Temporal point processes and scalable convolutional dictionary learning : a unified framework for m/eeg signal analysis in neuroscience

Cédric Allain

université
PARIS-SACLAY

Temporal Point Processes and Scalable Convolutional Dictionary Learning: A Unified Framework for M/EEG Signal Analysis in Neuroscience

Processus ponctuels temporels et apprentissage scalable de dictionnaires convolutionnels : un cadre unifié pour l'analyse des signaux M/EEG en neurosciences

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°580 : sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat: Informatique mathématique
Graduate School : Informatique et sciences du numérique
Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **Inria Saclay-Île-de-France** (Université Paris-Saclay, Inria), sous la direction d'**Alexandre Gramfort**, Directeur de Recherche, et le co-encadrement de **Thomas Moreau**, Chargé de Recherche

Thèse soutenue à Paris-Saclay, le 9 février 2024, par

Cédric Allain

Composition du jury

Membres du jury avec voix délibérative

Laurent Oudre Professeur, Université Paris Saclay	Président
Vincent Rivoirard Professeur, Paris Sorbonne Université	Rapporteur & Examinateur
Romain Tavenard Professeur, Université de Rennes 2	Rapporteur & Examinateur
Patricia Reynaud-Bouret Directrice de recherche, Université Côte d'Azur	Examinatrice

THESE DE DOCTORAT
NNT : 2024UPSAG008

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Processus ponctuels temporels et apprentissage scalable de dictionnaires convolutionnels : un cadre unifié pour l'analyse des signaux M/EEG en neurosciences

Mots-clefs : Apprentissage de dictionnaire convolutif (CDL) ; Processus ponctuels temporels (TPP) ; Inférence par noyau ; Neurosciences computationnelles ; M/EEG

Résumé : Dans le domaine de l'imagerie cérébrale non invasive, la magnéto- et l'électroencéphalographie (M/EEG) offrent un précieux aperçu des activités neuronales. Les données enregistrées consistent en des séries temporelles multivariées qui fournissent des informations sur les processus cognitifs et sont souvent complétées par des détails auxiliaires liés au paradigme expérimental, tels que l'horodatage des stimuli externes ou des actions entreprises par les sujets. En outre, l'ensemble des données peut inclure des enregistrements de plusieurs sujets, ce qui facilite les analyses en population.

Cette thèse de doctorat présente un nouveau cadre pour l'analyse des signaux M/EEG qui synergise l'Apprentissage Convolutif de Dictionnaire (CDL) et les Processus Ponctuels Temporels (TPP). Ce travail est divisé en deux composantes principales : les avancées en modélisation temporelle et le passage à l'échelle computationnelle. En matière de modélisation temporelle, deux nouveaux modèles de processus ponctuels sont introduits, accompagnés de méthodes d'inférence efficaces pour capturer les activités neuronales liées aux tâches. La méthode proposée d'Inférence Discrétisée Rapide pour les Processus de Hawkes (FaDIn) a également des implications pour des applications plus larges. De plus, ce travail aborde les défis computationnels de l'analyse des données M/EEG à grande échelle basée sur le CDL, en introduisant un nouvel algorithme robuste de CDL avec fenêtrage stochastique. Cet algorithme permet de traiter efficacement les signaux entachés d'artefacts ainsi que les études de population à grande échelle. Le CDL populationnelle a ensuite été utilisé sur le grand ensemble de données en libre accès Cam-CAN, révélant des aspects de l'activité neuronale liée à l'âge.

Title: Temporal Point Processes and Scalable Convolutional Dictionary Learning: A Unified Framework for M/EEG Signal Analysis in Neuroscience

Keywords: Convolutional Dictionary Learning (CDL); Temporal Point Processes (TPPs); Kernel Inference; Computational Neuroscience; M/EEG

Abstract: In the field of non-invasive brain imaging, Magnetoencephalography and Electroencephalography (M/EEG) offer invaluable insights into neural activities. The recorded data consist of multivariate time series that provide information about cognitive processes and are often complemented by auxiliary details related to the experimental paradigm, such as timestamps of external stimuli or actions undertaken by the subjects. Additionally, the dataset may include recordings from multiple subjects, facilitating population-level analyses.

This doctoral research presents a novel framework for M/EEG signal analysis that synergizes Convolutional Dictionary Learning (CDL) and Temporal Point Processes (TPPs). The work is segmented into two primary components: temporal modeling advancements and computational scalability. For temporal modeling, two novel point process models are introduced with efficient inference methods to capture task-specific neural activities. The proposed Fast Discretized Inference for Hawkes Processes (FaDIn) method also has implications for broader applications. Additionally, this work addresses the computational challenges of large-scale M/EEG data CDL-based analysis, by introducing a novel Stochastic Robust Windowing CDL algorithm. This algorithm allows to process efficiently artifact-ridden signals as well as large population studies. Population CDL was then used on the large open-access dataset Cam-CAN, shedding light on age-related neural activity.

