UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

PREDICCIÓN NO LINEAL EN LÍNEA DE SERIES DE TIEMPO MEDIANTE EL USO Y MEJORA DE ALGORITMOS DE FILTROS ADAPTIVOS DE KERNEL

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN ELÉCTRICA

MEMORIA PARA OPTAR AL TITULO DE INGENIERO CIVIL ELÉCTRICO

IVÁN ALONSO CASTRO OJEDA

PROFESOR GUÍA:
FELIPE TOBAR HENRÍQUEZ
PROFESOR CO-GUÍA:
JORGE SILVA SÁNCHEZ

MIEMBROS DE LA COMISIÓN:
MARCOS ORCHARD CONCHA
PABLO ZEGERS FERNÁNDEZ

SANTIAGO DE CHILE
2018

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN ELÉCTRICA
POR: IVÁN ALONSO CASTRO OJEDA
FECHA: 2018
PROF. GUÍA: FELIPE TOBAR HENRÍQUEZ

PREDICCIÓN NO LINEAL EN LÍNEA DE SERIES DE TIEMPO MEDIANTE EL USO Y MEJORA DE ALGORITMOS DE FILTROS ADAPTIVOS DE KERNEL

El modelamiento de series de tiempo es un problema transversal a diferentes áreas de ingeniería y ciencias. Este tópico, visto a través del foco de aprendizaje de máquinas o aprendizaje estadístico, se reduce a elegir un modelo de regresión que sea lo suficientemente flexible sin que sobreajuste al conjunto de entrenamiento y, por ende, permita generalizar. No obstante, la elección de modelos flexibles suele venir de la mano de poca interpretabilidad de los mismos, como por ejemplo en modelos con estructura tipo *caja negra*. Los modelos más flexibles son preferidos para problemas de alta complejidad, debido a su ajuste con mayor precisión a las observaciones. Más aún, el ajuste de los modelos predictivos es una componente crìtica para la regresión en línea aplicada a problemas reales.

Es por ello que se decide abordar el tema del aprendizaje en línea para series de tiempo no lineales a través de un modelo flexible, que extiende la teoría del filtrado adaptivo lineal, al caso no lineal, haciendo uso de transformación de espacio de características basadas en *kernel* reproductivos. Los objetivos de la investigación realizada son (i) presentar e interpretar el estimador de filtro de *kernel* adaptivo (KAF) al contexto de regresión no lineal de series de tiempo, (ii) extender, en términos de mejoras sobre el algoritmo y el ajuste de sus hiperparámetros, la aplicación estándar de KAF validada sobre series sintéticas y datos reales y (iii) acercar la interpretabilidad y aplicabilidad de los métodos KAF para usuarios, validando la mejora tanto en desempeño predictivo como en ajuste de modelos con las extensiones propuestas.

Para ello, este trabajo de investigación reúne los resultados principales de dos investigaciones previas, la primera enfocada en mejorar la predicción de KAF utilizando una entrada exógena de un sistema. En ese contexto se estudió el comportamiento de descarga de batería de ion-litio para una bicicleta eléctrica que utilizaba como entrada exógena mediciones de altitud derivadas a partir de coordenadas de geolocalización. El objetivo era caracterizar la posible dependencia oculta a través del descubrimiento automático de relevancia de las variables al momento de la predicción; para lo cual se usó un *kernel* Gaussiano de Determinación de Relevancia Automática (ARD). Por otro lado, la segunda investigación se centró en la validación de una metodología para la inicialización de KAF extendiendo el estimador a una variante probabilística para mejorar su desempeño y entrenamiento, proponiendo hibridar la estimación en línea adicionando un entrenamiento en *batch* que permite encontrar los hiperparámetros óptimos de la extensión propuesta. Adicionalmente, este enfoque permitió proponer un regularizador novedoso para abordar dos de los problemas más desafiantes de diseño según el estado del arte para KAF: el ajuste del hiperparámetro del *kernel* Gaussiano y el tamaño del diccionario usado por el estimador. La metodología fue validada tanto en datos sintéticos, específicamente para el caso del atractor caótico de Lorentz, como en datos reales, los cuales correspondieron a una serie de viento extraída a partir de mediciones de anemométro.

Ambos estudios mostraron resultados prometedores, acercando el uso de KAF a usuarios neófitos, tanto por las metodologías desarrolladas que quedan como guías metodológicas aplicadas, como por la interpretabilidad proporcionada a través de toda la investigación, caracterización y desarrollo del uso de KAF. Finalmente se dejan desafíos futuros con respecto a promover más aún la automatización con respecto a la selección de hiperparámetros del modelo, lo que culminaría con un desarrollo completamente adaptivo de estos métodos, vale decir, con intervención mínima del usuario en la selección de los hiperparámetros.

A mis padres

Agradecimientos

El logro de culminar una larga aventura que fue el insertarse en el mundo de investigación para lograr proponer soluciones y metodologías en el estado del arte, ha sido producto de un conjunto de factores y personas detrás. Agradezco profundamente a Felipe Tobar, con quien tuve la oportunidad de compartir todo este proceso bajo su supervisión, revisión y discusión de enfoques relativos a mi investigación. Debido a ello, este proceso fue muy enriquecedor en términos del aprendizaje obtenido tanto por el desarrollo de mi propia investigación, como la participación constante con el grupo de investigación GAMES-CMM. Gracias Felipe por mostrarme como realizar trabajos de excelencia y por la paciencia durante todo este proceso de aprendizaje hasta la culminación del mismo.

Por otro lado, quiero agradecer a Jorge Silva por la confianza puesta en mi formación y las herramientas que me proporcionó para desarrollar de manera autónoma durante estos últimos años esta travesía y gracias por el espacio dentro del grupo IDS-DIE, el cual fue rico en discusión académica y camaradería. A Marcos Orchard, por ser un eje de inspiración para mi como alumno de la carrera, tanto por su desempeño de excelencia, como su cercanía y preocupación. Agradezco la manera en que me motivaba siempre a explorar o dirigía a nuevas problemáticas relativas a las herramientas que iba desarrollando, buscando nuevos trabajos o problemáticas desafiantes. Y si, obviamente por el equipo que uso sobre el cual redacto estas palabras y que le haré entrega cerrado este proceso... Finalmente a Pablo Zegers por aceptar formar parte del proceso de evaluación y a su propuesta de que solucionemos problemas de realidad nacional, que me han hecho explorar nuevas posibilidades dentro del área de la salud.

A mi familia le agradezco todo su soporte por estos años, en particular al cariño que siempre me mostraron mis abuelos, a mi abuelo Juan que gracias a su ayuda tuve la tranquilidad económica para cursar mi pregrado gracias a su beca, y a mi abuela Gabriela que hubiese estado muy orgullosa de mis estudios en la Casa de Bello. No tengo manera de agradecer suficiente en lo que me resta de vida a mis padres, quienes siempre me han proporcionado las herramientas, tranquilidad y amor al momento de perseguir mis sueños y metas. A mi madre Angélica, que siempre ha estado a mi lado incondicionalmente, quien desde la infancia me ha inculcado valores, desafíos y respeto los otros, tanto de su capacidad profesional como humana. A mi padre Rodrigo por su preocupación, cariño y constancia con su trabajo durante estos años.

A mis amigos, del grupo más cercano del DIE "Los Opamps": Romina, Joaquín, Carolina, José R., José O., Christopher, Claudia, Ignacio, Gerardo y Sergio. En particular agradecer a Cristóbal por diversos trabajos que realizamos durante la carrera y finalmente el último

con que ganamos el mejor paper de la conferencia, ¡secos! También a Jacqueline y Cecilia por la buena onda y amenas conversaciones. A mis amigos de Plan Común, Ivana, Felipe G. y Natalia, con quienes he mantenido el contacto desde el comienzo de la carrera y se han convertido en personas entrañables. Y finalmente, agradecer al último par de Felipes, Santibañez por apoyarme profesionalmente en este último tiempo y mostrarme el mundo en la Facultad de Medicina y a Valle por nuestras salidas culinarias, amistad, videojuegos y acogimiento en su familia durante la travesía a sus tierras.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Glosario y Acrónimos

ARD Determinación de Relevancia Automática *(Automatic Relevance Determination)*.

ARX Modelo Autorregresivo con entrada Exógena *(Auto-Regressive with eXogenous input)*.

batch Entrenamiento por lote *(batch)*.

KAF Filtro de Kernel Adaptivo *(Kernel Adaptive Filter)*.

LMS Optimización de Mínimos Cuadrados*(Least Mean Square)*.

MAP Máximo A Posteriori.

MCMC Simulación estocástica con algoritmo de Montecarlo para Cadenas de Markov *(Markov Chain Montecarlo)*.

MKLMS Filtro de múltiples kernel con funcional LMS *(Multiple-Kernel Least Mean Squares)*.

MSE Error Cuadrático Medio *(Mean Square Error)*.

RKHS Espacio de Hilbert con Kernel Reproductor *(Reproducing Kernel Hilbert Space)*.

RLS Optimización de Mínimos Cuadrados Recursivos *(Recursive Least Squares)*.

SOC Estado de Carga *(State of Charge)*.

sparse Modelo con representación rala o eficiente en la cantidad de términos.

Declaración de autoría

Parte del trabajo que conforma esta tesis ha sido presentado previamente en

- **I. Castro**, F. Tobar. Improvement of voltage prediction combining altitude measurements using kernel least mean square algorithm. Presentado en *concurso de pósters EVIC*, 2016.
- F. Tobar, **I. Castro**, J. Silva, and M. Orchard. Improving battery voltage prediction in an electric bicycle using altitude measurements and kernel adaptive filters. *Pattern Recognition Letters*, 2017.
- **I. Castro**, C. Silva, and F. Tobar. Initialising kernel adaptive filters via probabilistic inference. In *Proc. of the IEEE International Conference on Digital Signal Processing*, 2017.
- **I. Castro**, F. Tobar. A practical guide for online learning with kernels. Presentado en *concurso de pósters EVIC*, 2017.

Capítulo 1

Introducción

En el área de aprendizaje de máquinas [14] se hace cada vez más necesario contar con modelos que puedan ajustarse a un comportamiento que es incierto a priori, debido al gran volumen de datos que continuamente se generan día a día y a la posibilidad de explotar dicha información en aplicaciones industriales y tecnológicas. Para ello, es necesario contar con modelos flexibles que puedan adaptarse en base a observaciones de algún fenómeno de interés, y que puedan propagar el aprendizaje obtenido para predecir su comportamiento futuro. Lo anterior es posible en el contexto de las señales temporales que cumplan la propiedad Markoviana [16], vale decir, que manifiesten una dependencia temporal a una cantidad finita de pasos en el pasado, sobre los cuales se encuentre caracterizado un modelo del sistema.

El estudio de las señales temporales o *series de tiempo* es fundamental en diversas áreas de la ingenería y ciencias. Ejemplos de éstas abarcan las series financieras [9], de clima [21], estado de carga de dispositivos [30], información de sensores [7], entre otras. Dependiendo de las características del fenómeno temporal a estudiar, existen distintas metodologías para modelar los datos obtenidos, reconociéndose dos enfoques principales: si se encuentra consolidado un conocimiento a priori del fenómeno, entonces el modelo descriptivo es denominado *fenomenológico*, el cual se desprende de leyes conocidas que gobiernan las observaciones. Por otro lado, en carencia de dicho conocimiento, o frente a un vasto volumen de datos, el enfoque preferido es realizar un modelo empírico basados en datos, el cual genera o adapta modelos que minimicen algún costo o riesgo empírico condicional a las observaciones.

Los modelos *data driven* requieren de muestras de la misma serie temporal para ajustar sus parámetros. Dicho ajuste es conocido como *entrenamiento del modelo* y sujeto a la disponibilidad de las observaciones se clasifica en *entrenamiento fuera de línea* u *offline* o *batch*, o en *entrenamiento en línea* u *online*. El entrenamiento *batch* consiste en utilizar un subconjunto de las muestras para ajustar el modelo, para posteriormente realizar predicción sobre datos no observados en el entrenamiento. Por otro lado, el entrenamiento en línea permite actualizar los parámetros del aprendizaje en la medida que nuevos datos son procesados.

La ventaja de construir un algoritmo con entrenamiento y predicción en línea radica en su adaptabilidad frente a diferentes escenarios de operación y generalidad en su implementación. De esta manera, la metodología de implementación se hace más independiente del problema a modelar. No obstante, esta flexibilidad al momento de modelar viene arraigada a considerar nuevos desafíos, tales como convergencia de los métodos usados para entrenar el modelo, selección de datos de entrenamiento, transformación de espacio de características, entre otros. Estos requerimientos vienen dados por el algoritmo a utilizar y, dependiendo de cada método, requieren mayor conocimiento por parte del usuario al momento de aplicarlos.

Uno de los modelos flexibles utilizado en el área de ingeniería eléctrica corresponde al filtro lineal adaptivo. Dicho modelo tiene su génesis en el área de procesamiento de señales [28], el cual extiende el modelo clásico del filtro lineal [1] para su actualización basada en un funcional de riesgo empírico, típicamente el error cuadrático medio o MSE. Para la actualización en tiempo real del modelo se utilizan métodos de optimización sobre el funcional de riesgo empírico como el método de gradiente descendente aplicado sobre los coeficientes del mismo.

[1] modelo que utiliza observaciones como entradas, aplica una transformación lineal a éstas (lo que se traduce en una ponderación de las entradas), la cual posteriormente se entrega como salida del modelo

Para el caso de esta investigación, el modelo de predicción en línea será construido a partir del algoritmo de *filtro de kernel adaptivo (KAF por sus siglas en inglés)*. KAF forma parte de los algoritmos que utilizan el truco del kernel, beneficiándose de la no linealidad otorgada por la función de kernel. Para conseguirlo, se modifica el dominio de entrada usando una función no lineal, de manera de permitir un ajuste no lineal manteniendo un modelo lineal en los parámetros.

Las interrogantes abiertas con respecto al uso de KAF están dadas por la selección del hiperparámetro del kernel, selección de datos relevantes a través de los cuales construir el estimador y el ajuste óptimo de parámetros relativos al modelo KAF. Múltiples investigaciones recientes han abordado dichos tópicos para los algoritmos KAF [4, 5, 23, 26, 29], proponiendo diferentes soluciones enfocadas usualmente a un algoritmo o variante aplicado a la mejora del estado del arte de dichos métodos.

1.1. Hipótesis

El objetivo de esta tesis es aportar en el desarrollo de KAF en pos de la automatización y acercamiento para usuarios no expertos, proponiendo soluciones que promuevan la aplicación completamente adaptiva de KAF para la regresión no lineal de series de tiempo. De manera más concreta, se abordan mejoras aplicadas en la selección del hiperparámetro del kernel y construcción eficiente del diccionario, mostrando resultados tanto en ejemplos sintéticos como datos reales a partir de la implementación estándar del algoritmo kernel con mínimos cuadrados (KLMS). Es por lo anterior, que se enuncia la siguiente hipótesis de este trabajo:

Es posible formular una variante de KAF que establezca un criterio de aprendizaje automático para el hiperparámetro del kernel para el caso Gaussiano. Dicha variante debería mejorar tanto la facilidad de implementación, evaluación y desempeño de los algoritmos KAF para el contexto de regresión no lineal de series de tiempo.

1.1.1. Objetivo General

El objetivo general de la tesis consiste en mejorar el desempeño para la regresión no lineal de series de tiempo a través del diseño de reglas de aprendizaje automático de parámetros para KAF.

