Universidad de Valladolid

PROGRAMA DE DOCTORADO EN MATEMÁTICAS

TESIS DOCTORAL:

Statistical Oscillatory Models to Solve Problems in Neuroscience

Presentada por Alejandro Rodríguez-Collado para optar al grado de Doctor por la Universidad de Valladolid

Dirigida por:

Dra. Cristina Rueda Sabater

Cristina Rueda Sabater, Catedrática de Universidad, certifica que la presente memoria ha sido realizada, bajo su dirección, por Alejandro Rodríguez-Collado, en el Departamento de Estadística e Investigación Operativa de la Universidad de Valladolid.

Valladolid, 29 de Marzo de 2022

Esta tesis va dedicada a mi Chispi, a quien debo mi forma de ser y mis logros.
Fuiste el lucero que necesitamos en la más óscura de las tempestades,
y serás el fundamento del brillante futuro aún por venir.

Jamás te olvidaré. Allá donde estés, gracias por todo peluche.

Alejandro Rodríguez-Collado

Agradecimientos

En primer lugar, querría dar las gracias a mi directora, Cristina, por el viaje que ha supuesto escribir esta tesis. Tuve mucha suerte de que me brindases la oportunidad de incorporarme a una línea de investigación tan interesante y eminentemente práctica. Trabajar junto a tí me ha servido para sentir lo que es tener fascinación y pasión absoluta por la investigación, así como para adquirir muchísimos conocimientos que, sin duda, no olvidaré jamás.

También querría agradecer el acogimiento que he recibido por parte de los compañeros del departamento de Estadística e Investigación Operativa de la Universidad de Valladolid. Como no podría ser de otra forma, quiero dar las gracias especialmente a Yolanda, Itziar, Jesús y Christian. Me hábeis facilitado enormemente el día a día. Asimismo, no hubiese emprendido esta aventura sin los ánimos que me dieron Araceli y José Antonio. Por otro lado, en estos dos últimos años he tenido la suerte de tener alumnos que no solo han aprendido de mí, sino que me han enseñado mucho. Para mí, esto siempre será fundamental para ser un buen docente.

Esta tesis ha sido posible gracias al apoyo de mi madre, Belén, y Alfon. Este largo camino hubiera sido inasequible sin vosotros, especialmente teniendo en cuenta los tiempos convulsos que hemos vivido. Esto tampoco hubiera sido posible sin Chispi, cuya compañía y cariño incondicional me han dado el tesón para luchar por mis sueños. Gracias por todo. También es fundamental para mí agradecer los ánimos que he recibido de mi padre, hermana, abuelo Jesús y abuela Julia. Por último, quiero dar las gracias a todos aquellos que hábeis estado a mi alrededor en estos dos años, tanto amigos como conocidos, y que hábeis hecho que dejase de lado, de vez en cuando, mi habitual forma de ser responsable.

¡Gracias por vuestro apoyo!

UNIVERSIDAD DE VALLADOLID

Resumen

PROGRAMA DE DOCTORADO EN MATEMÁTICAS

Statistical Oscillatory Models to Solve Problems in Neuroscience

por Alejandro Rodríguez-Collado

Un sistema oscilatorio es aquel en el que una partícula o conjunto de partículas vuelven a su estado inicial tras un periodo, y una oscilación es la variación repetitiva de una señal asociada a estos sistemas. Los sistemas oscilatorios regulan todo tipo de fenómenos estudiados en diversos campos. Por ejemplo, la luz de las estrellas en astronomía, las curvas de espectrometría en química experimental, la demanda eléctrica en ingeniería energética, entre otros muchos. En biomedicina, las señales oscilatorias surgen de forma ubicua: la actividad eléctrica del corazón, el flujo respiratorio de los pulmones, la activación de genes circadianos, así como la actividad eléctrica neuronal. Las señales oscilatorias han sido estudiadas por investigadores provenientes de muchas disciplinas, existiendo entre ellos diferencias significativas en cuanto a terminología, conceptos y métodos. El enfoque predilecto por los investigadores provenientes del campo de la comunicación es el basado en tiempo-frecuencia (Boashash, 2016), mientras que los neurocientíficos prefieren un enfoque físico basado en la descripción dinámica de los sistemas mediante ecuaciones diferenciales (Wigren, 2015; Pikovsky y Rosenblum, 2015; Ashwin, Coombes y Nicks, 2016). Por último, la perspectiva estadística, especialmente útil al observar señales con ruido, se ha empleado en cronobiología (Larriba et al., 2020).

El objeto de estudio de esta tesis es el análisis de las señales eléctricas neuronales, que registran los efímeros aumentos de voltaje que se dan en las neuronas, llamados curvas de potencial de acción o *spikes*. Los *spikes* se analizan de forma individual o como cadenas (también denominados *spike trains*). Los *spikes* transmiten información entre neuronas, y tanto su forma como su tasa de aparición caracterizan los diferentes tipos celulares, por lo que su estudio es crucial en neurociencia. El primer modelo que describió con precisión la dinámica neuronal fue el modelo de Hodgkin-Huxley (Hodgkin y Huxley, 1952), que asumía que las neuronas se comportan como circuitos eléctricos regidos por un sistema de ecuaciones diferenciales. Además, sirvió de inspiración para la creación de otros modelos biofisiologicamente realistas, como FitzHugh-Nagumo (Fitzhugh, 1961). Sin embargo, su naturaleza determinista, formulación compleja, falta de identificabilidad de sus parámetros y sensibilidad al ruido hicieron a los neurocientíficos buscar otras alternativas. En el presente, se emplean modelos más simples, como Hopfield (Abbott y Kepler, 1990), Izhikevich (Izhikevich, 2003) o la familia de modelos *leaky integrate-and-fire* (Teeter et al., 2018, y referencias en el mismo); y modelos basados en datos, como Brunton y Beyeler, 2019. Algunos retos que surgen de analizar los datos neuronales son los siguientes: (a) las señales suelen tener bastante ruido, especialmente si han sido tomadas de forma no-invasiva; (b) los *spikes* son altamente asimétricos, lo que imposibilita el uso de ciertos métodos; (c) las señales neuronales requieren ser caracterizadas con medidas de escala, localización y forma debido a la amplia variedad de patrones que exhiben; (d) si dos neuronas se disparan simultáneamente, las formas de sus *spikes* se superponen y su separación no es sencilla; (e) en los análisis en tiempo real, el alto volumen de datos restringe el uso de modelos complejos. Algunos avances recientes en neurociencia gracias a los modelos de *spikes* son los siguientes: entender el funcionamiento del sistema nervioso, principalmente mediante la definición de subtipos neuronales (Zeng y Sanes, 2017); análisis de los procesos de aprendizaje y memorización (Donato, Rompani y Caroni, 2013; Rey, Pedreira y Quiroga, 2015), estudio de la conectividad entre neuronas (Buzsáki, 2004), tratamiento de pacientes con epilepsia (Sharma et al., 2017), e incluso el desarrollo de interfaces cerebro-computadora de alta precisión (Moghaddasi et al., 2020).

En el ámbito de la neurociencia existen aún muchos problemas sin resolver. En esta tesis, nos centramos en dos: la clasificación en tipos celulares y el *Spike Sorting*. La clasificación en tipos celulares busca definir taxonomías jerárquicas celulares en base a características electrofisiológicas, morfológicas y genéticas (Zeng y Sanes, 2017). En el *Spike Sorting*, los *spikes* se clasifican de forma no supervisada, correspondiendo cada grupo a una neurona o tipo de neurona (Rey, Pedreira y Quiroga, 2015). En ambos casos, se requiere representar la forma del *spike* en términos de características. La extracción de características se ha hecho de muchas formas en la literatura: empleando mediciones directas sobre la señal (Caro-Martín et al., 2018; Gouwens et al., 2019), usando análisis en componentes principales (Quiroga, Nadasdy y Ben-Shaul, 2004; Ekanadham, Tranchina y Simoncelli, 2014; Veerabhadrappa et al., 2020) o usando modelos. Dentro del último grupo, algunos apuestan por modelos biofisiológicos (Teeter et al., 2018), descomposiciones de Fourier (Ghaderi, Marateb y Safari, 2018), wavelets (Wang et al., 2019), o modelos gaussianos (Souza et al., 2019). Además, otro problema desafiante en neurociencia es la eliminación del ruido de las señales, esencial especialmente en mediciones *in vivo* por su alto nivel de ruido. Las redes neuronales se utilizan predominantemente para resolver esta tarea (Lecoq et al., 2021; Sebastian et al., 2021).

La motivación fundamental de esta tesis es demostrar el potencial de modelos estadísticos oscilatorios para describir, analizar y generar conclusiones a partir de señales neuronales. La formulación paramétrica de los modelos empleados resulta sencilla a la vez que flexible, lo que los hace eficaces para caracterizar las formas de los *spikes*. En concreto, se ha demostrado su potencial en la clasificación en tipos celulares y el *Spike Sorting*.

Los modelos frequency modulated Möbius (FMM) se proponen en esta tesis como marco teórico general que genera un nuevo paradigma para formular y resolver una amplia gama de problemas en neurociencia. Los modelos han sido implementados en el lenguaje de programación R. El núcleo de esta metodología son las ondas FMM y los modelos FMM. Las ondas FMM son funciones no lineales paramétricas para describir oscilaciones en las que la fase se modela con transformaciones Möbius (véase Downs y Mardia, 2002; Kato, Shimizu y Shieh, 2008). Cada onda se define con cuatro parámetros básicos que miden amplitud, fase y forma. A partir de estos, se pueden definir paramétricamente otras medidas útiles en la práctica, como son los picos y los valles en las señales. Los FMM se definen como modelos de señal más error, en los que la señal es una suma de ondas FMM. El modelo monocomponente se introdujo en Rueda, Larriba y Peddada, 2019, donde se presenta un algoritmo para hallar los estimadores máximo verosímiles del modelo. Otro elemento fundamental en la metodología son las restricciones de orden entre parámetros, que integran información a priori en el modelo para incrementar su eficiencia e interpretabilidad fisiológica (Brunk, 1955; Barlow et al., 1972; Menéndez y Salvador, 1991; Rueda, Ugarte y Militino, 2016; Larriba et al., 2020). También se han resuelto diversos problemas mediante técnicas de machine learning o aprendizaje automático. Por un lado, se han incorporado características extraídas con modelos FMM a procedimientos estándares como el discriminante lineal, support vector machine o random forest (Hastie, Tibshirani y Jerome, 2009). Por otro lado, se ha desarrollado un procedimiento de clustering que combina la metodología FMM con modelos basados en mixturas (Ramsay y Silverman, 2005; Chamroukhi y Nguyen, 2019).

Esta tesis, que se presenta como un compendio de publicaciones, consta de cinco artículos científicos: tres publicados, uno aceptado y otro bajo revisión, existiendo entre ellos una cohesión temática. La exposición sigue el orden cronológico en el que se desarrollaron.