Remerciements

Avant toute chose, je tiens à exprimer mes sincères remerciements à mes directeurs de thèse, Alexandre Gramfort et Thomas Moreau. Votre soutien ininterrompu a été un pilier essentiel tout au long de cette aventure académique. Je vous suis en premier lieu reconnaissant d'avoir accepté de me prendre en stage au pied levé, suite à des changements de dernière minute si caractéristiques de la période Covid. Votre proposition de poursuivre en thèse qui s'en ait suivi a été pour moi une opportunité inestimable. Trois ans c'est long, et c'est sans aucun doute grâce à votre confiance, vos encouragements et votre grande humanité que j'ai pu surmonter mes moments de doute et continuer à avancer sur le projet. Vos conseils, votre pédagogie et votre rigueur m'ont fait progresser, tant d'un point de vue technique que méthodologique, et promis, je ferai toujours en sorte de m'assurer que la trottinette fonctionne avant d'envisager le 4x4. Merci encore pour toutes les connaissances que vous avez su me transmettre, notamment en neuroscience, domaine passionnant qui m'était totalement inconnu. La fabrique du savoir peut paraître lointaine, et je vous remercie une ultime fois de m'avoir fait découvrir ce monde particulier qu'est celui de la recherche. Je ressors de cette aventure grandi. En toute sincérité, je n'aurai pu envisager meilleur encadrement.

Je souhaite également adresser mes chaleureux remerciements aux membres du jury. Je tiens à exprimer ma gratitude envers les rapporteurs Vincent Rivoirard et Romain Tavenard pour leur lecture minutieuse de mon manuscrit et pour leurs retours très précis visant à en améliorer la qualité. Merci également à Laurent Oudre et à Patricia Reynaud-Bouret d'avoir accepté de faire partie de ce jury. Ce fut un honneur de vous présenter mes travaux et de répondre à vos questions.

Merci à toutes les personnes qui ont contribué de près ou de loin à cette thèse. In particular, I extend my heartfelt thanks to Lindsey Power and Timothy Bardouille from Dalhousie University (Halifax, Canada), for our year-long collaboration. I have gained invaluable insights into neuroscience research through our interactions, and I am now even more convinced of the potential for significant outcomes arising from the integration of our respective disciplines.

Bien entendu, je tiens à remercier les membres de l'équipe Parietal – désormais Mind/Soda – de l'Inria Saclay : qu'est-ce que ça aurait été sans vous ! Merci à Benoît d'avoir partagé cette aventure avec moi depuis les début en pleine période Covid, à deux dans un "bureau des stagiaires" – qui n'en ait désormais plus un – bien vide, merci aussi à Apolline pour les pauses jeux de société, à Théo pour m'avoir accompagné aux manifs, to Maria, thank you for spending the summer with me in the library while I was writing my manuscript[1], a Gabriela, con quien pude practicar mi español, merci bien bien sûr à tous les autres, Julia, Alexis[2], Alexandre[4], Florent, Thomas, Mathieu, Guillaume, Antoine, Louis, Raphaël, Ambroise, Matthieu, Bénédicte, Omar, Charlotte, Cédric, Lilian, Samuel, Julie, Félix, Léo, Jade, Virginie, Pierre-Louis, Nicolas, et tous les autres que j'oublie. Les escapades ponctuelles lors de conférences, les retraites, et les nombreux verres sur la montagne Sainte-Geneviève ont été des moments que j'ai particulièrement apprécié partager avec vous et qui n'ont fait qu'améliorer l'ambiance générale du labo. Merci aussi aux PI de l'équipe, Philippe Ciuciu, Bertrand Thirion, Gaël Varoquaux, Marine Le Morvan, Judith Abecassis et Demian Wassermann, de faire de cette équipe ce qu'elle est.

Merci surtout à mes ami.e.s qui m'accompagnent depuis de nombreuses années et qui ont été, et sont toujours, d'un soutien inestimable. Merci à Fanny pour cette amitié qui dure depuis presque deux décennies, Agatha et Claire, d'avoir ouvert la voie de la thèse. Merci aussi à mes amies de mes années universitaires toulousaines, Marilou, Léa, Clara, Margot et Molly, pour cette merveilleuse amitié malgré la distance qui s'installe progressivement (promis j'arrête de travailler pendant nos vacances !). Merci encore à mes amis de l'Ensae, Shu et Suzanne, avec qui j'ai pu vivre l'aventure de la thèse, mais également Alexis, Valentin, Clotilde, Juliette, Édith, Idriss et Aurélien. Merci aussi à Manon, je te souhaite tout mon courage pour ta fin de thèse.

