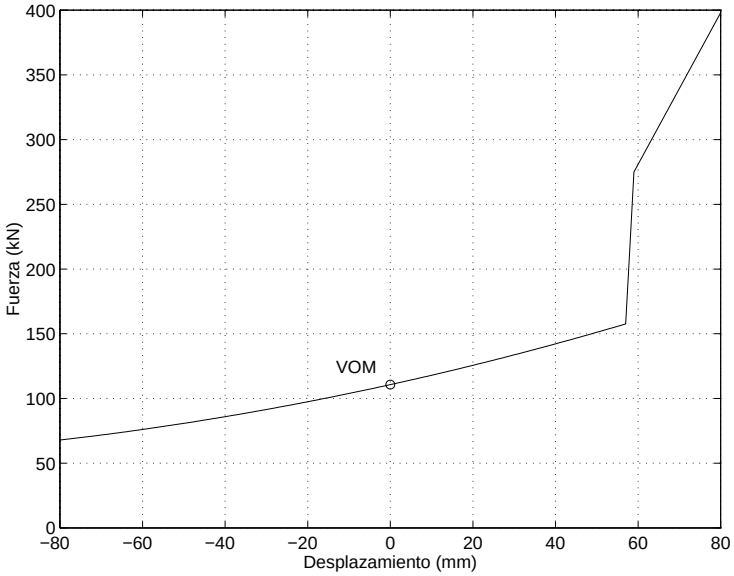


Figura 7.17: Curva característica de la suspensión secundaria

El único amortiguador del sistema de suspensión secundaria que afecta al movimiento vertical es el llamado **amortiguador vertical** (segundo dibujo de la figura 7.14, página 169). Este amortiguador tiene una característica de fuerza–velocidad bastante lineal mientras las velocidades sean bajas, como puede observarse en la figura 7.18.

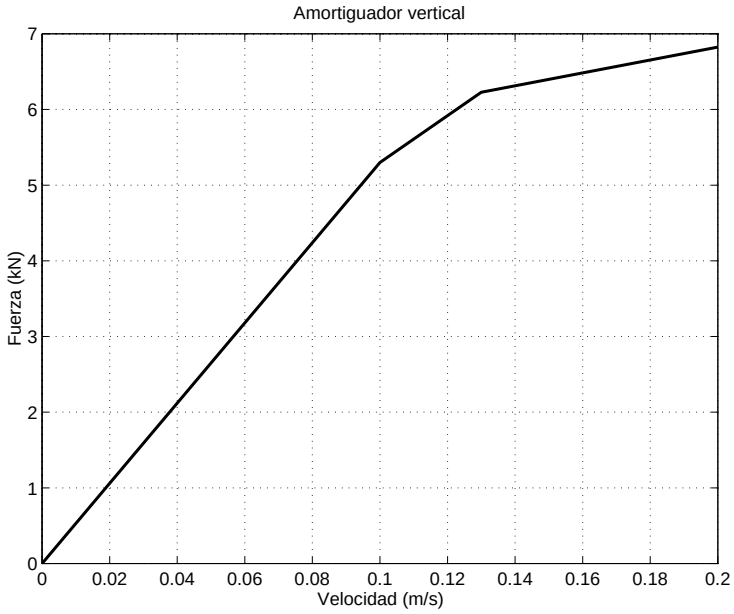


Figura 7.18: Curva característica del amortiguador vertical

Este comportamiento se describe matemáticamente de acuerdo al conjunto de ecuaciones 7.7.

$$\begin{aligned} f_{av} &= 53\ \dot{x} && |\dot{x}| \le 0{,}10 \\ f_{av} &= 30{,}9333\ \dot{x} + 2{,}2067 && 0{,}10 \le |\dot{x}| \le 0{,}13 \\ f_{av} &= 8{,}5143\ \dot{x} + 5{,}1211 && 0{,}13 \le |\dot{x}| \le 0{,}20 \end{aligned} \qquad \textbf{(7.7)}$$

donde f_{av} representa la fuerza (en kN) ejercida por el amortiguador y $\dot{x}$ es la velocidad de desplazamiento (en m/s).

7.3 Modelo de simulación

Se ha desarrollado un modelo del comportamiento dinámico de la unidad motriz que permite generar los datos de movimiento vertical en diferentes condiciones de funcionamiento. El modelo está basado en las características de los componentes fundamentales de los sistemas de suspensión descritos en el apartado 7.2 y en las masas reales de los trenes.

Para no complicar innecesariamente la estructura del modelo ni el proceso de simulación de datos, sólo se ha tenido en cuenta el movimiento vertical del coche con respecto a la vía. Por lo tanto no se utilizan los datos de momento de inercia de la caja ni de los bogies. Realmente sólo se produciría un movimiento vertical puro si todas las ruedas recibiesen exactamente la misma excitación y al mismo tiempo (sabiendo que las características de los elementos de suspensión y la distribución de masas son uniformes). En un caso real, todas las perturbaciones que actúan directamente sobre cualquier rueda se transmiten a todos los elementos de suspensión del tren por efecto de rotación. Sin embargo, la inercia a la rotación hace que la influencia de una perturbación se centre fundamentalmente en la rueda afectada.

El movimiento de cabeceo se produce, en condiciones normales de marcha, como consecuencia de perturbaciones verticales en las vías. Este movimiento puede verse realimentado ya que existe un retraso entre las excitaciones que afectan a las ruedas delanteras y a las traseras. Un pequeño movimiento de cabeceo provocado por un bache en las ruedas delanteras puede intensificarse de nuevo, dependiendo de la severidad del bache y de la velocidad del tren, cuando el bache afecte a las ruedas traseras.

La fuente principal de excitación de movimientos de balanceo son las imperfecciones de las vías en dirección perpendicular a la marcha. Este movimiento también puede verse realimentado si en algún momento se producen perturbaciones con la cadencia equivalente a la frecuencia de resonancia del movimiento de balanceo.

Tanto el movimiento de cabeceo como el de balanceo se manifiestan en forma de movimientos verticales de los sistemas de suspensión primario y secundario. Por lo tanto el análisis aislado del movimiento vertical es una simplificación razonable si se utiliza un amplio conjunto de funciones de excitación. Realmente sólo se podrían simular mejores conjuntos de datos si además de complicar el modelo del tren, con efectos de rotación, se conociese exactamente la ecuación tridimensional de la vía.

El modelo simplificado de la suspensión de la unidad motriz es el conjunto de masas, muelles y amortiguadores que se muestra en la figura 7.19.

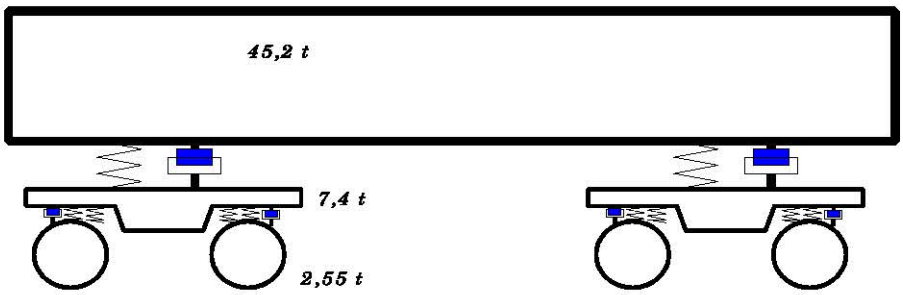


Figura 7.19: Modelo simplificado de la suspensión

Dado que el reparto de masas es uniforme y que sólo se va a considerar el movimiento vertical, el modelo puede reducirse al análisis de una sola rueda. Para llevar a cabo esta simplificación es necesario obtener las masas y rigideces equivalentes de todos los componentes. La tabla siguiente recoge los valores reales de masa de los componentes de la unidad motriz.

Tabla 7.3: Masas de la unidad motriz

Masa total VOM	70,2 t
Masa de la caja	45,2 t
Masa de un bogie con ejes	12,5 t
Masa de un eje (incluyendo frenos, reductora, etc)	2,55 t

De estos datos se pueden obtener los valores totales de masa soportada por cada sistema de suspensión. La tabla 7.4 muestra las masas totales y su reducción a una sola rueda de todos los componentes de la suspensión. También se incluye la manera de calcular la fuerza de cada componente de acuerdo a las ecuaciones del apartado 7.2.

La fuerza de sustentación de la suspensión primaria debida a la rigidez viene dada en función del desplazamiento por f_p (ecuación 7.4, p. 166). Esta ecuación está referida a una sola rueda ya que incluye el efecto de los dos muelles dobles, de los *silentblocs* y del tope de goma. La fuerza producida por el amortiguador anti–galope f_{ag} (ecuación 7.5, p. 168) también se aplica directamente ya que existe un amortiguador por cada rueda.

En el caso de la suspensión secundaria, existen dos balonas y dos amortiguadores verticales en cada bogie, por lo tanto las fuerzas de cada uno de estos elementos afectan en partes iguales a dos ruedas. La reducción a una sola rueda supone dividir por dos los valores de la fuerza producida por la balona f_s (ecuación 7.6, página 169) y la fuerza producida por el amortiguador vertical f_{av} (ecuación 7.7, página 173).

Tabla 7.4: Características de la suspensión

Suspensión primaria

Masa total sobre la suspensión primaria	60 t
Masa total sobre cada rueda (60/8)	7,5 t
Masa del bogie por cada rueda (7,4/4)	**1,85** t
Fuerza de origen elástico: f_p	
Rigidez media en vacío	1,36 kN/mm
Capacidad (flexibilidad)	**0,7353 10^{-6}** N/m
Fuerza de origen dinámico: f_{ag}	
Resistencia (amortiguamiento) nominal del amortiguador anti–galope ($\dot{x}$ =0,39 m/s)	21,09 kN/ms^{-1}
	21.090 N/ms^{-1}

Suspensión secundaria

Masa total sobre la suspensión secundaria	45,2 t
Masa sobre cada balona (45,2/4)	11,3 t
Masa equivalente por cada rueda (45,2/8)	**5,65** t
Fuerza que ejerce la balona: f_s	
Rigidez media en vacío	0,70 kN/mm
Fuerza del amortiguador vertical: f_{av}	
Resistencia (amortiguamiento) nominal	47,91 kN/ms^{-1}
Fuerza estática por cada rueda: $f_s/2$	
Rigidez media en vacío	0,35 kN/mm
Capacidad (flexibilidad)	**2,8571 10^{-6}** N/m
Fuerza dinámica por cada rueda: $f_{av}/2$	
Resistencia (amortiguamiento) nominal del amortiguador vertical ($\dot{x}$ =0,13 m/s)	23,96 kN/ms^{-1}
	23.960 N/ms^{-1}

El modelo de la unidad motriz, reducido a una sola rueda, es un sistema de dos grados de libertad formado por dos masas acopladas mediante muelles y amortiguadores (figura 7.20). Las masas equivalen al bogie (*Mb*) y a la caja (*Mc*); el muelle *Cp* y el amortiguador *Rp* representan el sistema de suspensión primaria, y el muelle *Cs* y el amortiguador *Rs* representan la suspensión secundaria. El modelo de bond–graph equivalente se obtiene de manera sencilla.

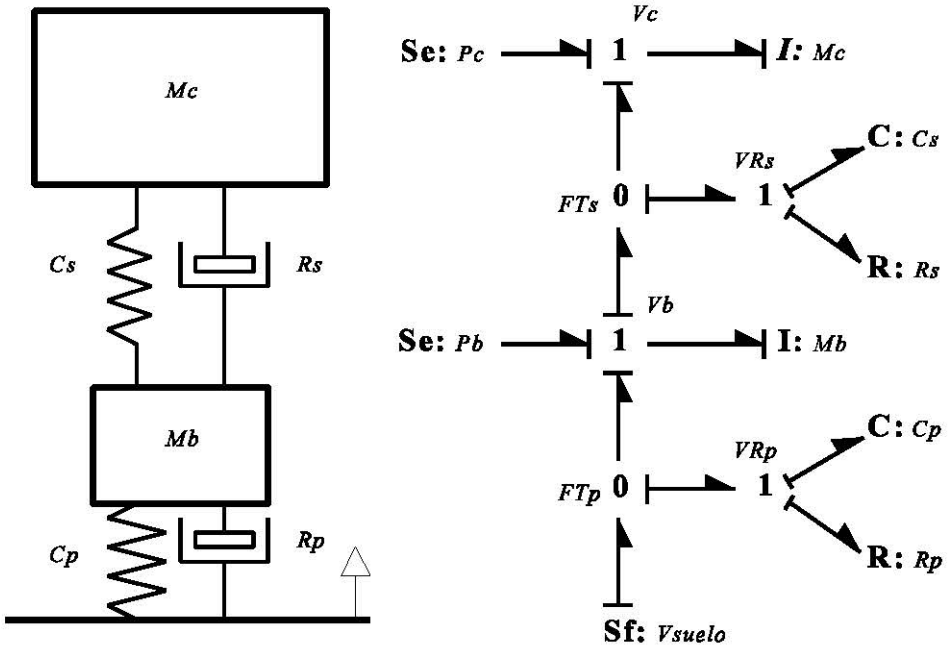


Figura 7.20: Esquema del modelo reducido a una rueda

En la figura se han incluido los parámetros asociados a los puertos **Se**, **I**, **C**, y **R**; en una primera aproximación pueden utilizarse los valores medios de las características (reducidos a una sola rueda) mostrados anteriormente en la tabla 7.4. Estas características, expresadas en el Sistema Internacional y de acuerdo a la nomenclatura de la figura 7.20, son las siguientes:

Mc	5.650	kg	Masa de la caja
Mb	1.850	kg	Masa del bogie
Cp	$0{,}7353e^{-6}$	m/N	Capacidad de la suspensión primaria
Rp	21.090	N/ms^{-1}	Resistencia de la suspensión primaria

Cs	2,8574e^{-6}	m/N	Capacidad de la suspensión secundaria
Rs	23.960	N/ms^{-1}	Resistencia de la suspensión secundaria
Pc	*Mc g*	N	Fuente de esfuerzo, peso de la caja
Pb	*Mb g*	N	Fuente de esfuerzo, peso del bogie

También se han incluido en la figura 7.20 los nombres de las variables asociadas a los puertos **1** y a los puertos **0**. Todas las flechas que llegan o salen de un puerto **1** tiene la misma velocidad y todas las que llegan o salen de un puerto **0** tienen la misma fuerza (ver apéndice A). La gran mayoría de los datos de flujo y esfuerzo del modelo quedan definidos con estas variables.

Vsuelo	[m/s]	Velocidad del suelo
FTs	[N]	Fuerza total en la suspensión secundaria
Ftp	[N]	Fuerza total en la suspensión primaria
Vb	[m/s]	Velocidad absoluta del bogie
Vc	[m/s]	Velocidad absoluta de la caja
VRs	[m/s]	Velocidad relativa de la suspensión secundaria
Vrp	[m/s]	Velocidad relativa de la suspensión primaria

Hay que tener en cuenta que en las simulaciones realizadas no se han utilizado los valores medios de las características de la suspensión sino las ecuaciones descritas en el apartado 7.2, lo que da lugar a un comportamiento no lineal que representa la realidad con mayor precisión. El modelo de Bond–graph es el mismo en cualquier caso, sólo cambiarán los parámetros de los puertos **C** y **R**.

También hay que tener en cuenta que durante el funcionamiento normal del tren se producen variaciones en la masa *Mc* debidas a la subida y bajada de viajeros. Aunque la máxima variación de carga supone una variación del parámetro *Mc* inferior al 30%, se han efectuado simulaciones con distintos valores para obtener datos más realistas.

7.3.1 Datos de funcionamiento normal

Los datos de funcionamiento normal se generan introduciendo diferentes funciones de excitación en el modelo de simulación. En el caso de disponer

de un tren instrumentado, la generación de datos en funcionamiento normal equivale a recoger información con el sistema de adquisición de datos haciendo circular el tren por distintos trazados y en diferentes condiciones de carga.

Las características del trazado quedan representadas por la ecuación de la vía que se introduce en el modelo mediante la variable *Vsuelo*, que es la única entrada del modelo de simulación. A partir de las perturbaciones que esta señal produce, el sistema de suspensión reacciona dando lugar a los datos instantáneos de posición de las dos masas que se analizan más adelante. En otro tipo de estudios teóricos y de simulaciones se suelen utilizar señales de entrada estándar (por ejemplo la función escalón) para analizar la respuesta de los modelos. En este caso no se pretende analizar la calidad del sistema de suspensión, ni su función de respuesta en frecuencia, ni el tiempo de establecimiento, etc. El objetivo es obtener una señal que represente lo mejor posible las condiciones normales de movimiento.

Al no disponer de datos reales del trazado de las vías, para este ejemplo se han utilizado ecuaciones de vía generadas aleatoriamente. Basándose en un generador aleatorio con distribución normal se han obtenido diferentes alturas de la vía en función del tiempo (o de la distancia recorrida equivalente). Sin embargo, dado que la vía no puede cambiar su altura bruscamente, se han filtrado las componentes de frecuencia alta. La figura 7.21 muestra la comparación entre una señal aleatoria pura y la misma señal filtrada, que es la que se utiliza como ecuación del perfil de la vía. Esta gráfica muestra la ecuación del perfil de la vía durante 5 s, lo que equivale a 100 m de distancia cuando el tren circula a una velocidad de 72 km/h. La manera de acondicionar las ecuaciones de vía ha sido aplicar un filtrado de tipo exponencial que elimine progresivamente las frecuencias altas. La figura 7.22 muestra el espectro entre 0 y 15 Hz de un conjunto de muestras generadas por este procedimiento. Puede observarse que para frecuencias superiores a 5 Hz las amplitudes de oscilación son ciertamente pequeñas (5 Hz es la frecuencia equivalente a oscilaciones de 1 m de longitud cuando el tren circula a una velocidad de 72 km/h).

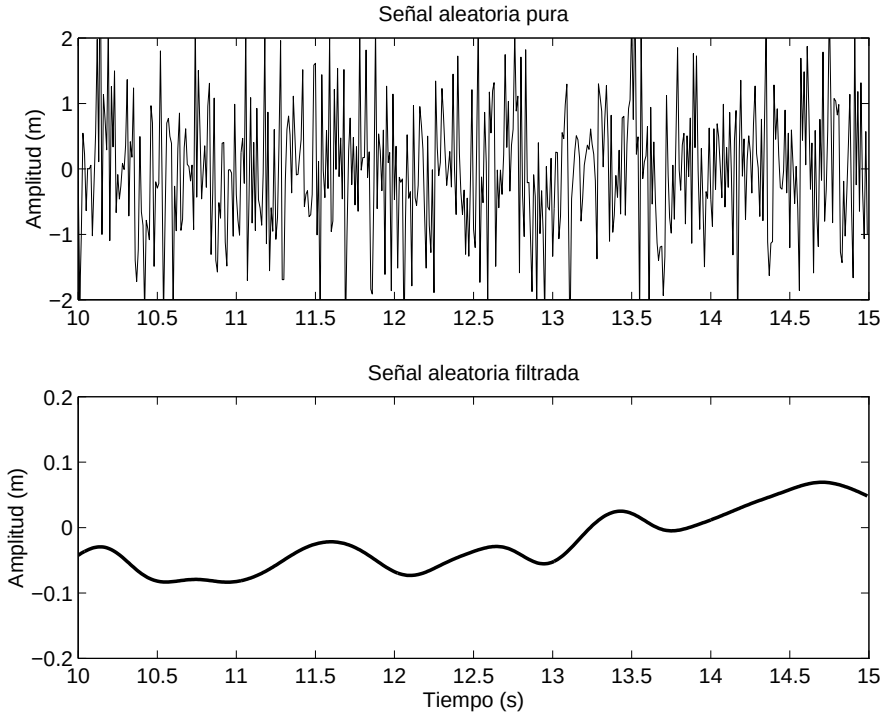


Figura 7.21: Ejemplo de ecuación de vía utilizado para simular datos

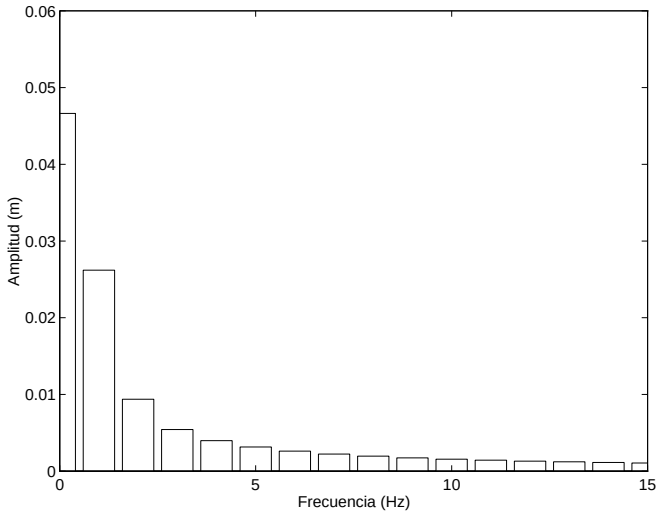


Figura 7.22: Ejemplo de un espectro en frecuencia de la vía

Por último se ha impuesto que la altura de la vía al principio de la simulación sea siempre cero. De esta manera se evitan discontinuidades al empezar la simulación, ya que los cálculos siempre comienzan en condiciones iniciales nulas de posición y velocidad.

Durante el funcionamiento normal, las características de los componentes de la suspensión (*Cp*, *Rp*, *Cs* y *Rs*) permanecen invariantes con el tiempo. Los valores de las masas *Mc* y *Mb* también se mantienen constantes durante cada simulación, ya que no se producen variaciones de masa durante la marcha del tren. Se han realizado simulaciones con distintos valores de masa, pero siempre partiendo de condiciones iniciales nulas de posición. Todos los desplazamientos están medidos con respecto a la posición de equilibrio y por lo tanto cuando se cambia la masa *Mc* entre dos simulaciones, cambian los valores de precarga de los muelles, pero las posiciones y velocidades iniciales son nulas. Por lo tanto cada simulación equivale a un tramo entre dos estaciones durante el cual la masa permanece constante. El resultado final es un conjunto de historias de datos que reflejan la reacción normal del sistema de suspensión del tren para diferentes condiciones de vía y diferentes condiciones de masa.

La figura 7.23 muestra un ejemplo de un pequeño conjunto de datos obtenidos por simulación. La línea en trazo grueso es la posición del suelo en función del tiempo, la línea fina continua es la posición del bogie respecto a la posición inicial y la línea discontinua es la posición de la caja. Puede observarse que el bogie sigue con bastante fidelidad las oscilaciones del suelo, aunque con un poco de retraso y con los habituales sobrepasos. En el movimiento de la caja también se observa un retraso en la respuesta a las perturbaciones producidas por el suelo y sobre todo que los desplazamientos son mayores aunque algo más lentos. En definitiva el sistema de suspensión filtra correctamente las perturbaciones introducidas por el suelo y hace que los viajeros oscilen verticalmente a frecuencias próximas a 1 Hz, que resultan más confortables aunque las amplitudes sean mayores.