1.1.2. Objetivos Específicos

- Diseñar una aplicación para el aprendizaje de relaciones no lineales que plantee el problema de múltiples canales a través del uso de KAF con variante ARD. Bajo este enfoque, abordar el problema de usar entrada exógena con ponderación automática por ARD.
- Desarrollar variante de KAF con ajuste de parámetros que consideren construcción del diccionario y transformación no lineal simultáneamente. Esto es, utilizando una extensión probabilística de KAF que propone un prior que aborda ambos problemas a la vez.
- Validar sobre datos sintéticos y reales el desempeño de las variantes adaptivas diseñadas. Los datos sintéticos consideran la serie del atractor caótico de Lorentz, mientras que los datos reales son de voltaje de descarga de bicicleta eléctrica y velocidad de viento medida por anemómetro.

1.1.3. Estructura de la tesis

El siguiente documento se organiza en 4 capítulos adicionales, en el Capítulo 2 se introducen los conceptos básicos para entender la solución del problema de regresión no lineal a través del uso de KAF, en el Capítulo 3 detalla la teoría y metodología relativa a las contribuciones generadas a partir de la teoría básica de KAF, el Capítulo 4 muestra la validación experimental de la teoría desarrollado en el capítulo de contribuciones, y el Capítulo 5 recapitula los resultados obtenidos en los experimentos e identifica los desafíos abiertos que se identificaron tanto en la investigación como desarrollo de mejoras.

Capítulo 2

Preeliminares

2.1. Definiciones, propiedades y conceptos fundamentales

A continuación se caracterizan nociones generales para comprender la teoría relativa a KAF a partir del concepto de *kernel* y los espacios funcionales que involucran, denotados como espacios de Hilbert con *kernel* reproductor (RKHS por sus siglas en inglés). En términos generales, los espacios de Hilbert corresponden a espacios vectoriales funcionales completos y normados; más aún, poseen un producto interno. Por otro lado, los RKHS exigen una condición adicional que define una función particular que caracteriza dicho espacio: la función de *kernel*.

Para entender como se relaciona la función de *kernel* con el espacio $\mathcal{H}_\mathcal{K}$ (espacio de Hilbert con *kernel* $\mathcal{K}$), se caracterizará dicho espacio a partir de la noción del funcional de evaluación.

Definición 2.1 *Funcional de evaluación*
Un funcional de evaluación δ_x se define como $\delta_x : \mathcal{H} \to \mathbb{R}$ tal que $\delta_x f := f(x) \quad \forall x \in \mathbb{X}$. En otras palabras, un funcional de evaluación toma una función del espacio funcional $\mathcal{H}$ y la evalúa sobre el dominio $\mathbb{X}$.

Para los RKHS, dicho funcional caracteriza completamente dichos espacios, como indica la siguiente definición.

Definición 2.2 *Espacio de Hilbert con Kernel Reproductor RKHS*
Un RKHS corresponde a un espacio de Hilbert funcional, sobre el cual el funcional de evaluación es continuo en el espacio $\mathcal{H}_\mathcal{K}$.

En otras palabras, la distancia entre funciones dentro de un RKHS es "suave" con respecto a su dominio. Por ejemplo, la función $x^3 + \frac{1}{50x}$ no pertenecería a un RKHS dada la discontinuidad producida al aproximarse a $x = 0$ por la izquierda o por la derecha, como lo muestra la Figura 2.1.

A partir del funcional de evaluación, se puede caracterizar la propiedad reproductora de manera concisa, la cual nos dirá como hacer regresión con *kernels* a traves del "truco del *kernel*".

Proposición 2.3 (ver [2]) *Propiedad Reproductora*
El funcional de evaluación $\delta_x f$ en $\mathcal{H}$ puede ser expresado como

$$\delta_x f = \langle f, \mathcal{K}(\cdot, x) \rangle \quad \forall f \in \mathcal{H}_\mathcal{K} \ , \tag{2.1}$$

donde $\mathcal{K}$ es denotado como el kernel reproductor de $\mathcal{H}_\mathcal{K}$ y $\langle \cdot, \cdot \rangle$ es el producto interno de $\mathcal{H}$

Básicamente, la propiedad reproductora permite caracterizar la naturaleza de esta función de *kernel* como elemento del producto interno de funciones del espacio $\mathcal{H}_\mathcal{K}$, cumpliendo con ser miembro del mismo espacio.

En [2] se muestra que el *kernel* reproductor es único, vale decir, cada RKHS posee su propio *kernel* reproductor y es único a dicho espacio. Adicionalmente, el *kernel* reproductor admite diferentes descomposiciones como producto interno entre distintas funciones. Aquí es

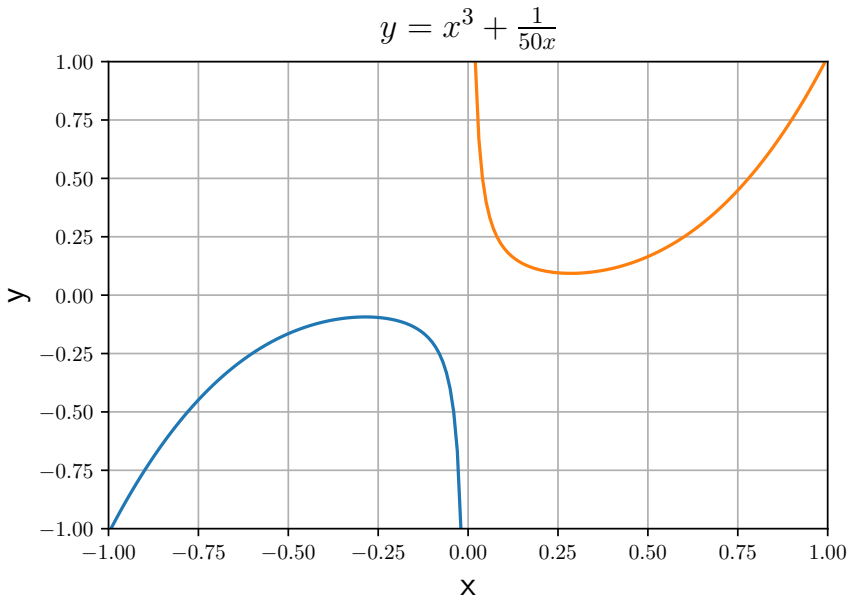

Figura 2.1: Ejemplo de función que no pertenece a un RKHS dada la discontinuidad de su recorrido.

donde se hace el nexo con los conceptos de aprendizaje de máquinas de *transformación de espacio de características* o cambio de mapa.

La transformación de espacio de características consiste en usar una función ϕ como mapa tal que dado el conjunto de entrada $\mathbb{X}$, se pase a otro espacio de representación, es decir aplicar $x \to \phi(x)$, sobre el cual las muestras posean un comportamiento conocido. Por ejemplo, se puede hacer regresión no lineal de datos usando los mismos principios de la regresión lineal, si es que se escoge una transformación de espacio de características que compense la no linealidad de los datos en el espacio transformado.

La manera en que el *kernel* utiliza el mapa de características es a través del mapa $\phi(x)$, sobre el cual el *kernel* es una función definida positiva que se construye como

$$\mathcal{K}(x, y) = \langle \phi(x), \phi(y) \rangle \ . \tag{2.2}$$

De esta construcción se origina el truco del *kernel*, ya que la función de *kernel* corresponde a una matriz de evaluación del *span* (rango o conjunto) de observaciones, definida positiva y simétrica. No se necesita conocer de manera explícita cual es el mapa de características ϕ que la induce, se delega esta labor en definir la función de *kernel* que cumpla las características antes mencionadas. Para emplear este truco, sólo basta diseñar algoritmos que puedan admitir dentro de su construcción esta forma cuadrática generada a partir de productos internos, extendiendo algoritmos lineales a versiones no lineales de los mismos.

Por otro lado, aún es necesario definir que la representación en base a *kernel* cumpla con algún criterio de optimalidad para el problema de regresión. Este problema es solventado usando el teorema del representante.

Teorema 2.4 (ver [19]) *Teorema del representante*
Sea $\mathbb{X}$ un conjunto no vacío y $\mathcal{K}$ un kernel definido positivo a valores reales sobre $\mathbb{X} \times \mathbb{X}$ con su correspondiente RKHS $\mathcal{H}_\mathcal{K}$. Dado el span de muestras de entrenamiento $\{\mathbf{x}_\mathrm{i}, \mathbf{y}_\mathrm{i}\}_{\mathrm{i}=1:N}$, además de una función estrictamente monotónica creciente a valores reales $g : [0, \infty) \to \mathbb{R}$, y una función de riesgo empírico arbitraria $E : (\mathbb{X} \times \mathbb{R}^2)^n \to \mathbb{R} \cup \{\infty\}$. Entonces, para cualquier $f^* \in \mathcal{H}_\mathcal{K}$ satisfaciendo

$$f^* = \underset{f \in \mathcal{H}_\mathcal{K}}{\arg\min}\ E((\mathbf{x}_1, \mathbf{y}_1, f(\mathbf{x}_1)), ..., (\mathbf{x}_N, \mathbf{y}_N, f(\mathbf{x}_N))) + g(||f||)\ , \tag{2.3}$$

f^* admite una representación de la forma

$$f^*(\cdot) = \sum_{\mathrm{i}=1}^{N} \alpha_\mathrm{i} \mathcal{K}(\cdot, \mathbf{x}_\mathrm{i})\ , \tag{2.4}$$

donde $\alpha_\mathrm{i} \in \mathbb{R},\ \forall \mathrm{i} = 1 : N$.

En otras palabras, el teorema del representante asegura encontrar la solución óptima para el modelo de regresión f dentro del span de observaciones de la señal, cuya forma corresponde a una combinación lineal de *kernels* evaluados sobre las observaciones.

Con todo lo anterior, se puede mejorar el modelo de combinador lineal adaptivo (mínimos cuadrados (LMS por sus siglas en inglés) o mínimos cuadrados recursivos (RLS por sus siglas en inglés) dependiendo del funcional de riesgo empírico[27]) y proponer una versión no lineal que permita tener control de cual es la relevancia de las observaciones históricas con respecto a nuevas muestras de la señal recibidas de manera secuencial.

2.1.1. Matriz de Gram

Cada función de *kernel* genera diferentes relaciones entre las observaciones evaluadas, sujeto funcionales que apliquen sobre las muestras e hiperparámetros que los caracterizan. Considerando la forma de producto interno y simetría de la transformación de mapa de características implícito en el *kernel*, como se mostró en la Ecuación (2.2), se puede generar una representación matricial que puede entenderse como una "huella digital" de la transformación, debido a la unicidad de las transformaciones de *kernel* en su respectivo RKHS.

Definición 2.5 (ver [2]) *Matriz de Gram*
Sea el vector de muestras $\{x_\mathrm{i}\}_{\mathrm{i}=1:N}$, el kernel $\mathcal{K}$. El elemento i, j-ésimo de la matriz de Gram queda definido por

$$G_{\mathrm{i}j} = \mathcal{K}(x_\mathrm{i}, x_j)\ . \tag{2.5}$$

Es decir, la matriz de Gram agrupa todas las evaluaciones del kernel sobre un conjunto de muestras.

2.2. Regresión y filtros adaptivos

2.2.1. Regresión con filtros adaptivos lineales

El problema de regresión consiste en utilizar información histórica de un proceso temporal o serie de tiempo, para realizar predicción sobre futuros comportamientos del proceso dados los datos actuales. Dada la generalidad del problema de regresión, es abordado por distintas disciplinas, como por ejemplo econometría, procesamiento de señales y aprendizaje de máquinas, sobre las cuales cada una desarrolla conceptos bajo la óptica de su propio arte para definir la regresión (por ejemplo filtro de respuesta finita al impulso con modelo autorregresivo).

Para el caso de procesamiento de señales, el modelo estándar usado por excelencia para regresión lineal[1], corresponde al filtro lineal. Este filtro está demostrado como la solución óptima para regresión lineal [8] y su extensión adaptiva a través del método de gradiente descendente [31]. Los filtros lineales adaptivos ajustan la combinación lineal de sus entradas, o coeficientes del filtro, a medida que recibe nuevas observaciones. La Figura 2.2 muestra el caso del filtro lineal adaptivo donde f corresponde a una combinación lineal de las entradas x ponderadas por los coeficientes α.

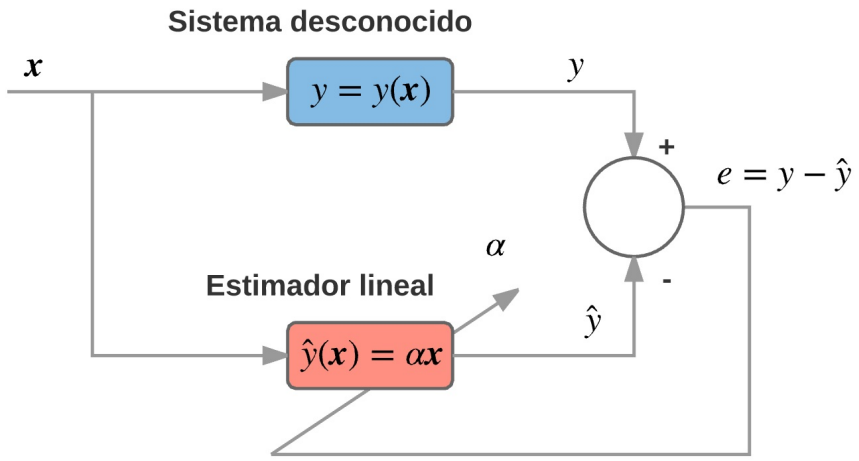

Figura 2.2: Esquema de filtro lineal adaptivo

En términos más formales, para la regresión con filtros adaptivos, se considera el conjunto de muestras como pares ordenados $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1:N} \in \mathbb{X} \times \mathbb{Y}$ donde $\mathbb{X} \subseteq \mathbb{R}^n$ es denotado como el conjunto de entrada de dimensionalidad n, e $\mathbb{Y} \subseteq \mathbb{R}^m$ es el conjunto de salida con dimensionalidad m. Particularmente para el problema de regresión $\mathbb{Y} = \mathbb{R}$, ya que se desea estimar la observación actual en base a observaciones anteriores, es decir $\mathbf{x}_i = x_{i-1}, x_{i-2}, ..., x_{i-d}$ e $\mathbf{y}_i = x_i$, en donde d es el orden del filtro, o cantidad de muestras autorregresivas del modelo.

[1]Regresión que asume que la relación histórica de las muestras corresponde a una combinación lineal de muestras pasadas.

El modelo autorregresivo entonces queda definido por

$$\hat{\mathbf{y}}_{\mathrm{i}} = f(\mathbf{x}_{\mathrm{i}}) \ , \tag{2.6}$$

donde la forma explícita de la función f se determina a través de la solución de la optimización del riesgo empírico, usualmente el error cuadrático medio (MSE por sus siglas en inglés) para caso determinístico y *máximo a posteriori* (MAP) para inferencia Bayesiana [6]:

$$\mathrm{MSE} = \sum_{\mathrm{i}=1}^{N} (\mathbf{y}_{\mathrm{i}} - \hat{\mathbf{y}}_{\mathrm{i}})^2 \ , \tag{2.7}$$

$$\mathrm{MAP} \propto p(\mathbf{y}_{\mathrm{i}}|f(\mathbf{x}_{\mathrm{i}}))p(f(\mathbf{x}_{\mathrm{i}})) \ . \tag{2.8}$$

En donde $p(\mathbf{y}_{\mathrm{i}}|f(\mathbf{x}_{\mathrm{i}}))$ corresponde a la probabilidad condicional de que los datos $\mathbf{y}_{\mathrm{i}}$ al modelo $f(\mathbf{x}_{\mathrm{i}})$.