Maman, Papa, Estelle, merci infiniment de m'avoir accompagné tout au long de ce – très – long parcours scolaire. C'est grâce à vous que je suis la personne que je suis aujourd'hui. Merci de m'avoir transmis le goût des sciences et de m'avoir guidé et encouragé dans tout ce que j'ai pu entreprendre. Merci.

Miguel, merci d'être chaque jour à mes côtés.

[1] I hope that my few attempts as a French teacher have helped you understand our (complicated) language, even if I often do not understand the rules myself.

Résumé en français des travaux de thèse

Les neurosciences forment un domaine interdisciplinaire où convergent la biologie, la psychologie et les mathématiques, cherchant à démystifier les mécanismes du cerveau et du système nerveux. Une composante centrale consiste en l'analyse des activités neuronales, capturées à l'aide de techniques sophistiquées telles que la magnétoencéphalographie (MEG) et l'électroencéphalographie (EEG). Ces méthodes non invasives exploitent respectivement les principes physiques des champs magnétiques et électriques pour fournir un aperçu des processus dynamiques du cerveau. Typiquement, plusieurs dizaines voire centaines de capteurs sont disposés autour de la tête d'un individu, permettant l'enregistrement des fluctuations du champ électrique ou magnétique à une résolution temporelle élevée, généralement de l'ordre de 1000 Hz, sur des périodes allant de quelques minutes à plusieurs heures.

Les données M/EEG, caractérisées par leur haute résolution temporelle, sont représentées sous forme de signaux de séries temporelles multivariées – un par capteur –, capturant les modèles complexes d'activité neuronale. Parallèlement, un composant tout aussi essentiel de notre recherche implique l'enregistrement minutieux des événements externes. Ces événements vont des stimuli contrôlés et des actions des participants dans des dispositifs expérimentaux aux interventions cliniques, telles que les injections de médicaments lors de procédures chirurgicales. L'intégration de ces événements avec les données M/EEG fournit un cadre complet pour comprendre l'activité cérébrale dans divers contextes.

La MEG et l'EEG ont joué un rôle crucial dans l'avancement de notre compréhension de divers processus cognitifs et neurologiques. À partir de ces signaux, nous pouvons extraire une multitude d'informations, y compris, mais sans s'y limiter, les oscillations neuronales, les potentiels évoqués – liés aux événements –, et les modèles de connectivité neuronale. Ces aperçus ont des implications profondes tant dans les contextes cliniques que de recherche, offrant des voies vers de nouvelles stratégies thérapeutiques et une compréhension plus approfondie de la fonctionnalité du cerveau.

Dans cette thèse, nous abordons deux défis principaux :

1. **Développement d'un cadre unifié pour l'analyse des dépendances temporelles.** Notre premier défi consiste à utiliser la décomposition en Dictionnaire Convolutionnel (CDL) pour transformer les signaux M/EEG en un flux d'événements. Nous visons à établir un cadre capable d'apprendre les dépendances temporelles entre les motifs neuronaux récurrents – appelés "atomes" – et les stimuli externes, modifiant ainsi la manière dont on modélise l'interaction entre l'activité neuronale et les influences externes.

2. **Extension de la décomposition CDL à l'échelle de populations.** Le second défi est d'élargir et de raffiner le processus de décomposition CDL pour une application à une population plus large. Cet élargissement est double. D'une part, nous visons à améliorer la vitesse et la robustesse du processus CDL, le rendant résistant à une variété de données aberrantes. Ces anomalies peuvent provenir de multiples sources, allant de facteurs endogènes tels que des défaillances de capteurs à des facteurs exogènes comme les mouvements des sujets. D'autre part, notre objectif est de développer une nouvelle méthodologie d'agrégation pour synthétiser les résultats CDL spécifiques à chaque sujet. Cette approche nous permettra de découvrir des effets au niveau de la population, offrant de nouvelles perspectives sur la généralisabilité et la variabilité des motifs neuronaux entre différents individus.