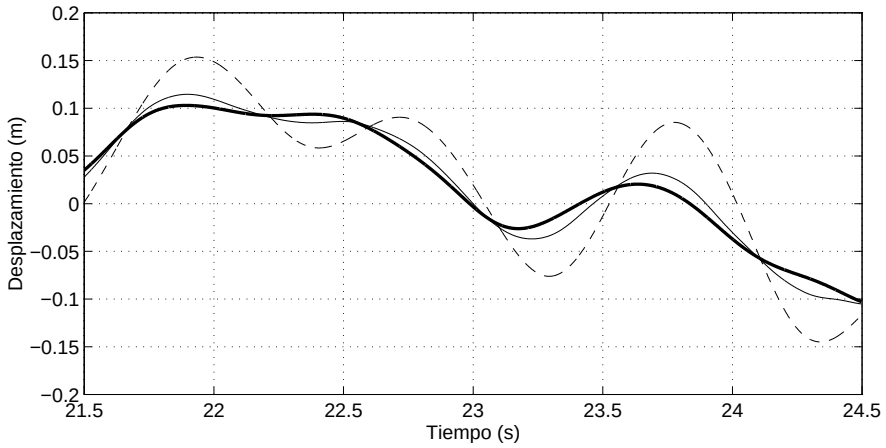


Figura 7.23: Simulación de las posiciones de la vía, del bogie y de la caja

Los datos recogidos por un tren instrumentado tendrían el gran valor de tratarse de datos reales, lo que puede incluir efectos no considerados en el modelo de simulación, sin embargo haría falta bastante tiempo para conseguir un amplio conjunto de condiciones de funcionamiento, especialmente si los tramos recorridos durante la adquisición de datos se repiten. Por el contrario, el método de generación de ecuaciones de la vía basado en la utilización de números aleatorios garantiza una gran riqueza de datos. Esto permite estudiar una gran variedad de situaciones diferentes y despeja dudas sobre una posible especialización del sistema de detección incipiente de fallos es el análisis de una situación concreta. Una sola trama de datos generada de esta manera proporciona información suficiente sobre el comportamiento del tren en multitud de situaciones (tramos buenos y malos a velocidades lentas y rápidas). A modo de ejemplo se muestra a continuación una de las situaciones especiales que se producen durante la simulación. Se trata de una situación extrema en la cual se produce una compresión tal en los sistemas de suspensión que provoca un choque contra los elementos elásticos de emergencia (tope de goma y suspensión de socorro). Esta situación da lugar a importantes no–linealidades en el comportamiento del sistema que dificultan la detección de fallos por los métodos tradicionales. Aunque estas situaciones no se presenten con mucha frecuencia en la realidad, es necesario considerarlas para poder realizar un ajuste realista del sistema de detección. Gracias a la generación de

ecuaciones de vía a partir de datos aleatorios se producen comportamientos extremos del sistema de suspensión en suficiente cuantía (sobre todo cuando se simulan datos de funcionamiento a plena carga). Se garantiza por lo tanto una gran riqueza de datos aunque el volumen de los mismos no sea muy grande.

La figura 7.24 muestra una gráfica de posiciones durante 4 s de funcionamiento que da lugar a una situación extrema. En este ejemplo se observa que la posición del suelo (línea gruesa) tiene unas oscilaciones a una frecuencia parecida a la frecuencia de resonancia de la suspensión

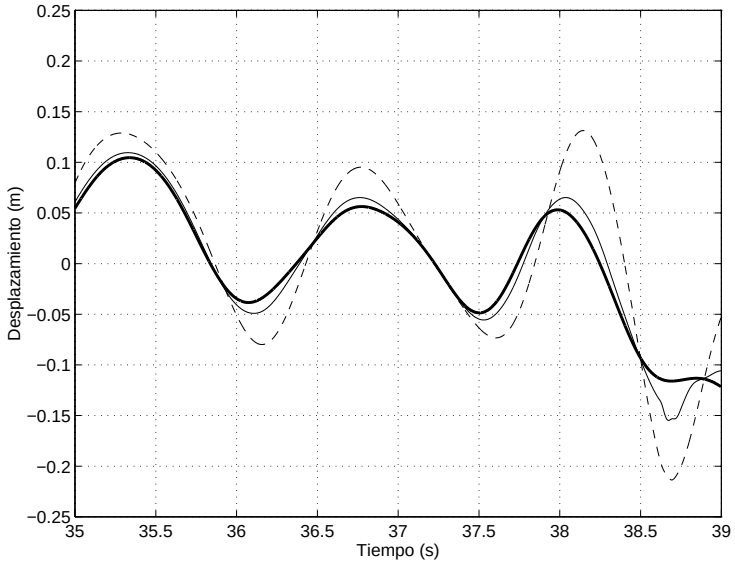


Figura 7.24: Situación extrema producida durante la simulación

secundaria. Como consecuencia se aprecia un aumento en la amplitud de las oscilaciones de la caja (línea discontinua) que finalmente da lugar a un choque de la suspensión. Este choque se produce durante el segundo 38, en un momento en el cual el suelo se estabiliza mientras la caja se encuentra bajando con bastante velocidad. A pesar de que la vía deja de bajar, las fuerzas producidas por los sistemas de suspensión no son capaces de frenar la gran inercia de la caja, que comprime la balona hasta el punto de chocar contra la suspensión de socorro. Tras el choque contra la suspensión de socorro, el bogie se ve empujado hacia abajo provocando un segunda discontinuidad al apoyar sobre el tope de goma.

En la figura 7.25 se aprecia con mayor claridad el momento en el que los sistemas de emergencia empiezan a actuar. La primera gráfica muestra las curvas de la fuerza elástica ejercida por la suspensión primaria (línea

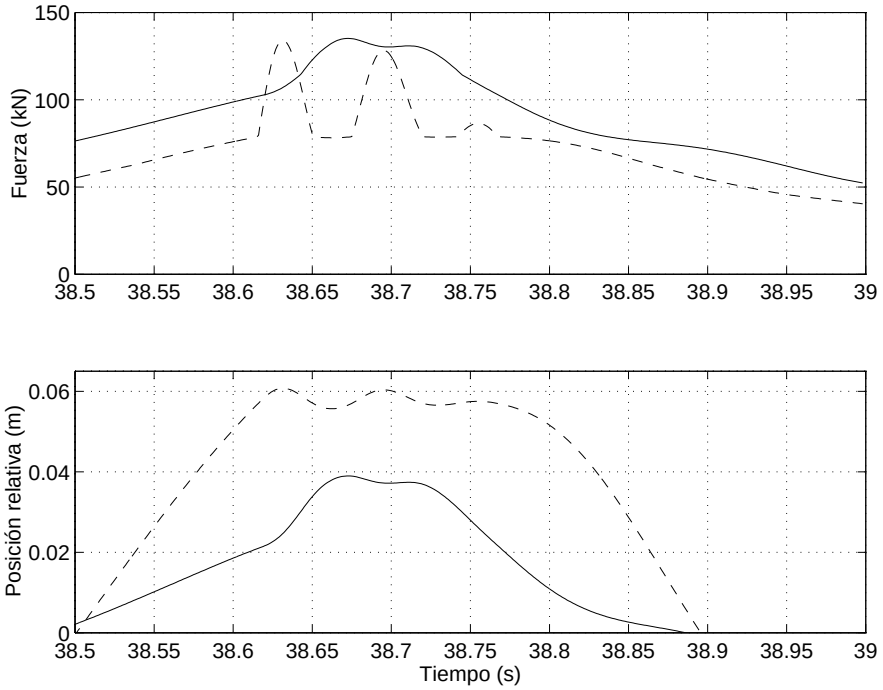


Figura 7.25: Fuerzas y desplazamientos durante el choque de la suspensión

continua) y secundaria (línea discontinua). La segunda gráfica muestra las posiciones relativas, o acortamientos, de ambos sistemas de suspensión. Cuando la suspensión secundaria se acorta en más de 57 mm entra en acción la suspensión de socorro, lo que da lugar a un brusco incremento de la fuerza. El tope de goma de la suspensión primaria actúa para desplazamientos relativos mayores de 30 mm, aumentando rápidamente la fuerza de sustentación desde 110 kN hasta 140 kN aproximadamente. Los pequeños rebotes que aparecen en la suspensión secundaria son consecuencia de la inercia del bogie y de no haber modelado el amortiguamiento que produce la lámina de goma de la suspensión de socorro.

7.3.2 Datos de degradación

Para simular datos de degradaciones se ha utilizado el mismo modelo de simulación y el mismo tipo de ecuaciones de vía que en el caso de datos de funcionamiento normal. La única diferencia ha sido la variación con el tiempo de las características de determinados componentes del sistema de suspensión. Estas variaciones permiten reproducir el comportamiento del tren durante el envejecimiento de los componentes hasta una situación en la que el fallo es evidente.

Realmente no existe información sobre el proceso real de degradación de los componentes de la suspensión de las unidades UT-450 y UT-451. Este tipo de información puede obtenerse mediante ensayos periódicos, pero suficientemente frecuentes, de los distintos componentes. Sería interesante analizar la relación existente entre la degradación de los componentes y el trabajo al que se han visto sometidos, tal y como se propone en [Sanz92] y [Kobbacy97]. Con estos datos se obtendría la auténtica curva de degradación en función de las condiciones de utilización del tren. Esta información se convertiría en otro criterio que ayudaría al sistema de detección de fallos en su proceso de diagnóstico.

También es posible estimar las características de los componentes sin tener que desmontarlos, mediante el análisis de los datos recogidos por trenes instrumentados en un trazado patrón o mediante ensayos específicos. Sin embargo, en el caso de disponer de trenes instrumentados que recojan datos del funcionamiento del sistema de suspensión, no sería necesario simular los datos de degradación sino que con el tiempo se iría creando una buena base datos de fallos del proceso.

Los datos de degradación que se han generado se centran en el envejecimiento de los amortiguadores, ya que son los elementos del sistema de suspensión que mayor deterioro suelen sufrir. Se ha simulado la degradación del amortiguador vertical que actúa sobre la suspensión secundaria. Para ello se ha definido una **curva general de degradación** (figura 7.26) que se utiliza para modificar las características de los amortiguadores con el tiempo. Inicialmente esta curva vale uno, pero a partir de un cierto momento disminuye progresivamente hasta cero. El punto en el que la curva empieza a disminuir, llamado **fin de la condición normal** (FCN en la curva de la figura 7.26), representa el momento en el

que aparece un fallo incipiente; este fallo evoluciona aumentando su influencia hasta la situación de fallo total en que la curva vale cero.

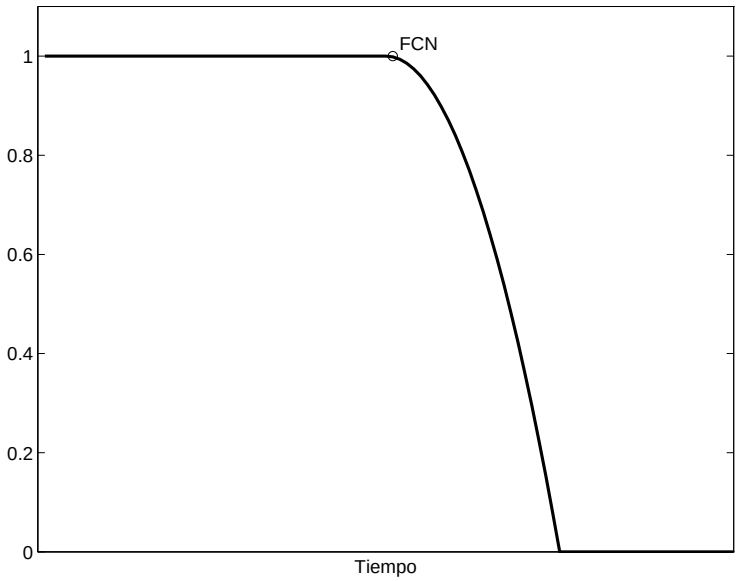


Figura 7.26: Curva general de degradación

La manera de obtener las **curvas de degradación** de los amortiguadores ha sido multiplicar su curva característica por la curva general de degradación. Inicialmente la curva característica no sufre ninguna alteración, ya que no existe fallo de ningún tipo. Cuando aparece el fallo, la característica velocidad–fuerza sufre una disminución (para igual velocidad la fuerza que se ejerce es menor), lo que equivale a una pérdida de eficacia del amortiguador. En la situación final, donde la curva general de degradación vale cero, la curva característica del amortiguador también se hace cero y por lo tanto el amortiguador no ejerce ningún efecto, equivale a haberlo eliminado del sistema. La figura 7.27 representa la curva de degradación del amortiguador anti–galope.

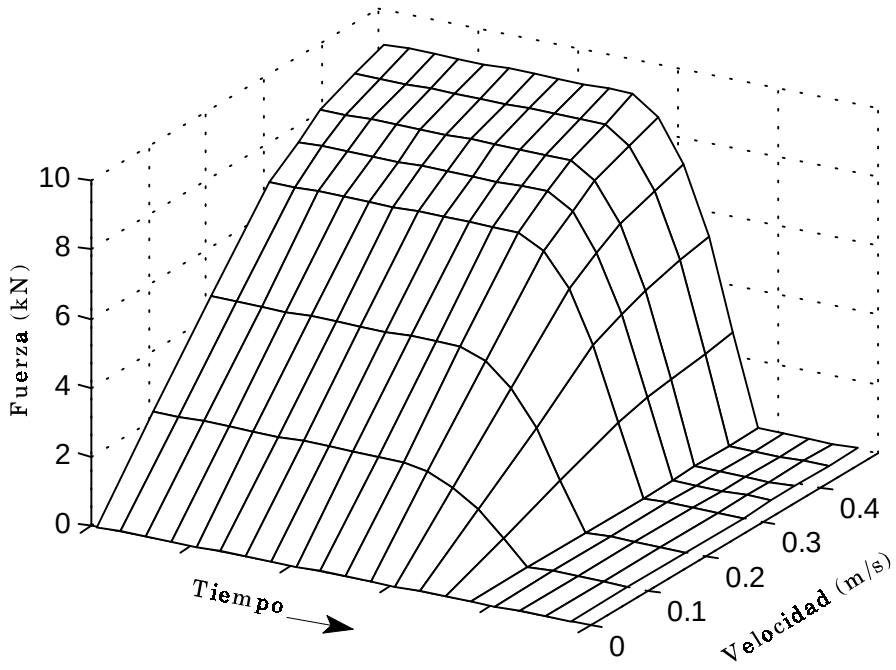


Figura 7.27: Curva de degradación del amortiguador anti–galope

La simulación de historias de fallos siempre resulta complicada ya que la forma de la curva general de degradación y su escala de tiempo son desconocidas y generalmente variables en la realidad. El procedimiento que finalmente se ha adoptado ha sido la aplicación de las curvas de degradación a un conjunto amplio de condiciones de vía y de masa. De esta manera se obtienen datos de funcionamiento normal y datos a todos los niveles de degradación, para diferentes condiciones de funcionamiento.

Cada simulación dura 100 s, lo que equivale a 2 km a una velocidad de 72 km/h. También se puede considerar que el tren circula a una velocidad diferente y por lo tanto la distancia equivalente cambia. Puesto que las oscilaciones de la vía están referidas al tiempo, su equivalencia en metros depende de la velocidad que se considere. Aunque el tiempo de simulación en cada caso no es muy largo, hay que tener en cuenta que las ecuaciones de vía generadas aleatoriamente proporcionan gran riqueza de información. Por otro lado la velocidad con que se produce la degradación del amortiguador durante las simulaciones tampoco es real, pero permite obtener información sobre el comportamiento del proceso en todos los

niveles de degradación. A pesar de que el sistema de detección no se basa en características relacionadas con la velocidad de degradación, se han simulado los datos utilizando diferentes curvas generales de degradación. Esto evita pensar que el sistema de detección resultante pueda estar especializado en una forma determinada de la curva.

Al aplicar cada curva de degradación se obtienen simultáneamente los datos de condición normal (antes de FCN), datos de situación de fallo (al final de la curva) y datos a todos los niveles de degradación; se llamará *training set* o historia de fallo al conjunto de datos completo. Se ha generado en total 90 *training sets*, que se utilizarán para ajustar y comprobar el sistema de detección de fallos. Son los correspondientes a 10 ecuaciones de vía, recorridas con 3 condiciones de masa y aplicando para cada caso 3 curvas de degradación.

7.4 Sistema de detección de fallos

El sistema de detección incipiente de fallos debe basarse en señales que puedan adquirirse de forma continua durante el funcionamiento del tren. Se podría pensar en un amplio conjunto de medidas que permitiesen conocer con facilidad el estado de salud de todos los componentes de la suspensión, sin embargo en un caso real no es posible disponer de todas las medidas (por problemas físicos de instalación de sensores y por limitaciones de presupuesto). Esta tesis propone una metodología para realizar un análisis sofisticado de las medidas de un proceso con objeto de aprovechar al máximo toda la información disponible. Por lo tanto en el ejemplo que aquí se describe no se hace un análisis de las medidas necesarias para obtener un diagnóstico completo ni se suponen conocidas muchas variables disponibles durante la simulación pero difícilmente medibles en la realidad (por ejemplo la fuerza ejercida por cada amortiguador). En adelante se va a suponer que la única variable disponible para realizar la detección incipiente de fallos es la distancia relativa entre la caja y la vía. Esta medida puede obtenerse con facilidad y de manera económica mediante la instalación de un sensor de desplazamiento (por infrarrojos, ultrasonidos u otra tecnología de medida sin contacto). Más adelante se demuestra que la información aportada por esta medida es suficiente para detectar los fallos simulados.

7.4.1 Estructura del modelo

El modelo matemático utilizado en el sistema de detección incipiente de fallos contempla la posibilidad de utilizar varias señales, y tanto sus valores más recientes como valores del pasado. El objetivo de este modelo es describir matemáticamente o mediante algoritmos el comportamiento del proceso, para poder estimar las variables de salida a partir de las variables de entrada. En el caso más general se considera que el proceso responde al esquema de la figura 7.28

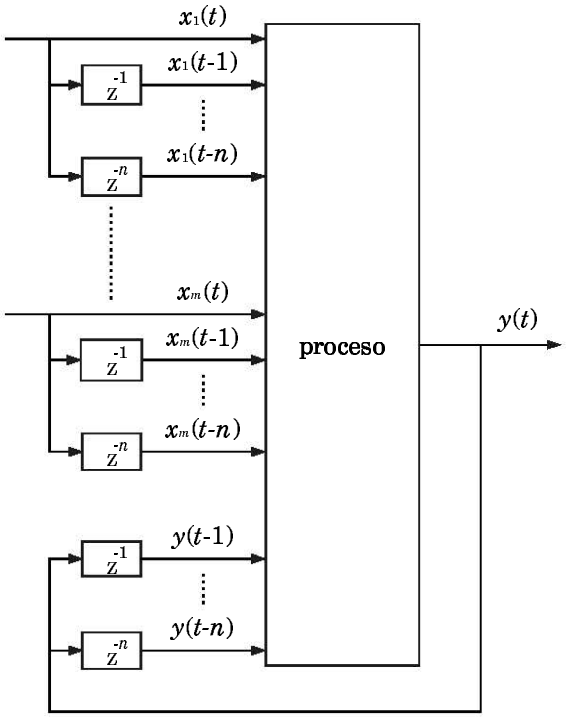


Figura 7.28: Estructura general del proceso

En el ejemplo que se presenta en este capítulo, sólo se dispone de una señal para analizar el comportamiento del proceso, por lo tanto el sistema queda reducido a una única variable de entrada y de salida. No existen medidas de otras variables del proceso y la posición relativa entre la caja y la vía se considera que sólo depende de los valores anteriores de la misma variable. La figura 7.28 queda simplificada en el esquema más sencillo de la figura 7.29, donde se muestra que la salida sólo es función de los valores anteriores de la misma señal. Sin embargo los datos se han simulado a

partir de la ecuación de vía; por lo tanto el modelo de simulación se ajusta más al esquema de la figura 7.28, donde $x_1(t)$ sería la ecuación de la vía, y(t) la posición relativa de la caja y las rigideces, amortiguamientos y masas formarían el conjunto de características del proceso. Pero ya que no es posible medir la posición absoluta de la vía, el modelo del sistema de detección no debe basarse en esta señal y es más correcto pensar en un esquema del proceso como el que aparece en la figura 7.29.

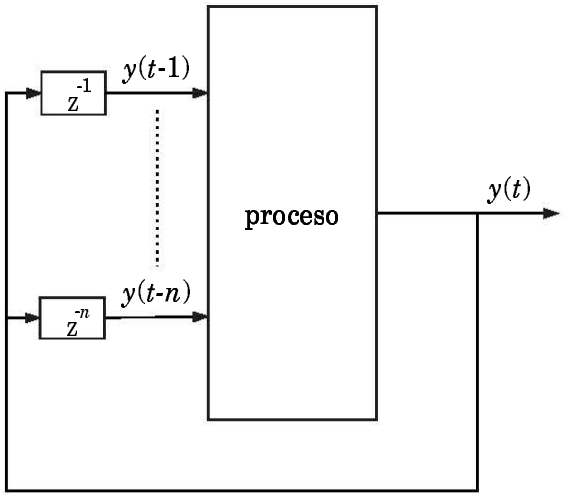


Figura 7.29: Esquema simplificado del proceso

El modelo matemático más sencillo que se ajusta a esta estructura del proceso es una combinación lineal de los valores del pasado $(y(t-1), \ldots, y(t-n))$ para estimar el valor actual de la variable $(y(t))$. Esto es un modelo auto–regresivo de orden n, donde n es el número de valores del pasado que se utilizan. El modelo para la estimación de la señal $y(t)$ se expresa matemáticamente de la siguiente manera:

$$\bar{y}(t) = a_1\, y(t-1) + a_2\, y(t-2) + \ldots + a_n\, y(t-n) \qquad \textbf{(7.8)}$$

Esta es la ecuación que se ha utilizado como modelo matemático del sistema de detección incipiente de fallos, siendo y la señal de la distancia relativa entre la caja y la vía. Se trata de un modelo lineal con n parámetros que puede expresarse en forma vectorial de la siguiente manera:

$$\bar{y}(t) = \boldsymbol{x}^{\mathrm{T}}(t) \cdot \theta \qquad \textbf{(7.9)}$$

donde θ es un vector columna formado por los n parámetros del modelo.

Dado que se trata de un sistema mecánico, sometido a grandes inercias, y que la variable considerada para el análisis es una medida de posición, nunca se producirán cambios bruscos en esta variable (no ocurriría lo mismo con una medida de fuerza ya que ésta sí puede cambiar bruscamente). Por lo tanto es esperable que una estimación basada en valores del pasado sea bastante precisa en este caso.