En el caso de la Ecuación (2.7) se minimiza el error de estimación y en el caso de la Ecuación (2.8) se maximiza la distribución posterior $p(f(\mathbf{x}_{\mathrm{i}})|\mathbf{y}_{\mathrm{i}})$ sujeta a la verosimilitud (primer término) que pondera la pertinencia de la elección de f sujeta a la estimación de la salida; y a la distribución a priori (segundo término), que caracterizan la elección de los parámetros del modelo f[2].

2.2.2. Extensión no lineal: KAF

Como se mostró en la Sección 2.1, en la Ecuación (2.2), el *kernel* se construye como el producto interno de un mapa características ϕ que viene implícito con la selección de un *kernel* dado. En particular, para los *kernels* de base radial [20] está demostrado que su descomposición en mapa de características corresponde a una suma infinita de términos (debido a la caracterización de serie de potencias de su argumento exponencial) y por tanto una función que recopila infinitas transformaciones no lineales. Esto provee una gran ventaja al momento de elegir un modelo con gran flexibilidad que queda determinado por la elección de los parámetros del *kernel* de base radial.

La Figura 2.3 muestra el efecto de modificar el espacio de características usando un vector de mapas no lineales h versus el uso de una función de *kernel*, la cual viene representada implícitamente por la propiedad reproductora por la función ϕ_x. Para este caso, el estimador queda construido de la misma forma que muestra la Ecuación (2.4) con $W = \alpha\phi_x$. De esta manera se construye un filtro no lineal adaptivo lineal en los parámetros (coeficientes α) que puede ser actualizado con la regla de gradiente descendente análogamente como se realizaba en el caso lineal. Más aún, gracias al teorema del representante, se sabe que la solución de este filtro es óptima.

[2]en el caso que MAP no se posean priors, se conoce como máxima verosimilitud y posee solución idéntica al MSE para distribuciones Gaussianas

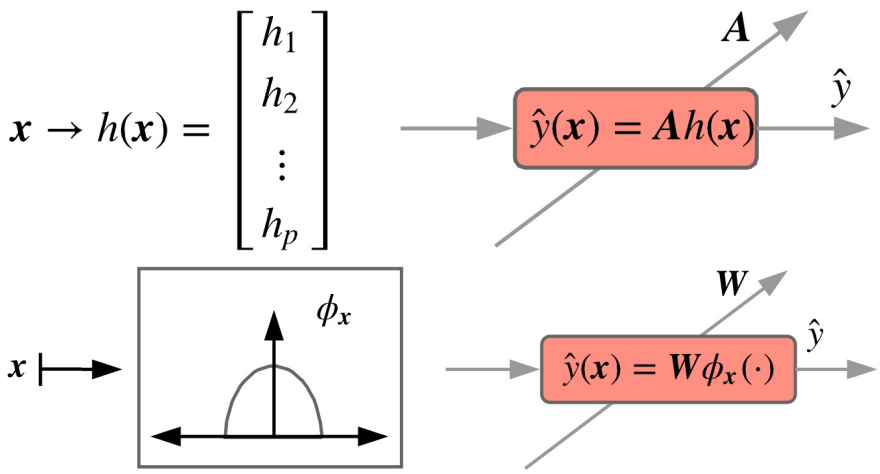

Figura 2.3: Modificación sobre el filtro adaptivo lineal. En la parte superior se muestra una transformación h que consiste de p transformaciones no lineales combinadas linealmente por A. En el inferior se aprecia el uso de mapa de características ϕ_x y el producto interno con $W = \alpha\phi_x$ lo cual es la transformación que realiza KAF permitiendo tener soporte de transformaciones no lineales infinito.

2.3. Implementación algorítmica de KAF

La Ecuación (2.4) entrega una forma cerrada para el estimador KAF. De dicha ecuación, se extraen los elementos principales del modelo KAF:

- Pesos (coeficientes del filtro).
- Observaciones relevantes a evaluar a través del *kernel*.
- Los métodos para seleccionar dichas observaciones.

Dado el carácter adaptivo y la entrada secuencial de muestras al filtro, se utiliza el método de gradiente descendente para recorrer el funcional de mínimos cuadrados del error de estimación para actualizar los coeficientes del filtro en línea. Con respecto al diccionario, se observa que a medida que aumenta la cantidad de observaciones, aumenta el tamaño del estimador, haciendo que éste no sea tratable para volúmenes de datos crecientes en el tiempo. Para mantener el algoritmo tratable, se utilizan heurísticas que permiten seleccionar de manera más eficiente las muestras que construirán el estimador. Este subconjunto de observaciones es conocido como diccionario, centros o vectores de soporte [3]. Los criterios que se utilizan para la selección del diccionario buscan reforzar una construcción *sparse* de éste, es decir, con elección eficiente de sus elementos.

[3] esto es debido a la cercanía que comparte el estimador KAF con el utilizado en Support Vector Regression

2.3.1. Criterios *sparse* para KAF

Se define el diccionaro como el conjunto de observaciones $\mathcal{D} = \{s_j\}_{j=1:m}$ con $m < N$, donde cada s_j es seleccionado en el instante i como la muestra x_i a través de un criterio de relevancia *sparse*.

Definición 2.6 (ver [17]) *Criterio de Novedad (Novelty Criterion)*
La muestra $\mathbf{x}_\mathrm{i}$ *es agregada al diccionario si*

$$\delta_1 = \max_{s_j \in \mathcal{D}} \|x_\mathrm{i} - s_j\| \ , \tag{2.9}$$

supera un umbral determinado previamente. Este criterio es entendido como a mayor distancia entre muestras en el espacio de entrada, mayor relevancia existe debido a la disimilitud medida a través de la norma de la diferencia. Por otro lado, si es que el error de estimación actual supera cierto umbral definido por

$$\delta_2 = \|\mathbf{y}_\mathrm{i} - \hat{\mathbf{y}}_\mathrm{i}\| \ . \tag{2.10}$$

Entonces se espera que el diccionario aún no posee la suficiente cantidad de elementos como para acotar el error de predicción, por lo tanto admite nuevos elementos.

Definición 2.7 (ver [18]) *Criterio de Coherencia (Coherence Criterion)*
La muestra $\mathbf{x}_\mathrm{i}$ *es agregada al diccionario si*

$$\delta_3 = \max_{s_j \in \mathcal{D}} \mathcal{K}(x_\mathrm{i}, s_j) \ , \tag{2.11}$$

es menor a un umbral determinado previamente. Este criterio relaciona directamente la similitud entre muestras en el espacio de características transformado por el kernel.

2.3.2. Kernel Gaussiano

Sea la muestra x_i, el diccionario $\mathcal{D} = \{s_j\}_{j=1:m}$ y el ancho de *kernel* σ. Se define el *kernel* Gaussiano de la siguiente manera:

$$\mathcal{K}(x_\mathrm{i}, s_j) = \exp\left(\frac{-(x_\mathrm{i} - s_j)^2}{\sigma^2}\right) \ . \tag{2.12}$$

Características del *kernel* Gaussiano:

- Función de base radial (RBF): las funciones de base radial se caracterizan por evaluar la distancia desde cierto centro y de esta manera actuando como métrica de distancia. Dicha construcción permite utilizarlas como medida de similitud entre muestras, por ejemplo mediante la evaluación de la magnitud de la distancia entre muestras.
- Dimensionalidad del espacio de características: realizando la expansión de serie de Taylor de la exponencial, se obtiene una serie de potencias de soporte infinito, lo cual interpreta la aplicación del *kernel* exponencial como equivalente a trabajar con una transformación de espacio de características de soporte infinito.

- Interpretación de límites: Debido a la exponenciación del cuadrado de la diferencia entre muestras, el recorrido del *kernel* va entre 0 y 1, para muestras idénticas y completamente disímiles respectivamente. La magnitud del ancho del *kernel* σ regula la relación espacial de olvido entre muestras. En el límite cuando $\sigma \to \infty$ equivale a colocar una función rectangular, ponderando todas las muestras de manera equivalente, efecto que actúa de la misma manera que $\{x_{\mathrm{i}} = s_j \ \ \forall s_j \in \mathcal{D}\}$. Si el límite de $\sigma \to 0$ es equivalente a colocar funciones δ de Dirac centradas en cada elemento del diccionario, lo que es análogo a considerar ninguna similitud entre muestras.

2.3.3. Efecto del escalamiento del hiperparámetro σ

Dependiendo de la selección del ancho del *kernel* hiperparámetro σ, se puede cambiar de manera significativa tanto la *sparsidad* del estimador, como la capacidad de generalización del mismo. En la Figura 2.4 se muestra la medida de similitud bajo distintos regímenes de σ, apreciándose que a medida que aumenta el ancho del *kernel* se aumenta la similitud entre muestras; por ejemplo, en la Figura 2.4, el valor 0,2 posee una similitud con respecto al centro 0 de 0,02 y 0,37 para $\sigma = 0{,}1$ y 0,2 respectivamente.

Si bien hacer crecer el ancho del *kernel* aumenta la similitud entre muestras y por tanto reduce el tamaño del diccionario, para valores muy elevados induce sobreajuste, ya que se fuerza artificialmente a tener pocos coeficientes del filtro, los cuales deben variar continuamente a medida que varía la señal. Es por lo anterior, que la elección de σ se considera como un problema crítico de diseño en los algoritmos KAF, ya que no existe a ciencia cierta una única regla que permita controlar el grado de generalización de la solución evitando el sobreajuste a medida que el tamaño de la misma crece.

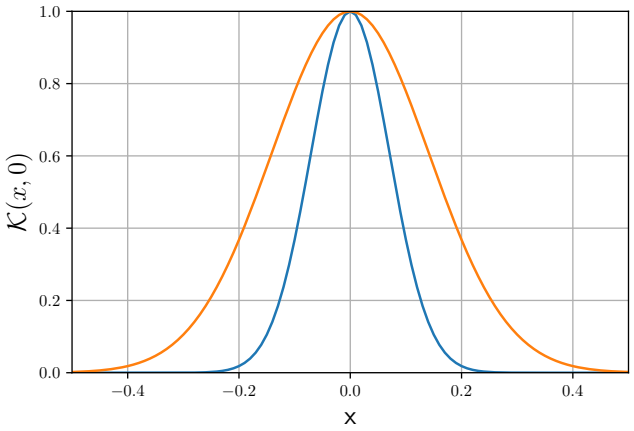

Figura 2.4: Efecto de escalamiento del ancho del *kernel* σ como medida de similitud entre muestras. Para valores mayores de σ, aumenta el valor de similitud con respecto al centro del *kernel*.

Capítulo 3

Contribuciones

3.1. Determinación de relación no lineal entre variables

En la Sección 2.1 se detalló como se construye el estimador autorregresivo de una señal a través de KAF. No obstante, surge la interrogante de si este modelo adaptivo puede mejorar su desempeño al utilizar variables exógenas, como los modelos *autorregresivos de entrada exógena* (ARX por sus siglas en inglés) [13]. El problema anterior adquiere mayor relevancia en el contexto de aprendizaje automático, ya que en el caso de existir una ganancia en la estimación, esto daría indicios de relaciones ulteriores existentes entre las variables, permitiendo hacer descubrimiento de patrones ocultos entre diferentes características de las observaciones. Para ello, se propone el estudio de la generalización del *kernel* Gaussiano para múltiples variables, en su variante de selección automática de parámetros (ARD por sus siglas en inglés).

3.1.1. Kernel Gaussiano anisotrópico para variable exógena (*kernel* ARD)

Como se introdujo en la Ecuación (2.12), el *kernel* gaussiano compara la similitud de las muestras x_i, s_j. Bajo la definición estándar, éstas muestras pueden ser multidimensionales, habiendo en cada dimensión un canal o señal distinta. No obstante, si se extiende directamente el *kernel* gaussiano de esta manera lo que se obtiene es una extensión isotrópica para cada una de las dimensiones involucradas. Esto está condicionado por el parámetro único de ancho de *kernel* que modula todos los canales. En la Figura 3.1 se aprecia la extensión isotrópica para el caso bidimensional.

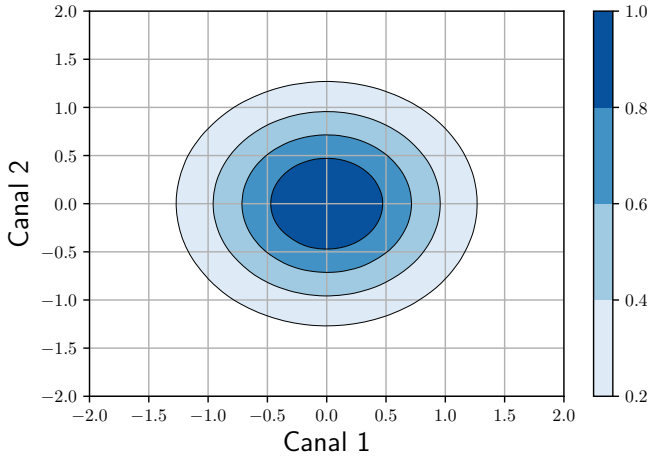

Figura 3.1: Kernel Gaussiano isotrópico para 2 canales en donde el ancho del *kernel* está fijado en 1 para ambos canales

La interpretación práctica de la extensión isotrópica es que cada canal pesa lo mismo al momento de ponderar la similitud del *kernel*. Para estimación con diferentes entradas, puede optimizarse la relevancia de cada canal al momento de construir el estimador *kernel*. Para realizar dicha optimización, se propone un enfoque inspirado en la Determinación de

Relevancia Automática ARD [15]. El proceso consiste en definir un vector de parámetros variables para cada una de las dimensiones del problema y determinar la combinación que maximice alguna medida de desempeño, de modo de caracterizar la relevancia individual de cada canal al momento de realizar la estimación. En la Figura 3.2 se muestra el peso de cada característica con peso mayor en el canal 1. Otra forma de entender el *kernel* ARD es como una normalización en términos de varianza unitaria, pero sin saber las estadísticas de cada canal, lo cual es aprendido por el *kernel*.

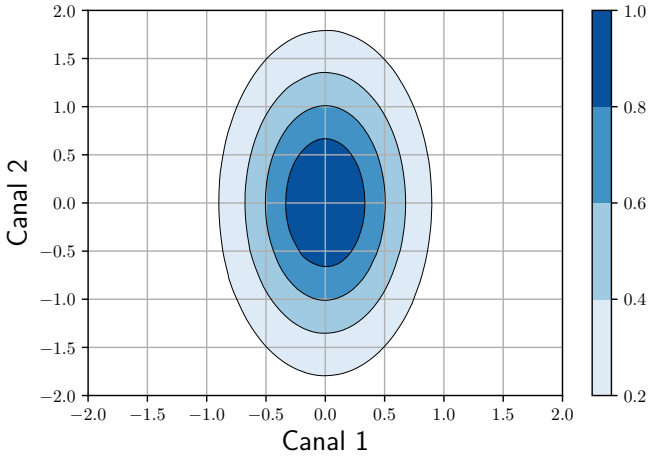

Figura 3.2: Kernel Gaussiano anisotrópico para 2 canales en donde el ancho del *kernel* para el canal 1 es de 0,5 y para el canal 2 de 2

Para x, y multivariables de dimensión n (cantidad de canales). El *kernel* Gaussiano ARD queda caracterizado por

$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{y})^T \Sigma^{-1} (\mathbf{x}-\mathbf{y}) \right) , \tag{3.1}$$

en donde Σ es la matriz de anchos de *kernel* para todos los canales. Si asumimos una estructura diagonal para $\Sigma = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & 0 & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n \end{bmatrix}$, entonces el *kernel* queda descrito por

$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = \exp\left(-\sum_{j=1}^{n} \frac{(x_j - y_j)^2}{2\sigma_j^2} \right) . \tag{3.2}$$

Bajo la representación anterior se puede entender que σ_j corresponde al ancho de *kernel* de la característica o canal j-ésimo. Para este *kernel* se entrenan los parámetros σ_j de modo de mejorar el estimador. Si para alguna característica $\sigma_j \to \infty$ entonces se descarta dicha característica de manera automática.