Rueda, Rodríguez-Collado y Larriba, 2021 establece el marco teórico fundamental que conducirá todas las contribuciones de la tesis. Los modelos FMM son presentados como una descomposición de amplitud y frecuencia modulada especialmente apta para señales oscilatorias. La contribución más relevante de este trabajo es la definición del modelo FMM multicomponente. Se propone un algoritmo *backfitting* para la estimación de los parámetros del modelo (Buja, Hastie y Tibshirani, 1989). La sencillez de la formulación paramétrica de los modelos FMM facilita, por un lado, la obtención de características relevantes para el estudio de las señales, como es la señal analítica o el espacio de estados (Wigren, 2015; Pikovsky y Rosenblum, 2015; Ashwin, Coombes y Nicks, 2016). Por otro lado, permite formular el modelo como un sistema de ecuaciones diferenciales. Por último, se muestra la capacidad del modelo para reproducir la dinámica neuronal y se compara con otros modelos previamente empleados en la literatura.

En Rodríguez-Collado y Rueda, 2021a se presenta un modelo paramétrico de formulación sencilla capaz de reproducir la dinámica neuronal del modelo de Hodgkin-Huxley, referente aún en neurociencia décadas después de su creación (Hodgkin y Huxley, 1952). El FMM_{ST} es un modelo para *spike trains* en el que cada *spike* se modela con dos componentes y se asume, mediante restricciones, que todos los *spikes* de la cadena tienen la misma forma. Ataja diversos problemas del modelo de Hodgkin-Huxley, como es su falta de robustez o identificabilidad de los parámetros (Marom, 2016). El modelo se valida con una amplia experimentación, lo que nos lleva a conclusiones interesantes como son relacionar la morfología y los flujos de iones de la neurona con la forma y tasa de aparición de los *spikes*, respectivamente.

El problema de la clasificación en tipos celulares se afronta en Rodríguez-Collado y Rueda, 2021b. En concreto, se define una taxonomía jerárquica para células de la corteza visual en ratones combinando características electrofisiológicas extraídas con el modelo FMM de la base de datos Allen cell types (Allen Institute for Brain Science, 2021) con características genéticas de Tasic et al., 2016. Se trata de la primera taxonomía genuinamente circular de la literatura, de tal forma que es una combinación de una herramienta visual con un procedimiento de clustering integrado en el que se define un orden circular a partir del análisis de componentes principales. Por último, se prueba la alta capacidad de los parámetros FMM para discriminar los diferentes tipos neuronales mediante procedimientos de aprendizaje automático.

En Rodríguez-Collado y Rueda, 2022, se propone un nuevo enfoque basado en los modelos FMM en el marco de los modelos de mixturas para resolver el problema del *Spike Sorting*. Se trata de un procedimiento de análisis de datos funcional y permite, entre otras cosas, comparar paramétricamente los patrones medios de los clusters. El artículo presenta un algoritmo esperanza-maximización para estimar los parámetros del modelo, siguiendo la metodología de Dempster, Laird y Rubin, 1977; Chamroukhi y Nguyen, 2019, así como un procedimiento para determinar de forma automática el número de clusters existente en los datos. El método se compara con otros enfoques seguidos en la literatura y la nueva propuesta resulta ser superior en términos de precisión, robustez e interpretabilidad.

En paralelo a estos trabajos, se desarrolló el paquete de software **FMM** en el lenguaje de programación R. Permite ajustar, explorar, visualizar y generar todo tipo de modelos FMM de forma sencilla. Se programó cuidando el estilo, siguiendo pautas como evitar la duplicidad del código, sacar partido a la programación orientada a objetos y optimizar los cálculos realizados mediante el cómputo paralelizado (R Core Team, 2020; Microsoft Corporation y Weston, 2020a; Microsoft Corporation y Weston, 2020b). Una vez que la primera versión completa y estable del paquete fue publicada en el Comprehensive R Archive Network, se presentó y detalló su uso en Fernández et al., 2022.

Esta tesis supone un desarrollo significativo de una metodología nueva y de alta aplicabilidad basada en las ondas FMM. Se ha implementado en un paquete software disponible de forma abierta y se ha demostrado su capacidad para resolver retos actuales en neurociencia, superando los resultados de métodos empleados habitualmente en la literatura. A partir de las conclusiones obtenidas en la tesis, se pueden definir muchas líneas futuras de investigación. Por una parte, quedan muchos problemas en neurociencia a los que dar respuesta. Algunos ejemplos son el estudio de la sincronización entre poblaciones neuronales (Aydore, Pantazis y Leahy, 2013), esencial para conocer los mecanismos que rigen el sistema nervioso, o el análisis de datos neuronales multicanal (Yger et al., 2018), en el que serán claves los modelos FMM multivariantes como el definido en Rueda et al., 2022 para el análisis de las doce derivaciones del electrocardiograma. Otra señal estrechamente asociada con algunos problemas no resueltos en neurociencia es el electroencefalograma, en el que se mide la actividad neuronal del cerebro. Más allá de la neurociencia, la versatilidad de la metodología FMM servirá para estudiar otros muchos problemas ligados a señales oscilatorias de diversos campos, como son la electrofisiología ocular, la electrocardiografía, o la química experimental, entre otros. Creemos que la metodología puede tener una importante repercusión en la medicina, mejorando la detección, tratamiento y control de enfermedades como la epilepsia, esquizofrenia, o demencia. Por otro lado, queda mucho trabajo por hacer desde un punto de vista metodológico, como es el desarrollo de procedimientos inferenciales o la creación de nuevos procedimientos para el análisis de datos funcionales en base al trabajo hecho en Rodríguez-Collado y Rueda, 2022. Evidentemente, esto implicará la implementación de nuevas funcionalidades en el paquete software **FMM**. Por último, se afrontará el reto de optimizar el tiempo de cómputo del algoritmo de estimación, usando lenguajes más eficientes, como C, así como programación GPU.

Contents

List of Figures

List of Tables

1. Introduction

1.1 Motivation

A system in which a particle or set of particles moves returning to its initial state after a certain period is an oscillatory system, and an oscillation is the repetitive variation of a signal or measure associated with the system. Oscillatory systems govern all kinds of phenomena in many different fields. For instance, the starlight patterns in astronomy, the spectrophotometry curves in experimental chemistry, the electricity demand in energy engineering, among many others. In biomedicine, oscillatory signals arise ubiquitously in: the electrical activity of the heart, the respiratory flow of the lungs, and circadian rhythms gene activation, as well as the electric neuronal activity. The study of oscillatory signals have been approached by researchers from many different disciplines, and differences often arise in terms of terminology, concepts, and methods. The time-frequency approach is preferred by researchers coming from the communication field (Boashash, 2016), whereas neuroscientists prefer a more physical focus by describing systems dynamics with ordinary differential equations (ODEs; Wigren, 2015; Pikovsky and Rosenblum, 2015; Ashwin, Coombes, and Nicks, 2016). Finally, the statistical perspective, suitable for real and noisy signals, have been preferably adopted in chronobiology (Larriba et al., 2020).

The focus of this thesis is the analysis of neuronal electric signals. Those signals register rapid voltage rises in a neuron lasting a few milliseconds called spikes or action potential curves, after which the voltage returns to the initial baseline level. Figure 1.1, modified from Rueda, Rodríguez-Collado, and Larriba, 2021, shows a typical spike and the three stages composing it. The spikes in a neuronal electric signal can either be analyzed separately or as succession (which is known as spike train). The spikes serve as the informational unit between neurons, and their firing rate and shape determine the cell's type, making them crucial for neuronal research. The first model that described accurately the dynamics of the spikes was the Nobel-awarded Hodgkin-Huxley (HH) model (Hodgkin and Huxley, 1952), which assumed that the neuronal activity behaves as an electric circuit governed by an ODE system. With the HH model as reference, many biophysically-realistic models were presented afterwards, such as the FitzHugh-Nagumo model (Fitzhugh, 1961). However, their deterministic nature, formulation complexity, lack of parameter identifiability, and noise susceptibility made neuroscientists search for alternative approaches. Nowadays, two main types of models are used: biophysically simplistic models, such as the Hopfield model (see Abbott and Kepler, 1990), the Izhikevich model (Izhikevich, 2003) or the leaky integrate-and-fire family of models (Teeter et al., 2018, and references therein), and data-driven models, such as Brunton and Beyeler, 2019. Some

challenges that arise from the analysis of neuronal spike data are the following: (a) signals have high levels of noise, especially if recorded non-invasively; (b) spikes are highly asymmetrical, making unsuitable some traditionally popular models; (c) neuronal signals display a wide variety of patterns which require a precise characterization of the waveforms, not just timing and amplitude; (d) if two neurons fire simultaneously, their spikes are recorded overlapping, and its separation is not trivial; (e) in real-time analyses, the high data volume limits the use of complex models. Some recent advances in neuroscience due to spike models are the following: understanding of the neuronal system functioning, mainly by defining neuronal subpopulations (Zeng and Sanes, 2017), comprehension of learning and memory processes (Donato, Rompani, and Caroni, 2013; Rey, Pedreira, and Quiroga, 2015), study of connectivity between close-by neurons (Buzsáki, 2004), treatment of epileptic patients (Sharma et al., 2017), and even development of high-accuracy brain-machine interfaces (Moghaddasi et al., 2020).

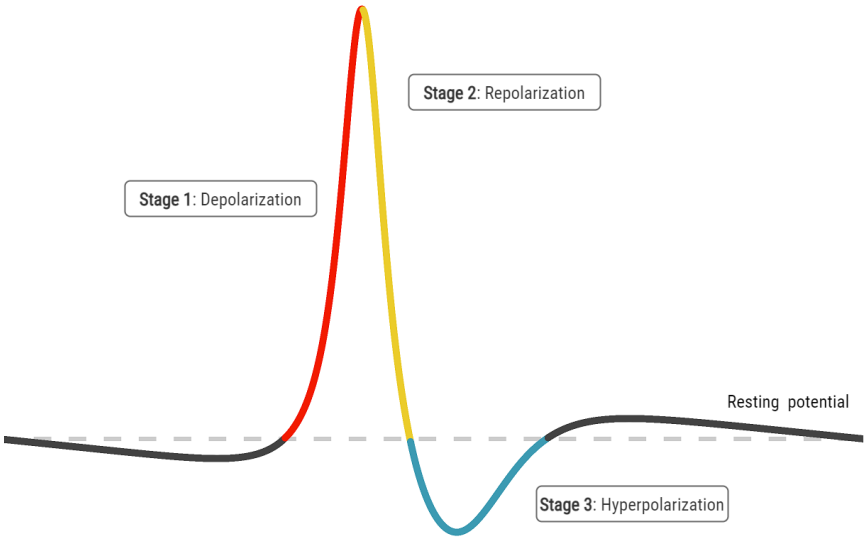

FIGURE 1.1: A typical neuronal spike and its stages.

Nowadays, there are still many open problems in the neuroscience literature. In this thesis, we deal with two problems: the cell-type classification and Spike Sorting. Cell-type classification entails the definition of hierarchical taxonomies of cells based on electrophysiological features, along with morphological and genetic features (Zeng and Sanes, 2017). Spike Sorting implies the identification of unlabelled neuronal signals corresponding to different neurons (Rey, Pedreira, and Quiroga, 2015). In both problems, the spike waveform is required to be adequately characterized in term of features. This has been conducted with many different approaches, such as landmark measurements directly in the spikes (Caro-Martín et al., 2018; Gouwens et al., 2019), using principal component analysis (PCA; Quiroga, Nadasdy, and Ben-Shaul, 2004; Ekanadham, Tranchina, and Simoncelli, 2014; Veerabhadrappa et al., 2020; Gouwens et al., 2020), or using features derived from a model. Within the latter group, some proposals include biophysical models (Teeter et al., 2018), Fourier models (Ghaderi, Marateb, and Safari, 2018), wavelets (Wang et al., 2019; Veerabhadrappa et al., 2020), or gaussian models (Souza et al., 2019), among others. Moreover, the data denoising is another challenging open problem in neuroscience, essential for *in vivo* recordings as high noise levels exist in data. This task is predominately solved with neuronal networks (Lecoq et al., 2021; Sebastian et al., 2021).