La selección del orden del modelo debe realizarse a partir del análisis de datos experimentales. El procedimiento de ajuste del sistema de detección de fallos puede encargarse de la selección del modelo más apropiado, si se incluye la variable n como uno de los parámetros a optimizar. Esta manera de seleccionar el tipo de modelo más apropiado es útil cuando la decisión no está clara por métodos tradicionales o cuando se quiere realizar la selección teniendo en cuenta el comportamiento del modelo en situaciones de fallo.

El objetivo de este ejemplo es mostrar las mejoras que se obtienen en la detección incipiente de fallos al ajustar un modelo dado mediante la metodología propuesta, y no tanto el mostrar las mejoras que se obtendrían al utilizar distintas técnicas de modelado. Por lo tanto se ha seleccionado el modelo de la ecuación 7.8 con un orden n fijo, obtenido a partir de diversos ajustes basados en situaciones de comportamiento normal.

Para obtener el orden del modelo se han realizado 7 ajustes por el método de mínimos cuadrados, en los que se han utilizado los datos de 12 historias de comportamiento normal. Los 7 ajustes corresponden a diferentes órdenes del modelo, desde orden 1 (qué sólo se basa en el último valor y no es capaz de modelar ninguna dinámica) hasta orden 7 (donde los valores más antiguos no aportan mucha información). Las 12 historias de comportamiento normal consideran distintas ecuaciones de vía y distintas condiciones de carga, con objeto de obtener un modelo general. La figura 7.30 muestra la evolución del error de ajuste con el orden del modelo. En este caso el error de ajuste es la suma de los cuadrados de los errores instantáneos para todos los datos utilizados. Este error nunca se hace cero debido al ruido asociado a la medida y a que la estructura del modelo no representa exactamente el comportamiento del sistema, ni siquiera en condiciones normales de funcionamiento.

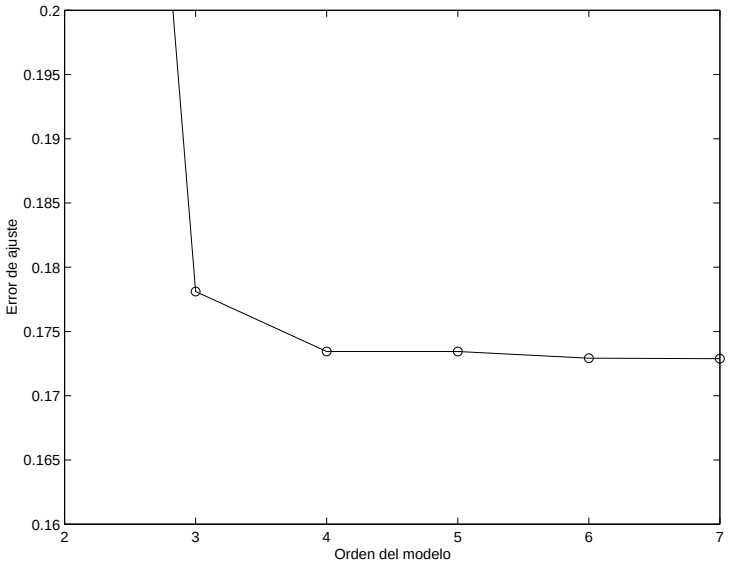


Figura 7.30: Evolución del error de ajuste con el orden del modelo

En esta gráfica se observa que para modelos de orden mayor que 4 no se obtiene una disminución significativa en el error de ajuste y por lo tanto los beneficios que se obtienen no compensan el incremento de complejidad del modelo. Además se corre el peligro de estar sobre–entrenando el modelo al especializarlo en un conjunto reducido de datos, lo que supone una pérdida de la capacidad de generalización, que daría lugar a mayores errores al considerar otros datos.

7.4.2 Atributos de fallo

En este ejemplo se han utilizado dos atributos de fallo, uno basado en el valor instantáneo de los residuos y otro basado en la tendencia de los mismos.

El primer atributo de fallo es el número de muestras consecutivas cuyos residuos superan un determinado umbral, al que se llamará u1. Este atributo depende por lo tanto de un parámetro que es el umbral de comparación. Cuanto mayor sea el valor del atributo, mayor será la probabilidad de que exista un fallo en el proceso ya que no sólo se ha detectado un aumento de los residuos sino que este sobrepaso se ha mantenido durante varias muestras. Este atributo de fallo puede evitar bastantes falsas alarmas como consecuencia de rápidas situaciones

transitorias debidas a cambios en las condiciones de trabajo que hacen aumentar los residuos durante pocas muestras.

El segundo atributo de fallo es el número de muestras consecutivas cuya tendencia de los residuos supera un determinado umbral, al que se llamará u2. La tendencia de los residuos se calcula como la pendiente de la recta de ajuste que considera los últimos valores instantáneos. Este atributo suele reaccionar con un cierto retraso ante un incremento en el nivel de los residuos. Esto hace que sea más robusto ante situaciones transitorias que no son fallos, pero más lento a la hora de detectar un fallo real.

El sistema de detección incipiente de fallos cuenta por lo tanto con dos nuevos parámetros, los umbrales u1 y u2, además de los cuatro parámetros del modelo. Como se verá más adelante la calidad del sistema de detección depende de los valores de todos estos parámetros, que deberán ajustarse apropiadamente.

7.4.3 Función de detección de fallos

La función de detección de fallos (*FDF*) se encarga de valorar de manera conjunta la información proporcionada por los dos atributos de fallos. Se ha elegido la siguiente expresión de tipo lógico:

$$FDF = AF1>1 \quad \text{AND} \quad AF2>1 \tag{7.10}$$

Esta función sólo avisa de la existencia de fallos cuando los dos atributos de fallo coinciden en la identificación de una anomalía. Puesto que los atributos de fallo son generalmente independientes, ya que utilizan distintas técnicas de detección o se basan en el análisis de datos de distinta naturaleza, la condición lógica tipo AND es bastante conservadora y da lugar a menos alarmas que una condición lógica tipo OR o una condición de tipo matemático (combinación lineal de atributos).

En este ejemplo, el atributo de fallo *AF1* está basado directamente en los residuos, mientras que *AF2* está basado en su tendencia. Sólo se detectará fallo cuando los dos criterios de detección coincidan, lo que permite que los dos atributos sean muy sensibles en su análisis particular ya que las anomalías que detecten aisladamente no serán consideradas fallos del proceso por la *FDF*. Por el contrario, si se hubiese elegido una función

de detección de fallos tipo OR los atributos de fallo no podrían ser tan sensibles en su análisis particular ya que darían lugar a muchas más alarmas, la mayoría falsas. Sin embargo hay que tener en cuenta que la selección del tipo de función de detección de fallos no es un punto crítico que afecte definitivamente a la eficacia final del sistema de detección de fallos, ya que el procedimiento de ajuste que se propone se encarga de ajustar convenientemente los parámetros de los atributos de fallo de manera que se adapten al tipo de función de detección seleccionado.

En otro tipo de aplicación, donde se definan atributos de fallo especializados cada uno en la detección de un tipo de fallo concreto del proceso, no sería aconsejable utilizar funciones de detección de fallos tipo AND ya que este tipo de función puede no detectar todos los modos de fallo. Como consecuencia se observaría que las mejoras que se obtiene al realizar el ajuste global del sistema de detección son menores que las que se obtienen al repetir el ajuste cambiando la ecuación de la función de detección de fallos. Como norma general siempre existe la posibilidad de ensayar distintos tipos de atributos de fallos y distintas funciones de detección de fallos ya que, una vez definida toda la estructura del sistema de detección y del procedimiento de ajuste, se pueden obtener fácilmente las gráficas que representan la eficacia del sistema de detección para cada una de las opciones.

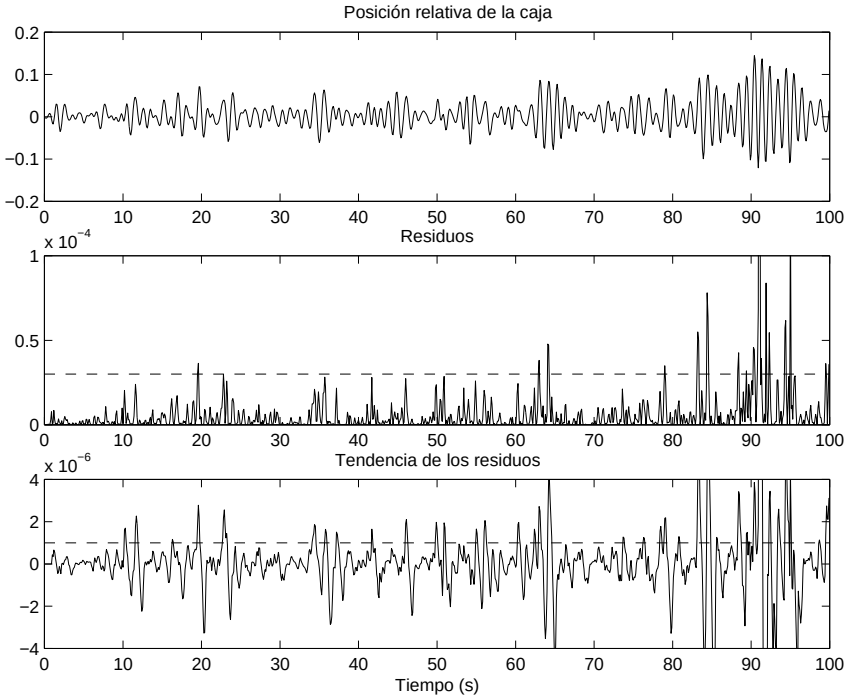


Figura 7.31: Señal del sensor, residuos y tendencia

Se describe a continuación el funcionamiento en continuo del sistema de detección de fallos para un trayecto durante el cual se produce una degradación (una de las 90 historias de fallo generadas mediante simulación). La figura 7.31 muestra tres gráficas en función del tiempo. La primera representa la posición relativa de la caja con respecto a la vía, que es una simulación de la señal proporcionada por el sensor en cada instante. A partir de 4 valores del pasado de esta señal (única señal del proceso en la que se basa el sistema de detección) se aplica el modelo matemático para obtener una estimación del valor instantáneo de la misma variable. Este valor se compara con el valor instantáneo de la primera gráfica y da lugar a un residuo. El valor que va tomando este residuo a lo largo del tiempo se ha representado en la segunda gráfica. Considerando los valores que adquieren los residuos se calcula su tendencia (según la ecuación 3.2, página 41) que indica si están cambiando o si permanecen estables. Los valores de la tendencia de los residuos se han representado en la tercera gráfica de la figura 7.31.

A partir de los valores de los residuos y de las tendencias se calculan los dos atributos de fallo y la función de detección de fallos, cuyos valores instantáneos se han representado en la figura 7.32. El primer atributo de

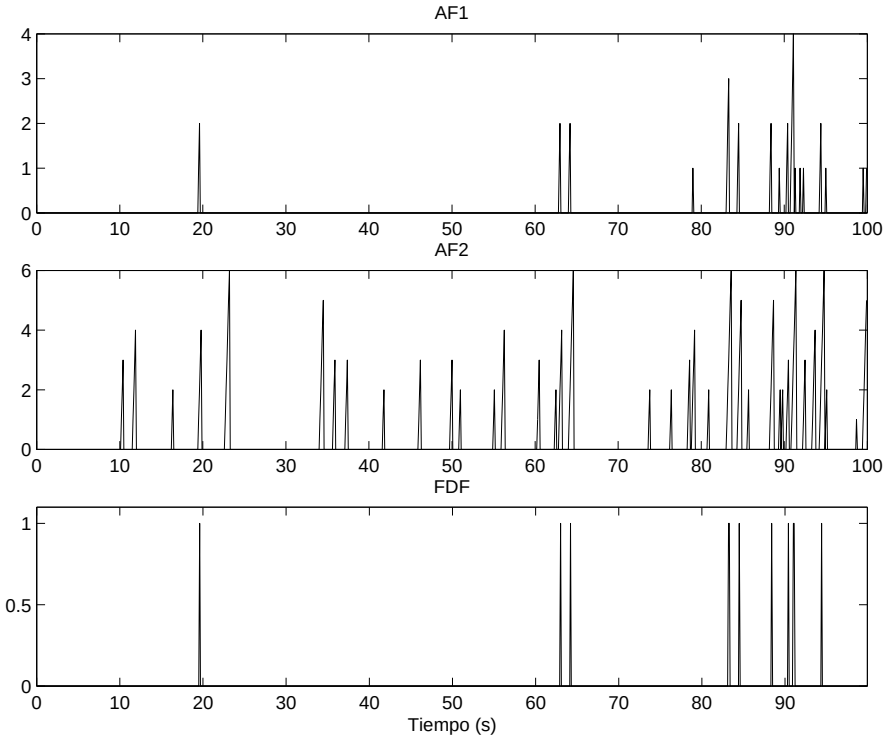


Figura 7.32: Valores de los atributos de fallo y de la función de detección de fallos

fallo se basa en la comparación del valor de los residuos con el umbral u1, que se ha fijado al valor 3e-5 para generar estas gráficas. Este umbral está representado mediante una línea discontinua en la segunda gráfica de la figura 7.31, donde se observa que en determinados puntos es superado por el valor de los residuos. El atributo de fallo *AF1* aumenta cuando los residuos superan el umbral, lo que se considera una anomalía. La evolución del primer atributo de fallo con el tiempo se ha representado en la primera gráfica de la figura 7.32. De manera análoga el atributo de fallo *AF2* contabiliza el número de sobrepasos de la tendencia de los residuos. En este caso la comparación se realiza en valor absoluto ya que este atributo detecta variaciones anormales en los residuos, pero no necesariamente de aumento sino variaciones de cualquier tipo. El umbral utilizado en la tendencia es

de 0.1e-5 y ha sido dibujado en línea discontinua en la tercera gráfica de la figura 7.31. Los valores del segundo atributo de fallo se han representado en la segunda gráfica de la figura 7.32, se observa que este atributo de fallo detecta un gran número de anomalías para el valor del umbral seleccionado. Sin embargo la función de detección de fallos (tercera gráfica de la figura 7.32) sólo considera fallos del proceso en aquellos instantes en que los dos atributos de fallo coinciden.

Los datos con los cuales han sido generadas las dos gráficas anteriores incluyen una degradación del amortiguador vertical de la suspensión secundaria que se inicia lentamente a partir del segundo 50. En la señal del sensor puede observarse que los movimientos verticales de la caja son mayores al final de la simulación, ya que la energía absorbida por el amortiguador es cada vez menor como consecuencia de la degradación. También se observa en la figura 7.31 que los residuos y la tendencia cambian en la segunda mitad de las gráficas y por lo tanto estas fuentes de información son útiles para detectar el fallo del amortiguador de manera incipiente si se realiza un análisis apropiado. En la tercera gráfica de la figura 7.32 se observa que las alarmas de la función de detección de fallos se producen en mayor número en la segunda mitad de la simulación, si bien existe una falsa alarma en el segundo 20 donde el amortiguador no se encuentra deteriorado.

Como resumen se muestra a continuación el esquema general del sistema de detección que depende en total de 6 parámetros: 4 parámetros del modelo y 2 parámetros de los atributos de fallo. Este sistema se aplica de forma continua y emite un diagnóstico de salud o fallo (cero o uno de la FDF) en cada instante de tiempo. El paso siguiente es ajustar todos los parámetros de sistema de detección de manera que se optimice la capacidad de detección incipiente de fallos.

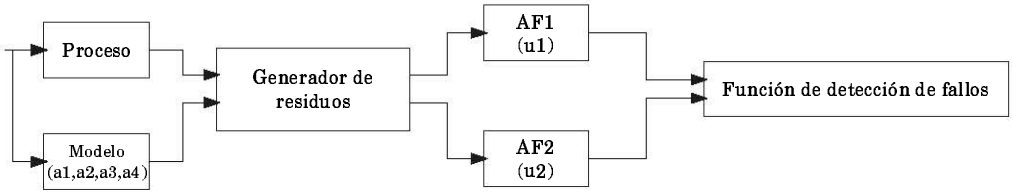


Figura 7.33: Esquema del sistema de detección

7.5 Ajuste del sistema de detección

Se ha definido un sistema de detección incipiente de fallos para monitorizar el estado de salud de la suspensión de trenes. Este sistema sólo utiliza una señal del proceso para realizar el análisis y cuenta con 6 parámetros internos. Los 4 parámetros del modelo se han ajustado inicialmente por el método de mínimos cuadrados utilizando un conjunto de datos de funcionamiento normal que cubre gran variedad de condiciones de funcionamiento. Por otro lado los parámetros de los atributos de fallo han sido seleccionados a partir del valor que adquieren los residuos en unos pocos ejemplos de datos. Finalmente se ha comprobado que el sistema de detección es capaz de detectar un proceso de degradación (figura 7.32). Esta gráfica aislada no garantiza sin embargo que el sistema sea capaz de detectar fallos en otras situaciones de degradación y de no detectarlos durante el funcionamiento normal en otras condiciones de trabajo. Por otro lado tampoco se garantiza que los parámetros seleccionados sean los más apropiados para el objetivo que se persigue, la detección incipiente.

Es el momento por tanto de realizar un ajuste global de este sistema de detección de manera que se especialice en la detección incipiente de fallos. Este ajuste estará basado en la utilización de datos de funcionamiento normal y de datos de situaciones degradación y de fallo.

Para poder realizar el ajuste es fundamental definir unas funciones que permitan valorar de manera objetiva la calidad del sistema de detección. Estas funciones, que fueron bautizadas con el nombre de atributos de detección, permiten comparar la eficacia de los diferentes sistemas de detección que se obtienen al variar los parámetros. Son por lo tanto las funciones que van a orientar al algoritmo de ajuste y que van a permitir comparar gráficamente los resultados obtenidos.

7.5.1 Atributos de Detección

Se han definido sólo dos atributos de detección en este ejemplo, que representan las dos cualidades fundamentales que se analizan en general en cualquier sistema de detección. El primer atributo, AD1, representa el

número de falsas alarmas que produce el sistema. Este atributo se obtiene haciendo funcionar el sistema de detección de fallos sobre un conjunto de *training sets* y contando el número de muestras de las zonas sin anomalía para las cuales la FDF vale 1. Para que este atributo de detección sea independiente del volumen de datos considerados, se calcula posteriormente el número medio de falsas alarmas dividiendo por el número de *training sets*.

El segundo atributo, AD2, representa el **tiempo medio de detección** del fallo, medido desde el instante en que comienza la degradación. Este atributo se obtiene haciendo funcionar el sistema de detección de fallos sobre un conjunto de *training sets* y calculando el número de muestras que existe desde que se inició la degradación en la simulación de datos hasta que se detecta el fallo por primera vez.

Estos atributos de detección dan una medida objetiva de la calidad del sistema de detección. Al estar basados en un gran conjunto de datos, que cubren diferentes condiciones de funcionamiento y velocidades de degradación, informan sobre el comportamiento global del sistema y no sobre el comportamiento en un caso particular. El mejor sistema de detección será aquel que haga mínimos ambos atributos de detección. En una representación multi–objetivo estos son los sistemas que se encuentran más cerca de los ejes y del origen de coordenadas.

7.5.2 Resultados del ajuste

El procedimiento de ajuste consiste en hacer funcionar el sistema de detección incipiente de fallos para diferentes *training sets*, calculando el valor de los atributos de detección en cada caso y finalmente el valor medio. Siempre se utiliza el mismo conjunto de *training sets*, pero el valor de los atributos de detección cambia en función de los parámetros utilizados.

La sensibilidad a los cambios en los umbrales queda clara en las figuras 7.34 y 7.35, que son en realidad una ampliación de una parte las figuras 7.31 y 7.32 mostradas anteriormente. Puede observarse que en determinados puntos de la gráfica de residuos y de la gráfica de tendencias se superan los umbrales dibujados en línea discontinua. Esto da lugar a las gráficas de atributos de fallo de la figura 7.35 y como consecuencia a dos alarmas en la función de detección de detección de fallos. Pero con pequeñas variaciones

en los valores de los umbrales cambiarían la forma de los atributos de fallo y de la función de detección de fallos.

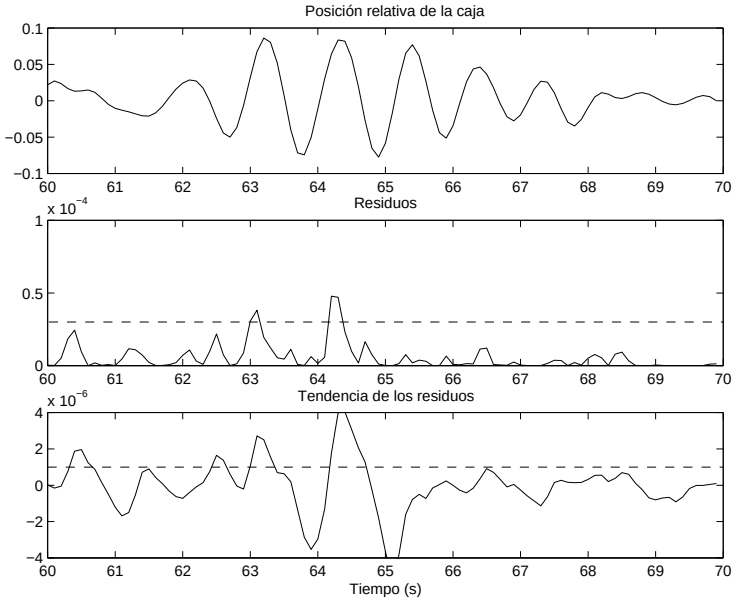


Figura 7.34: Señal del sensor, residuos y tendencia

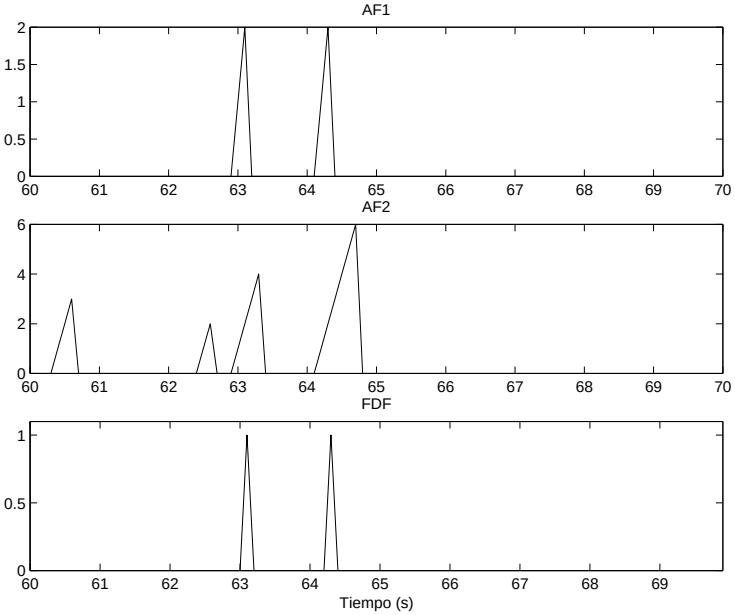


Figura 7.35: Atributos de fallo y función de detección de fallos

Para unos valores fijos en los parámetros del modelo matemático, pueden obtenerse las curvas que representan la variación de cada atributo de

detección en función de los dos parámetros de los atributos de fallo. La figura 7.36 representa el número medio de falsas alarmas en función de los valores de los umbrales. Puede observarse que para valores grandes de los umbrales no existen falsas alarmas, mientras que cuando los umbrales son pequeños el valor del atributo de detección se sale de la gráfica[1].