Otra forma de entender la extensión del *kernel* gaussiano de una dimensión a ARD, es la multiplicación de los *kernels* individuales por canal, de manera de lograr una condición de suma de en el argumento de éstos. La intuición detrás de la configuración escogida es la extensión natural usando un ancho distinto para cada canal. La ventaja de ello es la simple interpretación que tiene como ponderación individual de cada componente.

3.1.2. Optimización estocástica de hiperparámetros

El entrenamiento para el *kernel* ARD requiere de la evaluación del error de estimación para distintas configuraciones de parámetros para el conjunto de entrenamiento. A diferencia del enfoque clásico de búsqueda exhaustiva de validación cruzada, se plantea usar un conjunto de entrenamiento sobre el cual se realice una búsqueda avara para obtener la mejor combinación de parámetros. Utilizar un enfoque avaro se sustenta en la alta no linealidad presente en el error de estimación con respecto a los anchos de *kernel* para cada canal de la señal, volviéndose intratable computacionalmente a medida que crece n. Por ejemplo, si se considera una grilla de búsqueda con resolución g (cantidad de elementos a evaluar por parámetro), el funcional de ajuste debe evaluarse $\mathcal{O}(g^n)$ veces sobre el conjunto de entrenamiento; mientras que con una estrategia avara se requiere sólo $\mathcal{O}(2^n cr)$, con 2^n combinaciones para evolucionar el parámetro en la siguiente iteración, c la cantidad de simulaciones (evoluciones) del modelo por búsqueda y r cantidad de búsquedas. Se decide usar caminata aleatoria [16] por su capacidad de optimizar a través de heurísticas sin tener que derivar ni incorporar conocimiento previo del modelo y por su carácter aleatorio sobre el dominio de búsqueda, permitiendo inferir un óptimo con el promedio de las caminatas.

Por otro lado, no existe una restricción sobre la superficie de desempeño que rija el grado de sobreajuste del *kernel* ARD. Como se origina a partir de una composición de *kernels* Gaussianos para cada coordenada, hereda y complejiza el problema de sobreajuste para tamaños de ancho de *kernel* muy grandes relativos a la magnitud de cada canal, como se explicó en 2.3.2. Es por ello que se decide utilizar como condición inicial de las caminatas valores pequeños, preservando de esta manera soluciones que aporten de manera más efectiva a la generalización de la regresión sobre el conjunto test. Si bien este enfoque posee un reducido costo computacional, sufre de resultados subóptimos al explorar mínimos locales que posean vecindades mayores al salto aleatorio del método. Adicionalmente, para obtener un resultado más robusto dentro de la búsqueda se requiere una cantidad de búsquedas suficientes, con distintas condiciones iniciales, definir un paso apropiado de salto aleatorio (rango). Aún teniendo en cuenta todas las consideraciones anteriores de diseño, no hay certeza de obtener un óptimo global, por lo que se hace necesario refinar más dicho método. Esta mejora se expone posteriormente con el enfoque de inicialización de parámetros probabilística en la Sección 3.2.

3.1.3. Mejora propuesta sobre criterios sparse: seguimiento adaptivo del error

En la Sección 2.3.1 se introdujeron los criterios de novedad y coherencia. La mayor diferencia entre ambos criterios es la manera en que relacionan la distancia entre muestras, ya sea en el espacio de entrada o en el espacio de características (transformado), inclusive en [11] se muestra que para el caso de *kernels* RBF existe una correspondencia en los umbrales de ambos espacios. Adicionalmente el criterio de novedad introduce un seguimiento del error de estimación para determinar si es necesario agregar una nueva muestra cuando ya se tiene una estimación de cierta calidad de la señal. Según lo anterior, se decide utilizar la métrica de comparación de coherencia gracias a su rango acotado en la determinación del umbral

dependiente del *kernel* y no de la señal per se, agregando el criterio del seguimiento del error del criterio de novedad.

No obstante, si se desea tener una independencia completa del uso del rango dinámico de la señal, el criterio de seguimiento del error debe ser modificado. Más aún, si las señales a estimar poseen cruces por cero, cosa que se hace mucho más probable al normalizar canales. Al acercarse a cero, la variable a estimar podría alcanzar de manera artificial el umbral establecido para el seguimiento, provocando que se desechen elementos de diccionario alrededor de dicha zona, en detrimento de la representatividad del diccionario en esa vecindad. La variación sobre el seguimiento del error viene dada por la siguiente modificación de la Ecuación (2.10)

$$\delta_2^* = \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|\mathbf{y}_i \tag{3.3}$$
$$= \delta_2 \mathbf{y}_i \ , \tag{3.4}$$

en donde se aprecia que ahora se pondera de manera adaptiva la cota original δ_2 de la Ecuación (2.10) en función de $\mathbf{y}_i$. Con la modificación anterior, se logra transformar la cota del seguimiento del error para que sea robusta al acercarse a vecindades de ceros. De esta manera es posible obtener una cota adaptiva con respecto a la magnitud de la señal en conjunto con su error de estimación. En otras palabras, se mantiene la cota del seguimiento estándar, agregando el ajuste adaptivo que considera el valor de la señal actual para el nuevo umbral.

3.1.4. Mejora en el estimador: ajuste adaptivo de la tendencia

Dentro del estudio de señales, la hipótesis de estacionaridad asegura que el promedio, varianza y autocorrelación de una señal no cambian como función del tiempo. En términos del estudio de series de tiempo, lo anterior implica señales sin tendencia (componente de baja frecuencia marcada), usualmente homocedásticas (varianza constante) y autocorrelación constante sin estacionalidad (fluctuaciones periódicas). Debido al funcional de minímos cuadrados que minimizan los algoritmos KAF (varianza del error), estos están diseñados para tener un aprendizaje eficiente asumiendo estacionaridad.

Sin embargo, las señales de algunos sistemas reales poseen tendencias fuertemente marcadas y conocidas dentro de su área de aplicación. El efecto de la tendencia en la predicción genera una componente de baja frecuencia que no puede ser compensada de manera efectiva por el filtro (offset). Dicho offset puede ser aproximado localmente, pero a medida que avanza el tiempo necesita ser reajustado continuamente por el algoritmo. Esto genera diccionarios de mayor complejidad y conlleva un menor desempeño tanto en la estimación como el aprendizaje en comparación a un caso con tendencia compensada. El efecto de la tendencia puede ser apreciado en la Figura 3.3. Ejemplos de datos con tendencia son crecimiento de países, precio de acciones o descarga de dispositivos.

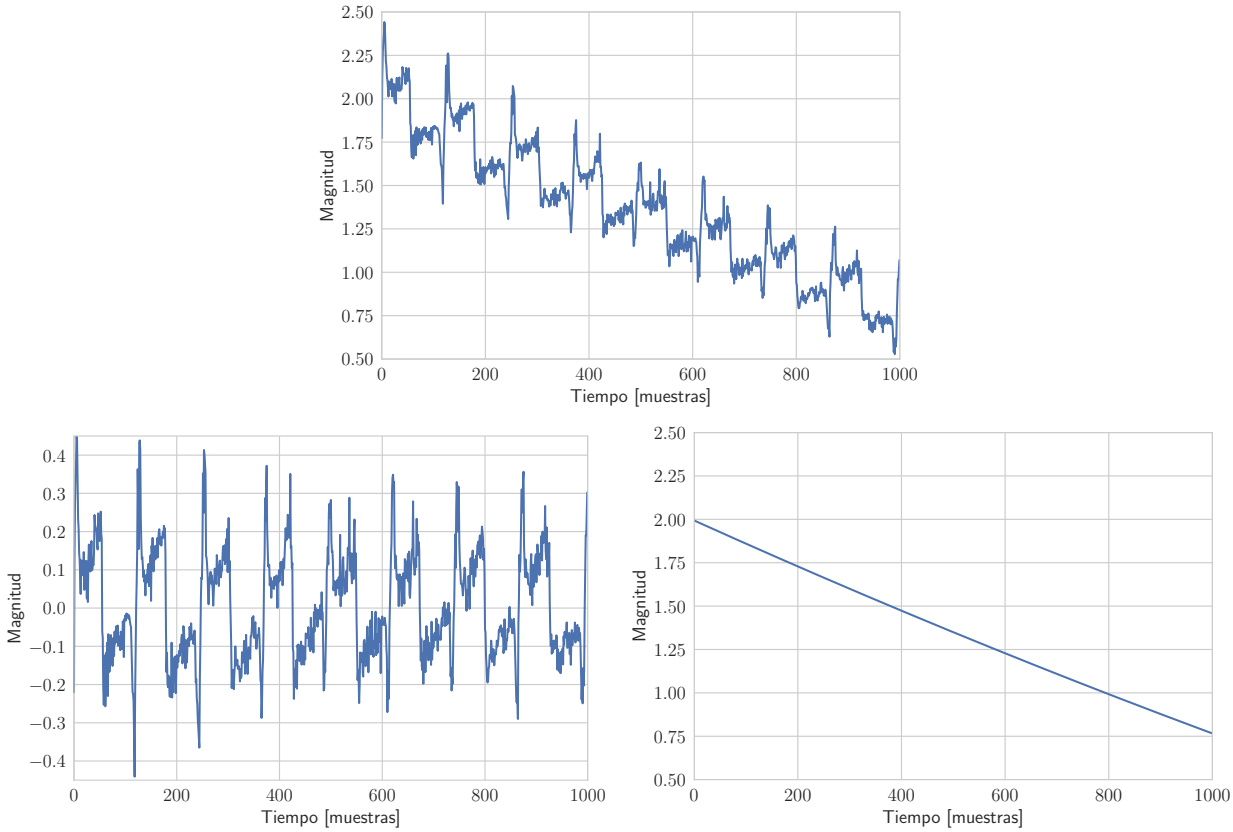

Figura 3.3: Ejemplo de extracción de tendencia en señal. En la parte superior se muestra una señal con una fuerte tendencia lineal, en la esquina inferior izquierda la señal con la tendencia extraída y en la esquina inferior derecha la tendencia lineal.

Se modificó el estimador KAF para agregar una componente para seguir la tendencia de señales como un filtro adaptivo lineal de primer orden

$$\hat{\mathbf{y}}_{\mathrm{i}}^{*} = \sum_{j=1}^{C} \alpha_j \mathcal{K}(x_{\mathrm{i}}, s_j) + \beta y_{\mathrm{i}-1} \ , \tag{3.5}$$

con el parámetro β ajustado también con el funcional de mínimos cuadrados y optimización por método de gradiente descendente. La ventaja de agregar este filtro lineal es que permite compensar componentes de baja frecuencia de manera simple, haciendo que el ajuste adaptivo del parámetro β compense la tendencia y que el residuo sea estimado como un proceso estacionario por el filtro KAF.

3.2. Inicialización probabilística para KAF [3]

Como se propuso en 3.1.2 una elección para el hiperparámetro del *kernel* Gaussiano puede ser determinado a través del uso de un conjunto de entrenamiento. El uso del conjunto de entrenamiento e hibridar el aprendizaje en línea del algoritmo puede ser mejorado significativamente si se utiliza un modelo para encontrar la solución óptima global en vez de buscar un

subóptimo. El costo que conlleva lo anterior es modelar el problema del entrenamiento en si mismo, de modo de resolverlo acabadamente para obtener la inicialización de los parámetros para KAF[1]. Para lo anterior, se proponen los siguientes requerimientos de diseño.

- Mejorar inicialización obtenida por el entrenamiento del parámetro por búsqueda estocástica a través de un modelo de inferencia Bayesiana.
- Generalizar el funcional de optimización de mínimos cuadrados a su variante probabilística de máxima verosimilitud.
- Agregar regularizadores para los parámetros a pre-entrenar, ya que la operación de algoritmos como KRLS requieren de regularización, inclusive KLMS que se autorregulariza no tiene garantías de esto para la solución obtenida por el pre-entrenamiento, por lo que es necesario plantear regularizadores para inicializar KAF en general.
- Considerar como componente central la relación entre el hiperparámetro del *kernel* y la construcción del diccionario, dada la importancia del escalamiento mostrada en 2.3.3.
- La utilización de MAP o MCMC para inicializar. Gracias a la definición de un modelo expresivo probabilístico para todos los parámetros del estimador KAF es posible utilizar técnicas de optimización basadas en muestreo de las distribuciones de los parámetros (simulación por cadenas de Markov o MCMC por sus siglas en inglés) o utilizar directamente los parámetros que maximizan la distribución posterior del modelo (MAP). La técnica de muestreo o la de optimización de la posterior son computacionalmente más demandantes pero permiten obtener una solución óptima.
- Uso de distribuciones a priori jerárquicas. Bajo el enfoque probabilístico se puede automatizar la selección de todos los parámetros del modelo, sujeto a una definición previa de tamaño de diccionario gracias al uso de distribuciones a priori jerárquicas. Esto permite alivianar la carga sobre el usuario para el diseño del estimador KAF.

3.2.1. Extensión a modelo de inferencia Bayesiana

Debido a la gran variabilidad e interdependencia entre (i) los pesos del estimador, (ii) el diccionario y (iii) el hiperparámetro del kernel, se hace necesario utilizar un funcional de optimización que asocie todas estas componentes simultáneamente. A modo de extender de manera atractiva el funcional de optimización de mínimos cuadrados a un enfoque probabilístico, se propone utilizar un modelo de máxima verosimilitud para el error de predicción y acoplarlo a distribuciones a priori sobre cada uno de las componentes que desean entrenarse del modelo; y de esta manera, generar un enfoque de máximo a posteriori para entrenar el estimador KAF.

- Modelo generativo: en vez de adaptar el entrenamiento con el conjunto subóptimo basado en el Teorema del Representante, se utiliza un modelo probabilístico que respeta el estimador KAF, agregando un término de ruido de observación, vale decir,

$$\hat{\mathbf{y}}_{\mathrm{i}} = \sum_{j=1}^{C} \alpha_j \mathcal{K}_{\sigma_k}(\mathbf{x}_{\mathrm{i}}, s_j) + \varepsilon_{\mathrm{i}} \ , \tag{3.6}$$

[1]Para distinguir el ajuste adaptivo clásico de KAF con el entrenamiento de inicialización, se le llamará a este último pre-entrenamiento

con C tamaño deseado del diccionario, $\varepsilon_{\mathrm{i}} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ el ruido de observación Gaussiano y σ_k el parámetro de ancho de *kernel* Gaussiano.