The principal motivation of this thesis is to demonstrate the potential of statistical oscillatory models to describe, analyze and generate conclusions from neuronal signals. The simple but rich parametric formulation of the models is an effective tool to characterize the spike waveforms. In particular, its contribution to cell-type classification and Spike Sorting has been shown.

1.2 Methodology

This thesis proposes a general statistical framework, the frequency modulated Möbius (FMM) models, which states a new paradigm to formulate and solve a wide range of problems in neuroscience. The implementation of the models has been done in R software.

The core of the methodology are the FMM waves, also called components, and the FMM models. FMM waves are non-linear parametric time functions describing a single oscillation. They are defined as follows:

$$W(t; A, \alpha, \beta, \omega) = A \cos\left(\beta + 2\arctan\left(\omega \tan\left(\frac{t-\alpha}{2}\right)\right)\right), \quad t \in [0, 2\pi), \tag{1.1}$$

where $A \in \Re^+$, $\alpha \in [0, 2\pi)$ measure the wave amplitude and phase location, respectively, and $\beta \in [0, 2\pi), \omega \in [0, 1]$ measure the shape. The phase of the wave is defined using a Möbius link (see Downs and Mardia, 2002; Kato, Shimizu, and Shieh, 2008), which provides flexibility to describe non-sinusoidal patterns. Figure 1.2, taken from Rodríguez-Collado and Rueda, 2022, shows waveforms patterns for different parametric configurations. Moreover, other important parameters of practical use, such as peak and trough times, can be derived directly from the model's basic set of parameters.

The FMM_m models are defined as signal plus error models, in which the signal is a sum of m FMM waves. The monocomponent model ($\text{m} = 1$) was introduced in Rueda, Larriba, and Peddada, 2019, where a two-step algorithm is designed for the parameters' maximum likelihood estimator (MLE). In the first step, α and ω are selected from a grid of values, whereas the rest of the parameters are derived by solving a least-squares problem. In the second step, the estimators are refined using Nelder-Mead optimization (Nelder and Mead, 1965). Another crucial element in the methodology is the incorporation of order restrictions on the model parameters, which integrate prior information to increase the model's efficiency, as well as to enhance its physiological interpretability (Brunk, 1955; Barlow et al., 1972; Menéndez and Salvador, 1991; Rueda, Ugarte, and Militino, 2016; Larriba et al., 2020).

In the presented works, certain problems have been solved using machine learning or statistical learning procedures. On the one hand, FMM parameters were incorporated as features in standard supervised learning methods, such as linear discriminant analysis (LDA), support vector machine (SVM), and random forest (RF; Hastie, Tibshirani, and Jerome, 2009; Fernandez-Delgado et al., 2014). On the other hand, a specific functional clustering procedure has been developed based on the FMM methodology and the mixture modeling framework (Ramsay and Silverman, 2005; Chamroukhi and Nguyen, 2019).

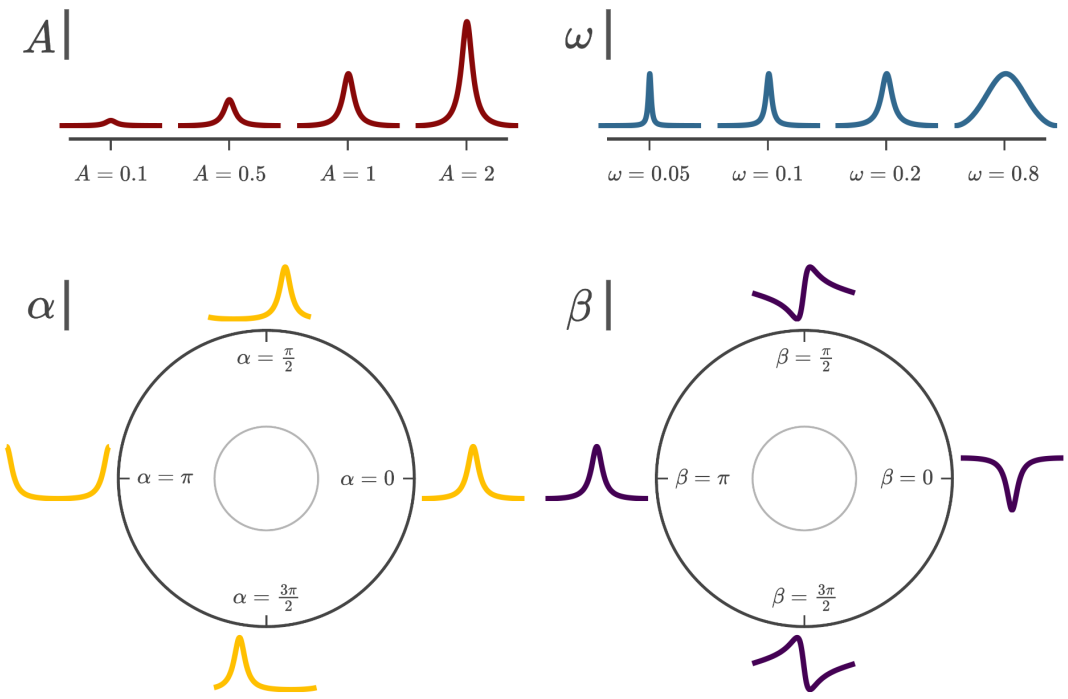

FIGURE 1.2: $W(t; A, \alpha, \beta, \omega)$ for various $(A, \alpha, \beta, \omega)$. Unless stated otherwise, the value of the parameters are $A = 1, \alpha = 0, \beta = \pi, \omega = 0.2$.

1.3 Publications in the compendium

This thesis is presented as a compendium of publications. Specifically, it consists of five scientific publications: three published, one in-press, and another under review. The methodological details, contributions and results of each of them are shown later in this document. An overview of the five papers included in the thesis can be seen in Figure 1.3, and a general outline of the contributions of the works is given on the next paragraph.

The thesis project began with the exhaustive study of Rueda, Larriba, and Peddada, 2019, which defined the monocomponent FMM model and served as an introduction to the current state of the art. Initially, two lines of work were defined: the study of the FMM models' theoretical properties and the applied analysis of neuronal signals with the FMM model. The former line of work originated Rueda, Rodríguez-Collado, and Larriba, 2021, which establishes the thesis' theoretical framework. It introduces the FMM_{m} model as an alternative amplitude frequency-modulated (AM-FM) decomposition for oscillatory signals. The model's flexible parametric formulation allows the estimation of interesting signal characteristics and the demonstration of several solid mathematical properties that describe the underlying oscillatory system's dynamics. The other line of work, in which single spikes and spike trains generated by the HH model were analyzed, resulted in Rodríguez-Collado and Rueda, 2021a. The main methodological contribution of the paper is the presentation of the FMM_{ST} model, a simple model with identifiable parameters that accurately represents the renowned HH model. It is validated by analyzing a wide range of simulated and real data. The results in these first two papers revealed the potential of the FMM models to analyze neuronal data, which motivated a more thorough

validation using real data. In particular, we addressed the problem of the cell-type classification. Rodríguez-Collado and Rueda, 2021b was initially outlined as an applied paper; however, it also entailed a highly compelling methodological contribution: the proposed neuronal taxonomy has a rigorous circular topology, the first in the literature. The final part of the thesis deals with the unsupervised classification of neuronal signals, addressing the problem of Spike Sorting. Rodríguez-Collado and Rueda, 2022 is both a methodological and applied contribution that presents a functional approach based on the FMM waves to analyze all kinds of oscillatory or quasi-oscillatory signals arising in multiple contexts. In parallel with all the works aforementioned, the **FMM** package was developed as an R package ready to fit, explore, visualize, and generate all kinds of FMM models. Following the package's publication in the Comprehensive R Archive Network (CRAN), Fernández et al., 2022 presented and detailed its use.

1.4 Contributions and results

The following section contains a detailed description of the contributions and results of each of the publications considered in this thesis.

To improve the readability of this section, the mathematical notation from the original contributions has been slightly modified, so that an homogeneous notation can be considered for the rest of this section.

In the following, let us assume that $\boldsymbol{t} = (t_1, ..., t_p)$, with $t_0 \leq t_1 < t_2 < ... < t_p \leq T$, and $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, $\boldsymbol{x}_i = (x_i(t_1), ..., x_i(t_p)), \forall i \in \{1, ..., n\}$ are the time points and observed signals, also called curves, respectively. To improve readability, the subscript i will be deleted whenever is possible. Without loss of generality, we assume that $t \in [0, 2\pi)$. Otherwise, consider $t' \in [t_0, T + t_0]$ with t_0 as the initial time value and T as the period, and transform the time points by $t = \frac{(t'-t_0)2\pi}{T}$.

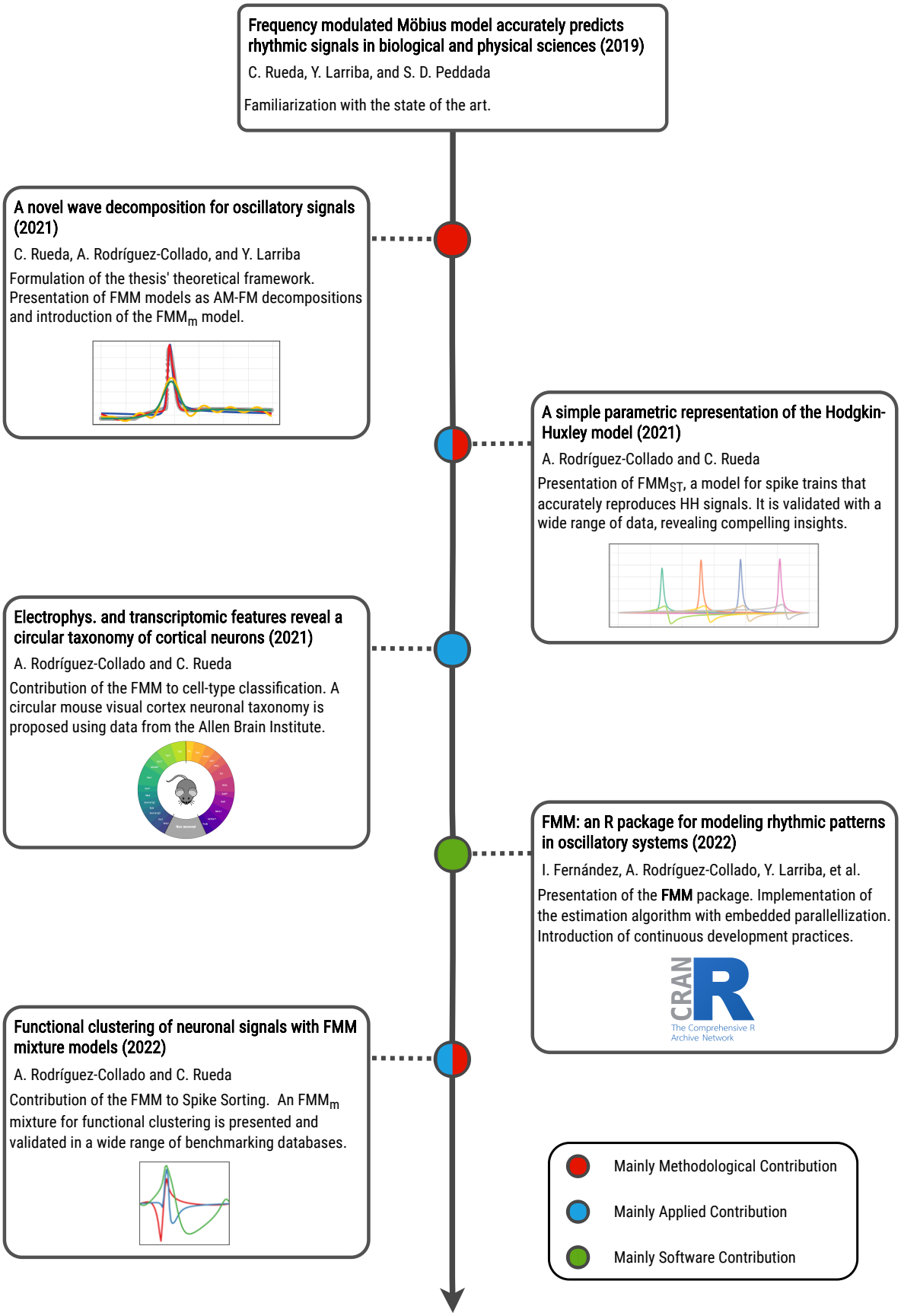

FIGURE 1.3: Timeline of the contributions included in the thesis.