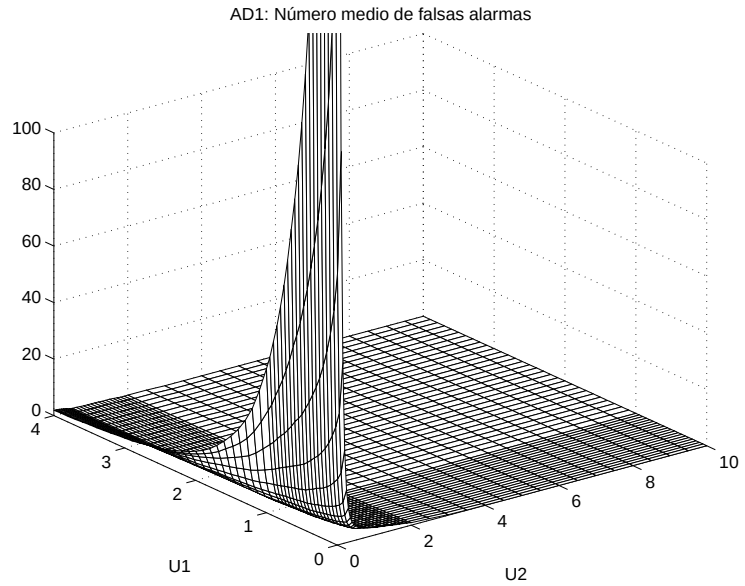


Figura 7.36: AD1 en función de los valores de los umbrales

La forma de la curva del tiempo de detección (figura 7.37) es más bien la inversa. Para valores altos de los umbrales el tiempo medio de detección crece hacia el valor que 500, que es el máximo valor posible para este atributo ya que desde que aparece la degradación hasta que se termina la simulación transcurren 500 muestras (cuando el sistema no ha detectado el fallo el tiempo de detección se fija en 500 muestras). Para valores pequeños de los umbrales, donde los atributos de fallo se hacen muy sensibles, el tiempo de detección tiende a cero ya que se detecta el fallo nada más iniciarse la degradación.

[1] Esta gráfica ha sido generada con un mallado más fino en las zonas conflictivas, con objeto de obtener mayor precisión donde se producen cambios más bruscos. Por lo tanto el tamaño de la proyección de los elementos que forman la retícula no es constante. Lo mismo ocurre en el resto de las gráficas de superficies que se muestran a continuación.

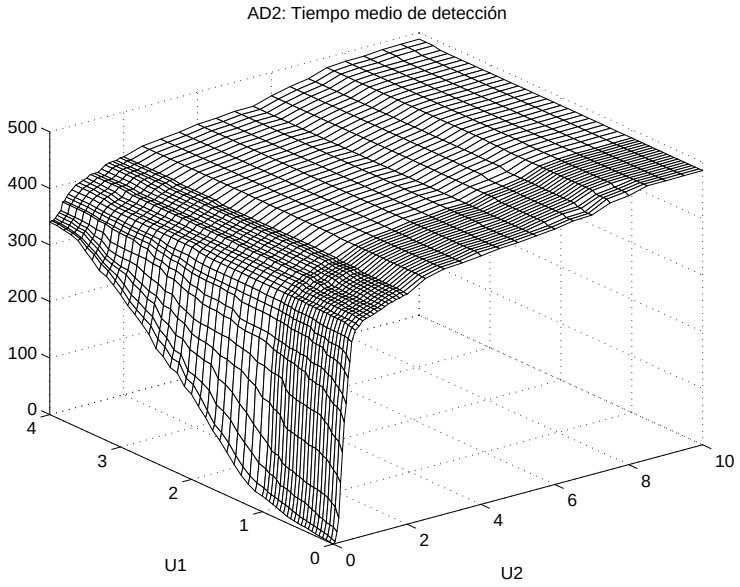


Figura 7.37: AD2 en función de los valores de los umbrales

En el caso de variaciones en los parámetros del modelo matemático, los cambios que experimenta el sistema de detección son más difíciles de predecir. En general puede decirse que una variación en los parámetros del modelo, con respecto a los obtenidos por el método de mínimos cuadrados, dará lugar a un aumento en la curva de residuos. Esto se debe a que el ajuste por mínimos cuadrados hace mínimo el valor medio de los residuos si estos se calculan como el cuadrado de la diferencia entre la estimación y la señal real. La figura 7.38 muestra tres ejemplos de curvas de residuos generadas para las mismas condiciones de carga y de vía pero utilizando distintos parámetros en el modelo. La primera gráfica corresponde a los parámetros obtenidos por mínimos cuadrados, y las otras dos a otros parámetros que difieren menos de un 10% de los anteriores. Puede observarse que el nivel medio de los residuos es mayor en las dos segundas gráficas, aunque los picos máximos no siempre coinciden.

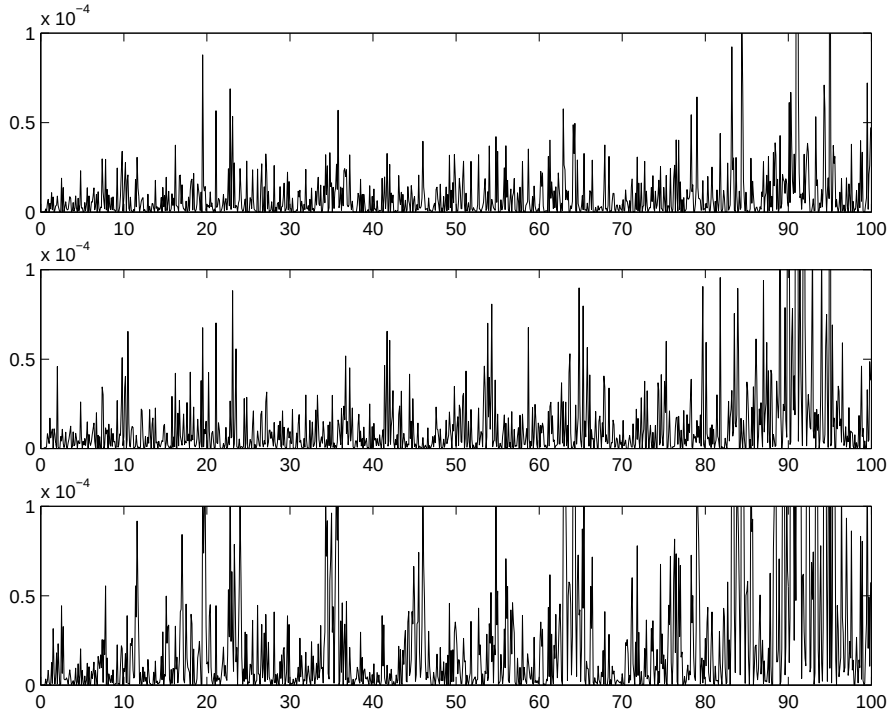


Figura 7.38: Consecuencia de la variación de los parámetros del modelo

Como resultado de los cambios en las curvas de residuos se alteran los resultados de la función de detección de fallos (figura 7.39). La primera curva no tiene ninguna alarma durante la primera mitad, que corresponde a funcionamiento normal, por lo tanto no tiene ninguna falsa alarma. La segunda gráfica presenta dos falsas alarmas aunque también el tiempo de detección se reduce, ya que la degradación que se inicia en el segundo 50 es detectada a los 3 segundos en lugar de esperar 14 como en el primer caso. En el tercer caso el número de falsas alarmas es claramente excesivo.

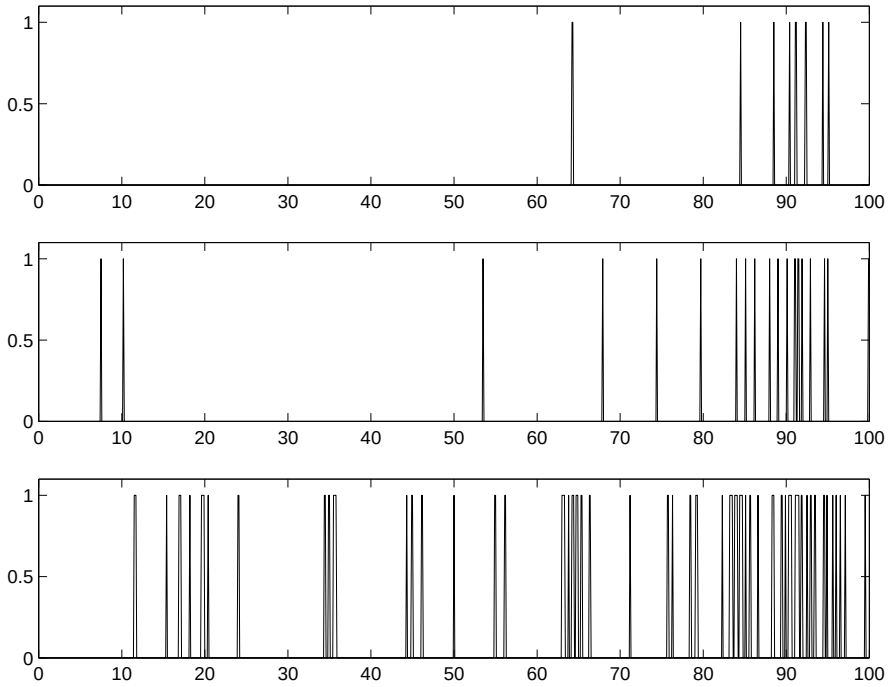


Figura 7.39: Consecuencia de la variación de los parámetros del modelo

El aumento de la curva de residuos no implica necesariamente un aumento del número de falsas alarmas, como podría pensarse a la vista de las gráficas anteriores, ya que el funcionamiento del sistema de detección depende en gran medida del tipo de análisis que se realice sobre esta curva de residuos. En el caso de utilizar atributos de fallo basados en la comparación del nivel de residuos con un umbral fijo, el número de falsas alarmas depende más de los valores máximos que adquieren los residuos puntualmente que del valor medio. Como se mostró en el capítulo 4, y más concretamente en la figura 4.6 (página 64), pueden existir conjuntos de parámetros diferentes de aquéllos que se obtienen por el método de mínimos cuadrados que dan lugar a un menor nivel máximo de los residuos.

Por otro lado hay que tener en cuenta que el número de falsas alarmas también depende de los parámetros de los atributos de fallo y por lo tanto si una variación de los parámetros del modelo va acompañada de una variación en los umbrales se puede mejorar el comportamiento del sistema de detección (y que al variar los parámetros del modelo también puede aumentar la sensibilidad a la detección de fallos). Más adelante se comprueba que el sistema de detección mejora sus cualidades al realizar una

optimización de los parámetros de acuerdo al esquema de ajuste que propone esta tesis.

La optimización se basa en la información proporcionada por los atributos de detección y se encarga de ajustar todos los parámetros del sistema de detección. La siguiente figura resume el procedimiento de ajuste aplicado a este ejemplo, donde sólo se utilizan dos atributos de fallo y dos atributos de detección. En función de los valores de los atributos de detección, el algoritmo de ajuste propone un nuevo conjunto de parámetros del sistema (a1, a2, a3, a4, u1 y u2). En este caso la función de detección de fallos es de tipo lógico y no tiene ningún parámetro.

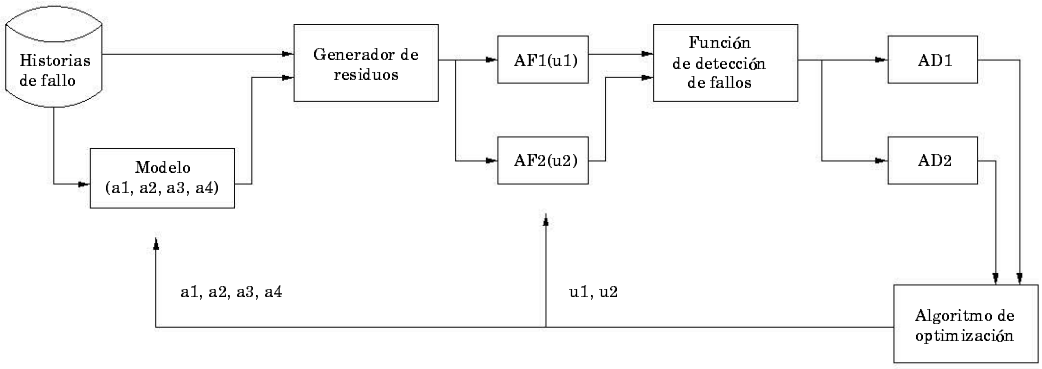


Figura 7.40: Esquema del procedimiento de ajuste aplicado al ejemplo

El algoritmo de ajuste es el encargado de valorar los atributos de detección y de proponer diferentes conjuntos de parámetros hasta encontrar aquellos que ofrezcan los mejores resultados. La evaluación de los atributos de detección no resulta en absoluto sencilla ya que se trata de cualidades con un comportamiento contradictorio. Este comportamiento puede comprobarse claramente en las figuras 7.41 y 7.42 que representan a los atributos de detección AD1 (en blanco) y AD2 (en gris) en función de los umbrales y para unos parámetros fijos en el modelo matemático. Ambas figuras representan la misma gráfica, sólo se ha modificado el punto de observación para proporcionar dos vistas diferentes. Dado que interesa disminuir simultáneamente ambos atributos de detección, la localización del punto óptimo de funcionamiento no está clara.

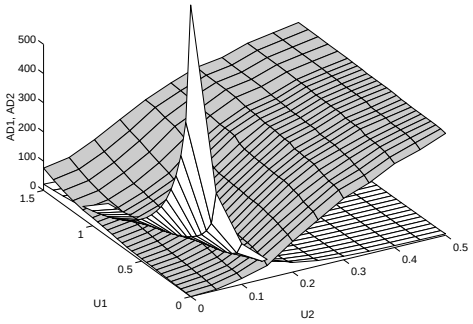


Figura 7.41:

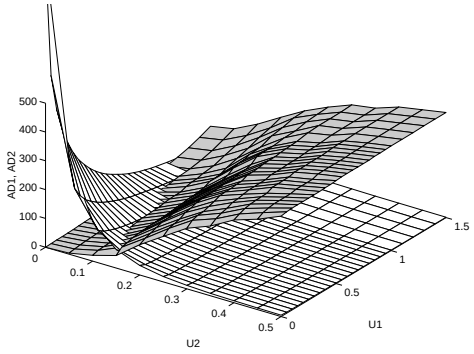


Figura 7.42:

Se ha considerado que la mejor manera de abordar este problema es realizar una optimización multi–objetivo para presentar como solución el conjunto de sistemas con los mejores valores de los dos atributos de detección. Este conjunto de sistemas se puede obtener realizando la optimización de varias combinaciones lineales de los dos atributos de detección donde se cambian sus pesos. Por ejemplo, cuando los pesos de los dos atributos de detección son iguales a la unidad, se trata de minimizar la función objetivo FO=AD1+AD2. Esta función objetivo tiene la representación gráfica de la figura 7.43 cuyo óptimo se encuentra cercano al origen, es decir para valores pequeños de los umbrales.

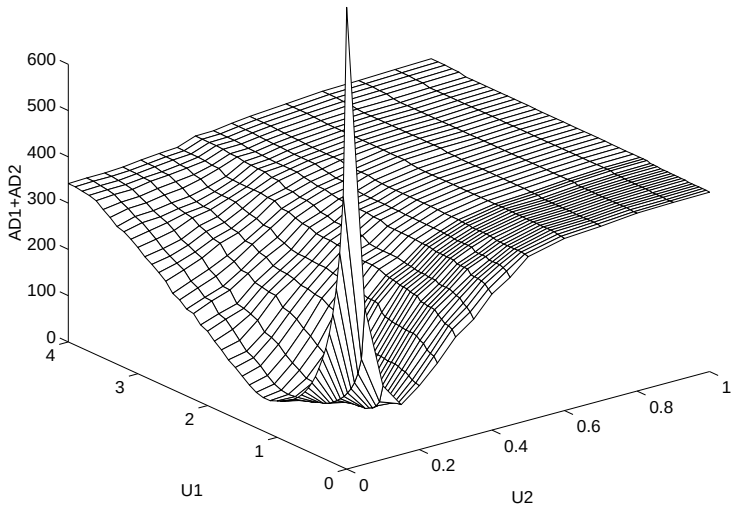


Figura 7.43: Tipo de función objetivo que se debe minimizar

La localización del mínimo se puede realizar utilizando cualquiera de los métodos de optimización por búsqueda directa descritos en el capítulo 5. En este ejemplo se ha utilizado el método simplex de optimización no lineal, por ser uno de los que mejor comportamiento tuvo en el estudio comparativo de métodos del capítulo 5. La figura 7.44 muestra una ampliación de la figura 7.43 en la que se encuentra el óptimo del problema (señalado mediante un punto •). Lógicamente esta solución sólo es válida para un conjunto fijo de parámetros del modelo y para el caso de dar igual importancia a los dos atributos de detección.

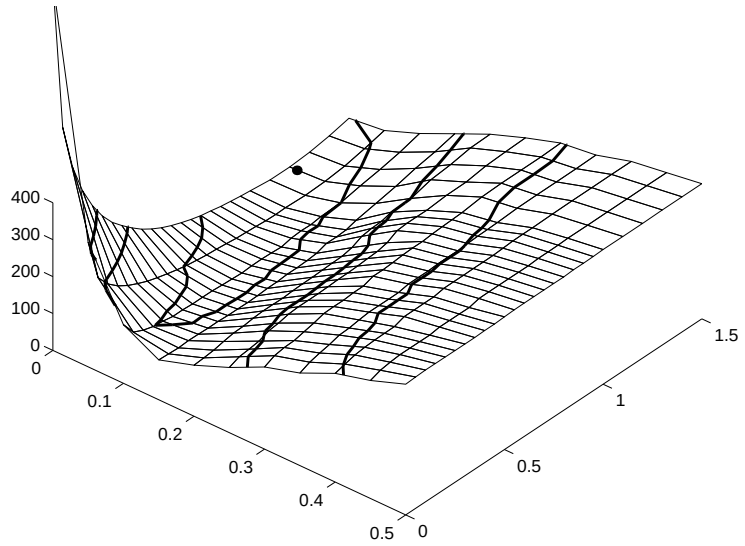


Figura 7.44: Valores óptimos de U1 y U2

El conjunto final de soluciones se obtiene realizando varias optimizaciones de este tipo, en las que se varían los pesos que establecen la importancia relativa entre los dos atributos de detección. Cada una de las optimizaciones da lugar a una posible solución que mejora en todos los atributos de detección al sistema con el que se inicia la búsqueda. Por último es necesario eliminar los puntos dominados, es decir aquellas soluciones que son peores en todos los atributos de detección que otras soluciones. Aunque el método de optimizar una combinación lineal de los atributos de detección variando los pesos permite obtener directamente los puntos de la hiper–superficie óptima [deCuadra90], sólo funciona cuando cada optimización localiza con éxito el mínimo global de la función objetivo. Pero si existen mínimos locales, como es generalmente el caso de problemas no lineales, pueden obtenerse puntos que no pertenecen realmente al

conjunto de soluciones óptimas y fácilmente se localiza otra solución que los domina. Por lo tanto el último paso en el ajuste del sistema de detección es eliminar los puntos dominados y dibujar la curva de soluciones óptimas (en caso de utilizar sólo dos atributos de detección).

La figura 7.45 muestra las curvas de soluciones óptimas que se obtienen al ajustar el sistema partiendo de un mismo punto de funcionamiento inicial. En la definición del punto inicial se han utilizado, para el modelo, los parámetros obtenidos en el ajuste por mínimos cuadrados, y para los umbrales, unos valores que parecían razonables tras el estudio de unas pocas historias de funcionamiento normal. Este es el procedimiento que normalmente se sigue para ajustar sistemas de detección que utilizan modelos matemáticos del proceso y se basan en el análisis de residuos. Se ha representado en la gráfica mediante un punto (•) este sistema de detección, tras obtener los valores de los dos atributos de detección aplicando el sistema a una historia de fallo. Partiendo de los parámetros de este sistema y trabajando con todas las historias de fallo se han llevado a cabo diversos ajustes que en todos los casos mejoran la situación de partida.

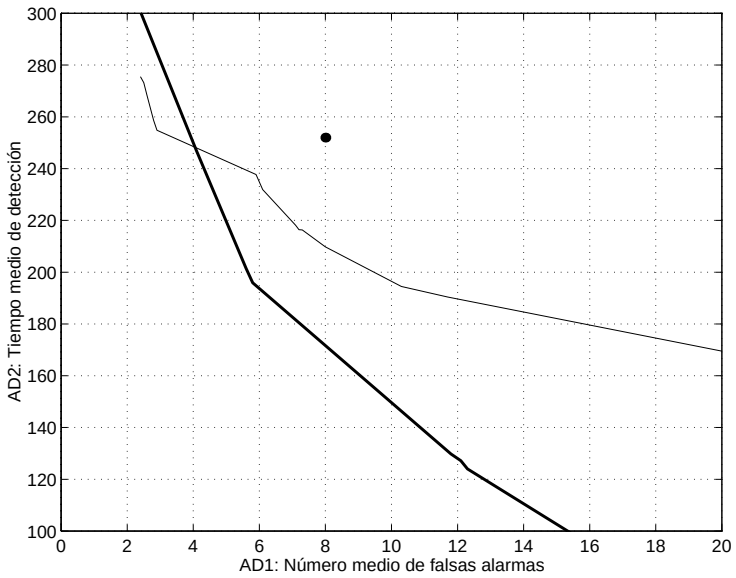


Figura 7.45: Resultados del ajuste

La línea fina representa el resultado del ajuste de los parámetros del modelo matemático del sistema de detección, dejando fijos los umbrales. Puede

observarse que, a pesar de no variar el resto del sistema de detección, es posible mejorar la calidad del sistema en cualquiera de los dos atributos de detección y en ambos simultáneamente. Esto demuestra que la metodología de ajuste que propone la tesis permite obtener mejores modelos que el obtenido por el método de mínimos cuadrados; es decir se consigue **especializar el modelo en detección incipiente de fallos**.

La línea gruesa representa el resultado de optimizar los parámetros de los atributos de fallo dejando fijos los parámetros del modelo. La gran mejora que se obtiene indica que los umbrales elegidos inicialmente no eran muy acertados. En este caso el sistema se beneficia más del ajuste de estos parámetros que del ajuste de los parámetros del modelo, esto se debe a que no se ha aplicado ninguna herramienta de ajuste (equivalente a los mínimos cuadrados en el caso del modelo) para obtener los valores iniciales de los umbrales.