- Ecuación de verosimilitud: para un set de valores observados o target del conjunto de entrenamiento $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_{N_t}\}$, la verosimilitud del modelo puede escribirse como una productoria, dado que el proceso $\mathbf{y}_{\mathrm{i}}$ está caracterizado como un proceso Markoviano del mismo orden d que el filtro presentado en la Sección 2.1:

$$p(\mathbf{Y}) = \prod_{\mathrm{i=d}}^{N_t} p(\mathbf{y}_{\mathrm{i}}|\mathbf{x}_{\mathrm{i}}) \tag{3.7}$$

$$= \prod_{\mathrm{i=d}}^{N_t} \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(- \frac{(\mathbf{y}_{\mathrm{i}} - \boldsymbol{\alpha}^T \mathcal{K}_{\sigma_k}(\mathbf{x}_{\mathrm{i}}, \mathcal{D}))^2}{2\sigma_\varepsilon^2} \right), \tag{3.8}$$

con $\boldsymbol{\alpha}_{\mathrm{i}} = [\alpha_1, ..., \alpha_{N_t}]$ como el vector de pesos en el instante i, $\mathcal{K}_{\sigma_k}(\mathbf{x}_{\mathrm{i}}, \mathcal{D})$ denota al vector de evaluación de *kernel* con respecto a la entrada $\mathbf{x}_{\mathrm{i}}$ y cada uno de los elementos del diccionario $\mathcal{D}$.

- Descripción de distribuciones a priori: gracias al modelo probabilístico, se puede simplificar la elección de parámetros del modelo realizando incluso la determinación automática de las desviaciones estándar de las distribuciones a priori regularizadoras a través de las distribuciones jerárquicas. Vale decir, los parámetros son obtenidos a través de una distribución a priori, la cual actúa como distribución candidata que restringe el dominio de búsqueda. Luego dichos parámetros muestreados de la primera distribución a priori se utilizan como hiperparámetros para muestrear parámetros de una segunda distribución de interés. Cada una de las etapas que usan una distribución inicial para muestrear parámetros de una distribución subyacente, se les conocen formalmente como etapas de modelamiento jerárquico Bayesiano [1].

  A continuación se describen las etapas de modelamiento jerárquico Bayesiano usadas:

  - Distribuciones de primera etapa: distribuciones normales positivas para σ_ε, σ_k, l_α y $l_\mathcal{D}$.
  - Distribuciones de segunda etapa:

$$p(\boldsymbol{\alpha}) = \frac{1}{\sqrt{2\pi l_{\boldsymbol{\alpha}}^2}} \exp\left(\frac{-\|\boldsymbol{\alpha}\|^2}{2l_{\boldsymbol{\alpha}}^2} \right) \tag{3.9}$$

$$p(\mathcal{D}) = \frac{1}{\sqrt{2\pi l_\mathcal{D}^2}} \exp\left(\frac{-\|\mathcal{K}_{\sigma_k}(\mathcal{D}, \mathcal{D})\|^2}{2l_\mathcal{D}^2} \right), \tag{3.10}$$

    donde los regularizadores para los coeficientes del conjunto están indexados por $\boldsymbol{\alpha}$, y la distribución a priori para la selección del diccionario en conjunto con el hiperparámetro del *kernel* por $\mathcal{K}_{\sigma_k}$ respectivamente. En la Sección 3.2.2 se ahonda en la explicación de la penalización utilizada para el diccionario.

3.2.2. Distribución a priori para selección de diccionario *sparse*

La creación de la distribución a priori para la regularización está inspirada en una mejora del concepto de coherencia [18]. Al poder entrenar *offline* fijando la cantidad de elementos

del diccionario y agregando la selección óptima de los mismos, se hace posible concebir una extensión de la coherencia, que considera utilizar toda la matriz de Gram al momento de seleccionar el diccionario. Como ahora trabajamos con la matriz de Gram, el concepto de máximo debería ser evaluado entre cada uno de los elementos de la matriz (excepto la diagonal).

Una posible extensión es asumir un umbral, para el cual cada uno de los elementos de la matriz de Gram sean menores a éste, no obstante complejiza de manera innecesaria la búsqueda de parámetros agregando una restricción. Decidiendo dejar el enfoque lo más automático posible, se deriva una métrica que posee el mismo impacto e inclusive tiene posibilidad de encontrar soluciones más expresivas e irrestrictas: usar la norma de la matriz de Gram como regularizador.

De la Ecuación (2.5) para el *kernel* Gaussiano, se conoce a priori la estructura de la matriz, siendo una diagonal de $1s$ y con elementos entre $[0, 1)$ fuera de ésta. En casos límites se tiene que la norma de la matriz va entre $[\sqrt{n}, n)$, donde el primer caso corresponde a elementos completamente distintos y el segundo a la matriz redundante con el mismo elemento n veces (o σ tendiendo a cero) respectivamente. Imponiendo el regularizador basado en minimizar la norma de la matriz de Gram se refuerza la búsqueda de un diccionario *sparse*, tanto para la selección del hiperparámetro, como para la elección de candidatos del diccionario. Las ventajas de este regularizador por sobre la elección heurística de un umbral corresponden a:

- Ser una expresión matemática cerrada y explícita a minimizar, en vez de una restricción de desigualdad para la evaluación del *kernel*. De esta manera se formaliza en términos de optimización el concepto de coherencia y se mejora al eliminar la elección manual del umbral, reemplazada por una búsqueda automática de diccionarios *sparse*.
- Al ser optimizado en conjunto con la verosimilitud del modelo se evita obtener soluciones que rompan la similitud mínima necesaria entre muestras (caso extremo de Gram como una identidad), ya que se necesita que la solución además de ser *sparse* sea el estimador óptimo para el conjunto de entrenamiento.
- Finalmente, optimizando este parámetro a la vez con las demás distribuciones a priori se asegura que los parámetros estén dentro de márgenes aceptables, evitando, por ejemplo, sobreajuste.

Para la investigación realizada, la optimización del modelo Bayesiano sobre el conjunto de entrenamiento tardaba del orden de un par de minutos, en un computador personal con características de procesador i7 y 16 Gb de RAM.

Capítulo 4

Experimentos

4.1. KAF aplicado a predicción con entrada exógena [24]

El objetivo de esta investigación es determinar la capacidad de KAF para encontrar relaciones no lineales de manera automática a medida que se reciben más muestras del sistema. Para ello, se analiza el caso de la descarga de una bicicleta eléctrica, monitoreando la variable de voltaje la batería de la misma.

4.1.1. Caso estudio: Estimación del voltaje de descarga de bicicleta eléctrica usando mediciones de altitud

La descarga de dispositivos electrónicos, el monitoreo de su carga remanente, o su nivel de autonomía corresponden a problemas recurrentes y transversales a diferentes ámbitos dada la gran masa de dispositivos alimentados por baterías existentes en la actualidad. En [30] se muestra como el estado de carga (SOC por sus siglas en inglés) es la variable central en la resolución de los problemas anteriores, siendo crítico tener un modelo establecido para el voltaje en borne de las baterías.

Inspirado en el estudio del aprendizaje automático para la autonomía de baterías de ion-litio, se trabajó con una serie de datos extraída de la operación de una bicicleta eléctrica. Otra interrogante era determinar que variable podía aportar en la estimación del voltaje, que fuese externa al sistema de la batería en si, vale decir, alguna variable ambiental o derivada de la operación según el recorrido que decidiese realizar un usuario.

El experimento tenía las siguientes características:

- Se acondicionó un motor acoplado a la bicicleta de 250$[W]$ RMS.
- Características del sistema en batería:
  - El paquete de batería compuesto por 30 celdas ion-litio, con capacidad nominal de 4000$[mAh]$ y voltaje 3,7$[V]$ cada una.
  - Las celdas estaban conectadas en una configuración de 10×3 para lograr valores de salida nominales de 37$[V]$ y 12$[A]$, obteniendo por tanto una potencia de salida de 444$[W]$.
  - La capacidad de la batería en $[Joules]$ era $37[V] \times 12[A] \times 3600[s] = 1598400[Joules]$.
- Mediciones de GPS:
  - Extrapolación de mediciones de altitud con información de mapas públicos a partir de coordenadas extraídas por GPS.
  - La precisión de los datos extrapolados por GPS poseen una precisión de hasta 11$[mm]$.
- Circuito recorrido:
  - Vueltas al circuito dentro de la elipse del parque, explanada del parque O'higgins de Santiago de Chile, como se muestra en la figura 4.1 con largo de 440$[m]$.
  - Velocidad promedio de 13,5$[km/h]$ con máxima de 16,6$[km/h]$, diferencia máxima de altitud medida de 10$[m]$ y la distancia total recorrida 29,4$[km]$ en un tiempo

de 2 horas 10 minutos.

- Extracción de datos:
  - Medición del voltaje en borne a través de un convertidor análogo-digital implementado en una placa Arduino con tasa de muestreo $1[Hz]$.
  - Información de GPS a través de smartphone con aplicación Strava. La tasa de muestreo corresponde a $1[Hz]$.
  - Serie de 3900 datos con 500 datos utilizados para el entrenamiento de los parámetros de KAF.

Debido a sincronización de datos entre ambas fuentes y limpieza en toma de muestras uniforme, se determinó reducir la secuencia de muestreo a $0{,}5[Hz]$, interpolando linealmente en ausencia de muestras en la grilla uniforme. En la figura 4.2 se puede observar un régimen periódico de comportamiento en los datos, dando las vueltas dentro del mismo circuito.

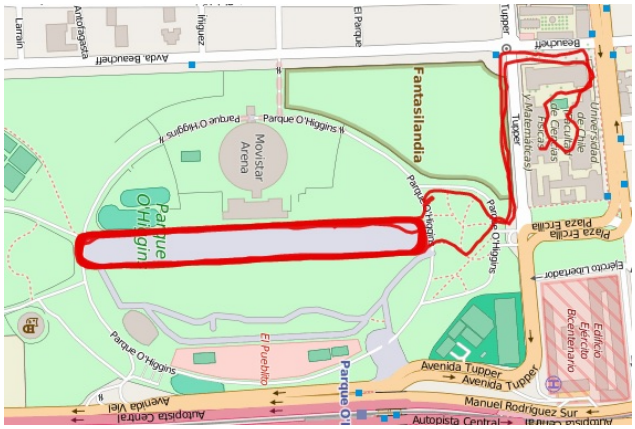

Figura 4.1: Mapa del circuito realizado en el experimento realizado para la bicicleta eléctrica con fines de caracterizar el voltaje de descarga de la misma.

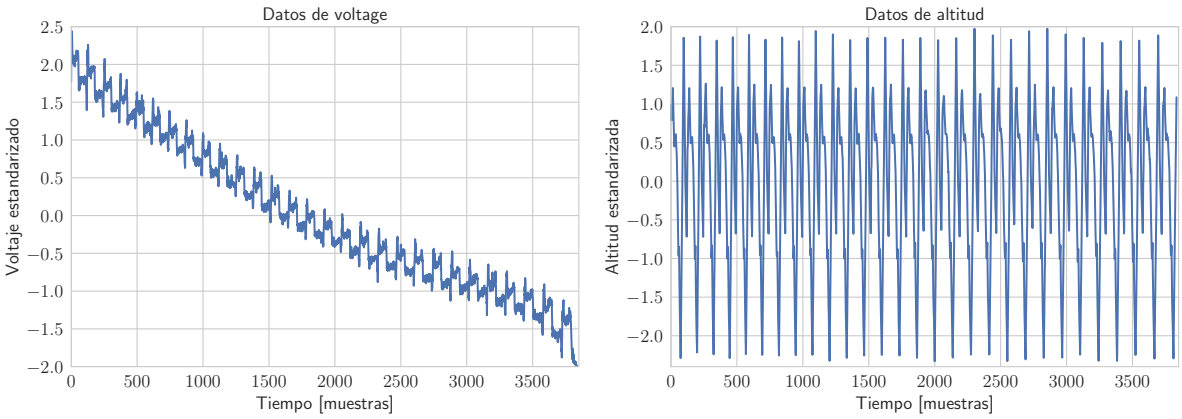

Figura 4.2: Señales de voltaje obtenida a través de Arduino (izquierda) y altitud a través de GPS de smartphone (derecha)

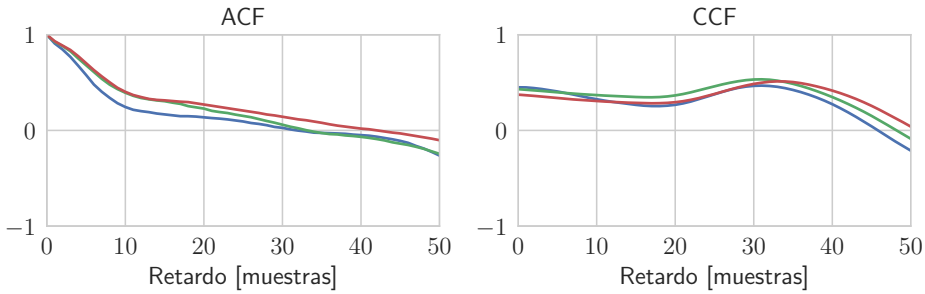

Figura 4.3: Funciones de autocorrelación y correlación cruzada para la señal de voltaje contra la señal de altitud.

4.1.2. Análisis exploratorio y entrenamiento de modelo

A partir de la figura 4.2, la primera inquietud que surge es si la periodicidad presentada entre ambas señales puede ser descrita como una relación lineal entre las variables. En caso afirmativo, bastaría utilizar un modelo lineal que combine las variables para la predicción. Para analizar la dependencia temporal a modelos de diferentes retardos, se realizó un análisis de correlación para la señal de voltaje, autocorrelación (ACF por sus siglas en inglés) y correlación cruzada (CCF por sus siglas en inglés) con la variable de altitud. Los coeficientes del filtro lineal están relacionados directamente a partir de los valores de las correlaciones, dependiendo el orden del filtro. La figura 4.3 muestra la ACF y la CCF, en donde debido a la no estacionareidad de las señales se tomaron 3 diferentes secciones de la señal para estimarla (separadas cada 3000 muestras). Para no contaminar este análisis con la tendencia de la señal de voltaje, se realizó un ajuste polinomial para extraer la tendencia y se trabajó sobre la señal sin tendencia. La caída rápida de la ACF y los valores bajos de la CCF no permiten concluir que exista una relación lineal tanto para un modelo autorregresivo, como para uno con entrada exógena.

A modo de estudiar la capacidad de KAF para encontrar automáticamente la relación no lineal entre el voltaje y la altitud para la predicción del voltaje en la descarga de la bicicleta, se realizó una comparación entre todas las posibles combinaciones de algoritmos de filtro adaptivos con respecto a las categorías de linealidad del modelo y uso de entrada de exógena. Los algoritmos han sido son denotados por LMS, KLMS, LMS-X y KLMS-X; los primeros modelos puramente autorregresivos lineal y no lineal, mientras que los segundos sus contrapartes con entrada exógena, indexada por el sufijo X, respectivamente. Adicionalmente se agregaron dos métodos adicionales para generalizar el desempeño de los filtros adaptivos contra otros algoritmos: el método de estimador persistente, que consiste en predecir manteniendo el valor de la variable según el sistema actual, y el filtro de Kalman, con su matriz de transición de estados entrenada, a través de MCMC, usando el mismo conjunto de entrenamiento que usa KAF.

Los experimentos realizados para la regresión fueron 3: la comparativa general de los 6 métodos como función del orden de los modelos (retardos) para la predicción a un paso, la modificación del estimador con una componente autorregresiva lineal para los modelos KAF (como fue presentado en la ecuación 3.5) y finalmente la estimación a un horizonte de 20 pasos.