Développement d'un cadre unifié pour l'analyse des dépendances temporelles

Ainsi, le premier défi est de développer un cadre unifié visant à élucider les relations complexes entre les stimuli externes et les réponses neuronales, avec un minimum de traitement des données et d'intervention experte. Au cœur de ce cadre se trouve l'analyse de données de séries temporelles neuronales, riches en formes d'ondes de signaux prototypiques connues sous le nom d'"atomes". Ces atomes invariants par translation, essentiels dans la recherche clinique et cognitive, sont extraits pour comprendre la chronologie et l'occurrence des événements neuronaux. Les méthodes traditionnelles, telles que le "moyennage d'époques" – *epoch averaging* en anglais, c'est-à-dire le moyennage de segments temporels "centrés" autour d'un événement particulier –, échouent souvent à saisir les nuances de ces réponses synchronisées en raison de légères déviations temporelles.

En réponse à cela, cette thèse tire parti de l'apprentissage de dictionnaire convolutifs (CDL ; *Convolutional Dictionary Learning*), spécialement adapté aux principes physiques sous-jacents aux signaux électrophysiologiques, comme décrits par les équations de Maxwell. Le CDL offre une approche efficace et non supervisée pour l'extraction de motifs dans les signaux électrophysiologiques. Le modèle représente les données comme des combinaisons linéaires parcimonieuses – *sparse* – de convolutions entre atomes du dictionnaire et codes invariants par décalage, présentant ainsi une nouvelle représentation basée sur les événements des dynamiques temporelles. Ces atomes sont définis par leurs caractéristiques spatiales et temporelles et peuvent correspondre à diverses activités physiologiques – par exemple les battements de cœur ou les clignements d'yeux – ou à des réponses neuronales à des stimuli externes tels que des indices auditifs ou visuels.

Les processus ponctuels temporels (TPP ; *Temporal point processes*) fournissent un cadre statistique idéalement adapté pour modéliser ces activations d'événements discrets. Historiquement utilisés en neurosciences pour modéliser les enregistrements monocellulaires et les trains de spikes neuronaux – *neural spike trains* –, ces processus n'ont cependant pas directement abordé l'interaction entre les déclenchements de stimuli déterministes et les activations aléatoires d'atomes.

Pour combler cette lacune, cette thèse introduit deux modèles : les processus ponctuels temporels dirigés (DriPP ; *Driven Temporal Point Processes*) et l'inférence discrétisée rapide (FaDIn ; *Fast Discretized Inference*). DriPP étend les capacités du CDL en reliant les occurrences d'événements à des conditions ou des tâches expérimentales spécifiques, en utilisant un modèle statistique novateur qui connecte les fonctions d'intensité des processus ponctuels aux événements de stimulation. Cette approche est soutenue par un algorithme efficace d'expectation-maximisation (EM), montrant des résultats prometteurs dans la révélation de réponses neuronales évoquées et induites. FaDIn, quant à lui, s'attaque aux défis inhérents à l'inférence des TPP, en particulier avec les processus auto-excitants de Hawkes, en introduisant un solveur rapide basé sur le gradient. Cette méthode améliore considérablement la précision et l'efficacité dans la modélisation des motifs induits par les stimuli dans les signaux cérébraux, améliorant notamment les estimations de latence.

Collectivement, DriPP et FaDIn contribuent à un cadre analytique unifié pour les données M/EEG, offrant des méthodes robustes pour identifier et interpréter les motifs temporels influencés par les stimuli externes.

Extension de la décomposition CDL à l'échelle de populations

Le second défi se concentre sur l'évolution du CDL, passant d'une application individuelle à un protocole robuste au niveau de la population. L'objectif final est d'établir une méthodologie basée sur le CDL capable de générer un dictionnaire commun de motifs neuronaux applicables à un ensemble de sujets.

L'utilisation actuelle du CDL est principalement limitée à des sujets individuels, avec des contraintes computationnelles posées par les algorithmes existants, tels que ceux du package `Python AlphaCSC`, qui représentent un frein significatif pour les études en population. De plus, les enregistrements expérimentaux contiennent souvent des artefacts provenant de diverses sources, y compris des défaillances de capteurs (facteurs endogènes) ou des mouvements des sujets (facteurs exogènes). Ces anomalies, si elles ne sont pas identifiées et éliminées, peuvent compromettre l'intégrité de l'analyse. En outre, la variabilité spatiale des motifs neuronaux, influencée par les morphologies cérébrales individuelles, présente un défi dans le transfert des atomes appris d'un sujet à un autre.