La optimización global de todos los parámetros da lugar a una curva que domina a las dos anteriores. Esta curva representa el conjunto de soluciones óptimas del sistema de detección; y cada uno de los puntos que la forman corresponde a un conjunto de parámetros. La decisión final sobre la importancia relativa que se da a cada atributo de detección permite localizar directamente el conjunto de parámetros que hacen óptimo el sistema de detección incipiente de fallos para esas especificaciones. Las curvas que resultan de la optimización conjunta de todos los parámetros se muestran en la figura 7.46.

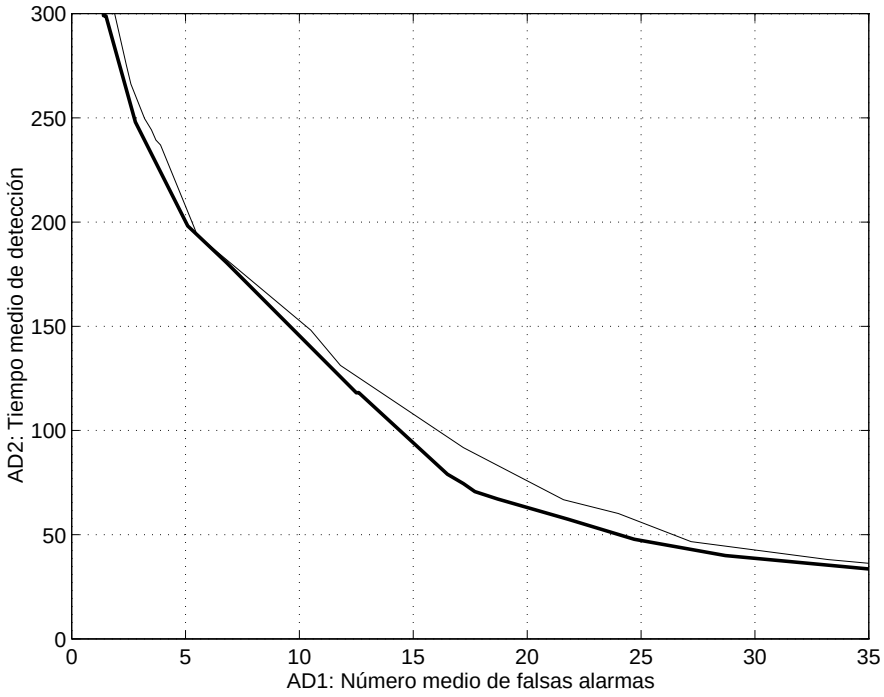


Figura 7.46: Resultados del ajuste global de todos los parámetros

Aunque la figura 7.44 (que muestra la suma de AD1 y AD2 en función de los valores de los umbrales) parece mostrar una función objetivo sin discontinuidades y fácil de optimizar, la realidad es que esta función vista desde más cerca contiene infinidad de mínimos locales que dificultan la búsqueda de la solución. Esta situación que es inherente a este tipo de problema, tal como se explica en el apartado 5.4.1 (página 107), hace que la obtención de la auténtica curva óptima no sea sencilla. También en el capítulo 5 se describen técnicas específicas para solucionar este problema de los mínimos locales y maneras de realizar el ajuste eficientemente. En este ejemplo se han aplicado técnicas de optimización por partes para mejorar la velocidad del ajuste global y para permitir optimizar de manera independiente los parámetros del modelo y los umbrales. También se ha aplicado un barrido de variables, que permite localizar la zona en la que se encuentra el mínimo global antes de iniciar el algoritmo de optimización, en el ajuste de los umbrales (segunda etapa de la optimización por partes).

La línea fina de la figura 7.46 representa los resultados de realizar las optimizaciones globales del sistema de detección de manera que la segunda

etapa de la optimización parte siempre de unos umbrales fijos para cada conjunto de parámetros del modelo que propone la primera etapa de la optimización. Por el contrario, la línea gruesa se obtiene como resultado de realizar un barrido de los umbrales antes de realizar la segunda optimización. Puede observarse que esta técnica de eliminación de mínimos locales mejora los resultados del ajuste.

Como resumen final, se muestra a continuación (figura 7.47) una gráfica donde se recogen todos los resultados de las diferentes optimizaciones del sistema. En línea fina se han representado los resultados de las optimizaciones parciales del problema; es decir, el resultado de optimizar sólo los parámetros del modelo o sólo los parámetros de los atributos de fallo. En línea gruesa se ha dibujado el resultado final, que se obtiene por medio del ajuste global de todos los parámetros. Puede observarse que la curva en línea gruesa domina en todos sus puntos a las líneas finas.

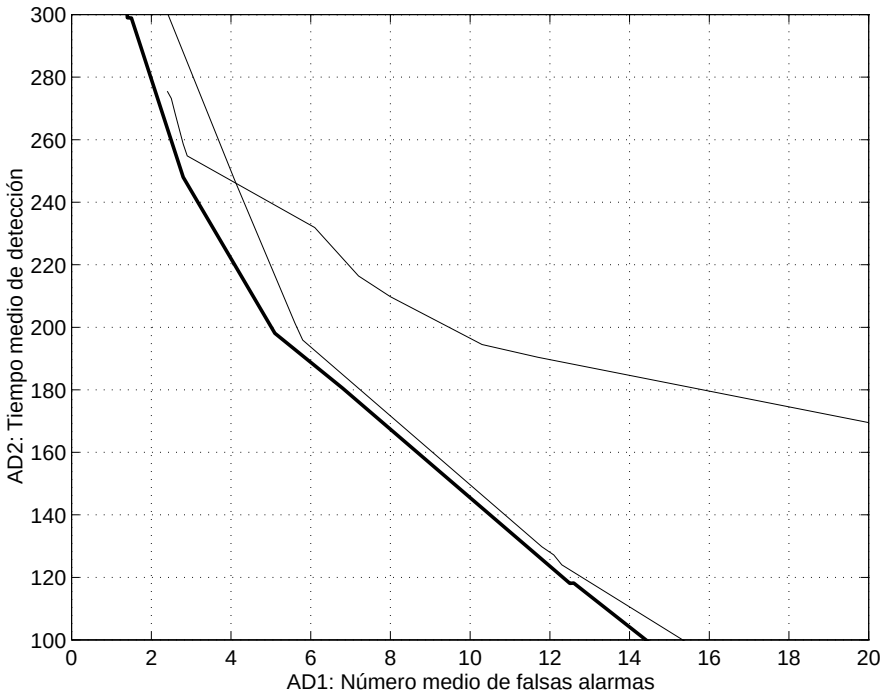


Figura 7.47: Comparación entre los distintos ajustes

7.6 Conclusiones

En este capítulo se ha presentado el ejemplo completo de un sistema real y se han descrito todos los pasos necesarios para definir un sistema de detección incipiente de fallos. Tras la descripción del proceso se ha explicado la manera utilizada para simular datos de comportamiento normal y datos de diferentes estados de degradación. Aunque el volumen de datos no es muy grande, la forma en que han sido generados garantiza una gran riqueza de condiciones de funcionamiento. Sería interesante sin embargo poder comprobar el funcionamiento del sistema con datos reales obtenidos mediante trenes instrumentados.

A continuación se ha descrito paso a paso la definición completa de un sistema de detección de acuerdo a la estructura general propuesta en esta tesis (modelo matemático, atributos de fallo y función de detección de fallos). Todo el sistema de detección se ha basado en una sola señal del proceso y a pesar de ello se ha demostrado la capacidad de detectar fallos de manera incipiente. La utilización de un mayor número de señales permite detectar fallos con mayor facilidad, sin embargo este ejemplo ha demostrado que un análisis adecuado de la información permite detectar los fallos sin necesidad de utilizar muchas señales.

Por último se ha definido la estructura del procedimiento de ajuste que se propone en la tesis y que se basa en la utilización de datos de funcionamiento normal y de degradación y en medir la eficacia del sistema de detección mediante los atributos de detección. Se ha demostrado que la calidad del sistema de detección es sensible a las variaciones de cualquiera de sus parámetros. También se ha demostrado que el ajuste global de todos los parámetros permite mejorar las características en cuanto a detección incipiente de fallos. Esto supone un gran avance frente a la metodología de ajuste tradicional donde los parámetros del modelo se fijan de acuerdo al método de mínimos cuadrados (que no obtiene los mejores modelos para detección de fallos) y los parámetros de los atributos de fallo se ajustan sin criterios objetivos y sin evaluar el comportamiento del sistema durante procesos de degradación.

Dado que el ajuste de los parámetros supone la optimización de un problema complicado, no se puede garantizar que un determinado algoritmo localice la mejor solución que existe. Pero, una vez establecida toda la

estructura del sistema de detección y del procedimiento de ajuste, es fácil mejorar aún más el sistema probando diferentes algoritmos de optimización para cada módulo. Como ejemplo se ha mostrado la mejora que se obtiene al utilizar uno de los métodos de eliminación de mínimos locales propuestos en el capítulo 5 para este tipo de problemas de optimización.

Capítulo 8

Conclusiones y futuros desarrollos

8.1 Conclusiones y aportaciones de la tesis

Esta tesis se ha dedicado al análisis de los sistemas de detección de fallos en procesos industriales con objeto de mejorar sus prestaciones.

La detección de fallos se basa en el análisis de datos representativos del estado de funcionamiento del proceso. La frecuencia con que deben recogerse estos datos depende de la dinámica del proceso y de la velocidad con que pueden aparecer los fallos. La disminución de los precios de los equipos para adquisición de datos en continuo permite actualmente montar sensores permanentes y sistemas de monitorización, lo que proporciona información suficiente como para realizar la detección de los fallos de manera incipiente.

Los sistemas de detección de fallos más avanzados utilizan modelos matemáticos para distinguir entre situación normal y situación de fallo, ya que en el análisis aislado de señales resulta difícil distinguir entre fallos y cambios de las condiciones de funcionamiento (ver apartado 2.2). Estos modelos se utilizan para estimar la respuesta del proceso en función de un conjunto de medidas que sean representativas de las condiciones de trabajo. Se distinguen dos tipos básicos de sistemas de detección incipiente de fallos: aquéllos que utilizan modelos matemáticos de parámetros fijos y aquéllos basados en modelos adaptativos (cuyos parámetros se reajustan en cada iteración). En el primer caso se puede utilizar cualquier estructura de modelado (modelos basados en fundamentos físicos, redes neuronales, modelos estadísticos...) y la detección se basa en un análisis de los residuos que se obtienen como discrepancia entre las estimaciones del modelo y las correspondientes señales reales del proceso. En el segundo caso resulta más adecuado utilizar modelos matemáticos basados en fundamentos físicos, donde los parámetros representan características de los componentes del proceso. Estos parámetros se reajustan utilizando algoritmos recursivos de estimación de parámetros (ver capítulo 6) de manera que una variación en las características físicas de un componente del proceso, se refleja en la variación de los correspondientes parámetros del modelo. En este caso los residuos se utilizan para reajustar el modelo y la detección de fallos se basa en el análisis de la evolución de sus parámetros.

La primera aportación fundamental de esta tesis ha sido la definición de una **estructura general del sistema de detección incipiente de fallos**. Esta estructura es aplicable a cualquier tipo de proceso y permite utilizar las dos técnicas de detección de fallos descritas anteriormente. De acuerdo a esta estructura, el sistema de detección de fallos utiliza un "modelo matemático" y un "generador de residuos", que proporcionan en cada instante de muestreo una información en forma de residuos o de parámetros. A continuación esta información es analizada mediante "atributos de fallo", que son funciones o algoritmos independientes encargados de buscar indicios de fallo. Finalmente la "función de detección de fallos" evalúa de manera conjunta todos los atributos de fallo utilizados. Esta evaluación conjunta puede realizarse en forma de combinación lineal, función de tipo lógico o función basada en lógica borrosa, y permite obtener el diagnóstico definitivo de salud o de fallo. La figura 8.1 representa el

esquema general del sistema de detección incipiente de fallos. En este esquema se identifican todos los bloques que componen el sistema de detección y el tipo de información con que trabajan.

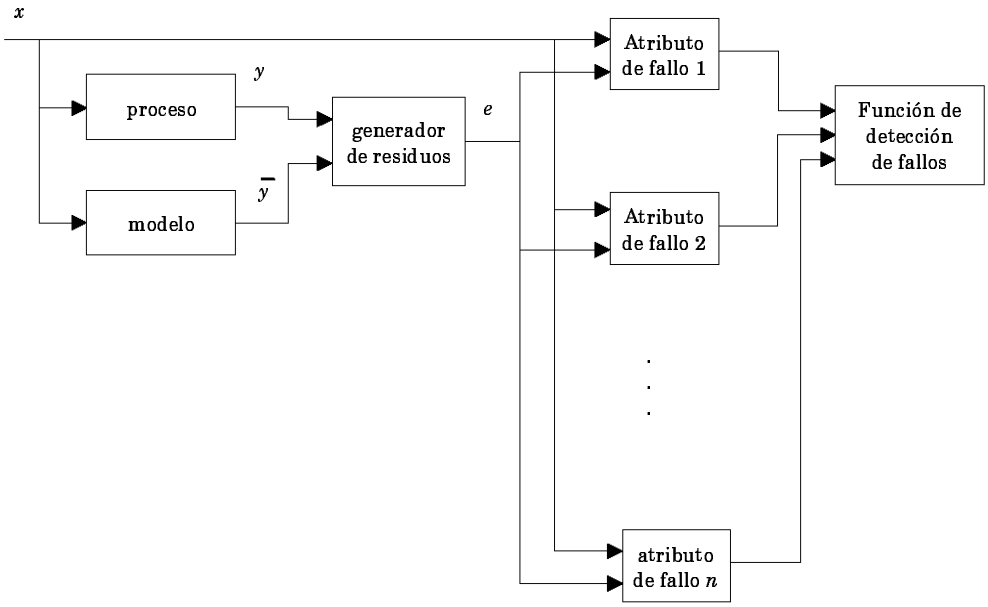


Figura 8.1: Esquema general del sistema de detección de fallos

La utilización de varios atributos de fallo permite aplicar diversas técnicas de análisis de forma simultánea. La función de detección de fallos permite definir la importancia relativa de cada atributo de fallo en una aplicación concreta y se encarga de emitir un único diagnóstico.

Además de la estructura del sistema de detección incipiente de fallos, también se ha definido un **procedimiento de ajuste** que marca la diferencia fundamental con otros sistemas de detección basados en modelos. Este procedimiento de ajuste constituye la segunda aportación fundamental de la tesis para mejorar la calidad de los sistemas de detección incipiente de fallos.

El ajuste de los modelos matemáticos de un proceso suele realizarse aplicando técnicas de mínimos cuadrados, lo que da lugar a modelos cuyas predicciones tienen un comportamiento óptimo durante el funcionamiento normal del proceso. Sin embargo los modelos ajustados por este procedimiento pueden no ser óptimos para el objetivo final del sistema, que es la detección incipiente de fallos. El procedimiento de ajuste que se

propone en la tesis se basa en analizar realmente la calidad del sistema de detección. Esta medida de calidad se obtiene mediante la evaluación del comportamiento del sistema ante diferentes situaciones de funcionamiento normal y de funcionamiento durante procesos de degradación. Se contempla la posibilidad de utilizar varias medidas de la capacidad de detección del sistema, a las que se ha llamado "atributos de detección".

Resulta difícil evaluar de manera conjunta varios atributos de detección, ya que cada uno se refiere a una característica concreta del sistema y la mejora de uno de ellos típicamente perjudica a otros (por ejemplo, tiempo de detección y número de falsas alarmas). En esta tesis se propone utilizar **técnicas de optimización multi–objetivo** para obtener el conjunto de soluciones óptimas para cualquier combinación de atributos de detección. Esto permite definir la importancia de los atributos de detección con posterioridad al ajuste o incluso variar dinámicamente (en función de otros criterios) el comportamiento del sistema de detección seleccionando distintos conjuntos óptimos de parámetros.

La utilización de los atributos de detección permite valorar el comportamiento global del sistema de detección de fallos y por lo tanto permite comparar su eficacia cuando se cambian los parámetros. Esta información es la que utiliza el algoritmo de optimización para determinar el conjunto de parámetros que realmente consigue maximizar la eficacia en la detección.

En el sistema de detección de fallos existen parámetros asociados a cada módulo: parámetros del modelo, parámetros de los atributos de fallo (por ejemplo el valor de los umbrales) y parámetros de la función de detección de fallos (por ejemplo los pesos asociados a los atributos de fallo). En esta tesis se ha propuesto, como aportación fundamental, realizar un **ajuste global de todos los parámetros del sistema de detección de fallos** basándose en la información proporcionada por los atributos de detección. Además este ajuste considera todos los conjuntos de datos (*training sets*) de comportamiento normal, en diferentes condiciones de funcionamiento, y de degradación, para diferentes modos de fallo. La figura 8.2 representa el esquema del procedimiento de ajuste propuesto en esta tesis. En el esquema aparecen todos los módulos que forman parte del sistema de

detección, con sus correspondientes parámetros, y los atributos de detección que se utilizan para dirigir el ajuste.

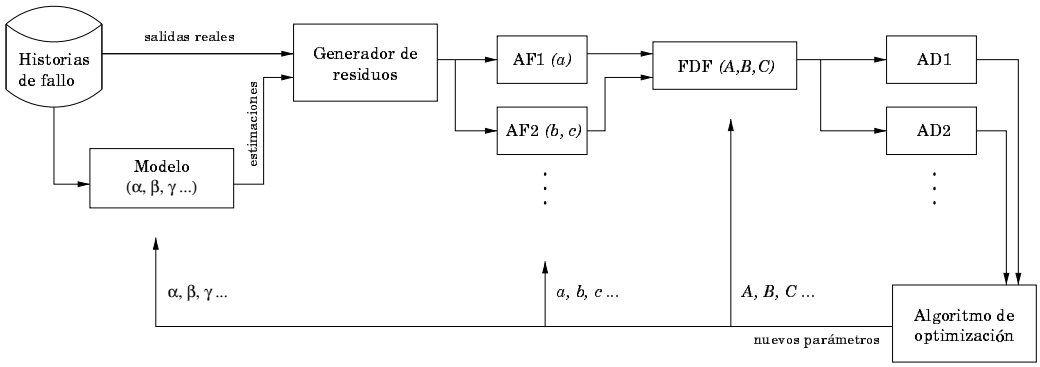


Figura 8.2: Esquema del procedimiento de ajuste

El ajuste global del sistema de detección incipiente de fallos no resulta sencillo, fundamentalmente porque la relación existente entre los atributos de detección y los parámetros del sistema no puede expresarse matemáticamente en forma directa. Por otro lado se trata de una optimización que normalmente presenta mínimos locales, debido a los cambios discretos que experimentan los atributos de detección ante mínimas variaciones en los parámetros (ver apartado 5.4.1). Además, la dimensión del problema de optimización puede ser grande, ya que habrá un gran número de parámetros si el modelo matemático del proceso es complejo o si se utilizan varios atributos de fallo. Otra aportación de esta tesis ha sido un **estudio comparativo de las técnicas de optimización de búsqueda directa** y el desarrollo de técnicas específicas para mejorar el proceso de ajuste. Estos estudios han contribuido decisivamente a la aplicabilidad del procedimiento de ajuste a cualquier tipo de proceso.

Se ha estudiado la convergencia de los métodos de optimización por búsqueda directa pero analizando la robustez general del método, más que la velocidad de convergencia en un caso concreto. Este estudio ha permitido identificar métodos de optimización que aparentemente realizan muy pocas iteraciones para la minimización de funciones no lineales pero que se vuelven muy lentos cuando cambian ligeramente las condiciones iniciales. Dado que no se conoce la forma de la función objetivo a optimizar, lo más

apropiado es utilizar un método que nunca tenga problemas de convergencia aunque no sea en promedio el más rápido.

Para **aliviar el problema de los mínimos locales** se proponen dos procedimientos: realizar un análisis discreto de la función objetivo antes de iniciar la optimización, o introducir un filtrado que suavice la función objetivo eliminando gran parte de los mínimos locales. Ambas técnicas requieren mayor tiempo de cálculo, pero permiten obtener mejores resultados. En el ajuste del caso ejemplo (capítulo 7) se ha aplicado un muestreo de los parámetros previo a la optimización, lo que ha permitido localizar el mínimo global en cada caso. La figura 8.3 muestra la mejora que se ha obtenido al aplicar este procedimiento frente a los resultados que se obtuvieron inicialmente para el ajuste global de los parámetros.

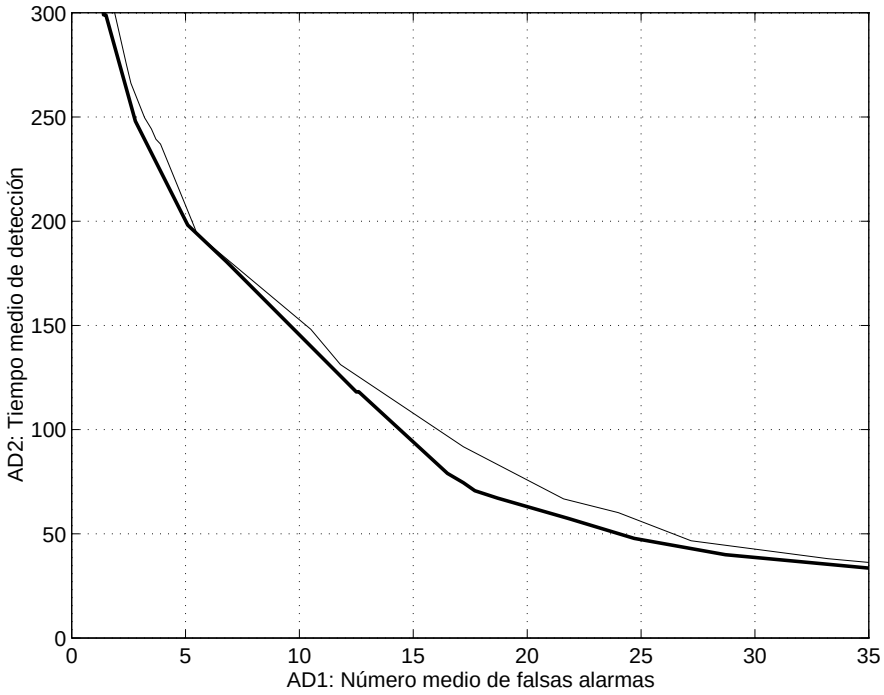


Figura 8.3: Resultados del ajuste global de todos los parámetros

Un planteamiento que sí ha contribuido a reducir drásticamente el tiempo del ajuste ha sido la división del problema en dos etapas independientes y la **optimización de los parámetros por partes**. Las dos etapas se identifican claramente gracias a la estructura del sistema de detección incipiente de fallos. La primera etapa está formada por el modelo y el generador de residuos y sólo depende de los parámetros del modelo o

de los parámetros del algoritmo de ajuste recursivo (en el caso de modelos adaptativos). La segunda etapa está formada por los atributos de fallo, la función de detección de fallos y los atributos de detección. Los cálculos de la primera etapa son los que más tiempo requieren, especialmente si el modelo es complicado, mientras que los módulos de la segunda etapa se resuelven rápidamente. Por lo tanto, para un conjunto fijo de parámetros del modelo es relativamente fácil obtener el valor óptimo del resto de los parámetros. La optimización general sólo varia los parámetros del modelo, y una segunda optimización anidada se encarga de calcular el resto de los parámetros en cada iteración.