Para la regresión con KAF, se usaron dos métodos distintos de entrenamiento. Para el caso puramente autorregresivo, se realizó una búsqueda de grilla para el hiperparámetro del *kernel* Gaussiano en el intervalo $(0, 20)$ con una resolución de 0,05. Para el caso con entrada exógena, se utilizó el kernel Gaussiano en su variante ARD con 2 canales, descrito en la Ecuación (3.2). El modelo fue entrenado usando la metodología descrita en 3.1.2, realizando la búsqueda estocástica con los siguientes parámetros: inicialización de anchos de *kernel* para cada canal entre $(0, 6)$, $(0, 1)$ como intervalo para saltos por perturbación estocástica, 200 pasos en cada búsqueda y 10 búsquedas estocásticas. La Figura 4.4 muestra el resultado del entrenamiento para KLMS con *kernel* ARD.

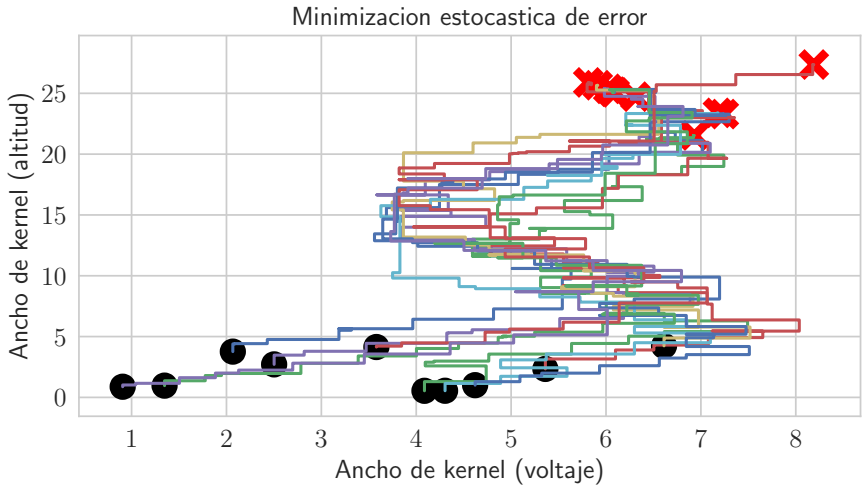

Figura 4.4: Optimización estocástica usando caminata aleatoria para encontrar los parámetros del *kernel* ARD para las variables de voltaje y altitud. Los puntos negros corresponden a condiciones iniciales y las cruces rojas los puntos alcanzados en cada simulación.

4.1.3. Resultados regresión a un paso

Para contrastar el desempeño predictivo de las diversos enfoques comparados con KAF, se utiliza la métrica de RMSE definida por:

$$RMSE = 100\frac{\sum_{t=1}^{T}(v_t - \hat{v}_t)^2}{\sum_{t=1}^{T} v_t^2} \ , \tag{4.1}$$

en donde la señal a estimar es el voltaje v_t y el estimador $\hat{v}_t$ en el instante t. Los métodos comparados se resumen en la Tabla 4.1, en donde el algoritmo con *kernel* ARD corresponde a KLMS-X.

A continuación se muestran los resultados de la comparativa de algoritmos para predecir a un paso la medición de voltaje de la batería de la bicicleta eléctrica. Se observa de la Figura 4.5 que los algoritmos *kernel* superan a sus contrapartes lineales para altos órdenes del filtro. Por otro lado se observa que el algoritmo de filtro adaptivo lineal posee un desempeño sistemáticamente superior al lineal con entrada exógena, probando empíricamente que la relación entre voltaje y altitud es no lineal. Lo opuesto ocurre con el algoritmo *kernel* con entrada

Tabla 4.1: Lista de algoritmos utilizados en el experimento de la descarga de voltaje de la bicicleta eléctrica, caracterizados por la linealidad y el uso de entrada exógena.

Algoritmo	Linealidad	Entrada exógena
LMS	Lineal	No
LMS-X	Lineal	Si
KLMS	*Kernel*	No
KLMS-X	*Kernel*	Si
Kalman	Lineal	Si

exógena, el cual termina superando inclusive al filtro de Kalman. La Tabla 4.2 muestra el desempeño comparativo para el orden 30 del experimento de predicción de descarga de la batería. Se fija el orden 30 para el análisis debido que a las 30 muestras ya se captura la periodicidad de la señal, vale decir, se realiza una vuelta del circuito.

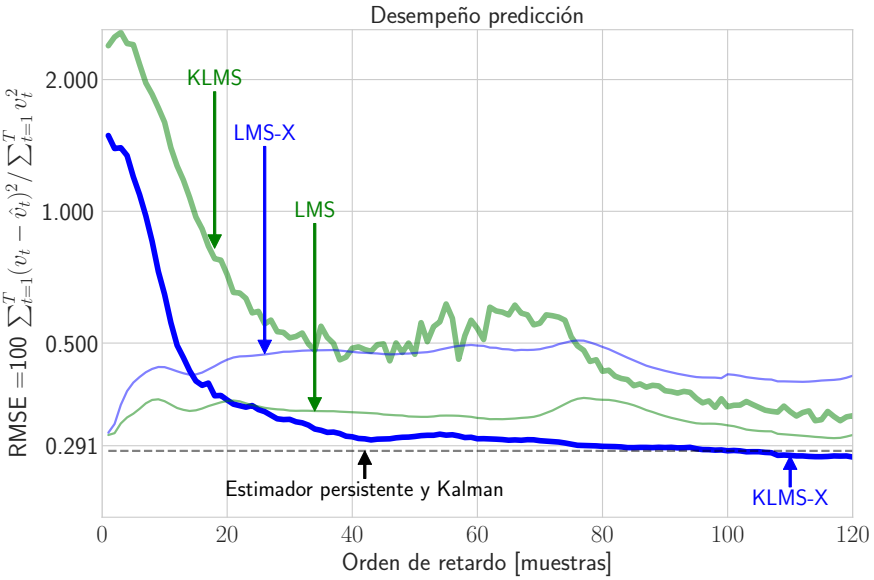

Figura 4.5: Desempeño de los algoritmos *kernel* y lineales considerados para la predicción a diferentes órdenes de modelo de la señal de voltaje (error cuadrático relativo RMSE).

Tabla 4.2: Desempeño y complejidad de los algoritmos considerados para la predicción a un paso de la señal de voltaje a 30 retardos.

	KLMS-X	KLMS	LMS-X	LMS	Kalman
RMSE	3.29e-1	5.12e-1	4.79e-1	3.5e-1	2.84e-1
Time	0.73 [s]	0.68 [s]	0.22 [s]	0.15 [s]	0.92[s]
#SV	136	107	N/A	N/A	N/A

4.1.4. Resultados compensación lineal adaptiva de tendencia

La Figura 4.6 muestra la mejora del desempeño de los algoritmos *kernel* cuando se compensa adaptivamente como filtro lineal la tendencia, dejando el residuo más cercano a la

hipótesis de estacionaridad. Lo anterior es debido a que para el aprendizaje de la tendencia puede ser ajustada como una componente de baja frecuencia que puede ser ajustada en cada iteración como un modelo autorregresivo de orden 1. De esta manera, si la tendencia es compensada correctamente, el algoritmo *kernel* se encarga de aprender una señal de media cero en vez de tener que compensar consecutivamente con nuevos elementos del diccionario el cambio en el rango dinámico de la señal.

Gracias a lo anterior, se aprecia una mejora significativa del desempeño de los algoritmos *kernel* para todos los órdenes del filtro, reduciendo hasta en dos órdenes de magnitud el error de los demás algoritmos. Se aprecia adicionalmente que el algoritmo *kernel* con entrada exógena requiere una menor cantidad de retardos en el modelo para obtener un mejor desempeño de estimación, cosa que el algoritmo puramente autorregresivo compensa para órdenes superiores. Por otro lado, la Tabla 4.3 muestra el desempeño comparativo para el orden 30 del experimento de ajuste adaptivo de tendencia. Con esta tabla, se puede tener una noción del error predictivo en función de la complejidad y costo computacional de los algoritmos.

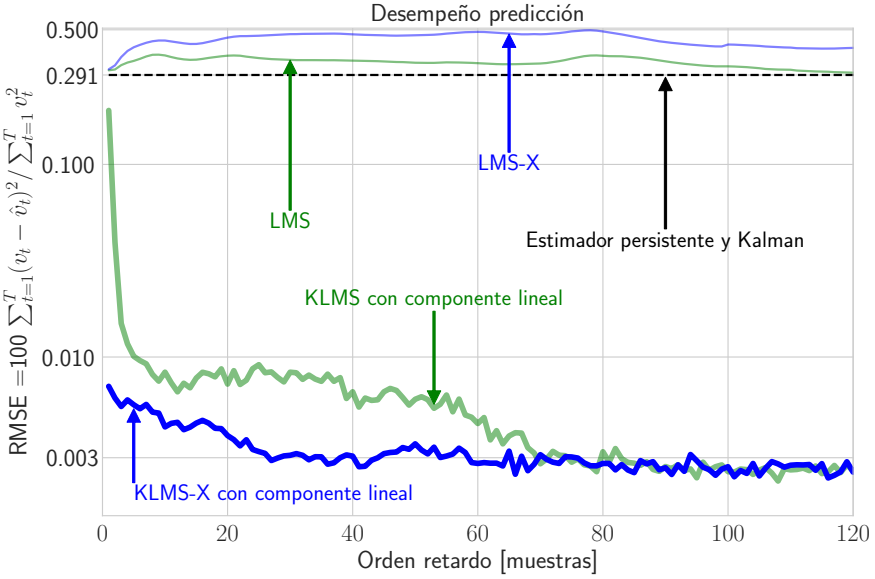

Figura 4.6: Desempeño de los algoritmos *kernel* con componente lineal para tendencia y lineales considerados para la predicción a diferentes órdenes de modelo de la señal de voltaje (error cuadrático relativo RMSE).

Tabla 4.3: Desempeño y complejidad de los algoritmos considerados para la predicción a un paso de la señal de voltaje a 30 retardos en donde los algoritmos *kernel* incluyen la parte lineal para compensar la tendencia

	KLMS-X	KLMS	LMS-X	LMS	Kalman
RMSE	3.65e-3	9.46e-3	4.79e-1	3.5e-1s	2.84e-1
Tiempo	0.94 [s]	0.67 [s]	0.22 [s]	0.15 [s]	0.92[s]
#SV	273	112	N/A	N/A	N/A

4.1.5. Prueba predicción a N pasos con tendencia compensada

Finalmente se realiza un experimento para verificar la capacidad de generalización, en términos predictivos, del modelo que se obtiene del filtro *kernel* con entrada exógena y tendencia compensada. Para ello se decide verificar la robustez que el algoritmo posee al ampliar el horizonte de predicción. En dicho contexto se realiza el experimento mostrado en la Figura 4.7, en donde se compara la versión original de KLMS con las mejoras propuestas consideradas por el uso de la entrada exógena de la señal, en conjunto de la compensación adaptiva de la tendencia.

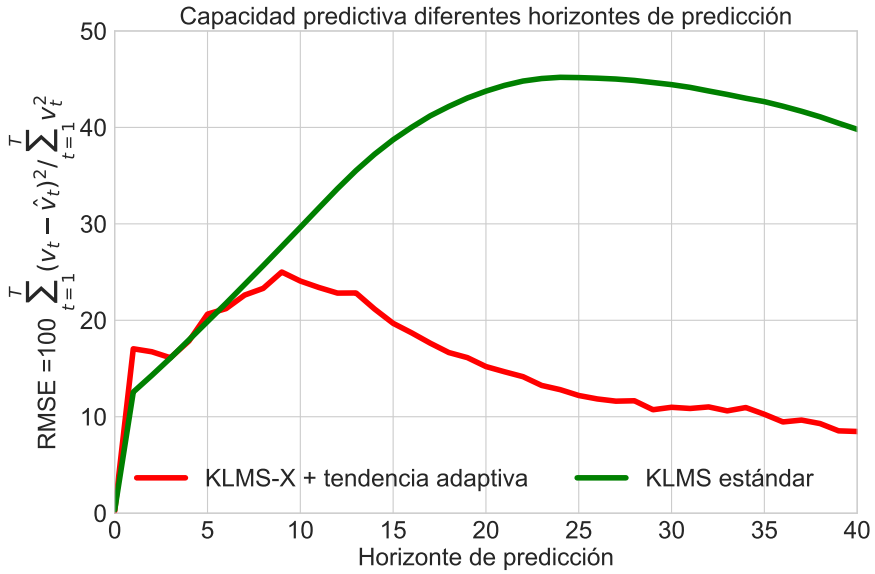

Figura 4.7: Predicción con diferentes horizontes de la implementación estándar de KLMS puramente autorregresiva sobre la serie de voltaje, en contraste con KLMS-X con la tendencia ajustada de manera adaptiva. Se aprecia la mejora en la capacidad predictiva al utilizar una entrada exógena en la medida que aumenta el horizonte de predicción.

4.2. Inicialización probabilística de KAF aplicado a datos sintéticos y reales [3]

Como se mencionó en la Sección 3.2 el entrenamiento de los hiperparámetros queda como un problema abierto con respecto a la elección óptima global en base al conjunto de entrenamiento. Más aún, dicho enfoque sólo considera el error de predicción en el entrenamiento usando distintos hiperparámetros, siguiendo el funcionamiento estándar de KAF con los criterios *sparse* para la construcción del diccionario y la actualización del filtro en base al gradiente descendente para los pesos. Debido a que se está utilizando un conjunto para realizar un entrenamiento *batch* del algoritmo, se puede explotar de manera más eficiente e incluir parámetros adicionales a optimizar, como los elementos del diccionario y los pesos que ponderan cada uno de ellos. A continuación se muestran ejemplos sintéticos y reales para la validación experimental de la metodología propuesta.

4.2.1. Validación experimental: Atractor caótico de Lorentz y predicción de velocidad de viento

El atractor caótico de Lorentz [12] corresponde a un sistema caótico no lineal que usualmente se utiliza para evaluar el desempeño de variantes de algoritmos adaptivos, ya que corresponde a un sistema que se genera a partir de 3 canales acoplados que generar un comportamiento dinámico que modela un fluido sometido a diferente temperatura entre fases. Dada la gran complejidad del modelo fenomenológico (ecuaciones diferenciales acopladas) es de interés probar algoritmos adaptivos para lograr aprender un modelo autorregresivo dadas las observaciones del mismo sistema, lo cual es posible después de un tiempo de observación. Para dicho ajuste, son conocidos los parámetros de KLMS [10], por lo que se vuelve una base comparativa para diseñar una variante. Según la metodología mostrada en 3.2, se utiliza el canal de convección de la serie caótica de Lorentz (eje x), para evaluar el desempeño del aprendizaje probabilístico para diferentes tamaños de entrenamiento.

La Figura 4.8 muestra que a medida que se aumenta el rango dinámico de las observaciones, mejora el desempeño en la predicción. Cabe destacar que esta prueba inicial entrena el modelo con el *batch* de entrenamiento y posteriormente mantiene sus parámetros constantes para realizar la predicción a un paso, es decir, no actualiza usando mínimos cuadrados posterior al entrenamiento. Este resultado valida el uso de la metodología para la aplicación de datos sintéticos, ya que la combinación de parámetros logran obtener una solución óptima en base al rango dinámico del conjunto de entrenamiento. La Figura 4.9 muestra que adicionalmente la relación entre los elementos del diccionario sigue la estructura esperada teóricamente para la matriz de Gram, cumpliendo los objetivos propuestos para la inicialización de parámetros. De esa manera, se promueve el validar la metodología sobre datos de un sistema real.