A novel wave decomposition for oscillatory signals (Rueda, Rodríguez-Collado, and Larriba, 2021)

Rueda, Rodríguez-Collado, and Larriba, 2021 establishes the theoretical foundation that conducts all the thesis' contributions. The dynamical information of an observed signal has traditionally been inferred with AM-FM decompositions such as Fourier. However, alternative decompositions have been proposed due to traditional ones falling short when the analyzed signals are not composed of harmonic functions (Kowalski, Meynard, and Wu, 2018; Lin, Su, and Wu, 2018; Sandoval and Leon, 2018, among others). This work presents the FMM as a novel AM-FM decomposition suitable for oscillatory or quasi-oscillatory signals in which amplitude is constant, and frequency is modelled using a Möbius transformation. In particular, the paper's main contribution is the introduction of the multicomponent FMM model of order m, denoted by FMM_m and defined as an additive m-component signal plus error model as follows:

$$x(t) = M + \sum_{J=1}^{m} W(t; A_J, \alpha_J, \beta_J, \omega_J) + e(t), \tag{1.2}$$

verifying $\pi \leq \alpha_1 \leq \alpha_2 \leq ... \leq \alpha_{m-1} \leq \alpha_m \leq \alpha_1 \leq \pi$, $A_1 = \max_{1\leq J\leq m} A_J$, and $e(t_j) \sim \mathcal{N}(0, \sigma^2), \forall j \in \{1, ..., p\}$. Identifiability is guaranteed by the αs and As restrictions.

A MLE algorithm is proposed for the parameters estimation. From a theoretical point of view, it is a backfitting algorithm as introduced in Buja, Hastie, and Tibshirani, 1989. Estimations are done by fitting repeatedly FMM_1 models to the residue until a stop criterion is attained.

The contribution of this paper is mainly methodological, as many theoretical properties of the FMM models are introduced. The analytic signal (Gabor, 1946), assumed to be the complex signal underlying oscillatory processes, can be easily derived for FMM_m signals thanks to its simple parametric formulation. This significantly facilitates the study of other key elements for signals analysts, such as the phase space, in which all possible states of a system are represented, or the instantaneous frequency, the derivative of the signal's phase (Wigren, 2015; Pikovsky and Rosenblum, 2015; Ashwin, Coombes, and Nicks, 2016). Many authors consider that the instantaneous frequency is physically interpretable only if it increases with time (Winfree, 2001; Deng et al., 2016; Oprisan, 2017). This is guaranteed in the case of the dominant component, the first FMM wave in cases in which its amplitude is expected to be much larger. It identifies the most prominent biological process existing in the signal. Moreover, it is particularly useful in practice when the observed data is noisy, as it provides an approximate reconstruction of the signal and its parameters estimators are robust. Finally, the ODE representation of the FMM_1 signals is presented, which describes the underlying system dynamically in time. It is derived by solving an inverse problem, in which ODEs governing the system are unknown and must be reconstructed from data (as in Brunton, Proctor, and Kutz, 2016 or Kang, Liao, and Liu, 2021, among many others). The associated ODE system to an FMM_m signal can be derived using the previous result and its additive structure, as done for instance in Wigren and Söderström, 2005. More details about the aforementioned results can be found in Section III.B) New Theoretical Properties of Rueda, Rodríguez-Collado, and Larriba, 2021.

The FMM_m model is validated with a simulated experiment and neuronal signals from the Allen cell types database, a repository of intracellular electrophysiological neuronal recordings (Allen Institute for Brain Science, 2021). In the former case, FMM_m signals have been used to simulate the waveforms shown in Figure 1.4, originating from different works (Aksenova et al., 2003; Lewandowska et al., 2015; Caro-Martín et al., 2018; Trainito et al., 2019; Raghavan, Fee, and Barkhaus, 2019). The high flexibility of the FMM_m signals is shown, along with the potential of the FMM parameters to characterize and discriminate the waveforms, and the high robustness of the parameters estimators. In the analysis of Allen cell types data, the FMM models attain outstanding goodness of fit, clearly superior to competing decompositions. Finally, the capacity of the FMM parameters to discriminate different neuronal types is briefly explored, foreshadowing future works in cell-type classification.

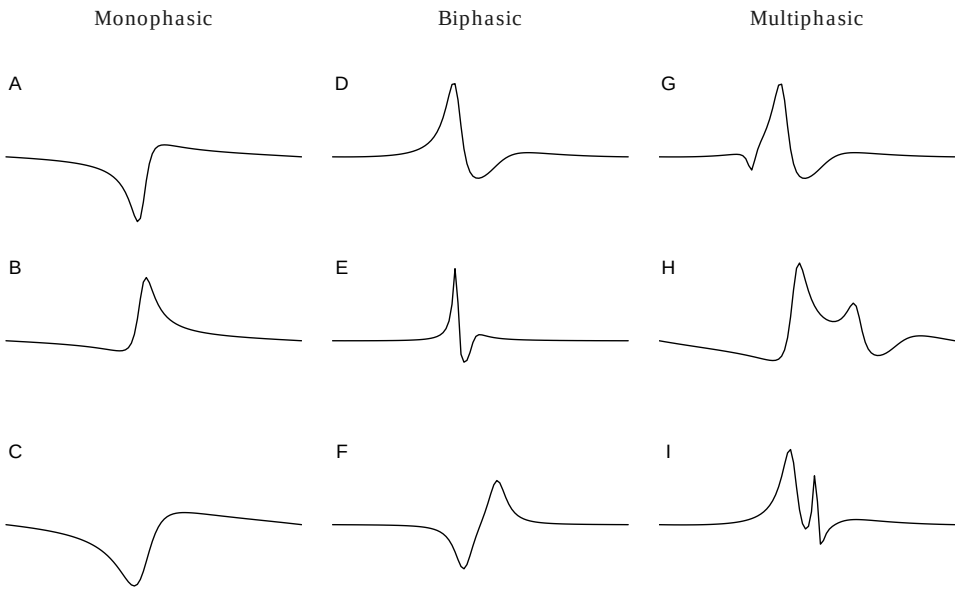

FIGURE 1.4: Different spike waveforms.

A simple parametric representation of the Hodgkin-Huxley model (Rodríguez-Collado and Rueda, 2021a)

Rodríguez-Collado and Rueda, 2021a presents the FMM_{ST} model, which reproduces accurately the signals generated by the Nobel-awarded HH model (Hodgkin and Huxley, 1952). It assumes that the electric activity of the neuron behaves like a circuit governed by an ODE system with four state variables and more than twenty-five parameters. Due to successfully replicating the electrophysiological activity of many organisms, the HH model remains to be a reference within neuroscience and has served to inspire several models (Fitzhugh, 1961; Abbott and Kepler, 1990; Izhikevich, 2003, among many others). However, it lacks identifiability, as various parameter configurations can lead to the same observed signal; and it also lacks robustness, as minor manipulations of the values of the parameters can change its output completely (Marom, 2016). In order to address these issues, alternative parametrizations have been proposed for the HH. The pair $(\mathcal{S}, \mathcal{K})$ is a popular bidimensional reparametrization which captures the neurons' excitability due to its structure and its ion fluxes respectively (Ori, Marder, and Marom, 2018).

The FMM_{ST} model is a particular multicomponent FMM model with restrictions on the parameters. It assumes that signals are a succession of equally shaped spikes, being each one described by two FMM components. Let s denote the number of spikes in a signal, which is assumed known. The FMM_{ST} model is defined as follows:

$$x(t) = M + \sum_{S=1}^{s} W(t; A_S^A, \alpha_S^A, \beta_S^A, \omega_S^A) + W(t; A_S^B, \alpha_S^B, \beta_S^B, \omega_S^B) + e(t), \tag{1.3}$$

verifying $A_S^A < A_S^B; \forall S \in \{1, ..., s\}$, $\pi \leq \alpha_1^A \leq \alpha_1^B \leq \alpha_2^A \leq \alpha_2^B ... \leq \alpha_S^A \leq \alpha_S^B \leq \alpha_1^A \leq \pi$, and $e(t_j) \sim \mathcal{N}(0, \sigma^2), \forall j \in \{1, ..., p\}$, as well as the restrictions

$$\beta_1^A = \beta_S^A, \;\; \beta_1^B = \beta_S^B, \;\; \omega_1^A = \omega_S^A, \;\; \omega_1^B = \omega_S^B, \quad \forall S \in \{2, ..., s\}. \tag{1.4}$$

The restrictions on the βs and ωs represent the equal spike shapes and provide physiologically interpretable solutions. Furthermore, depending on the application at hand, additional restrictions may be imposed in order to use a simpler model.

The model is validated with a broad simulated experiment and real data. In the first case, 5000 signals were simulated with the HH model corresponding to a large variety of parameters configurations. The attained prediction accuracy of the FMM_{ST} model is high, as can be seen in Figure 1.5. Moreover, the model parameters describe spike waveforms differences between neuron types, and can be used to predict the main HH parameters with machine learning methods (Ori, Marder, and Marom, 2018; Silva, 2014; Marom, 2016). In addition to the goodness of fit, the generalized degrees of freedom of the procedures has been calculated in order to measure the underlying complexity (Ye, 1998). The SVM obtained the best predictions and highest complexity, whereas the feature relevance measures provided by tree-based methods provide compelling insights: $\mathcal{S}$ is related to the spike shape, whereas $\mathcal{K}$ is strongly associated with the numbers of spikes.

This work also analyzed the squid giant axon signals from the SGAMP database (Paydarfar, Forger, and Clay, 2006), similar to those that originally inspired the HH model definition. Interestingly enough, differences are found between these signals and those generated by the HH model. In particular, evidence indicates that the former have simpler, more symmetrical waveforms than the latter. More details can be read in the Results Section of Rodríguez-Collado and Rueda, 2021a.