También se propone **realizar los cálculos de forma distribuida**, para obtener la curva de soluciones óptimas en la representación multi–objetivo. Dado que es necesario realizar el ajuste del sistema para distintos pesos relativos entre los atributos de detección y que estos cálculos son independientes, resulta apropiado realizarlos en paralelo. Contando con una red de n estaciones de trabajo Unix y distribuyendo las tareas de manera uniforme se pueden obtener tiempos de cálculo n veces menores. En el caso del ajuste del sistema de detección de fallos en la suspensión de trenes se han utilizado hasta 8 estaciones de trabajo simultáneamente, lo que ha permitido reducir el tiempo de cálculo a la sexta parte comparado con haber utilizado sólo el ordenador más rápido para todos los cálculos (y suponiendo que otros usuarios no accedan a dicho ordenador durante este tiempo).

Por último, otra de las aportaciones de esta tesis ha sido **analizar las limitaciones que pueden tener los procedimientos de ajuste recursivo de parámetros** para realizar detección de fallos de manera incipiente (capítulo 6). Estas limitaciones pueden desaconsejar la utilización de estos métodos en determinados procesos (por ejemplo si las condiciones de trabajo son muy estables). Por otro lado la detección de un fallo viene acompañada de la identificación del componente afectado, siempre que se utilicen modelos basados en fundamentos físicos, lo que simplifica el proceso de diagnóstico y permite valorar la gravedad de la anomalía.

8.2 Aplicaciones

Todos los datos sobre el comportamiento de **baterías en un circuito eléctrico** que se muestran en el capítulo 2 fueron obtenidos experimentalmente. Se obtuvieron varias historias de fallos en diversos experimentos y con estos datos se demostró [Palacios97] la posibilidad de obtener mejores resultados en la detección incipiente de fallos al utilizar conjuntos de parámetros diferentes a los obtenidos por el método de mínimos cuadrados. En la figura 8.4 (tomada de [Palacios97]) se muestra una gráfica multi–objetivo donde se reflejan las prestaciones de un sistema de detección cuando se varían los parámetros del modelo. En esta gráfica el punto (×) corresponde al sistema que utiliza los parámetros obtenidos por el método de mínimos cuadrados. Puede observarse que existen otros conjuntos de parámetros que dan lugar a sistemas mejores en ambos atributos de detección (como por ejemplo el marcado con un punto ●). En este caso no se aplicó ningún algoritmo de optimización sino que la variación de parámetros se realizó uniformemente alrededor de los valores obtenidos por el método de mínimos cuadrados.

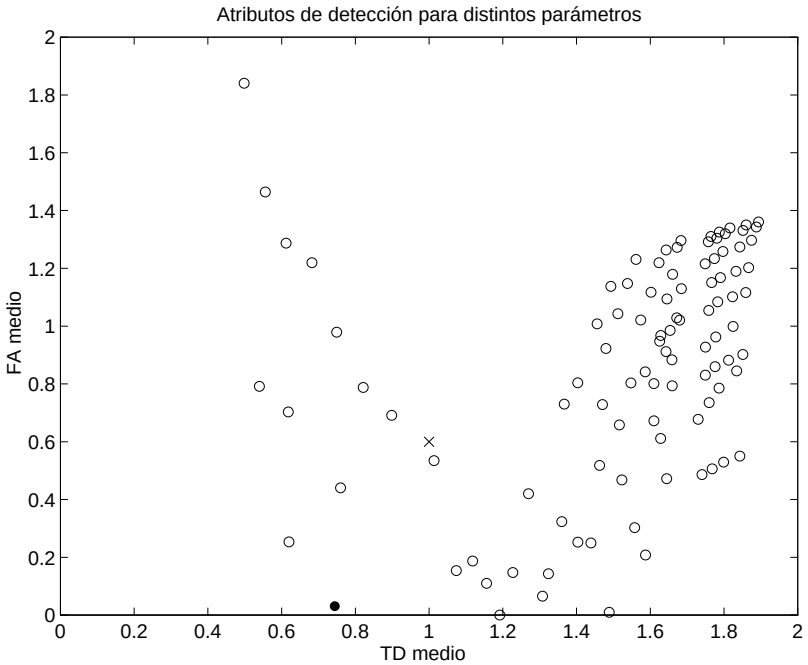


Figura 8.4: Falsas alarmas y tiempo de detección al utilizar distintos parámetros en el modelo

Una segunda aplicación ha sido el análisis de la **respuesta dinámica de la suspensión de un coche** (capítulo 4). Se ha desarrollado un modelo del sistema de suspensión que permite simular la respuesta ante un escalón de altura variable. Utilizando estas curvas de respuesta y suponiendo conocida la distancia al suelo, se ha definido un modelo matemático aplicable a la detección de fallos. Este modelo fue ajustado utilizando distintos métodos clásicos: suma de errores cuadráticos, suma de errores absolutos y error máximo. Los resultados demuestran que los criterios utilizados para el ajuste afectan a la eficacia del sistema de detección de fallos. También se justifica que el método de mínimos cuadrados no sea el procedimiento más adecuado para ajustar los modelos que se van a utilizar en detección incipiente de fallos.

Por último se ha aplicado toda la metodología para definir un sistema capaz de detectar fallos en el **sistema de suspensión de un tren**. Se ha buscado un proceso general, donde se produzcan efectos no lineales (tan comunes en la práctica) y donde exista desconocimiento del funcionamiento real del proceso. El ejemplo seleccionado para el capítulo 7 es un caso real donde la aplicación de técnicas de detección incipiente de fallos ayuda decisivamente al diagnóstico. Este caso ha sido analizado cuidadosamente, considerando multitud de condiciones de funcionamiento (diferentes trazados, diferentes condiciones de masa y diferentes velocidades de degradación), efectos no lineales en los elementos de suspensión, condiciones extremas de funcionamiento (choque de los elementos de suspensión) y ruido asociado a la señal del sensor. El modelo detallado del sistema se ha utilizado para simular las condiciones de funcionamiento con la máxima precisión. Sin embargo el modelo utilizado para el sistema de detección se ha basado en una sola medida en tiempo real, lo que dificulta la detección, para demostrar la utilidad de la metodología incluso en situaciones con información limitada.

Inicialmente se ajustó el modelo por el método de mínimos cuadrados y se utilizaron unos parámetros para los atributos de fallo que parecían razonables tras el análisis de un conjunto de historias de fallo. Utilizando el sistema de detección así definido pueden calcularse los atributos de detección y puede dibujarse sobre el diagrama multi–objetivo el punto correspondiente al sistema.

La figura 8.5 muestra la mejora que se obtiene partiendo del sistema de detección inicial, señalado con un punto (●), al aplicar la optimización de los parámetros del modelo (línea fina) o la optimización de los parámetros de los atributos de fallo (línea gruesa). En ambos casos se pueden obtener mejoras simultáneas en los dos atributos de detección considerados, lo que supone que el sistema se especializa en la detección incipiente de fallos.

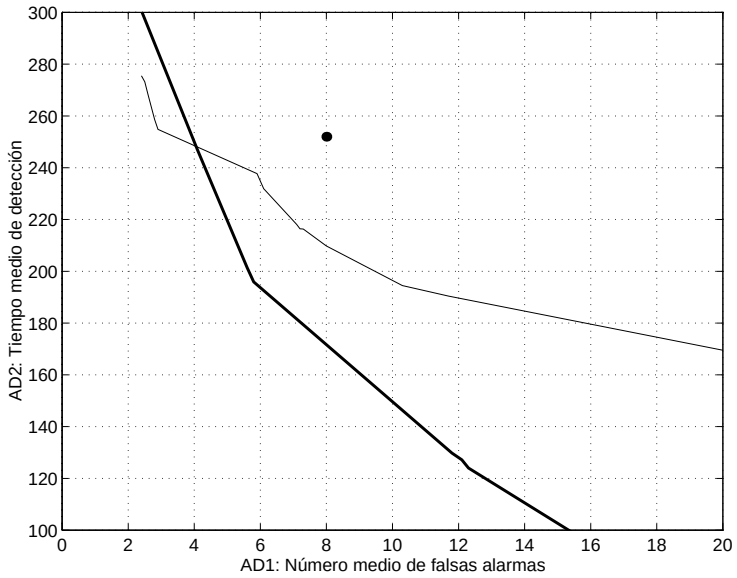


Figura 8.5: Resultados del ajuste

Como resultado final se ha realizado una optimización conjunta de todos los parámetros del sistema de detección, tal y como se propone en esta tesis. La gráfica de la figura 8.6 muestra los resultados que se obtienen al ajustar los parámetros del modelo y de los atributos de fallo de manera independiente y de manera global. El ajuste global de todos los parámetros siempre da lugar a una curva del sistema de detección que domina a todos los sistemas obtenidos mediante optimización parcial.

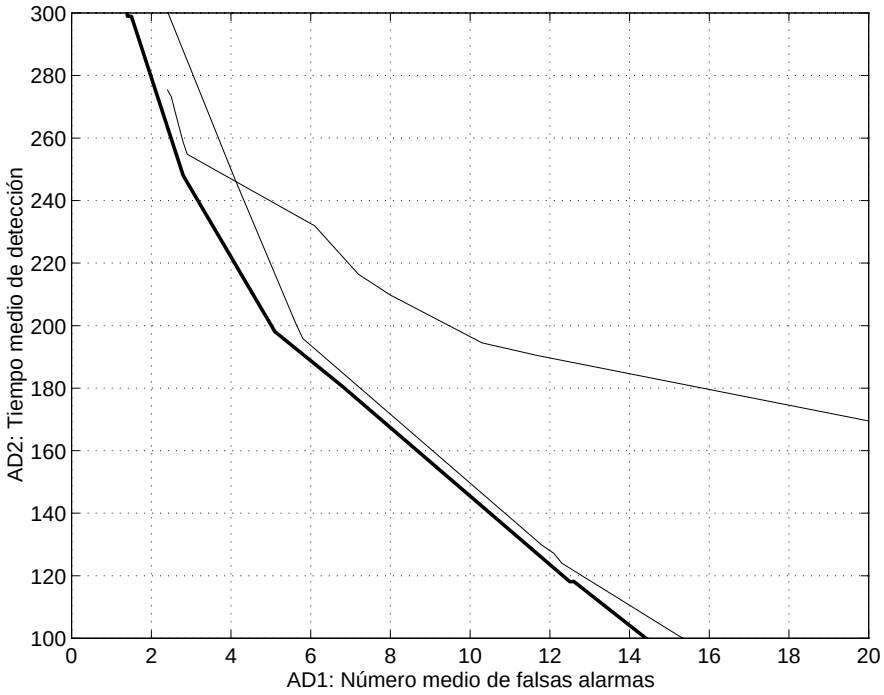


Figura 8.6: Comparación entre los distintos ajustes

8.3 Futuros desarrollos

Todos los datos del comportamiento del tren fueron simulados. A pesar del esfuerzo realizado para conseguir unos datos lo más realistas posible, no cabe duda que sería de gran interés poder disponer de una amplia **base de datos experimental** del funcionamiento normal y de situaciones de degradación de un proceso donde aplicar la metodología. Dado el éxito obtenido con datos simulados, se espera poder instrumentar un tren, dentro de las actividades de un proyecto de investigación en curso, y obtener datos reales.

En el capítulo 6 se ha presentado en detalle la manera de implantar un sistema de detección incipiente de fallos basado en el análisis de la evolución de los parámetros. También se han analizado las limitaciones que pueden tener estos sistemas de detección basados en algoritmos de estimación

recursiva. La mejor manera de comprobar estas limitaciones sería dibujar sobre el diagrama multi–objetivo las gráficas correspondientes a los sistemas basados en las dos técnicas de detección consideradas: detección basada en análisis de residuos y detección basada en análisis de parámetros. Un estudio muy interesante sería **comparar los sistemas basados en ambas técnicas de detección** para diferente procesos industriales y teniendo en cuenta el tipo de funcionamiento.

El modelo matemático utilizado en el sistema de detección de fallos del ejemplo práctico ha sido un modelo auto–regresivo de 4 parámetros. El número de parámetros se decidió por técnicas clásicas, basándose en el error del ajuste por mínimos cuadrados en función del orden del modelo y utilizando sólo datos de comportamiento normal. Sin embargo, sería interesante basar la decisión del orden del modelo, al igual que el valor de los parámetros, en los atributos de detección y no en el error de ajuste durante el funcionamiento normal. La decisión del orden del modelo en base a los atributos de detección se realiza comparando las curvas de los sistemas de detección óptimos (obtenidas para cada orden del modelo) sobre la gráfica multi–objetivo. Es posible asociar un parámetro del sistema al orden del modelo, dejando que el procedimiento de ajuste obtenga el valor óptimo del orden automáticamente. Pero resulta más interesante poder **comparar gráficamente el comportamiento de los sistemas de distinto orden** en una manera análoga (aunque orientada a la detección de fallos) a la selección basada en la curva del error de ajuste.

En la misma línea que la comparación de los órdenes del modelo estaría la **comparación de diferentes técnicas de modelado**. Es difícil definir qué técnica de modelado resulta más apropiada para cada tipo de proceso y cuando se realizan estudios sobre distintas técnicas únicamente se analiza el grado de fidelidad con el que la estimación representa el comportamiento real. Cuando el modelo va a ser utilizado para detección incipiente de fallos, la comparación de técnicas de modelado debe realizarse atendiendo a los atributos de detección, tras el ajuste completo de todos los parámetros (incluidos los parámetros de los atributos de fallo y de la función de detección de fallos) para cada técnica. Lo mismo ocurre con la comparación entre modelos adaptativos y modelos de parámetros fijos. Generalmente los primeros obtienen mejores estimaciones que los modelos equivalentes sin el algoritmo de reajuste, y sin embargo las limitaciones que pueden tener

para la detección incipiente de fallos (descritas en el capítulo 6) pueden dar lugar a peores atributos de detección.

Por consiguiente, una vez definido en esta tesis el procedimiento general que permite especializar cualquier modelo para la detección incipiente de fallos, el siguiente trabajo a realizar sería la determinación de la técnica de modelado y del orden del modelo más apropiados en cada tipo de proceso.

Apéndice A

Modelado de sistemas mediante Bond–Graphs

A.1 Introducción

La definición del modelo dinámico, basado en fundamentos físicos, de un proceso industrial es muy difícil, sobre todo cuando existe un gran número de componentes. La utilización de diagramas Bond–Graph facilita el desarrollo de modelos complejos al permitir trabajar con descripciones gráficas del proceso, que resulta más intuitivo que trabajar con ecuaciones diferenciales y multitud de variables. Además resulta sencillo añadir o quitar componentes del modelo, lo que permite obtener un modelo sencillo rápidamente y luego añadir efectos y componentes menos importantes para refinar el comportamiento (especialmente en el campo del diseño). También es posible desarrollar modelos de distintos subsistemas ya que luego resulta sencillo conectarlos y obtener todos los efectos de la influencia mutua.

Este apéndice tiene por objeto introducir las técnicas de modelado mediante Bond–Graph [Karnopp90] [Thoma90] [Xia93] y explicar los modelos utilizados en los capítulos 4 y 6.

A.2 Notación de los Bond–Graphs

El modelo de un sistema queda representado mediante un gráfico formado por enlaces (*bonds*) que interconectan nodos. Los enlaces representan flujos de potencia, y los nodos representan componentes o interconexiones de elementos del sistema.

A.2.1 Enlaces

Los enlaces se representan mediante flechas con media punta (⟶) que indican el sentido positivo del flujo de potencia. Cada enlace tiene asociadas dos variables cuyo producto representa la potencia instantánea que se transmite por el enlace, es decir la potencia que se transmite hacia un subsistema o hacia un componente. Estas variables se denominan de manera general **esfuerzo** $e(t)$ y **flujo** $f(t)$. La interpretación física de las variables esfuerzo y flujo depende del tipo de modelo, ya que la notación de los Bond–Graph es común a todas las ramas de la ingeniería. La siguiente tabla muestras en significado en cada caso:

Tabla A.1: Significado de las variables esfuerzo y flujo en distintas ramas de la ingeniería

Dominio	Esfuerzo	Flujo
Mecánica de translación	fuerza (N)	velocidad (m/s)
Mecánica de rotación	par (N m)	velocidad angular (rad/s)
Mecánica de fluidos	presión (Pa)	caudal (m^3/s)
Electrotécnia	tensión (V)	corriente (A)

A.2.2 Elementos básicos

Existen una serie de elementos que permiten representar el comportamiento de componentes básicos. Estos elementos sólo tienen un **puerto**, o punto de conexión, y por lo tanto sólo tienen un enlace con el resto del sistema. Toda la potencia que reciben o ceden al sistema fluye por este enlace. Estos elementos de un puerto se representan mediante las letras I, R, C, Se y Sf. El flujo de energía suele digujarse de acuerdo a la figura A.1.

Figura A.1: Elementos de un puerto

Esta manera representación se debe a que los elementos I, R y C suelen considerarse consumidores de energía mientras que Se y Sf son generadores de energía. El elemento I (*inertia*) acumula energía en forma de flujo, R (*resistor*) degrada energía y C (*capacitor)* acumula energía en forma de esfuerzo. Por otro lado Se (*source*) es un generador de esfuerzo y Sf es un generador de flujo.

Cada uno de estos elementos tiene un comportamiento esencialmente distinto y por lo tanto cada elemento impone una ligadura entre esfuerzo y flujo diferente. La tabla A.2 muestra la relación entre esfuerzo y flujo que imponen estos elementos y algunos ejemplos de componentes característicos dentro de cada materia.

Tabla A.2: Elementos de un puerto

	I	R	C	Se	Sf
Relación e—f	$e=\mathrm{I}\cdot\dot{f}$	$e=\mathrm{R}\cdot f$	$e=\frac{1}{C}\int f\,dt$	$e=\mathrm{Se}$	$f=\mathrm{Sf}$
Mec. Translación	masa	amortiguador	muelle		
Mec. Rotación	inercia		eje a torsión		
Mec. Fluidos		válvula	depósito		
Electrotecnia	bobina	resistencia	condensador	fuente de tensión	fuente de corriente

A.2.3 Elementos con dos puertos

Sólo son necesarios 2 tipos básicos de elementos de 2 puertos. En ellos se producen transformaciones de esfuerzo y de flujo pero no se acumula, ni se degrada, ni se genera energía; por lo tanto la potencia que entra al elemento es igual a la que sale.

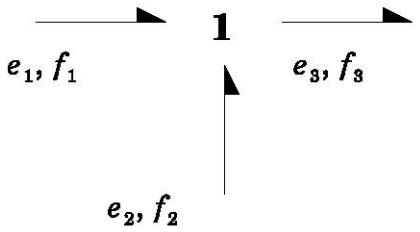
Estos dos elementos son el transformador (*transformer*), que se representa con TF y el girador (*gyrator*), que se representa mediante GY. Los ejemplos más conocidos son el transformador y el giróscopo. Estos elementos cumplen las siguientes ecuaciones:

$$\begin{array}{lll} \text{TF} & e_1\,f_1=e_2\,f_2 & e_1=m\cdot e_2\ ,\quad m\cdot f_1=f_2 \\ \text{GY} & e_1\,f_1=e_2\,f_2 & e_1=r\cdot f_2\ ,\quad r\cdot f_1=e_2 \end{array} \qquad \textbf{(A.1)}$$

A.2.4 Elementos multi–puerto

Son las uniones tipo 1 y tipo 0, que se utilizan para interconectar elementos o subsistemas y formar el Bond–Graph del proceso. Al igual que los elementos TF y GY, no acumulan energía.

En las uniones tipo 1 se cumple que todos los flujos son iguales y por lo tanto se establece el equilibrio de esfuerzos de acuerdo a las direcciones de potencia definidas. Por ejemplo:



$$f_1=f_2=f_3$$

$$e_1 + e_2 = e_3$$

Figura A.2: Unión tipo 1

En mecánica de translación representa el equilibrio dinámico de fuerzas que se aplican a igual velocidad. Por ejemplo, varias fuerzas que se aplican sobre el mismo componente indeformable.

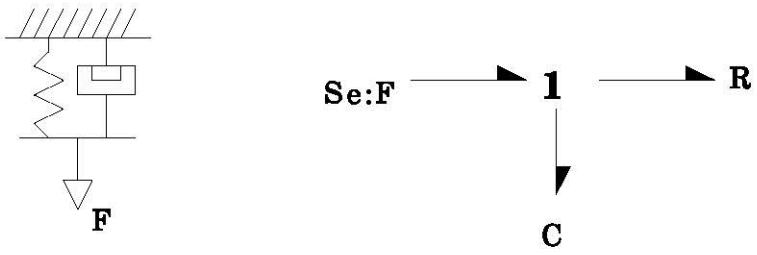
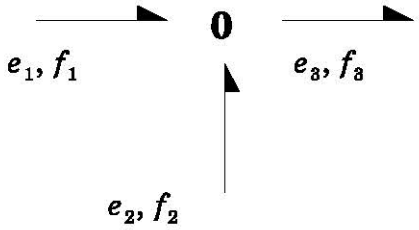


Figura A.3: Ejemplo mecánico de la unión tipo 1

En electrotecnia, la unión tipo 1 representa la igualdad de corriente con equilibrio de tensiones; es decir, la ley de tensiones de Kirchhoff aplicada en una malla. Los elementos eléctricos conectados en serie pueden experimentar diferentes caídas de tensión, de manera que la suma de todas las tensiones de la malla se anula, pero la corriente que los atraviesa es la misma.

Por el contrario, en las uniones tipo 0 se cumple que todos los esfuerzos son iguales y por lo tanto se establece un equilibrio de flujos. En este caso la representación gráfica y las ecuaciones de equilibrio son las siguientes:



$$e_1 = e_2 = e_3$$
$$f_1 + f_2 = f_3$$

Figura A.4: Unión tipo 0

En mecánica de translación la unión tipo 0 representa la compatibilidad geométrica para la misma fuerza y distintas velocidades. Es por ejemplo un sistema formado por un muelle y un amortiguador en serie, cada punto pude moverse a una velocidad diferente (pero de manera que la elongación total sea la suma de las elongaciones de cada elemento) pero la fuerza que soporta cada elemento es la misma.