Para la validación sobre datos reales se utilizó una medición de anemómetro obtenida en [22] como parte del desafío *PHM 2011 Data Challenge* de la PHM Society, el cual contenía como base de datos las mediciones de viento obtenidas a través de arreglos de anemómetros. La Figura 4.10 muestra una de las series de viento del conjunto de entrenamiento de la base de datos, la cual se utilizará para evaluar la metodología de inicialización.

El experimento consistió en comparar el desempeño de KLMS en su implementación estándar en contraste con una inicialización de 270 muestras, lo cual corresponde a un poco más del 10 % de la señal completa. Para comparar de manera justa, se decide comparar el MSE de ambas versiones posterior al conjunto de entrenamiento, activando el funcionamiento de filtro adaptivo para el algoritmo inicializado. Cabe destacar que el algoritmo inicializado construye su diccionario en base a candidatos aleatorios que optimicen el modelo MAP, en vez de seleccionar datos observados previamente, por lo que se necesita fijar una cantidad de elementos deseados en el diccionario en vez de obtenerlos en la marcha. Adicionalmente se ajustan los parámetros de selección del diccionario de KLMS estándar para que la cantidad de elementos sea igual a la determinada por el pre-entrenamiento.

La Figura 4.11 muestra el desempeño comparativo entre ambas versiones. Se observa que el pre-entrenamiento mejora el ajuste de la señal sobre las zonas incluidas en el rango dinámico sobre el cual se seleccionaron los parámetros[1], promoviento una reducción del MSE global.

[1] se evidencia por ejemplo que el ajuste es similar para el valle ocurrido antes del dato 2000, por un nuevo

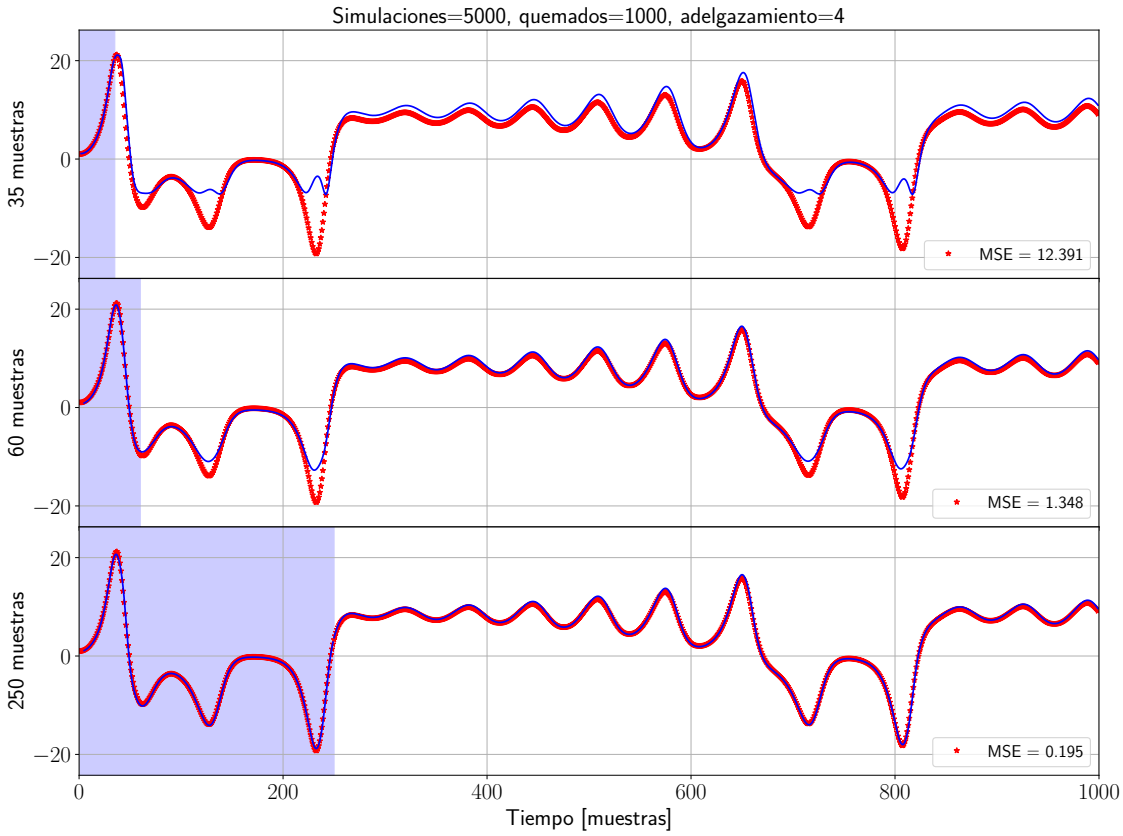

Figura 4.8: Predicción de la serie de Lorentz para la dimensión X usando el modelo de Ecuación (3.6) con parámetros fijos para diferentes subconjuntos de datos. El área azul muestra los datos usados para el pre-entrenamiento en cada caso. La cantidad de muestras elegidas para el pre-entrenamiento, corresponden a secciones del rango dinámico de la señal, 35 al primer pico, 60 al primer valle y 250 para todo el rango dinámico observado.

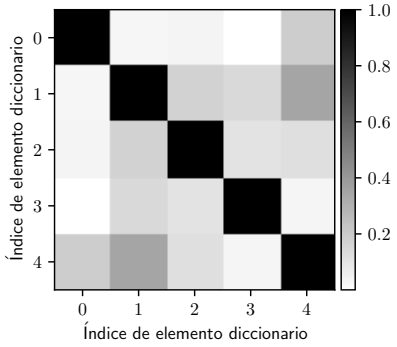

Figura 4.9: Matriz de Gram con tamaño 5 para el caso de 250 datos para pre-entrenamiento. Con ella se ve el grado de disimilitud entre los elementos del diccionario.

Adicionalmente la figura 4.12 muestra que el diccionario pre-entrenado es más *sparse* que el obtenido a través de las muestras, verificando el funcionamiento esperado para la distribución a priori del parámetro *sparse* en la Ecuación (3.10). Como garantía adicional se analiza cuan

comportamiento no observado en el entrenamiento

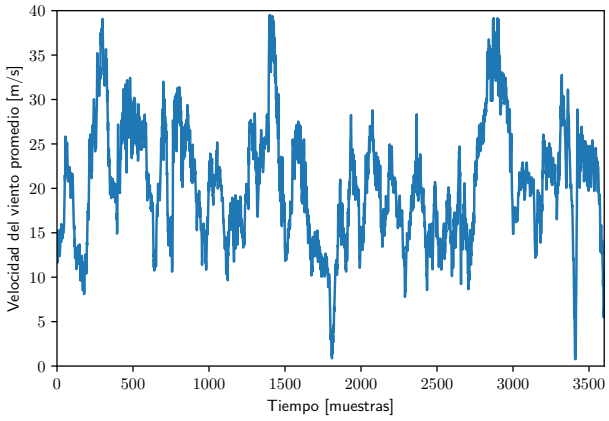

Figura 4.10: Datos correspondientes a conjunto de entrenamiento para el desafío *PHM 2011 Data Challenge*, consiste en medición promedio del viento a través de ventanas móviles de medición.

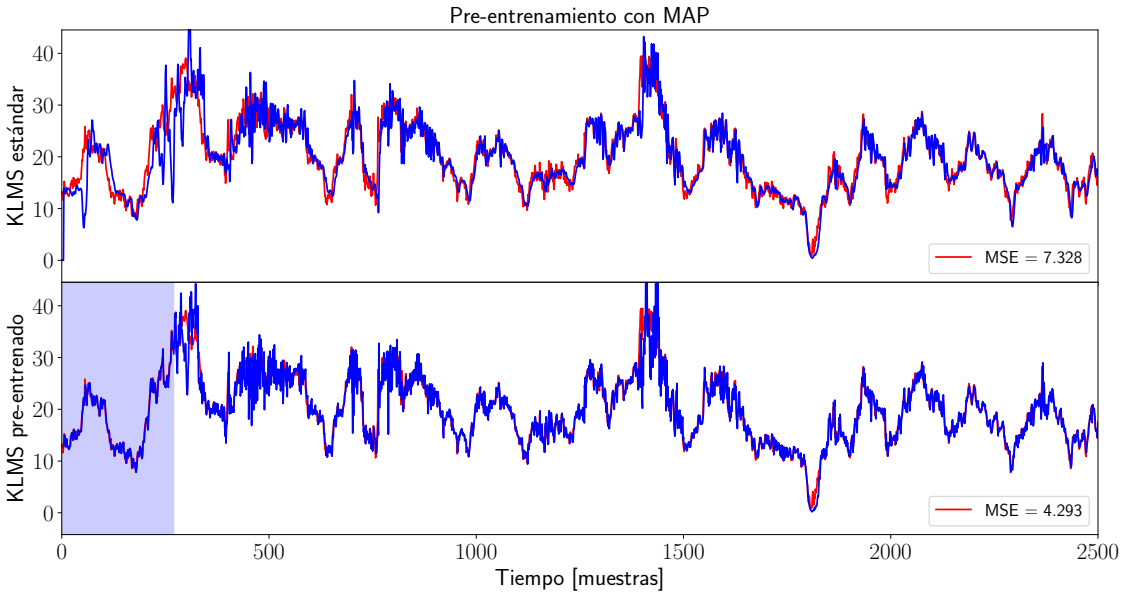

Figura 4.11: Estimación de serie de viento: Estándar KLMS (arriba) contra el propuesto pre-entrenado KLMS con 270 muestras (abajo). El área sombreada indica el período de pre-entrenamiento y el MSE fue calculado después del tiempo 270 para una comparación justa. Se observa que el MSE es reducido, validando el uso de la metodología de inicialización de parámetros.

sparse son las soluciones para diferentes regímenes de diccionario, como lo muestra la Figura 4.13. Se observa que sistemáticamente se llegan a soluciones *sparse*, independiente de la cantidad de elementos en el diccionario.

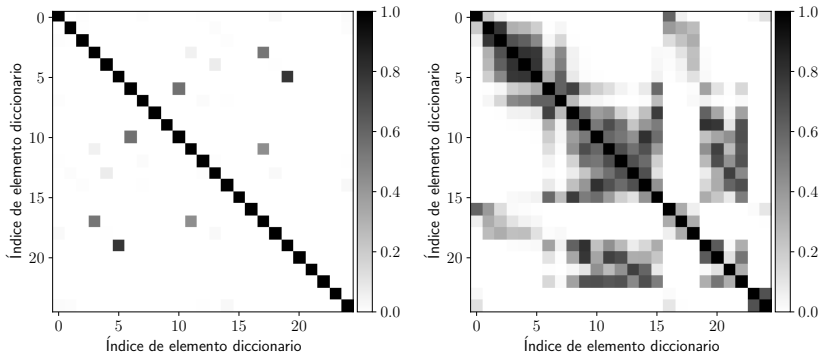

Figura 4.12: Matrices de Gram del pre-entrenado (izquierda) y estándar (derecha) diccionario de KLMS para la señal de viento. El propuesto pre-entrenamiento para KLMS obtiene un diccionario mucho más *sparse* que su contraparte estándar evidenciado por una matriz de Gram más cercana a una diagonal.

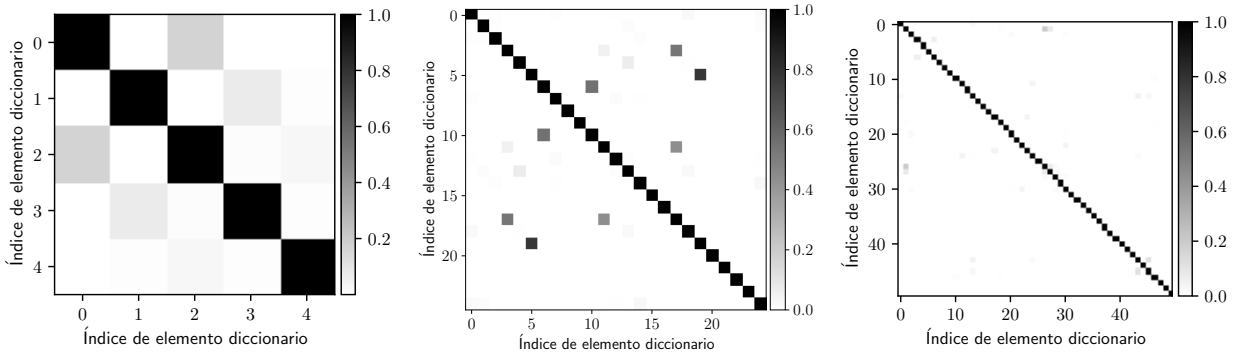

Figura 4.13: Matrices de Gram del pre-entrenamiento para KLMS usado en señal de viento usando tamaños de diccionario $5, 25$ y 50 (de izquierda a derecha). En todos los casos la matriz de Gram posee una estructura cercana a diagonal gracias a la distribución a priori que induce *sparsidad* (Ecuación (3.10)). Con esto se demuestra la consistencia del entrenamiento *sparse* propuesto para distintos tamaños de diccionario fijados con anterioridad.

Capítulo 5

Discusión de resultados y propuesta para KAF completamente adaptivo

5.1. Discusión

A partir de todo el desarrollo teórico y validación experimental sobre la aplicación y mejora de KAF, se logró avanzar en la caracterización automática de los parámetros del estimador en base a un conjunto de entrenamiento. Lo anterior posee un gran valor al momento de acercar el uso de estos algoritmos a usuarios que no son expertos en su uso, para su posterior aplicación para la estimación de señales no lineales estacionarias. Lo anterior se reflejó en rendimientos superiores en contraste con la implementación estándar, además de propuestas para estandarizar la selección de parámetros. Con lo último, se logra disminuir la intervención del usuario, aligerar la curva de aprendizaje en el uso de los métodos y reducir tiempo en ajuste manual de parámetros.

Los resultados presentados en la Sección 4.1 muestran que el *kernel* ARD permite descubrir la relación no lineal entre las variables, sustentado por el rendimiento superior obtenido al aplicar KAF con entrada exógena con respecto a los otros métodos. No solo se mejora el resultado en términos de la predicción a un paso, sino que se desarrollan todavía más las capacidades de KAF al incluir la compensación adaptiva para la tendencia de las señales y mostrando que el uso de una señal adicional hace más robusta la predicción para proyección sobre horizontes de predicción más amplios. Por otro lado, los resultados de la Sección 4.2 muestran consistentemente como se mejora la predicción al inicializar KAF con la metodología Bayesiana propuesta, tanto en términos del MSE, como lo *sparse* de los diccionarios obtenidos en la inicialización. Estos resultados reflejan como afecta el tamaño del conjunto de entrenamiento utilizado al momento de captar la dinámica de la señal y cómo se puede establecer un algoritmo de tamaño de diccionario fijo y obtener bases *sparse* de diferente tamaño dependiendo de los requerimientos del usuario.