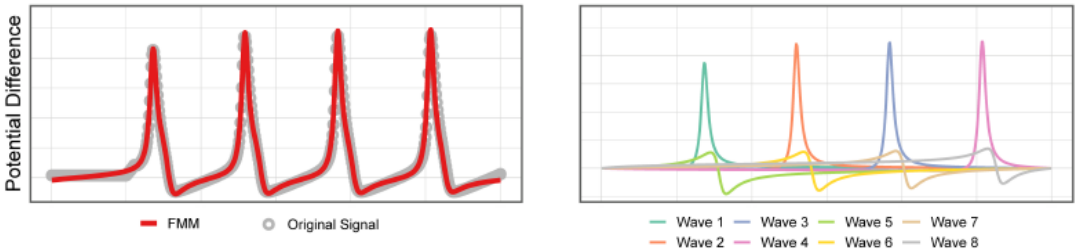

FIGURE 1.5: Neuronal signal simulated with the HH model and the estimated FMM_{ST} signals in red (left). Waves of the fitted FMM_{ST} model (right).

Electrophysiological and transcriptomic features reveal a circular taxonomy of cortical neurons (Rodríguez-Collado and Rueda, 2021b)

The problem of cell-type classification is dealt with in Rodríguez-Collado and Rueda, 2021b. A taxonomy for neuronal cells in the mouse visual cortex is derived with electrophysiological features extracted with an FMM_3 model from the Allen cell types database, and transcriptomic features from Tasic et al., 2016. It is shown in Figure 1.6. Its most original characteristic is the circular topology, as it combines a circular visualization tool (commonly used to represent genomic data) and an integrated clustering approach that uses the circular order defined with the two first principal components. The dissimilarity measure is a circular distance and the location of clusters in the circle is derived from the clusters location in the two first principal components plane. Thus, this is not just a visualization tool, but a genuine circular taxonomy. It is defined hierarchically: neurons are broadly classified based on their neurotransmitter into either GABAergic (right side) or glutamatergic (left), whereas further subclassifications are based on the expression of gene markers or Cre lines (Tremblay, Lee, and Rudy, 2016; Zeng and Sanes, 2017). The proposal mostly goes in agreement with the literature (Gouwens et al., 2019; Tasic et al., 2016), and locates certain Cre lines for the first time in a taxonomy. Moreover, a Shiny app has been developed to illustrate the spike differences between Cre lines (Chang et al., 2020). It can be accessed through https://alexarc26.shinyapps.io/median_ap_profile_by_cre_line/.

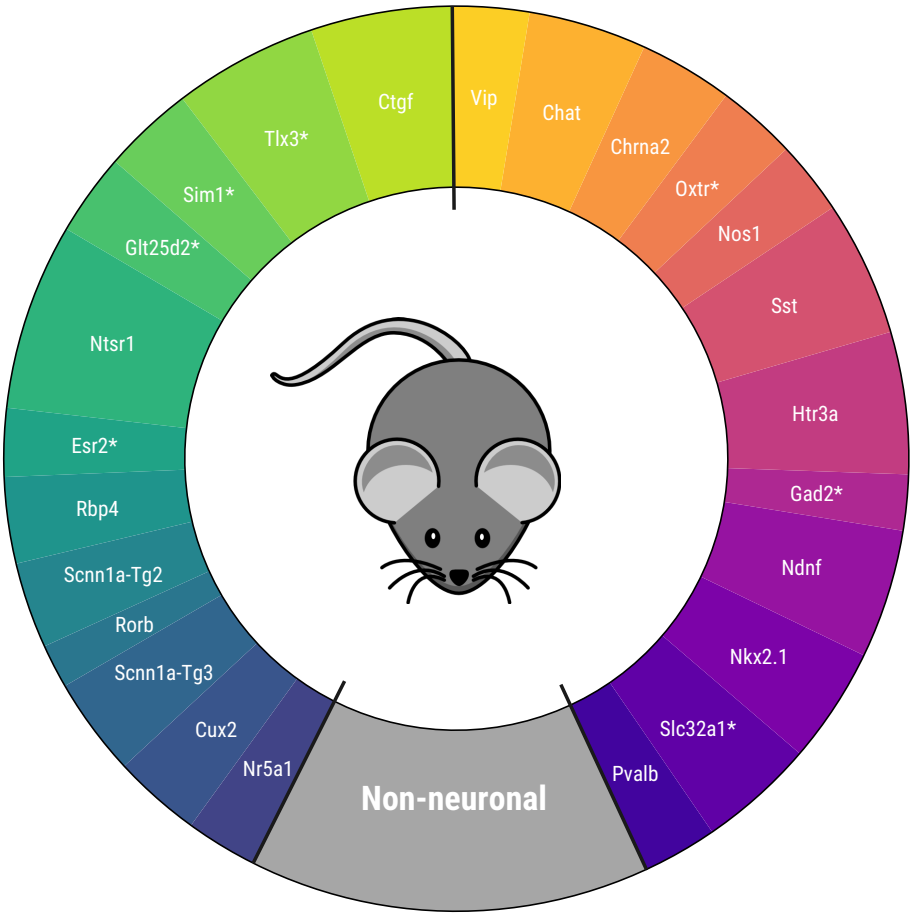

FIGURE 1.6: Proposed circular mouse cortical cell taxonomy.

The taxonomy was validated using FMM-derived features to discriminate different cell subclasses, defined with the expressed Cre lines. In particular, various machine learning methods were used, ranging from LDA to "black boxes" such as SVM and model averaged neural network (AvNNet), and including RF and gradient boosting decision trees (GBDT; Fernandez-Delgado et al., 2014; Zhang et al., 2017). The classification problem was tackled in different stages, depending on the used features, subclasses and discarded outliers. Table 1.1 presents the results of the final stage, with the left side showing the cross-validated accuracies and kappa statistics -which measures the improvement over a random classifier-, and the right side the AvNNet's confusion matrix. High discrimination accuracy is attained, and the feature relevance measures, provided by RF and GBDT, remark that the differences between subclasses are related to the shape of the repolarization and depolarization phases, captured mainly by the β parameter of the first FMM component.

TABLE 1.1: Cross-validated accuracy and kappa statistic of the different machine learning methods in the discrimination of the subclasses (left), and cross-validated confusion matrix of the AvNNet classifier (right) in stage (C+).

	Accuracy	Kappa
LDA	80.9 %	0.706
RF	89.4 %	0.841
GBDT	89.1 %	0.837
SVM	91.2 %	0.867
AvNNet	**91.3** %	**0.868**

		True class			
		Pvalb+	Sst+	Vip+	Glut.
Prediction	Pvalb+	93.1 %	14.3 %	5.3 %	0.2 %
	Sst+	6.5 %	79.5 %	4.1 %	1.2 %
	Vip+	0.0 %	3.1 %	82.4 %	1.7 %
	Glut.	0.3 %	3.1 %	8.2 %	96.9 %

FMM: an R package for modeling rhythmic patterns in oscillatory systems (Fernández et al., 2022)

Fernández et al., 2022 presents the **FMM** package, programmed in R and available in the CRAN at https://cran.r-project.org/package=FMM. It provides the functions required to fit, explore and visualize the results of mono- and multicomponent FMM models, including restricted multicomponent models. The functionality of the package revolves around the S4 FMM object. In addition, it provides the functions to simulate data with the FMM model, and example data from chronobiology, electrocardiography and neuroscience.

The **FMM** package has been developed with rigorous object-oriented programming, and following guidelines provided by widespread references (Wickham and Bryan, 2019; R Core Team, 2021). Moreover, several coding styles recommendations were adopted. In particular, code duplicity was avoided, objects and functions were semantically named, loops were substituted by vectorized operations wherever possible, comments were limited in number to indicate just meaningful insights, used functions from other packages were explicitly imported, and detailed documentation of the package was written. Besides, the package allows creating the visualizations with **ggplot2** (Wickham, 2016), and implements an embedded paralellized procedure for the model fitting that reduces significantly the computation times using the packages **parallel**, **doParallel**, and **foreach** (R Core Team, 2020; Microsoft Corporation and Weston, 2020a; Microsoft Corporation and Weston, 2020b).

It is important to remark that this work granted the research team a continuous development culture, in such a way that software is continuously improved and developed by adding new functionality and fixing possible bugs that arise. A Github repository was created to make the **FMM** package code open-source, available in https://github.com/alexARC26/FMM. Some good practices incorporated into the development cycle are the following: differentiation between stable and development versions of the package, issue-tracking with Github, use of Git as version-control tool, code-documentation syncing with package **roxygen2** (Wickham et al., 2020), unit testing with package **testthat** (Wickham, 2011), as well as code profiling and benchmarking with packages **profvis** and **microbenchmark** (Chang, Luraschi, and Mastny, 2020; Mersmann, 2019).

Functional clustering of neuronal signals with FMM mixture models (Rodríguez-Collado and Rueda, 2022)

Rodríguez-Collado and Rueda, 2022 presents the contribution of the FMM methodology to Spike Sorting. Motivated to solve this problem, we propose a novel model-based approach within the mixture modeling framework for clustering oscillatory functional data, the FMM mixture (MixFMM). The novel approach is based on the MixFMM_m model, defined as an FMM_m mixture:

$$\boldsymbol{x}(\boldsymbol{t}) = \sum_{k=1}^{K} \gamma_k (M^k + \sum_{J=1}^{m} W(\boldsymbol{t}; A_J^k, \alpha_J^k, \beta_J^k, \omega_J^k)) + \boldsymbol{e}_k(\boldsymbol{t}), \tag{1.5}$$

where $\gamma_1, ..., \gamma_K$, with $\gamma_k > 0$ and $\sum_{k=1}^{K} \gamma_k = 1$, are the mixture proportions, and $\boldsymbol{e}_k(\boldsymbol{t}) \sim \mathcal{N}_p(0, \sigma_k^2 \boldsymbol{I}_p)$, $1 \leq k \leq K$, with $\boldsymbol{I}_p$ being the $p \times p$ identity matrix. Besides, K is assumed known and corresponds to the number of clusters in our application, while $W(\boldsymbol{t}; A_J^k, \alpha_J^k, \beta_J^k, \omega_J^k) = (W(t_1; A_J^k, \alpha_J^k, \beta_J^k, \omega_J^k), ..., W(t_p; A_J^k, \alpha_J^k, \beta_J^k, \omega_J^k)), 1 \leq k \leq K, 1 \leq J \leq m$. Note that this paper's notation has been changed to be homogeneous with the rest of the works. The simple but rich parametric formulation facilitates handling the two main sources of variation in functional data (*amplitude* and *phase variation*), as well as the definition of the mean waveforms or patterns of the clusters (Marron et al., 2015; Park and Ahn, 2017; Claeskens, Devijver, and Gijbels, 2021). Moreover, these can be easily compared in terms of their parameters.

An expectation-maximization algorithm is proposed to solve the MixFMM model's parameter estimation, following the methodology in Dempster, Laird, and Rubin, 1977; Chamroukhi and Nguyen, 2019. Two steps are alternated until convergence is attained: the E-Step computes the expectation of the complete-data log-likelihood given the observed data and the current model parameters, whereas in the M-Step, the parameters are updated to maximize the expectation of the complete-data log-likelihood. We propose to initialize several times randomly in order to avoid local minima. Several numeric studies in the paper show how the proposal reaches a good solution in practice. Finally, the paper also proposes a likelihood-based method to select the number of clusters, similar to those in Birgé and Massart, 2007; Bouveyron, Côme, and Jacques, 2015; García, García-Ródenas, and Gómez, 2015.