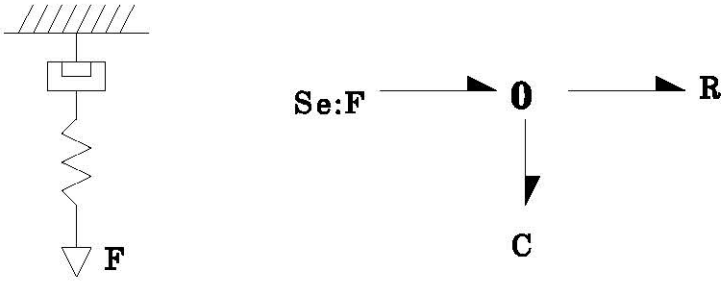


Figura A.5: Ejemplo mecánico de la unión tipo 0

En electrotecnia, la unión tipo 0 representa la ley de corrientes de Kirchhoff aplicada a un nudo. Todos los elementos que se conectan al mismo nudo adquieren igual tensión, mientras que las corrientes se dividen de tal forma que su suma se anula.

Como puede comprobarse en los ejemplos anteriores, las uniones tipo 1 o tipo 0 determinan la manera en que están conectados los elementos del sistema. Dos sistemas formados por los mismos elementos admiten múltiples configuraciones dependiendo del tipo de montaje. Es necesario

conocer perfectamente de qué manera están acoplados los componentes para poder definir el modelo.

A.3 Ejemplos de sistemas de suspensión

En el capítulo 4 y en el capítulo 7 se han mostrado dos ejemplos de sistemas de suspensión. El primer caso se trata del sistema de suspensión de un coche y es un sistema formado por dos masa, dos muelles y un amortiguador. Las masas corresponden a la rueda y al coche, un muelle representa el comportamiento elástico del neumático y el otro es parte de la suspensión, y el amortiguador también es parte de la suspensión. Este sistema tiene la siguiente representación:

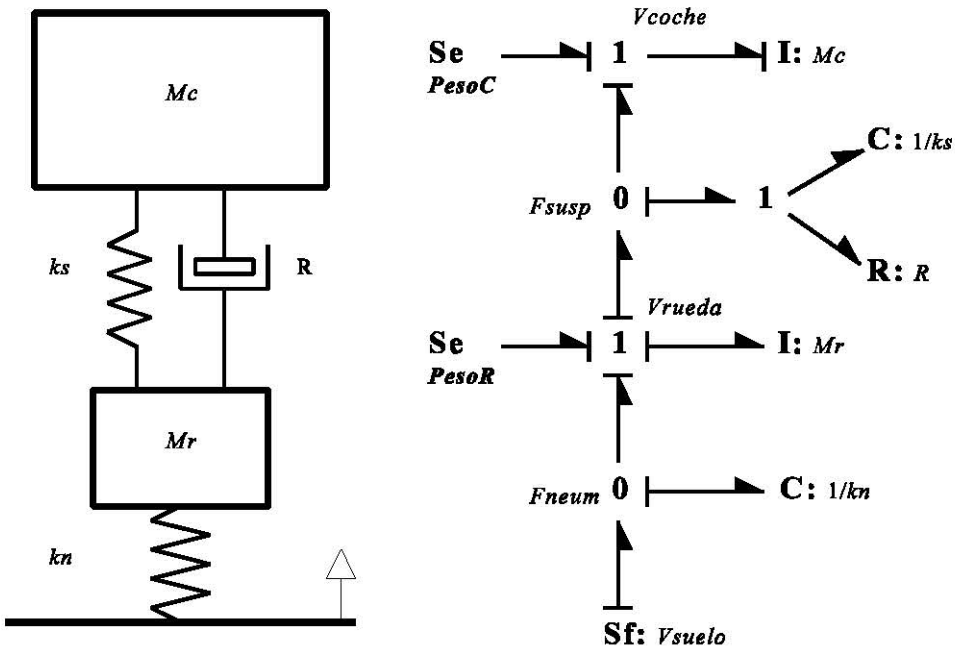


Figura A.6: Esquema del sistema de suspensión

Las fuentes de este sistema son los pesos de las masas (2 fuentes de esfuerzo, Se) y las variaciones del suelo (fuente de flujo, Sf).

Las dos uniones 1 de la vertical principal del Bond–Graph representan las velocidades de las masas, en ellas se establece el equilibrio de todas las fuerzas que actúan sobre las mismas.

Los ceros representan las fuerzas del neumático y de la suspensión. En este caso la fuerza que transmite la suspensión hacia arriba y hacia abajo es la misma, mientras que la velocidad de las dos masas puede ser diferente. La diferencia de velocidades queda absorbida en el conjunto muelle—amortiguador. El conjunto muelle—amortiguador sufre las mismas elongaciones; es decir el muelle y el amortiguador se mueven a igual velocidad aunque reparten los esfuerzos de manera diferente.

El segundo ejemplo es bastante parecido, se trata del sistema de suspensión de un tren. La mayor diferencia es que en este caso existen dos amortiguadores, uno en la suspensión primaria y otro en la suspensión secundaria. El esquema del sistema es el siguiente:

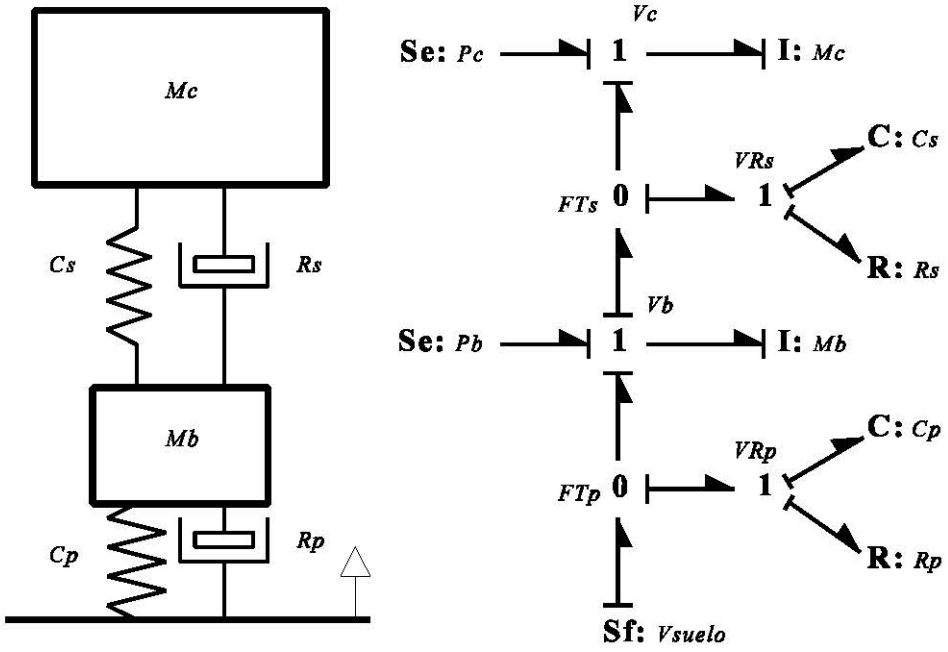


Figura A.7: Esquema de la suspensión del tren

Puede observarse que la única diferencia del modelo de Bond–Graph es la sustitución del muelle de abajo (elemento C) por el subsistema formado por un muelle y un amortiguador en paralelo; es decir, por un elemento C y un

elemento R conectador por una unión tipo 1. Generalmente, cualquier otra modificación de los elementos de la suspensión no supone más que algunos pequeños cambios en los enlaces y la incorporación de algún 1 o algún 0, teniendo cuidado para identificar si el elemento actúa en seria o en paralelo.

La obtención de las ecuaciones matemáticas del modelo tampoco supone ninguna dificultad. En cada unión tipo 1 se establece una ecuación de equilibrio de fuerzas y en cada unión tipo 0 otra de equilibrio de velocidades. Seguidamente se establecen las ecuaciones propias de cada elemento (descritas en la tabla A.2).

A partir de la representación en Bond–Graph es sencillo escribir un pequeño programa para realizar la simulación del sistema. Varios programas permiten realizar simulaciones a partir de modelos Bond–Graph: DRAM [DRAM], ENPORT [ENPORT] y el utilizado en este caso TUTSIM [TUTSIM].

Apéndice B

Bibliografía

B.1 Clasificación por temas

B.1.1 Detección de fallos, monitorización y diagnóstico

[Basseville83]
Michèle Basseville, Albert Benveniste
"Design and comparative study of some sequential jump detection algorithms for digital signals"
IEEE Transactions on Acoustic, Speech and Signal processing. Vol 31, No. 3, June 1983, p. 521-535

[Chow84]
Edward Y. Chow, Alan S. Willsky
"Analytical redundancy and the design of robust failure detection systems"
IEEE Transactions on Automatic Control, Vol AC-29, No. 7, p. 603-614

[Constantinescu]
Raluca F. Constantinescu, Peter D. Lawrence, Phillip G. Hill, Terrence S. Brown
"Model-Based fault diagnosis of a two-stroke diesel Engine"
IEEE International Conference on System Man and Cybernetics. Vancouver oct/95. Vol 3, p. 2216-2220, 1995

[Crowley90]
Thomas H. Crowley
"Automated diagnosis of large power transformers using adaptive model-based monitoring"
M.S. thesis. Massachusetts Institute of Technology. Cambridge MA, U.S.A. June 1990

[Dasgupta93]
Abhijit Dasgupta
"Failure mechanism models for cyclic fatigue"
IEEE Transactions on Reliability. Vol 42, No. 4, p. 548-555. 1993

[Filbert92]
D.Filbert
"Technical diagnosis and testing of electric motors in quality assurance."
8th International IMEKO symposium on Technical diagnostics. Dresden, Germany. Sep/92

[Frank90]
Paul M. Frank
"Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy. A survey and some new results"
Automatica. Vol 26, No. 3, p. 459-474, 1990

[Gertler88]
Janos J. Gertler
"Survey of model-based failure detection and isolation in complex plants"
IEEE Control Systems Magazine, Dec/88. p. 3-11, 1988

[Gertler91]
J. Gertler
"Analytical redundancy methods in fault detection and isolation. Survey and synthesis"
IFAC Symposium. Fault detection, supervision and safety for technical processes. Baden-Baden, Germany 1991. p.9-21

[Gertler93]
J.J. Gertler, M. Costin, Xiaowen Fang, R. Hira, Z. Kowalczuk, Qiang Luo
"Model-Based on-board fault detection and diagnosis for automotive engines"
Control Engineering Practice, Vol 1, No. 1, p. 3-17, 1993

[Gertler95]
Janos Gertler, Mark Costin, Xiaowen Fang, R. Hira, Zdzislaw Kowalczuk, Moid Kunwer, Ramin Monajemy
"Model based diagnosis for automotive engines -Algorithm development and testing on a production vehicle"
IEEE Transactions on Control Systems Technology. Vol 3, No.1, March 1995, p.61-69

[Gustafsson92]
Fredrik Gustafsson
"Estimation of discrete parameters in linear systems"
PhD Thesis. Linköping University. Linköping, Sweden, 1992.
ISBN 91-7870-876-1

[Isermann84]
Rolf Isermann
"Process fault detection based on modeling and estimation methods. A survey"
Automatica. Vol. 20, No. 4, p. 387-404, 1984

[Isermann93]
Rolf Isermann
"Fault diagnosis of machines via parameter estimation and knowledge processing. Tutorial Paper"
Automatica. Vol 29, No. 4, p. 815-835, 1993

[Joussellin95]
A. Joussellin
"Diagnosis of faults in EDF power plants: From monitoring to diagnosis"
Collection de notes internes de la direction des etudes et recherches. EDF 95NB00034

[Kirtley96]
James L. Kirtley, Wayne H. Hagman, Bernard C. Lesieutre, Mary Jane Boyd, E.Paul Warren, Hsiu P. Chou, Richard D. Tabors
"Monitoring the health of power transformers"
IEEE Computer Applications in Power, Jan 1996, p.18-23

[Kobbacy97]
K.A.H.Kobbacy, B.B.Fawzi, D.F.Percy, H.E.Ascher
"A full history proportional hazards model for preventive maintenance scheduling"
Quality and reliability engineering international. Vol 13, p. 187-198, 1997

[Lavalle92]
Juan C. Lavalle, Rafael Collantes, Miguel A. Sanz, Rafael Palacios
"SEDIMAHE: Sistema experto para el diagnóstico de máquinas herramientas"
X Congreso Nacional de Ingeniería Mecánica, Madrid, ESPAÑA. Anales de ingeniería mecánica. Año 9, No.2, sep/92, p.151-153

[Lavalle93]
Juan C. Lavalle, Rafael Collantes, Miguel A. Sanz, Rafael Palacios
"SEDIMAHE: An expert system to help the maintenance of machine tools"
Maintenance, Volume 8, Number 3, September 1993. Pages 10–14

[Maffezzoni92]
C. Maffezzoni
"Issues in modelling and simulation of power plants"
IFAC Symposium. Control of Power Plants and Power Systems. Munich, Germany 1992.

[Mehra71]
R.K. Mehra, J. Peschon
"An innovations approach to fault detection and diagnosis in dynamic systems"
Automatica, Vol 7, p. 637-640, 1971

[Min90]
Paul S. Min
"Validation of controller inputs in electronically controlled engines"
Proceedings of the 1990 American Control Conference. San Diego, May/90. Vol 3, p. 2887-2890, 1990

[Moré77]
Jorge J. Moré
"The Levenberg-Marquardt algorithm: Implementation and theory"
Proceedings of the biennial conference on numerical analysis, p. 105-116. Published by Springer-Varleg, Berlin, Germany 1977

[Muñoz96]
Antonio Muñoz San Roque
"Aplicación de técnicas de redes neuronales artificiales al diagnóstico de procesos industriales"
Tesis Doctoral. Universidad Pontificia Comillas, 1996

[Palacios97]
Rafael Palacios
"Detección incipiente de fallos: aplicación a la detección de la descarga de una batería en un circuito eléctrico"
Mantenimiento. No 108, oct/97, p.33-40

[Panossian97]
H.V.Panossian, W.D.Ewing
"Real-Time failure detection algorithm for the Space Shuttle main engine"
IEEE Control Systems. Vol 17, No. 4, p. 16-23, Aug/97

[Piety88]
K.R. Piety, E.F. Pardue, J.R. Cain, E.P. Phillips, R.H. Greene
"Periodic vibration monitoring: Utility experience"
Electric Power Research Institute, EPRI CS-5517, March 1988

[Pouliezos89]
A. Pouliezos, G. Stavrakakis, C. Lefas
"Fault detection using parameter estimation"
Quality and reliability engineering international. Vol 5, p. 283-290. 1989

[Reason95]
John Reason
"On-Line transformer monitoring. Special Report"
Electrical world, Oct 1995, p.19-26

[Rizzoni91]
Giorgio Rizzoni, Paul S. Min
"Detection of sensor failures in automotive engines"
IEEE Transactions on Vehicular Technology. Vol 40, No. 2, may 1991, p. 487-500

[Sanz92]
Miguel Angel Sanz Bobi
"Metodología de mantenimiento predictivo basada en análisis espectral y temporal de la historia de equipos industriales y enfoque de su aplicación a un sistema experto."
Tesis Doctoral. Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid, 1992

[Sanz93-1]
Miguel A. Sanz, Aurelio García-Cerrada, Rafael Palacios, José Villar, José Rolan, Alfonso Luengo, Angel M. Alonso, Juan C. Burgos, Francisco Fernández
"TRAFES: Expert system for diagnosis of power transformers"
Cigré:Symposium de Berlin, Abril 1993. Artículo 110-20

[Sanz93-2]
Miguel A. Sanz, Aurelio García-Cerrada, Rafael Palacios, José Villar, José Rolan, Alfonso Luengo, Angel M. Alonso, Juan C. Burgos, Francisco Fernández
"TRAFES: Sistema experto de diagnóstico de transformadores de potencia"
3as. Jornadas Hispano-Lusas. Barcelona, Jul/93

[Sanz94]
M.A. Sanz, J.I. Pérez-Arriaga, J.L. Serrano-Carbayo, M.E. Ortiz-Alfaro, J.J. Alba, A. Doménech, M.J. Villamediana, J. González-Huerta, J.J. Fernández-Martínez
"Control and diagnosis of water chemistry in the water-steam and water make-up in a fossil fuelled power plant"
Electrical Power & Energy Systems. Vol 16, No.4, 1994, p.251-258

[Tugnait82]
Jitendra K. Tugnait
"Detection and estimation for abruptly changing systems"
Automatica, Vol 18, p. 607-615, 1982

[Young81]
Peter Young
"Parameter estimation for continuous-time models. A survey"
Automatica, Vol 17, p. 23-39, 1981

B.1.2 Optimización

[Bandler69]
John W. Bandler, Patrick A. MacDonald
"Optimization of microwave networks by Razor Search"
IEEE Transactions on Microwave theory and techniques, vol MTT-17, no.8, August 1969, p.552-562

[Bertsekas79]
Dimitri P. Bertsekas
"Notes on nonlinear programming and discrete-time optimal control"
Department of electrical engineering and computer science. M.I.T., july 79

[deCuadra90]
Fernando de Cuadra García
"El problema general de la optimización de diseño por ordenador: Aplicación de técnicas de ingeniería del conocimiento."
Tesis Doctoral. Universidad Pontificia Comillas, 1990

[Farin96]
Gerald E. Farin
"Curves and surfaces for computer-aided design"
Academic Press Inc. 4th edition 1996. ISBN 0-122-49054-1

[Gill81]
Philip E. Gill, Walter Murray, Margaret H. Wright
"Practical optimization."
Academic Press, 1981. ISBN 0-12-283952-8

[Hansen92]
Eldon Hansen
"Global optimization using interval analysis"
Marcel Dekker, 1992. ISBN 0-8247-8696-3

[Hooke & Jeeves 61]
Robert Hooke, T.A. Jeeves
"'Direct Search' solution of numerical and statical problems"
Journal of the ACM, vol 8, 1961, p.212-229

[Luenberger84]
David G. Luenberger
"Linear and nonlinear programming. Second edition"
Addison-Wesley publishing company, 1984. ISBN 0-201-15794-2

[Mat1]
"Matlab Reference Guide"
The MathWorks, Inc. 1992

[Mat2]
Andrew Grace
"Optimization toolbox for use with Matlab"
The MathWorks, Inc. 1990

[Murray72]
W. Murray, C.G. Broyden, R. Fletcher, M.J.D. Powell, W.H. Swann
"Numerical methods for unconstrained optimization"
Academic Press, 1972. ISBN 0-12-512250-0

[Nelder & Mead 64]
J.A. Nelder, R. Mead
"A simplex method for function minimization"
Computer Journal, vol 7, 1964, p.308-313

[Nocedal92]
Jorge Nocedal
"Theory of algorithms for unconstrained optimization"
Acta Numerica, vol 1, p.199-242. Cambridge University Press, 1992

[Pike86]
Ralph W. Pike
"Optimization for engineering systems"
Van Nostrand Reinhold company, 1986. ISBN 0-442-27581-1

[Polyak87]
Boris T. Polyak
"Introduction to optimization"
Optimization Software Inc, 1987. ISBN 0-911575-14-6

[Powell64]
M.J.D. Powell
"An efficient method of finding the minimum of a function of several variables without calculating derivatives."
Computer Journal, vol 7, 1964, p.155-162

[Prudnikov93]
I.M. Prudnikov
"The method of global optimization of a funcion and estimation of the speed of its convergence"
Automatica and Remote Control, vol 54, No. 12, p. 1785-1793, 1993

[Reklaitis83]
G.V. Reklaitis, A. Ravindran, K.M. Ragsdell
"Engineering Optimization. Methods and applications."
John Wiley and sons, 1983. ISBN 0-471-05579-4

[Rosenbrock60]
H.H. Rosenbrock
"An automatic method for finding the greatest or least value of a function"
Computer Journal, vol 3, 1960, p.175-184

[Schweppe86]
F.C. Schweppe, H.M. Merril
"Multiple attribute trade-off analysis"
Electric Power Research Institute, EPRI RP 2537, April 1986

[Shanno90]
David F. Shanno
"Recent advances in numerical techniques for large-scale optimization"
Chapter 7 of Neural networks for control. M.I.T. Press, 1990.
ISBN 0-262-13261-3

[Swann]
W.H. Swann
"Report on the development of a new direct search method of optimization"
ICI Ltd. Central Instrument Research Laboratory. Research note 64/3

B.1.3 Computación

[Golub89]
Gene H. Golub, Charles F. Van Loan
"Matrix computations. Second edition"
The Johns Hopkins University Press, 1989. ISBN 0-8018-3772-3

[Marquardt63]
D. Marquardt
"An algorithm for Least-squares estimation of nonlinear paramenters"
SIAM Journal of applied mathematics, vol 11, p. 431-441, 1963.

[Press92]
William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery
"Numerical recipes in C. The art of scientific computing. Second edition"
Cambridge University Press, 1992. ISBN 0-521-43108-5

B.1.4 Señales, sistemas y control

[Basseville86]
M.Basseville, A.Benveniste
"Detection of abrupt changes in signals and dynamical systems"
Springer-Verlag, 1986. ISBN 0-387-16043-4

[Basseville87]
Michèle Basseville, Albert Benveniste, Georges Moustakides, Anne Rougée
"Detection and diagnosis of changes in the eigenstructure of nonstationary multivariable systems"
Automatica. Vol 23, No. 4, p. 479-489, 1987

[Basseville88]
Michèle Basseville
"Detecting changes in signals and systems. A survey"
Automatica. Vol 24, No. 3, p. 309-326, 1988

[Basseville93]
Michèle Basseville, Igor V. Nikiforov
"Detection of abrupt changes. Theory and applications"
Prentice-Hall, Inc. 1993. ISBN 0-13-126780-9

[Hsia77]
T.C. Hsia
"System identification. Least-squares methods."
D.C. Heath and Company, 1977. ISBN 0-669-99630-0

[Ludeman87]
Lonnie C. Ludeman
"Fundamentals of digital signal processing"
John Wiley & Sons, 1987. ISBN 0-471-61306-1

[Oppenheim83]
Alan V. Oppenheim, Alan S. Willsky, Ian T. Young
"Signals and systems"
Prentice Hall, 1983. ISBN 0-13-811175-8

[Sastry89]
Shankar Sastry, Marc Bodson
"Adaptive Control. Stability, convergence and robustness"
Prentice HAll, 1989. ISBN 0-13-004367-2

[Söderström89]
Torsten Söderström, Petre Stoica
"System Identification"
Prentice Hall, 1989. ISBN 0-13-881236-5

[Vélez-Reyes88]
Miguel Vélez-Reyes
"Speed and parameter estimation for induction machines"
MS Thesis. Massachusetts Institute of Technology. Cambridge MA, U.S.A. May 1988.

[Vélez-Reyes92]
Miguel Vélez-Reyes
"Decomposed algorithms for parameter estimation"
Ph.D. thesis. Massachusetts Institute of Technology. Cambridge MA, U.S.A. Sep 1992

[Wellstead91]
P.E. Wellstead, M.B. Zarrop
"Self-tuning systems. Control and signal processing"
John Wiley & Sons, 1991. ISBN 0-471-93054-7

B.1.5 Lógica borrosa y redes neuronales

[Dubois91]
Didier Dubois, Henri Prade
"Fuzzy sets in approximate reasoning, Part 1: Interference with possibility distributions"
Fuzzy sets and systems, vol 40, p.143-202, 1991.