Recapitulando sobre lo anterior, se mostró que los algoritmos KAF pueden ser utilizados en conjunto del *kernel* ARD para encontrar la relevancia de señales de múltiples canales al momento de realizar regresión. Esto se debe a la flexibilidad y capacidad de regresión universal de KAF dadas observaciones de algún fenómeno. El costo sobre el que se debe incurrir es la hibridación de un algoritmo aplicado en un contexto netamente adaptivo, es utilizar un conjunto de datos para realizar la búsqueda de los parámetros ARD. Posteriormente se aplicó el concepto de la hibridación del algoritmo formulando un funcional probabilístico basado en inferencia Bayesiana. Con la formulación Bayesiana se logró encontrar una inicialización óptima a diccionario fijo para la posterior utilización de KAF, en donde se destaca la propuesta de la distribución a priori que regula tanto el ancho del *kernel* como la elección de elementos del diccionario que no corresponde a muestras del batch.

Enumerando las mejoras para las reglas *sparse* existentes para KAF, se consolida una variante adaptiva del criterio de novedad para el seguimiento del error adaptivo en señales, adicionado a su uso combinado con el criterio de coherencia. Más aún, se desarrolla un funcional de regularización que extiende el criterio de coherencia para inicializar el diccionario de KAF junto al hiperparámetro del *kernel*, mostrando resultados promisorios dentro de datos sintéticos y reales.

Sin embargo, queda todavía una pregunta abierta para cerrar de manera adecuada una metodología adaptiva para el uso de KAF, y esta corresponde a la propuesta de una regla

adaptiva para la evolución del hiperparámetro del *kernel* posterior al entrenamiento batch (lo que impacta adicionalmente en la selección del diccionario, como se ha visto en secciones anteriores). Esta variante ha sido explorada en [4, 5] extendiendo el método de gradiente para el caso del hiperparámetro y mostrando propiedades de convergencia del error ya sea estacionario o instantáneo.

A continuación se aborda el problema del hiperparámetro adaptivo extendiendo el concepto del la inicialización probabilística a elementos de una ventana móvil para variar el parámetro del *kernel* en función de dicha ventana.

5.2. Implementación adaptiva a través de MCMC

La figura 5.1 muestra el ajuste para la señal de viento utilizando una ventana móvil de tamaño 25, un diccionario 10 y otros parámetros. Se observa que este enfoque logra mejorar aún más el ajuste de la señal que aplicando sólo la inicialización de la Sección 3.2. Esta mejora es corroborada no sólo en la forma más apegada de la estimación a la curva original, sino que minimiza aún más el MSE alcanzando un valor de 1,494, que es una reducción considerable de los casos mostrados en la figura 4.11. Sin embargo, el costo computacional de aplicar el algoritmo de MCMC por cada ventana de tiempo asciende a $\mathcal{O}(wrN)$, con w el tamaño de la ventana, r la cantidad de simulaciones de MCMC y N la cantidad de datos a predecir, lo cual se vuelve intratable para una aplicación de KAF en tiempo real. Cabe destacar que el costo en términos de la ventana no se traduce directamente en un costo de operaciones, sino más bien en el largo de los objetos en memoria sobre los cuales se minimiza el error.

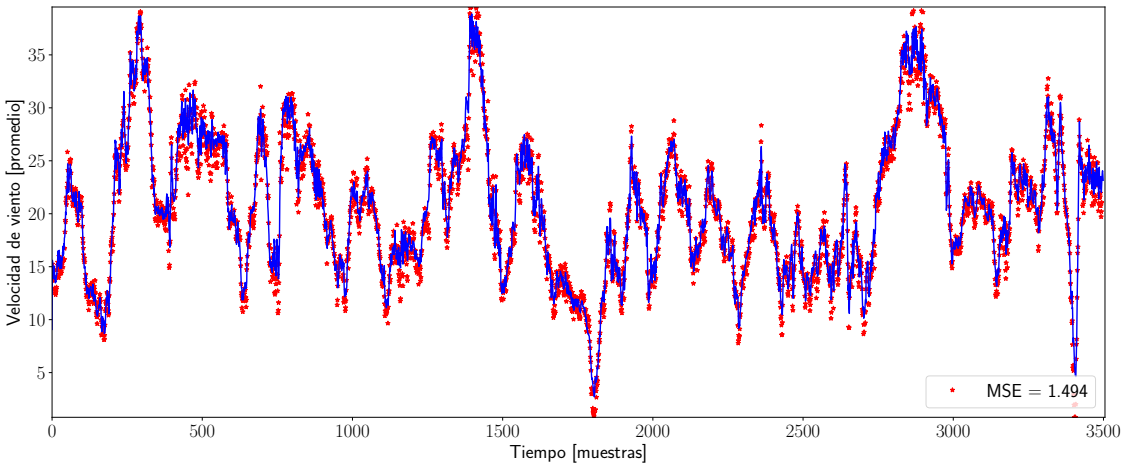

Figura 5.1: Versión de KAF completamente adaptiva aplicada a la serie de viento usando una ventana deslizante. Se logra reducir aún más el error de predicción, a costo de ejecutar MCMC por cada segmento de la ventana deslizante.

Es por lo anterior que se presentan consideraciones de diseño junto a posibles opciones para extender el funcionamiento de KAF en la actualización adaptiva del ancho del *kernel*.

5.3. Bases para el desarrollo de regla adaptiva para σ

Los algoritmos que trabajan con propuestas de regla adaptiva para σ carecen de condiciones de optimalidad en términos de la base o variación del diccionario para la estimación en cada instante, vale decir, no hay certeza en como cambia la transformación de espacio de características al variar el hiperparámetro del *kernel*. Para remediar esta carencia, se propone revisitar la teoría fundacional de los algoritmos KAF y derivar un desarrollo de regla adaptiva en base a RKHS.

5.3.1. Noción de diseño en base a RKHS

En [25], se muestra como propuesta adicional al desarrollo de MKLMS, la concepción de los vectores de espacios RKHS que conforman un super conjunto en dónde pueden coexistir distintas funciones de *kernel* a la vez. El concepto anterior se utiliza en el contexto de MKLMS para demostrar que existe una forma natural de obtener el estimador de mínimos cuadrados para combinaciones lineales de estos *kernel*, amparados por la propiedad de suma de *kernels* induce nuevos *kernels* [2]. Adicionalmente se muestra como se modifica el aprendizaje de gradiente descendente para el caso multi-*kernel*, realizando la combinación de los *kernels* de distintos espacios.

En términos de implementación se reconoce dos tópicos que MKLMS propone de extensión: la cantidad de *kernels* es fija a priori, y finita, vale decir, se debe seleccionar un conjunto discreto de funciones *kernel* a priori para que sean utilizadas en la regresión, de modo de explotar propiedades conocidas de estas funciones para adaptar distintos regímenes de la señal. Por otro lado, propone una forma análoga de colapsar MKLMS a KLMS, forzando que en cada iteración de predicción se fuerce a que la combinación de *kernels* colapse al uso de uno en particular, de esa manera generando un KLMS con *kernel* variante en el tiempo.

5.3.2. Requisitos de diseño de regla adaptiva

En base a la investigación realizada, se identifican los siguientes requerimientos para la regla adaptiva

- Preservación o caracterización de distancia entre muestras en el espacio transformado por los *kernels*
- Promover elección más *sparse* de los elementos del diccionario
- Búsqueda sobre un dominio continuo de hiperpárametros para la transformación *kernel*
- Que el fundamento teórico de la regla derive en las propiedades que preserve el estimador para la regresión

5.3.3. Extensiones factibles

Considerando los requisitos de la regla adaptiva, se propone explorar los siguientes algoritmos o enfoques con el fundamento base

- Proyección sobre *kernel* más relevante (por ejemplo *Principal Component Analysis* (PCA), *Matching Pursuit* (MP)). Este foco consiste en utilizar un método para seleccionar cuál es la base más relevante en el espacio de los *kernels*. La idea es inspirarse en algoritmos que pueden seleccionar de manera automática las componentes más relevantes dentro de una base de soluciones o características. Para un caso inspirado en PCA podría ser cambiar la métrica de la varianza por la dispersión de los elementos del diccionario sobre conjuntos de datos a cierta distancia relativa (grado de similitud dentro de una bola radial en el espacio de los *kernel*). Para el caso de MP es más directo, podría ser considerar los residuos de proyectar sobre los *kernels* que tengan más peso al momento de hacer la regresión de manera de seleccionar centros relevantes. La idea de seleccionar elementos relevantes del diccionario, podría ser utilizada para proyectar el diccionario sobre una de esas transformaciones relevantes, o ir agrupando los elementos existentes bajo elementos más representativos haciendo crecer el ancho del *kernel*.
- Método de diccionario fijo con transformación de multi a mono *kernel* (penalizar fuertemente los demás *kernels*, ir haciéndolos cero). Este enfoque consiste en modificar el criterio de optimización del entrenamiento a un funcional que penaliza binariamente el uso del *kernel* de mayor relevancia, modificando la actualización del aprendizaje del gradiente descendente por un método que incluya restricciones de penalización, como gradiente proyectado. Este enfoque conlleva una mayor complejidad en verificar la factibilidad y diseñar algún funcional de optimización que sea estable en operación en línea.

Capítulo 6

Conclusión

Se logró aplicar KAF de manera exitosa para la regresión no lineal en casos reales y sintéticos. Gracias a un proceso de aprendizaje en el uso de estos algoritmos se identificaron requerimientos de diseño clave sobre los cuales proponer contribuciones tanto algorítmicas como de optimización. Este trabajo se extendió tanto por heurísticas de implementación, como son la modificación de seguimiento del error o la búsqueda avara de caminata aleatoria, como por desarrollo teórico de modificación de funcional para el entrenamiento.

Se logró contribuir en el desarrollo del uso de KAF para aplicaciones de regresión no lineal, automatizando la determinación de los parámetros del modelo, de manera de hacerlos más aplicables a casos generales. A partir de las explicaciones y fundamentos de los desarrollos expuestos, se puede utilizar también como guía metodológica para el uso de KAF para usuarios neófitos.

Dentro de las contribuciones directas del trabajo realizado se destacan dos. Primero, el descubrimiento automático de relaciones no lineales a través de los resultados obtenidos con el algoritmo KLMS-X. Segundo, la proposición de una distribución a priori que induce una diccionario *sparse* a través del pre entrenamiento de los parámetros del modelo generativo de KAF.

En base a la totalidad de la investigación realizada, se concluye que es factible proponer variantes de KAF que utilicen criterios de aprendizaje automáticos para los hiperparámetros en el caso de usar *kernel* Gaussiano que mejoran el desempeño predictivo en regresión no lineal de series de tiempo.

Quedan desafíos pendientes en pos de completar la automatización y conseguir un framework completamente adaptivo para KAF. La gran contribución que sigue siendo un problema abierto es la determinación adaptiva del hiperparámetro del *kernel*. Se propone como trabajo futuro mantener el enfoque génesis de KAF que está basado fuertemente en RKHS para mantener garantías en la evolución de la transformación de *kernel*. Es necesario modelar dicho problema teniendo en consideración tanto el fundamento teórico de como están variando los espacios de característcas, como el matiz algorítmico de la relación del ancho del *kernel* con respecto a la similitud de las muestras y construcción del diccionario.

Capítulo 7

Bibliografía

[1] G. M. Allenby, P. E. Rossi, and R. E. McCulloch. Hierarchical bayes models: a practitioners guide. 2005.

[2] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[3] I. Castro, C. Silva, and F. Tobar. Initialising kernel adaptive filters via probabilistic inference. In *Proc. of the IEEE International Conference on Digital Signal Processing*, pages 1–5, Aug 2017.

[4] B. Chen, J. Liang, N. Zheng, and J. C. Príncipe. Kernel least mean square with adaptive kernel size. *Neurocomputing*, 191(Supplement C):95 – 106, 2016.

[5] H. Fan, Q. Song, and S. B. Shrestha. Kernel online learning with adaptive kernel width. *Neurocomputing*, 175(Part A):233 – 242, 2016.

[6] Dani Gamerman and Hedibert F Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press, 2006.

[7] G. Goel and D. Hatzinakos. Ensemble empirical mode decomposition for time series prediction in wireless sensor networks. In *2014 International Conference on Computing, Networking and Communications (ICNC)*, pages 594–598, Feb 2014.

[8] S. Haykin. *Adaptive filter theory*. Pearson Education India, 2008.

[9] L. Lin. A new method of financial time series prediction. In *2010 International Conference on Educational and Information Technology*, volume 1, pages V1–239–V1–241, Sept 2010.

[10] W. Liu, J. C. Principe, and S. Haykin. *Kernel Adaptive Filtering: A Comprehensive Introduction*. Wiley, 2010.

[11] W. Liu, J. C. Principe, and S. Haykin. Kernel adaptive filtering: a comprehensive

introduction, 2011.

[12] E. N. Lorenz. Deterministic nonperiodic flow. *Journal of the atmospheric sciences*, 20(2):130–141, 1963.

[13] E. Mosca and G. Zappa. Arx modelling of controlled armax plants and its application to robust multipredictor adaptive control. In *1985 24th IEEE Conference on Decision and Control*, pages 856–861, Dec 1985.

[14] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[15] R. M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996.

[16] J. R. Norris. *Markov chains*. Number 2. 1998.

[17] J. Platt. A resource-allocating network for function interpolation. *Neural Computation*, 3(2):213–225, 1991.

[18] C. Richard, J. C. M. Bermudez, and P. Honeine. Online prediction of time series data with kernels. *IEEE Transactions on Signal Processing*, 57(3):1058–1067, March 2009.

[19] B. Schölkopf, R. Herbrich, and J. Smola, A. *A Generalized Representer Theorem*, pages 416–426. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[20] B. Schölkopf, K-K. Sung, C. J. C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*, 45(11):2758–2765, Nov 1997.

[21] E. Soares, V. Mota, R. Poucas, and D. Leite. Cloud-based evolving intelligent method for weather time series prediction. In *Proc. of the IEEE International Conference on Fuzzy Systems*, pages 1–6, July 2017.

[22] The Prognostics and Health Management Society. Condition monitoring of anemometers, 2011.

[23] F. Tobar. Improving sparsity in kernel adaptive filters using a unit-norm dictionary. In *Proc. of the IEEE International Conference on Digital Signal Processing*, pages 1–5, Aug 2017.

[24] F. Tobar, I. Castro, J. Silva, and M. Orchard. Improving battery voltage prediction in an electric bicycle using altitude measurements and kernel adaptive filters. *Pattern Recognition Letters*, 2017.

[25] F. Tobar, S-Y. Kung, and D. Mandic. Multikernel least mean square algorithm. *IEEE Trans. on Neural Networks and Learning Systems*, 25(2):265–277, 2014.

[26] S. Van Vaerenbergh, M. Lazaro-Gredilla, and I. Santamaria. Kernel recursive least-squares tracker for time-varying regression. *IEEE Transactions on Neural Networks and*

Learning Systems, 23(8):1313–1326, Aug 2012.

[27] V. N. Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[28] M. Vetterli, J. Kovačević, and V. K. Goyal. *Foundations of signal processing*. Cambridge University Press, 2014.

[29] W. Wang, J. Zhao, H. Qu, B. Chen, and J. C. Principe. An adaptive kernel width update method of correntropy for channel estimation. In *Proc. of the IEEE International Conference on Digital Signal Processing*, pages 916–920, July 2015.

[30] N. Watrin, B. Blunier, and A. Miraoui. Review of adaptive systems for lithium batteries State-of-Charge and State-of-Health estimation. In *IEEE Transportation Electrification Conference and Expo (ITEC)*, pages 1–6, 2012.

[31] B. Widrow, J. M. McCool, M. G. Larimore, and C. R. Johnson. Stationary and nonstationary learning characteristics of the lms adaptive filter. *Proceedings of the IEEE*, 64(8):1151–1162, Aug 1976.