The MixFMM approach has been validated in ten benchmarking datasets, including a wide variety of spike waveforms, noise levels, and recording situations as seen in Figure 1.7 (Quiroga, 2009; Quiroga, 2019; Allen Institute for Brain Science, 2021).

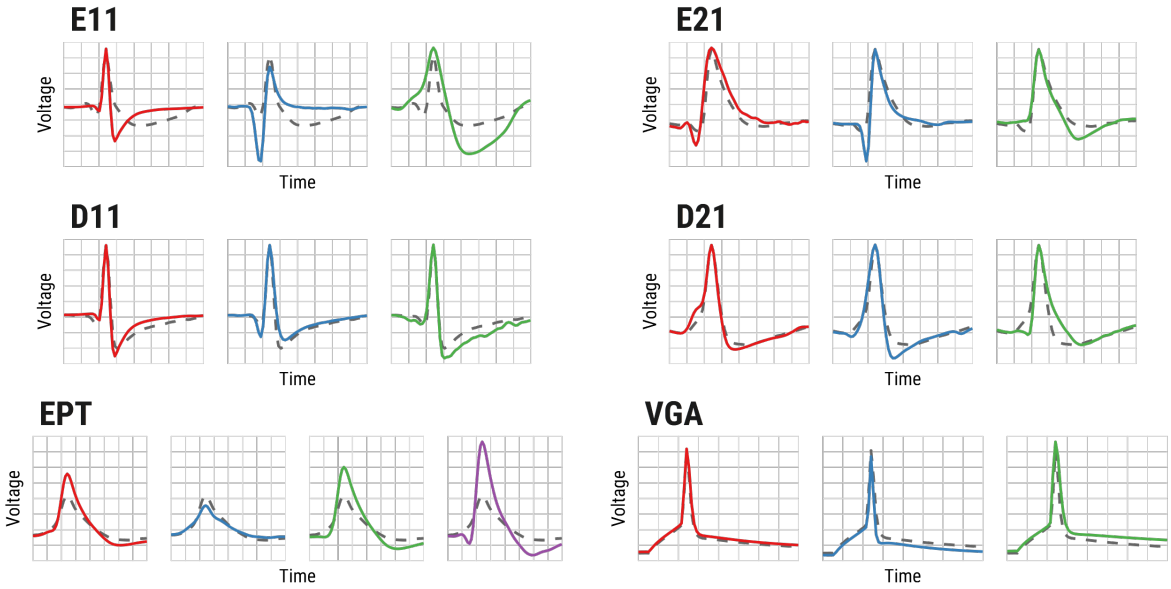

FIGURE 1.7: Spike means by class (solid lines) and global means (dashed) of datasets.

The results are compared with two traditionally used approaches in Spike Sorting, PCA plus either k-means (PCA+KM) or gaussian mixture models (PCA+GM) using indexes based on the datasets' associated groundtruth and internal cluster metrics. The MixFMM approach achieves clearly superior results in terms of accuracy, interpretability, robustness, cluster cohesion and separability. Table 1.2 shows the accuracy of the MixFMM approach against competitors. More details can be found in the Numerical Studies Section of Rodríguez-Collado and Rueda, 2022.

TABLE 1.2: Estimated accuracy for each dataset and method. The best result per dataset has been highlighted.

	Accuracy									
	E11	**E12**	**E21**	**E22**	**D11**	**D12**	**D21**	**D22**	**EPT**	**VGA**
PCA+KM	0.992	0.888	0.875	0.563	0.665	0.524	0.720	0.540	0.589	0.617
PCA+GM	0.960	0.958	0.851	0.347	0.905	0.510	0.640	**0.613**	0.680	0.710
MixFMM$_1$	0.992	0.966	0.936	0.700	0.895	0.530	0.705	0.575	**0.874**	**0.755**
MixFMM$_3$	**0.994**	**0.976**	**0.966**	**0.761**	**0.941**	**0.541**	**0.743**	0.606	**0.874**	0.746

1.5 Conclusions

This section summarizes the original contributions made within this thesis and discusses future research lines derived from its conclusion.

The thesis entails a significant statistical contribution: it lays the foundations of a novel, generally applicable methodology based on the FMM waves, which are simple yet rich parametric time functions. This thesis proved its potential to solve problems in neuroscience, outperforming recent methods proposed in the literature and highlighting compelling previously undiscovered neuronal insights. An R package of the presented methodology has been developed and is openly available in CRAN.

The contribution of this thesis to neuroscience is quite remarkable. The novel methodology is able to faithfully reproduce the electric neuronal dynamics, accurately characterizing the spike waveforms in terms of interpretable parameters. On the one hand, an equivalent but simple representation of the renowned HH model has been proposed, overcoming previous limitations such as its lack of robustness and parameter identifiability. Moreover, it stressed relevant notions about the neuron excitability: the shape of the spikes is related to the neuron's morphology, whereas the spikes' firing rate to the neuron's ion fluxes. On the other hand, the novel methodology is a significant contribution to the problems requiring spike characterization. Within cell-type classification, the first genuinely circular neuronal taxonomy has been defined. The taxonomy was the first to place some Cre lines and had relevant differences regarding other proposals in the literature, being the most significant one the placement of the Pvalb-positive neurons. These neurons, key in the learning and memorization processes, have been found to have similar characteristics to non-neuronal cells in the nervous system, such as glial cells. Many authors argue that cell-type taxonomies will be the key to completely understanding the nervous system's functioning. In the case of the Spike Sorting problem, a new model-based approach has been proposed, and the attained results are superior to traditionally used methods in terms of accuracy, interpretability, and robustness. It is important to remark that correct Spike Sorting in data like EPT has served to control and treat patients with epilepsy. Spike Sorting is also essential to study connectivity patterns between neurons and the development of brain-machine interfaces.

This thesis significantly extended the FMM methodology and proved its potential to solve challenging problems in neuroscience. Its conclusion discloses many highly compelling lines of future work. On the one hand, many other interesting problems in neuroscience could be addressed, such as the study of the neuronal activities synchronization (Aydore, Pantazis, and Leahy, 2013) and the analysis of multichannel neuronal recordings (Yger et al., 2018). In the former problem, measurements of the interaction between neuronal populations are used to uncover insights into the functioning of the nervous system. In the latter problem, specific procedures can be defined by either combining the FMM-extracted information of each channel or with multivariate FMM models, as those introduced in Rueda et al., 2022 to analyze ECG multi-lead signals. The FMM methodology could be used to handle many other challenges that arise from the analysis of the electroencephalogram, which records the firing of the brain neurons. In addition, many other problems associated with oscillatory signals of different fields can be approached with the FMM methodology, such as visual electrophysiology, electrocardiography, experimental chemistry, among others. We believe that the methodology can have significant implications in human health, improving the detection, treatment, and control of diseases like epilepsy, schizophrenia, or dementia. On the other hand, there still is much methodological work to develop, specially inferential procedures. Furthermore, the work in Rodríguez-Collado and Rueda, 2022 opens up the possibility of defining novel functional data analysis approaches, while the resolution of many applied problems will clearly require addressing new methodological questions. All of these entail the continuous development of the **FMM** package. Finally, we have the challenge of reducing the computational time of the estimation algorithm. It could be sped up by implementing the parameter search in a more computationally efficient programming language, such as C, and using GPU-based computing.

2. Publications

This section provides the main bibliographic data of the articles that are part of this thesis. Specifically, it consists of five scientific publications: three published, one in-press, and another under review.

The contributions are listed according to publication date.

2.1 Rueda, Rodríguez-Collado, and Larriba (2021)

A novel wave decomposition for oscillatory signals (Rueda, Rodríguez-Collado, and Larriba, 2021).
Cristina Rueda, **Alejandro Rodríguez-Collado**, and Yolanda Larriba. A novel wave decomposition for oscillatory signals. *IEEE Transactions on Signal Processing*, vol. 69, pages 960–972, 2021. DOI: 10.1109/TSP.2021.3051428.

Impact factor (2020): 4.931. Q1 (41/273) in "Engineering, electrical & electronic".

Abstract: Oscillatory systems arise in the different science fields. Complex mathematical formulations with differential equations have been proposed to model the dynamics of these systems. While they have the advantage of having a direct physiological meaning, they are not useful in practice as a result of the parameter adjustment complexity and the presence of noise. In this paper, a novel decomposition approach in AM-FM components, competing with Fourier and other decompositions is presented. Several interesting theoretical properties are derived including the ordinary differential equations describing the signal. Furthermore, the usefulness in real practice is demonstrated to analyse signals associated to neuron synapses and by addressing other questions in neuroscience.

2.2 Rodríguez-Collado and Rueda (2021a)

A simple parametric representation of the Hodgkin-Huxley model (Rodríguez-Collado and Rueda, 2021a).
Alejandro Rodríguez-Collado and Cristina Rueda. A simple parametric representation of the Hodgkin-Huxley model. *PLOS ONE*, vol. 16, number 7, page 1-19, 2021. DOI: 10.1371/journal.pone.0254152.

Impact factor (2020): 3.240. Q2 (26/72) in "Multidisciplinary sciences".

Abstract: The Hodgkin-Huxley model, decades after its first presentation, is still a reference model in neuroscience as it has successfully reproduced the electrophysiological activity of many organisms. The primary signal in the model represents the membrane potential of a neuron. A simple representation of this signal is presented in this paper. The new proposal is an adapted Frequency Modulated Möbius multicomponent model defined as a signal plus error model in which the signal is decomposed as a sum of waves. The main strengths of the method are the simple parametric formulation, the interpretability and flexibility of the parameters that describe and discriminate the waveforms, the estimators' identifiability and accuracy, and the robustness against noise. The approach is validated with a broad simulation experiment of Hodgkin-Huxley signals and real data from squid giant axons. Interesting differences between simulated and real data emerge from the comparison of the parameter configurations. Furthermore, the potential of the FMM parameters to predict Hodgkin-Huxley model parameters is shown using different Machine Learning methods. Finally, promising contributions of the approach in Spike Sorting and cell-type classification are detailed.

2.3 Rodríguez-Collado and Rueda (2021b)

Electrophysiological and transcriptomic features reveal a circular taxonomy of cortical neurons (Rodríguez-Collado and Rueda, 2021b).
Alejandro Rodríguez-Collado and Cristina Rueda. Electrophysiological and transcriptomic features reveal a circular taxonomy of cortical neurons. *Frontiers in Human Neuroscience*, vol. 15, page 684950, 2021. DOI: 10.3389/fnhum.2021.684950.

Impact factor (2020): 3.169. Q2 (27/77) in "Psychology".