[Klir88]
George J. Klir, Tina A. Folger
"Fuzzy sets, uncertainty and information"
Prentice Hall International Editions, 1988. ISBN 0-13-345638-2

[Terano91]
Toshiro Terano, Kiyji Asai, Michio Sugeno
"Fuzzy systems theory and its applications"
Academic Press Inc, 1991. ISBN 0-12-685245-6

[Cox92]
Earl Cox
"Fuzzy fundamentals"
IEEE Spectrum, Oct/92, p.58-61

[Hammerstrom93]
Dan Hammerstrom
"Working with neural networks"
IEEE Spectrum, Jul/93, p.46-53

[Sjöberg95]
Jonas Sjöberg
"Non-Linear System Idenfication with Neural Networks"
Dept. of Electrical Engineering. Linköping University, Sweden. 1995. ISBN 91-7871-534-2

B.1.6 Baterías eléctricas

[Powers95]
Robert A. Powers
"Batteries for low power electronics"
Proceedings of the IEEE, vol 83, no 4, april 1995, p.687-93

[Riezenman95]
Michael J. Riezenman
"The search for better batteries"
IEEE Spectrum. May/95, p.51-56

[Smith95]
Robert L. Smith Jr.
"Control batteries: Power system life savers"
IEEE Industrial Applications Magazine. Nov-Dic/95, p.18-25

B.1.7 Dinámica de vehículos

[Dugoff70]
Howard Dugoff, P.S. Fancher, Leonerd Segel
"An analysis of tire traction properties and their influence on vehicle dynamic performance"
1970 International Automobile Safety Conference Compendium. Society of Automotive Engineers. Paper 700377

[Dugoff71]
Howard Dugoff, Leonerd Segel, R.D. Ervin
"Measurement of vehicle response in severe braking and steering maneuvers"
Society of Automotive Engineers, 1971. Paper 710080

[Elbeheiry96]
E.M. Elbeheiry, D.C. Karnopp
"Optimal control of vehicle random vibration with constrained suspension deflection"
Journal of sound and vibration. 1996, vol 189, No. 5, pages 547-564

[Padovan94]
J.Padovan, P.Padovan
"Modelling tyre performance during anti-lock braking"
Tire Sci. Technology. Vol 33, No. 3, pag 182-204, 1994.

[Ray95]
Laura R. Ray
"Nonlinear state and tire force estimation for advanced vehicle control"
IEEE Transactions on control systems technology. Vol 3, No. 1, p. 117-124. March 1995

[Srinivasa80]
R. Srinivasa, R.R. Gunter, J.Y. Wong
"Evaluation of the performance of anti-lock brake systems using laboratory simulation techniques."
International Journal of Vehicular Design. Vol 1, No. 5, pag.467-485, 1980.

[Wellstead97]
P.E. Wellstead
"Analysis and redesign of an antilock brake system controller"
IEE Proceedings Control Theory and Applications. Vol 144, No. 5, pp. 413-426. Sep/1997

B.1.8 Bond–Graph

[DRAM]
"DRAM: Dynamic response of articulated mechanisms"
Mechanical dynamics, Inc. Ann Arbor, Michigan, USA

[ENPORT]
"The ENPORT reference manual"
Rosencode Associates, Inc. Lansing, Michigan, USA

[Karnopp90]
Dean C. Karnopp, Donald L. Margolis, Ronald C. Rosenberg
"System dynamics. A unified approach"
John Wiley & Sons, Inc. Second edition 1990. ISBN 0-471-62171-4

[Thoma90]
Jean U. Thoma
"Simulation by Bondgraphs. Introduction to a graphical method"
Springer-Verlag, 1990. ISBN 0-387-51640-9

[TUTSIM]
"TUTSIM on IBM-PC computer. Users Manual"
Meerman Automation. The Netherlands. ISBN 90-73117-03-8

[Xia93]
S. Xia, D.A. Linkens, S. Bennett
"Automatic modelling and analysis of dynamic physical systems using qualitative reasoning and bond graphs"
Intelligent systems engineering. Vol 2, No. 3, p. 201-212. Autumn 1993

B.2 Orden alfabético de referencias

[Bandler69]
John W. Bandler, Patrick A. MacDonald
"Optimization of microwave networks by Razor Search"
IEEE Transactions on Microwave theory and techniques, vol MTT-17, no.8, August 1969, p.552-562

[Basseville83]
Michèle Basseville, Albert Benveniste
"Design and comparative study of some sequential jump detection algorithms for digital signals"
IEEE Transactions on Acoustic, Speech and Signal processing. Vol 31, No. 3, June 1983, p. 521-535

[Basseville86]
M.Basseville, A.Benveniste
"Detection of abrupt changes in signals and dynamical systems"
Springer-Verlag, 1986. ISBN 0-387-16043-4

[Basseville87]
Michèle Basseville, Albert Benveniste, Georges Moustakides, Anne Rougée
"Detection and diagnosis of changes in the eigenstructure of nonstationary multivariable systems"
Automatica. Vol 23, No. 4, p. 479-489, 1987

[Basseville88]
Michèle Basseville
"Detecting changes in signals and systems. A survey"
Automatica. Vol 24, No. 3, p. 309-326, 1988

[Basseville93]
Michèle Basseville, Igor V. Nikiforov
"Detection of abrupt changes. Theory and applications"
Prentice-Hall, Inc. 1993. ISBN 0-13-126780-9

[Bertsekas79]
Dimitri P. Bertsekas
"Notes on nonlinear programming and discrete-time optimal control"
Department of electrical engineering and computer science. M.I.T., july 79

[Chow84]
Edward Y. Chow, Alan S. Willsky
"Analytical redundancy and the design of robust failure detection systems"
IEEE Transactions on Automatic Control, Vol AC-29, No. 7, p. 603-614

[Constantinescu]
Raluca F. Constantinescu, Peter D. Lawrence, Phillip G. Hill, Terrence S. Brown
"Model-Based fault diagnosis of a two-stroke diesel Engine"
IEEE International Conference on System Man and Cybernetics. Vancouver oct/95. Vol 3, p. 2216-2220, 1995

[Cox92]
Earl Cox
"Fuzzy fundamentals"
IEEE Spectrum, Oct/92, p.58-61

[Crowley90]
Thomas H. Crowley
"Automated diagnosis of large power transformers using adaptive model-based monitoring"
M.S. thesis. Massachusetts Institute of Technology. Cambridge MA, U.S.A. June 1990

[Dasgupta93]
Abhijit Dasgupta
"Failure mechanism models for cyclic fatigue"
IEEE Transactions on Reliability. Vol 42, No. 4, p. 548-555. 1993

[deCuadra90]
Fernando de Cuadra García
"El problema general de la optimización de diseño por ordenador: Aplicación de técnicas de ingeniería del conocimiento."
Tesis Doctoral. Universidad Pontificia Comillas, 1990

[DRAM]
"DRAM: Dynamic response of articulated mechanisms"
Mechanical dynamics, Inc. Ann Arbor, Michigan, USA

[Dubois91]
Didier Dubois, Henri Prade
"Fuzzy sets in approximate reasoning, Part 1: Interference with possibility distributions"
Fuzzy sets and systems, vol 40, p.143-202, 1991.

[Dugoff70]
Howard Dugoff, P.S. Fancher, Leonerd Segel
"An analysis of tire traction properties and their influence on vehicle dynamic performance"
1970 International Automobile Safety Conference Compendium. Society of Automotive Engineers. Paper 700377

[Dugoff71]
Howard Dugoff, Leonerd Segel, R.D. Ervin
"Measurement of vehicle response in severe braking and steering maneuvers"
Society of Automotive Engineers, 1971. Paper 710080

[Elbeheiry96]
E.M. Elbeheiry, D.C. Karnopp
"Optimal control of vehicle random vibration with constrained suspension deflection"
Journal of sound and vibration. 1996, vol 189, No. 5, pages 547-564

[ENPORT]
"The ENPORT reference manual"
Rosencode Associates, Inc. Lansing, Michigan, USA

[Farin96]
Gerald E. Farin
"Curves and surfaces for computer-aided design"
Academic Press Inc. 4th edition 1996. ISBN 0-122-49054-1

[Filbert92]
D.Filbert
"Technical diagnosis and testing of electric motors in quality assurance."
8th International IMEKO symposium on Technical diagnostics. Dresden, Germany. Sep/92

[Frank90]
Paul M. Frank
"Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy. A survey and some new results"
Automatica. Vol 26, No. 3, p. 459-474, 1990

[Gertler88]
Janos J. Gertler
"Survey of model-based failure detection and isolation in complex plants"
IEEE Control Systems Magazine, Dec/88. p. 3-11, 1988

[Gertler91]
J. Gertler
"Analytical redundancy methods in fault detection and isolation. Survey and synthesis"
IFAC Symposium. Fault detection, supervision and safety for technical processes. Baden-Baden, Germany 1991. p.9-21

[Gertler93]
J.J. Gertler, M. Costin, Xiaowen Fang, R. Hira, Z. Kowalczuk, Qiang Luo
"Model-Based on-board fault detection and diagnosis for automotive engines"
Control Engineering Practice, Vol 1, No. 1, p. 3-17, 1993

[Gertler95]
Janos Gertler, Mark Costin, Xiaowen Fang, R. Hira, Zdzislaw Kowalczuk, Moid Kunwer, Ramin Monajemy
"Model based diagnosis for automotive engines -Algorithm development and testing on a production vehicle"
IEEE Transactions on Control Systems Technology. Vol 3, No.1, March 1995, p.61-69

[Gill81]
Philip E. Gill, Walter Murray, Margaret H. Wright
"Practical optimization."
Academic Press, 1981. ISBN 0-12-283952-8

[Golub89]
Gene H. Golub, Charles F. Van Loan
"Matrix computations. Second edition"
The Johns Hopkins University Press, 1989. ISBN 0-8018-3772-3

[Gustafsson92]
Fredrik Gustafsson
"Estimation of discrete parameters in linear systems"
PhD Thesis. Linköping University. Linköping, Sweden, 1992. ISBN 91-7870-876-1

[Hammerstrom93]
Dan Hammerstrom
"Working with neural networks"
IEEE Spectrum, Jul/93, p.46-53

[Hansen92]
Eldon Hansen
"Global optimization using interval analysis"
Marcel Dekker, 1992. ISBN 0-8247-8696-3

[Hooke & Jeeves 61]
Robert Hooke, T.A. Jeeves
"'Direct Search' solution of numerical and statical problems"
Journal of the ACM, vol 8, 1961, p.212-229

[Hsia77]
T.C. Hsia
"System identification. Least-squares methods."
D.C. Heath and Company, 1977. ISBN 0-669-99630-0

[Isermann84]
Rolf Isermann
"Process fault detection based on modeling and estimation methods. A survey"
Automatica. Vol. 20, No. 4, p. 387-404, 1984

[Isermann93]
Rolf Isermann
"Fault diagnosis of machines via parameter estimation and knowledge processing. Tutorial Paper"
Automatica. Vol 29, No. 4, p. 815-835, 1993

[Joussellin95]
A. Joussellin
"Diagnosis of faults in EDF power plants: From monitoring to diagnosis"
Collection de notes internes de la direction des etudes et recherches. EDF 95NB00034

[Karnopp90]
Dean C. Karnopp, Donald L. Margolis, Ronald C. Rosenberg
"System dynamics. A unified approach"
John Wiley & Sons, Inc. Second edition 1990. ISBN 0-471-62171-4

[Kirtley96]
James L. Kirtley, Wayne H. Hagman, Bernard C. Lesieutre, Mary Jane Boyd, E.Paul Warren, Hsiu P. Chou, Richard D. Tabors
"Monitoring the health of power transformers"
IEEE Computer Applications in Power, Jan 1996, p.18-23

[Klir88]
George J. Klir, Tina A. Folger
"Fuzzy sets, uncertainty and information"
Prentice Hall International Editions, 1988. ISBN 0-13-345638-2

[Kobbacy97]
K.A.H.Kobbacy, B.B.Fawzi, D.F.Percy, H.E.Ascher
"A full history proportional hazards model for preventive maintenance scheduling"
Quality and reliability engineering international. Vol 13, p. 187-198, 1997

[Lavalle92]
Juan C. Lavalle, Rafael Collantes, Miguel A. Sanz, Rafael Palacios
"SEDIMAHE: Sistema experto para el diagnóstico de máquinas herramientas"
X Congreso Nacional de Ingeniería Mecánica, Madrid, ESPAÑA. Anales de ingeniería mecánica. Año 9, No.2, sep/92, p.151-153

[Lavalle93]
Juan C. Lavalle, Rafael Collantes, Miguel A. Sanz, Rafael Palacios
"SEDIMAHE: An expert system to help the maintenance of machine tools"
Maintenance, Volume 8, Number 3, September 1993. Pages 10–14

[Ludeman87]
Lonnie C. Ludeman
"Fundamentals of digital signal processing"
John Wiley & Sons, 1987. ISBN 0-471-61306-1

[Luenberger84]
David G. Luenberger
"Linear and nonlinear programming. Second edition"
Addison-Wesley publishing company, 1984. ISBN 0-201-15794-2

[Maffezzoni92]
C. Maffezzoni
"Issues in modelling and simulation of power plants"
IFAC Symposium. Control of Power Plants and Power Systems. Munich, Germany 1992.

[Marquardt63]
D. Marquardt
"An algorithm for Least-squares estimation of nonlinear paramenters"
SIAM Journal of applied mathematics, vol 11, p. 431-441, 1963.

[Mat1]
"Matlab Reference Guide"
The MathWorks, Inc. 1992

[Mat2]
Andrew Grace
"Optimization toolbox for use with Matlab"
The MathWorks, Inc. 1990

[Mehra71]
R.K. Mehra, J. Peschon
"An innovations approach to fault detection and diagnosis in dynamic systems"
Automatica, Vol 7, p. 637-640, 1971

[Min90]
Paul S. Min
"Validation of controller inputs in electronically controlled engines"
Proceedings of the 1990 American Control Conference. San Diego, May/90. Vol 3, p. 2887-2890, 1990

[Moré77]
Jorge J. Moré
"The Levenberg-Marquardt algorithm: Implementation and theory"
Proceedings of the biennial conference on numerical analysis, p. 105-116. Published by Springer-Varleg, Berlin, Germany 1977

[Muñoz96]
Antonio Muñoz San Roque
"Aplicación de técnicas de redes neuronales artificiales al diagnóstico de procesos industriales"
Tesis Doctoral. Universidad Pontificia Comillas, 1996

[Murray72]
W. Murray, C.G. Broyden, R. Fletcher, M.J.D. Powell, W.H. Swann
"Numerical methods for unconstrained optimization"
Academic Press, 1972. ISBN 0-12-512250-0

[Nelder & Mead 64]
J.A. Nelder, R. Mead
"A simplex method for function minimization"
Computer Journal, vol 7, 1964, p.308-313

[Nocedal92]
Jorge Nocedal
"Theory of algorithms for unconstrained optimization"
Acta Numerica, vol 1, p.199-242. Cambridge University Press, 1992

[Oppenheim83]
Alan V. Oppenheim, Alan S. Willsky, Ian T. Young
"Signals and systems"
Prentice Hall, 1983. ISBN 0-13-811175-8

[Padovan94]
J.Padovan, P.Padovan
"Modelling tyre performance during anti-lock braking"
Tire Sci. Technology. Vol 33, No. 3, pag 182-204, 1994.

[Palacios97]
Rafael Palacios
"Detección incipiente de fallos: aplicación a la detección de la descarga de una batería en un circuito eléctrico"
Mantenimiento. No 108, oct/97, p.33-40

[Panossian97]
H.V.Panossian, W.D.Ewing
"Real-Time failure detection algorithm for the Space Shuttle main engine"
IEEE Control Systems. Vol 17, No. 4, p. 16-23, Aug/97

[Piety88]
K.R. Piety, E.F. Pardue, J.R. Cain, E.P. Phillips, R.H. Greene
"Periodic vibration monitoring: Utility experience"
Electric Power Research Institute, EPRI CS-5517, March 1988

[Pike86]
Ralph W. Pike
"Optimization for engineering systems"
Van Nostrand Reinhold company, 1986. ISBN 0-442-27581-1

[Polyak87]
Boris T. Polyak
"Introduction to optimization"
Optimization Software Inc, 1987. ISBN 0-911575-14-6

[Pouliezos89]
A. Pouliezos, G. Stavrakakis, C. Lefas
"Fault detection using parameter estimation"
Quality and reliability engineering international. Vol 5, p. 283-290. 1989

[Powell64]
M.J.D. Powell
"An efficient method of finding the minimum of a function of several variables without calculating derivatives."
Computer Journal, vol 7, 1964, p.155-162

[Powers95]
Robert A. Powers
"Batteries for low power electronics"
Proceedings of the IEEE, vol 83, no 4, april 1995, p.687-93

[Press92]
William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery
"Numerical recipes in C. The art of scientific computing. Second edition"
Cambridge University Press, 1992. ISBN 0-521-43108-5

[Prudnikov93]
I.M. Prudnikov
"The method of global optimization of a funcion and estimation of the speed of its convergence"
Automatica and Remote Control, vol 54, No. 12, p. 1785-1793, 1993

[Ray95]
Laura R. Ray
"Nonlinear state and tire force estimation for advanced vehicle control"
IEEE Transactions on control systems technology. Vol 3, No. 1, p. 117-124. March 1995

[Reason95]
John Reason
"On-Line transformer monitoring. Special Report"
Electrical world, Oct 1995, p.19-26

[Reklaitis83]
G.V. Reklaitis, A. Ravindran, K.M. Ragsdell
"Engineering Optimization. Methods and applications."
John Wiley and sons, 1983. ISBN 0-471-05579-4

[Riezenman95]
Michael J. Riezenman
"The search for better batteries"
IEEE Spectrum. May/95, p.51-56

[Rizzoni91]
Giorgio Rizzoni, Paul S. Min
"Detection of sensor failures in automotive engines"
IEEE Transactions on Vehicular Technology. Vol 40, No. 2, may 1991, p. 487-500

[Rosenbrock60]
H.H. Rosenbrock
"An automatic method for finding the greatest or least value of a function"
Computer Journal, vol 3, 1960, p.175-184

[Sanz92]
Miguel Angel Sanz Bobi
"Metodología de mantenimiento predictivo basada en análisis espectral y temporal de la historia de equipos industriales y enfoque de su aplicación a un sistema experto."
Tesis Doctoral. Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid, 1992

[Sanz93-1]
Miguel A. Sanz, Aurelio García-Cerrada, Rafael Palacios, José Villar, José Rolan, Alfonso Luengo, Angel M. Alonso, Juan C. Burgos, Francisco Fernández
"TRAFES: Expert system for diagnosis of power transformers"
Cigré:Symposium de Berlin, Abril 1993. Artículo 110-20

[Sanz93-2]
Miguel A. Sanz, Aurelio García-Cerrada, Rafael Palacios, José Villar, José Rolan, Alfonso Luengo, Angel M. Alonso, Juan C. Burgos, Francisco Fernández
"TRAFES: Sistema experto de diagnóstico de transformadores de potencia"
3as. Jornadas Hispano-Lusas. Barcelona, Jul/93

[Sanz94]
M.A. Sanz, J.I. Pérez-Arriaga, J.L. Serrano-Carbayo, M.E. Ortiz-Alfaro, J.J. Alba, A. Doménech, M.J. Villamediana, J. González-Huerta, J.J. Fernández-Martínez
"Control and diagnosis of water chemistry in the water-steam and water make-up in a fossil fuelled power plant"
Electrical Power & Energy Systems. Vol 16, No.4, 1994, p.251-258

[Sastry89]
Shankar Sastry, Marc Bodson
"Adaptive Control. Stability, convergence and robustness"
Prentice HAll, 1989. ISBN 0-13-004367-2

[Schweppe86]
F.C. Schweppe, H.M. Merril
"Multiple attribute trade-off analysis"
Electric Power Research Institute, EPRI RP 2537, April 1986

[Shanno90]
David F. Shanno
"Recent advances in numerical techniques for large-scale optimization"
Chapter 7 of Neural networks for control. M.I.T. Press, 1990.
ISBN 0-262-13261-3

[Sjöberg95]
Jonas Sjöberg
"Non-Linear System Idenfication with Neural Networks"
Dept. of Electrical Engineering. Linköping University, Sweden. 1995.
ISBN 91-7871-534-2

[Smith95]
Robert L. Smith Jr.
"Control batteries: Power system life savers"
IEEE Industrial Applications Magazine. Nov-Dic/95, p.18-25

[Söderström89]
Torsten Söderström, Petre Stoica
"System Identification"
Prentice Hall, 1989. ISBN 0-13-881236-5

[Srinivasa80]
R. Srinivasa, R.R. Gunter, J.Y. Wong
"Evaluation of the performance of anti-lock brake systems using laboratory simulation techniques."
International Journal of Vehicular Design. Vol 1, No. 5, pag.467-485, 1980.

[Swann]
W.H. Swann
"Report on the development of a new direct search method of optimization"
ICI Ltd. Central Instrument Research Laboratory. Research note 64/3

[Terano91]
Toshiro Terano, Kiyji Asai, Michio Sugeno
"Fuzzy systems theory and its applications"
Academic Press Inc, 1991. ISBN 0-12-685245-6

[Thoma90]
Jean U. Thoma
"Simulation by Bondgraphs. Introduction to a graphical method"
Springer-Verlag, 1990. ISBN 0-387-51640-9

[Tugnait82]
Jitendra K. Tugnait
"Detection and estimation for abruptly changing systems"
Automatica, Vol 18, p. 607-615, 1982

[TUTSIM]
"TUTSIM on IBM-PC computer. Users Manual"
Meerman Automation. The Netherlands. ISBN 90-73117-03-8

[Vélez-Reyes88]
Miguel Vélez-Reyes
"Speed and parameter estimation for induction machines"
MS Thesis. Massachusetts Institute of Technology. Cambridge MA, U.S.A. May 1988.

[Vélez-Reyes92]
Miguel Vélez-Reyes
"Decomposed algorithms for parameter estimation"
Ph.D. thesis. Massachusetts Institute of Technology. Cambridge MA, U.S.A. Sep 1992

[Wellstead91]
P.E. Wellstead, M.B. Zarrop
"Self-tuning systems. Control and signal processing"
John Wiley & Sons, 1991. ISBN 0-471-93054-7

[Wellstead97]
P.E. Wellstead
"Analysis and redesign of an antilock brake system controller"
IEE Proceedings Control Theory and Applications. Vol 144, No. 5, pp. 413-426. Sep/1997

[Xia93]
S. Xia, D.A. Linkens, S. Bennett
"Automatic modelling and analysis of dynamic physical systems using qualitative reasoning and bond graphs"
Intelligent systems engineering. Vol 2, No. 3, p. 201-212. Autumn 1993

[Young81]
Peter Young
"Parameter estimation for continuous-time models. A survey"
Automatica, Vol 17, p. 23-39, 1981