Abstract: The complete understanding of the mammalian brain requires exact knowledge of the function of each neuron subpopulation composing its parts. To achieve this goal, an exhaustive, precise, reproducible, and robust neuronal taxonomy should be defined. In this paper, a new circular taxonomy based on transcriptomic features and novel electrophysiological features is proposed. The approach is validated by analysing more than 1850 electrophysiological signals of different mouse visual cortex neurons proceeding from the Allen Cell Types database. The study is conducted on two different levels: neurons and their cell-type aggregation into Cre lines. At the neuronal level, electrophysiological features have been extracted with a promising model that has already proved its worth in neuronal dynamics. At the Cre line level, electrophysiological and transcriptomic features are joined on cell types with available genetic information. A taxonomy with a circular order is revealed by a simple transformation of the first two principal components that allow the characterization of the different Cre lines. Moreover, the proposed methodology locates other Cre lines in the taxonomy that do not have transcriptomic features available. Finally, the taxonomy is validated by Machine Learning methods which are able to discriminate the different neuron types with the proposed electrophysiological features.

2.4 Fernández et al. (2022)

FMM: an R package for modeling rhythmic patterns in oscillatory systems (Fernández et al., 2022).
Itziar Fernández, **Alejandro Rodríguez-Collado**, Yolanda Larriba, Adrián Lamela, Christian Canedo, Cristina Rueda. FMM: an R package for modeling rhythmic patterns in oscillatory systems. *The R Journal*, in-press, 2022.

Impact factor (2020): 3.984. D1 (12/125) in "Statistics & probability".

Abstract: This paper is dedicated to the R package FMM which implements a novel approach to describe rhythmic patterns in oscillatory signals. The frequency modulated Möbius (FMM) model is defined as a parametric signal plus a gaussian noise, where the signal can be described as a single or a sum of waves. The FMM approach is flexible enough to describe a great variety of rhythmic patterns. The FMM package includes all required functions to fit and explore single and multi-wave FMM models, as well as a restricted version that allows equality constraints between parameters representing a priori knowledge about the shape to be included. Moreover, the FMM package can generate synthetic data and visualize the results of the fitting process. The potential of this methodology is illustrated with examples of such biological oscillations as the circadian rhythm in gene expression, the electrical activity of the heartbeat and neuronal activity.

2.5 Rodríguez-Collado and Rueda (2022)

Functional clustering of neuronal signals with FMM mixture models (Rodríguez-Collado and Rueda, 2022).
Alejandro Rodríguez-Collado and Cristina Rueda. Functional clustering of neuronal signals with FMM mixture models. *Preprint*, available in Arxiv. https://arxiv.org/abs/2203.03588. Submitted, 2022.

Abstract: The identification of unlabelled neuronal electric signals is one of the most challenging open problems in neuroscience, widely known as Spike Sorting. Motivated to solve this problem, we propose a model-based approach within the mixture modeling framework for clustering oscillatory functional data called MixFMM. The core of the approach are the FMM signals, non-linear additive parametric time functions. These functions are flexible enough to describe different oscillatory patterns and simple enough to be estimated efficiently. Concretely, specific model parameters describe the waveforms' phase, amplitude, and shape. A mixture of FMM models is defined using gaussian errors, and an EM algorithm is proposed for parameters' estimation. In addition, the approach includes methods for the number of clusters selection. Spike Sorting has received considerable attention in the literature, and different functional clustering approaches have traditionally been considered. We compare those approaches with the MixFMM in a broad collection of datasets, including benchmarking simulated and real data. The results are evaluated using a selection of indexes, quantifying within-group cohesion and inter-group separation. These indexes confirm the outstanding performance of the MixFMM across all datasets, and significant improvements against competitors are attained in certain scenarios.

Bibliography

L. F. Abbott and T. B. Kepler (1990). "Model neurons: from Hodgkin-Huxley to Hopfield". In: *Statistical Mechanics of Neural Networks*, pp. 5–18.

T. Aksenova, O. Chibirova, O. Dryga, I. Tetko, A. Benabid, and A. Villa (2003). "An unsupervised automatic method for sorting neuronal spike waveforms in awake and freely moving animals". In: *Methods* 30, pp. 178–187.

Allen Institute for Brain Science (2021). "Allen cell types database". Available in https://celltypes.brain-map.org/data.

P. Ashwin, S. Coombes, and R. Nicks (2016). "Mathematical frameworks for oscillatory network dynamics in neuroscience". In: *The Journal of Mathematical Neuroscience* 6, p. 2.

S. Aydore, D. Pantazis, and R. Leahy (2013). "A note on the phase locking value and its properties". In: *Neuroimage* 74, pp. 231–244.

R. Barlow, D. Bartholomew, J. Bremner, and H. D. Brunk (1972). "Statistical inference under order restrictions: the theory and application of isotonic regression". First edition. Wiley New York.

L. Birgé and P. Massart (2007). "Minimal penalties for Gaussian model selection". In: *Probability Theory and Related Fields* 138, pp. 33–73.

B. Boashash (2016). "Time-frequency signal analysis and processing: a comprehensive reference". Second edition. Elsevier Science.

C. Bouveyron, E. Côme, and J. Jacques (2015). "The discriminative functional mixture model for a comparative analysis of bike sharing systems". In: *The Annals of Applied Statistics* 9.4, pp. 1726–1760.

H. Brunk (1955). "Maximum likelihood estimates of monotone parameters". In: *The Annals of Mathematical Statistics* 26.4, pp. 607–616.

B. W. Brunton and M. Beyeler (2019). "Data-driven models in human neuroscience and neuroengineering". In: *Current Opinion in Neurobiology* 58, pp. 21 –29.

S. L. Brunton, J. L. Proctor, and J. N. Kutz (2016). "Discovering governing equations from data by sparse identification of nonlinear dynamical systems". In: *Proceedings of the National Academy of Sciences* 113.15, pp. 3932–3937.

A. Buja, T. Hastie, and R. Tibshirani (1989). "Linear smoothers and additive models". In: *The Annals of Statistics* 17, pp. 453 –510.

G. Buzsáki (2004). "Large-scale recording of neuronal ensembles". In: *Nature Neuroscience* 7, pp. 446–51.

C. R. Caro-Martín, J. M. Delgado-García, A. Gruart, and R. Sánchez-Campusano (2018). "Spike sorting based on shape, phase, and distribution features, and K-TOPS clustering with validity and error indices". In: *Scientific Reports* 8.1, pp. 1–28.

F. Chamroukhi and H. D. Nguyen (2019). "Model-based clustering and classification of functional data". In: *WIREs Data Mining and Knowledge Discovery* 9.4, e1298.

W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson (2020). "Shiny: web application framework for R". R package version 1.5.0. URL: https://CRAN.R-project.org/package=shiny.

W. Chang, J. Luraschi, and T. Mastny (2020). "profvis: interactive visualizations for profiling R code". R package version 0.3.7. URL: https://CRAN.R-project.org/package=profvis.

G. Claeskens, E. Devijver, and I. Gijbels (2021). "Nonlinear mixed effects modeling and warping for functional data using B-splines". In: *Electronic Journal of Statistics* 15.2, pp. 5245 –5282.

A. P. Dempster, N. Laird, and D. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–22.

Z. Deng, S. Arsenault, L. Mao, and J. Arnold (2016). "Measuring synchronization of stochastic oscillators in biology". In: *Journal of Physics: Conference Series*. Vol. 750. IOP Publishing, p. 012001.

F. Donato, S. Rompani, and P. Caroni (2013). "Parvalbumin-expressing basket-cell network plasticity induced by experience regulates adult learning". In: *Nature* 504, pp. 272–276.

T. D. Downs and K. V. Mardia (2002). "Circular regression". In: *Biometrika* 89.3, pp. 683–697.

C. Ekanadham, D. Tranchina, and E. P. Simoncelli (2014). "A unified framework and method for automatic neural spike identification". In: *Journal of Neuroscience Methods* 222, pp. 47–55.

M. Fernandez-Delgado, E. Cernadas, S. Barro, and D. Amorim (2014). "Do we need hundreds of classifiers to solve real world classification problems?" In: *Journal of Machine Learning Research* 15, pp. 3133–3181.

I. Fernández, A. Rodríguez-Collado, Y. Larriba, A. Lamela, C. Canedo, and C. Rueda (2022). "FMM: an R package for modeling rhythmic patterns in oscillatory systems". In: *The R Journal*. In-press.

R. Fitzhugh (1961). "Impulses and physiological states in theoretical models of nerve membrane." In: *Biophysical journal* 1 6, pp. 445–66.

D. Gabor (1946). "Theory of communication. Part 1: the analysis of information". In: *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering* 93.26, pp. 429–441.

M. L. L. García, R. García-Ródenas, and A. G. Gómez (2015). "K-means algorithms for functional data". In: *Neurocomputing* 151, pp. 231–245.

P. Ghaderi, H. Marateb, and M.-S. Safari (2018). "Electrophysiological profiling of neocortical neural subtypes: a semi-supervised method applied to in vivo whole-cell patch-clamp data". In: *Frontiers in Neuroscience* 12.

N. Gouwens, S. Sorensen, J. Berg, C. Lee, T. Jarsky, J. Ting, S. Sunkin, D. Feng, C. Anastassiou, E. Barkan, K. Bickley, N. Blesie, T. Braun, K. Brouner, A. Budzillo, S. Caldejon, T. Casper, D. Castelli, P. Chong, and C. Koch (2019). "Classification of electrophysiological and morphological neuron types in the mouse visual cortex". In: *Nature Neuroscience* 22, 1182–1195.

N. W. Gouwens, S. A. Sorensen, F. Baftizadeh, A. Budzillo, B. R. Lee, T. Jarsky, L. Alfiler, K. Baker, E. Barkan, K. Berry, D. Bertagnolli, K. Bickley, J. Bomben, T. Braun, K. Brouner, T. Casper, K. Crichton, T. L. Daigle, R. Dalley, R. A. de Frates, N. Dee, T. Desta, S. D. Lee, N. Dotson, T. Egdorf, L. Ellingwood, R. Enstrom, L. Esposito, C. Farrell, D. Feng, O. Fong, R. Gala, C. Gamlin, A. Gary, A. Glandon, J. Goldy, M. Gorham, L. Graybuck, H. Gu, K. Hadley, M. J. Hawrylycz, A. M. Henry, D. Hill, M. Hupp, S. Kebede, T. K. Kim, L. Kim, M. Kroll, C. Lee, K. E. Link, M. Mallory, R. Mann, M. Maxwell, M. McGraw, D. McMillen, A. Mukora, L. Ng, L. Ng, K. Ngo, P. R. Nicovich, A. Oldre, D. Park, H. Peng, O. Penn, T. Pham, A. Pom, Z. Popović, L. Potekhina, R. Rajanbabu, S. Ransford, D. Reid, C. Rimorin, M. Robertson, K. Ronellenfitch, A. Ruiz, D. Sandman, K. Smith, J. Sulc, S. M. Sunkin, A. Szafer, M. Tieu, A. Torkelson, J. Trinh, H. Tung, W. Wakeman, K. Ward, G. Williams, Z. Zhou, J. T. Ting, A. Arkhipov, U. Sümbül, E. S. Lein, C. Koch, Z. Yao, B. Tasic, J. Berg, G. J. Murphy, and H. Zeng (2020). "Integrated morphoelectric and transcriptomic classification of cortical GABAergic cells". In: *Cell* 183.4, pp. 935–953.

T. Hastie, R. Tibshirani, and F. Jerome (2009). "The elements of statistical learning: data mining, inference, and prediction". Second Edition. Springer.

A. Hodgkin and A. Huxley (1952). "A quantitative description of membrane current and its application to conduction and excitation in nerve". In: *The Journal of Physiology* 117.4, pp. 500–544.

E. M. Izhikevich (2003). "Simple model of spiking neurons". In: *IEEE Transactions on Neural Networks* 14.6, pp. 1569–1572.

S. H. Kang, W. Liao, and Y. Liu (2021). "Ident: identifying differential equations with numerical time evolution". In: *Journal of Scientific Computing* 87.1, pp. 1–27.

S. Kato, K. Shimizu, and G. S. Shieh (2008). "A circular-circular regression". In: *Statistica Sinica* 18.2, pp. 633–645.

M. Kowalski, A. Meynard, and H.-T. Wu (2018). "Convex optimization approach to signals with fast varying instantaneous frequency". In: *Applied and Computational Harmonic Analysis* 44.1, pp. 89–122.

Y. Larriba, C. Rueda, M. Fernández, and S. D. Peddada (2020). "Order restricted inference in chronobiology". In: *Statistics in Medicine* 39.3, pp. 265–278.

J. Lecoq, M. Oliver, J. H. Siegle, N. Orlova, and C. Koch (2021). "Removing independent noise in systems neuroscience data using DeepInterpolation." In: *Nature Methods* 18, 1401–1408.

M. K. Lewandowska, D. J. Bakkum, S. B. Rompani, and A. Hierlemann (2015). "Recording large extracellular spikes in microchannels along many axonal sites from individual neurons". In: *PLOS ONE* 10.3, pp. 1–24.

C. Lin, L. Su, and H. Wu (2018). "Wave-shape function analysis". In: *Journal of Fourier Analysis and Applications* 24.2, pp. 451–505.

S. Marom (2016). "Emergence and maintenance of excitability: kinetics over structure". In: *Current Opinion in Neurobiology* 40, pp. 66 –71.

J. S. Marron, J. Ramsay, L. M. Sangalli, and A. Srivastava (2015). "Functional data analysis of amplitude and phase variation". In: *Statistical Science* 30.4, pp. 468–484.

J. Menéndez and B. Salvador (1991). "Anomalies of the likelihood ratio tests for testing restricted hypothesis". In: *The Annals of Statistics* 19.2, pp. 889–898.

O. Mersmann (2019). "microbenchmark: accurate timing functions". R package version 1.4-7. URL: https://CRAN.R-project.org/package=microbenchmark.

Microsoft Corporation and S. Weston (2020a). "doParallel: foreach parallel adaptor for the 'parallel' Package". R package version 1.0.16. URL: https://CRAN.R-project.org/package=doParallel.

Microsoft Corporation and S. Weston (2020b). "foreach: provides foreach looping construct". R package version 1.5.1. URL: https://CRAN.R-project.org/package=foreach.

M. Moghaddasi, M. Aliyari Shoorehdeli, Z. Fatahi, and A. Haghparast (2020). "Unsupervised automatic online spike sorting using reward-based online clustering". In: *Biomedical Signal Processing and Control* 56, p. 101701.

J. Nelder and R. Mead (1965). "A simplex method for function minimization". In: *The Computer Journal* 7.4, pp. 308–313.

S. Oprisan (2017). "A consistent definition of phase resetting using Hilbert transform". In: *International Scholarly Research Notices* 2017.

H. Ori, E. Marder, and S. Marom (2018). "Cellular function given parametric variation in the Hodgkin and Huxley model of excitability". In: *Proceedings of the National Academy of Sciences* 115.35, E8211–E8218.

J. Park and J. Ahn (2017). "Clustering multivariate functional data with phase variation". In: *Biometrics* 73.1, pp. 324–333.

D. Paydarfar, D. B. Forger, and J. R. Clay (2006). "Noisy inputs and the induction of on–off switching behavior in a neuronal pacemaker". In: *Journal of Neurophysiology* 96.6, pp. 3338–3348.

A. Pikovsky and M. Rosenblum (2015). "Dynamics of globally coupled oscillators: progress and perspectives". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 25.9, p. 097616.

R. Q. Quiroga (2009). "Simulated datasets". Available in https://leicester.figshare.com/articles/dataset/Simulated_dataset/11897595.

R. Q. Quiroga (2019). "Human single-cell recording". Available in https://leicester.figshare.com/articles/Dataset_Human_single-cell_recording/11302427/1.

R. Q. Quiroga, Z. Nadasdy, and Y. Ben-Shaul (2004). "Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering". In: *Neural Computation* 16.8, pp. 1661–1687.

R Core Team (2020). "R: a language and environment for statistical computing". R Foundation for Statistical Computing.

R Core Team (2021). "Writing R extensions". Available in https://cran.r-project.org/doc/manuals/R-exts.html.

M. Raghavan, D. Fee, and P. Barkhaus (2019). "Generation and propagation of the action potential". In: *Handbook of clinical neurology*. Vol. 160. Elsevier, pp. 3–22.

J. Ramsay and B. W. Silverman (2005). "Functional data analysis". Second edition. Springer Series in Statistics.

H. G. Rey, C. Pedreira, and R. Q. Quiroga (2015). "Past, present and future of spike sorting techniques". In: *Brain Research Bulletin* 119, pp. 106–117.

A. Rodríguez-Collado and C. Rueda (2021a). "A simple parametric representation of the Hodgkin-Huxley model". In: *PLOS ONE* 16.7, pp. 1–19.

A. Rodríguez-Collado and C. Rueda (2021b). "Electrophysiological and transcriptomic features reveal a circular taxonomy of cortical neurons". In: *Frontiers in Human Neuroscience* 15, p. 684950.

A. Rodríguez-Collado and C. Rueda (2022). "Functional clustering of neuronal signals with FMM mixture models". Preprint. Available in https://arxiv.org/abs/2203.03588.

C. Rueda, M. Ugarte, and A. Militino (2016). "Checking unimodality using isotonic regression: an application to breast cancer mortality rates". In: *Stochastic Environmental Research and Risk Assessment* 30.4, pp. 1277–1288.

C. Rueda, Y. Larriba, and S. D. Peddada (2019). "Frequency modulated Möbius model accurately predicts rhythmic signals in biological and physical sciences". In: *Scientific Reports* 9.1, pp. 1–10.

C. Rueda, A. Rodríguez-Collado, I. Fernández, C. Canedo, M. D. Ugarte, and Y. Larriba (2022). "A unique cardiac electrophysiologic model". Preprint. Available in https://arxiv.org/abs/2202.03938.

C. Rueda, A. Rodríguez-Collado, and Y. Larriba (2021). "A novel wave decomposition for oscillatory signals". In: *IEEE Transactions on Signal Processing* 69, pp. 960–972.

S. Sandoval and P. de Leon (2018). "The instantaneous spectrum: a general framework for time-frequency analysis". In: *IEEE Transactions on Signal Processing* 66.21, pp. 5679–5693.

J. Sebastian, M. Sur, H. A. Murthy, and M. Magimai-Doss (2021). "Signal-to-signal neural networks for improved spike estimation from calcium imaging data". In: *PLOS Computational Biology* 17.3, pp. 1–19.

N. K. Sharma, C. Pedreira, M. Centeno, U. J. Chaudhary, T. Wehner, L. G. França, T. Yadee, T. Murta, M. Leite, S. B. Vos, S. Ourselin, B. Diehl, and L. Lemieux (2017). "A novel scheme for the validation of an automated classification method for epileptic spikes by comparison with multiple observers". In: *Clinical Neurophysiology* 128.7, pp. 1246–254.

J. Silva (2014). "Slow inactivation of Na+ channels". In: *Voltage Gated Sodium Channels*, pp. 33–49.

B. Souza, V. Lopes dos Santos, J. Bacelo, and A. Tort (2019). "Spike sorting with Gaussian mixture models". In: *Scientific Reports* 9, p. 3627.

B. Tasic, V. Menon, T. N. Nguyen, S. Kim, T. Jarsky, Z. Yao, B. Levi, L. Gray, S. Sorensen, T. Dolbeare, D. Bertagnolli, J. Goldy, N. Shapovalova, S. Parry, C. Lee, K. Smith, A. Bernard, L. Madisen, S. Sunkin, and H. Zeng (2016). "Adult mouse cortical cell taxonomy revealed by single cell transcriptomics". In: *Nature Neuroscience* 19, 335–346.

C. Teeter, R. Iyer, V. Menon, N. Gouwens, D. Feng, J. Berg, A. Szafer, N. Cain, H. Zeng, M. Hawrylycz, C. Koch, and S. Mihalas (2018). "Generalized leaky integrate-and-fire models classify multiple neuron types". In: *Nature Communications* 9.1, pp. 1–15.

C. Trainito, C. von Nicolai, E. Miller, and M. Siegel (2019). "Extracellular spike waveform dissociates four functionally distinct cell classes in primate cortex". In: *Current Biology* 29.18, pp. 2973–2982.

R. Tremblay, S. Lee, and B. Rudy (2016). "GABAergic interneurons in the neocortex: from cellular properties to circuits". In: *Neuron* 91.2, pp. 260–292.

R. Veerabhadrappa, M. Ul Hassan, J. Zhang, and A. Bhatti (2020). "Compatibility evaluation of clustering algorithms for contemporary extracellular neural spike sorting". In: *Frontiers in Systems Neuroscience* 14, p. 34.

P. K. Wang, S. H. Pun, C. H. Chen, E. A. McCullagh, A. Klug, A. Li, M. I. Vai, P. U. Mak, and T. C. Lei (2019). "Low-latency single channel real-time neural spike sorting system based on template matching". In: *PLOS ONE* 14.11, pp. 1–30.

H. Wickham (2011). "testthat: get started with testing". In: *The R Journal* 3, pp. 5–10.

H. Wickham (2016). "ggplot2: elegant graphics for data analysis". Second Edition. Springer - Verlag New York.

H. Wickham and J. Bryan (2019). "R packages". Second edition. O'Reilly.

H. Wickham, P. Danenberg, G. Csárdi, and M. Eugster (2020). "roxygen2: in-line documentation for R". R package version 7.1.1. URL: https://CRAN.R-project.org/package=roxygen2.

T. Wigren (2015). "Model order and identifiability of non-linear biological systems in stable oscillation". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 12.6, pp. 1479–1484.

T. Wigren and T. Söderström (2005). "A second order ODE is sufficient for modelling of many periodic signals". In: *International Journal of Control* 78.13, pp. 982–996.

A. Winfree (2001). "The geometry of biological time". First edition. Vol. 12. Springer Science & Business Media.

J. Ye (1998). "On measuring and correcting the effects of data mining and model selection". In: *Journal of The American Statistical Association* 93, pp. 120–131.

P. Yger, G. L. Spampinato, E. Esposito, B. Lefebvre, S. Deny, C. Gardella, M. Stimberg, F. Jetter, G. Zeck, S. Picaud, J. Duebel, and O. Marre (2018). "A spike sorting toolbox for up to thousands of electrodes validated with ground truth recordings in vitro and in vivo". In: *eLife* 7, e34518.

H. Zeng and J. Sanes (2017). "Neuronal cell-type classification: challenges, opportunities and the path forward". In: *Nature Reviews Neuroscience* 18, pp. 530 –546.

C. Zhang, C. Liu, X. Zhang, and G. Almpanidis (2017). "An up-to-date comparison of state-of-the-art classification algorithms". In: *Expert Systems with Applications* 82, pp. 128–150.