Pour relever ces défis, la première contribution significative est le développement du CDL stochastique et robuste par fenêtrage – *Stochastic Windowing and Robust Convolutional Dictionary Learning*. Cette approche vise à surmonter les limitations computationnelles rencontrées dans l'analyse de données de séries temporelles étendues et à gérer la qualité variable des mesures. En mettant en œuvre le fenêtrage stochastique combiné à la computation sur GPU et à la différentiation automatique de `PyTorch`, le processus devient plus efficace sur le plan computationnel. De plus, un mécanisme de détection des anomalies en cours d'apprentissage est intégré, améliorant ainsi la robustesse de la CDL contre les anomalies de données.

Le second développement clef implique l'application du CDL au niveau de la population. En utilisant une approche basée sur les données sur un grand jeu de données en libre accès (Cam-CAN), couplée à un algorithme d'agrégation spécialisé, ce travail révèle la relation complexe entre la performance de tâches et le vieillissement dans les caractéristiques spatiotemporelles des rafales transitoires neuromagnétiques – *neuromagnetic transient bursts*. Cette analyse au niveau de la population révèle des tendances liées à l'âge dans les niveaux d'activation de types spécifiques de rafales, offrant de nouvelles perspectives sur l'évolution de l'activité cérébrale humaine au cours de la vie.

Collectivement, ces avancées améliorent l'application du CDL dans l'analyse des données M/EEG, en surmontant les défis computationnels et en étendant la portée du CDL aux études en population.

Publications

International Publications

Conferences

- Cédric Allain, Alexandre Gramfort, and Thomas Moreau. DriPP: Driven point processes to model stimuli induced patterns in M/EEG signals. *International Conference on Learning Representations*, 2022

- Guillaume Staerman, Cédric Allain, Alexandre Gramfort, and Thomas Moreau. FaDIn: Fast Discretized Inference for Hawkes Processes with General Parametric Kernels. *International Conference on Machine Learning*, 2023

Journals

- Lindsey Power, Cédric Allain, Thomas Moreau, Alexandre Gramfort, and Timothy Bardouille. Using convolutional dictionary learning to detect task-related neuromagnetic transients and ageing trends in a large open-access dataset. *NeuroImage*, 267:119809, 2023

Other

National conferences

- Oral presentation for DriPP [Allain et al., 2022], Groupe de Recherche et d'Études de Traitement du Signal et des Images (Gretsi; Nancy, France, September 2022)

Abstract-only international conferences

- Poster presentation for DriPP [Allain et al., 2022], Organization for Human Brain Mapping (OHBM; Glasgow, Scotland, June 2022)

Participation to events and awards

- Junior Conference on Data Science and Engineering (JDSE; Université Paris-Saclay, France, September 2021), « Best Presentation » award for DriPP [Allain et al., 2022]
- Biomag 2022 - Dementia screening challenge (Birmingham, England, June 2022), winning solution in MEG signal processing in order to screen dementia and mild cognitive impairment with Apolline Mellot and Benoît Malézieux

Contents

Remerciements i

Résumé en français des travaux de thèse iii

Publications vii

List of Figures xii

List of Tables xv

Notations xvii

Introduction 1

I General Background 5

1 Background on Neuroscience and Neurophysiological Signals 9
- 1.1 Biology of neuroscience . . . 10
  - 1.1.1 The brain, the central organ of the neural system . . . 10
  - 1.1.2 Origin of the neurophysiological signals . . . 15
- 1.2 Pre-processing and classical analysis in neurosciences . . . 21
  - 1.2.1 Filtering and time-frequency analysis . . . 24
  - 1.2.2 Independent Component Analysis . . . 27
  - 1.2.3 Segmentation and epoching . . . 29

2 Background on Dictionary Learning 33
- 2.1 Motivation and sparse representation . . . 34
  - 2.1.1 Sparsity and the Lasso . . . 35
  - 2.1.2 Lasso and its optimization . . . 36
- 2.2 Dictionary Learning: mathematical formulation and optimization problem . . . 38
- 2.3 Convolutional Dictionary Learning . . . 41
- 2.4 Convolutional Dictionary Learning in neuroscience . . . 44
  - 2.4.1 Rank-1 constraint . . . 46

3 Background on Temporal Point Processes **49**
3.1 Definitions . . . 50
3.2 Temporal point processes . . . 51
3.2.1 Poisson process and likelihood function . . . 54
3.3 Hawkes processes . . . 55
3.4 Goodness of fit . . . 59

II Temporal Modeling and Inference in M/EEG Signals: A Point Process Approach **61**

4 DriPP: Driven Point Processes to Model Stimuli Induced Patterns in M/EEG Signals **65**
4.1 Mathematical formulation . . . 68
4.2 Parameters inference with an EM-based algorithm . . . 70
4.3 Experiments . . . 78
4.3.1 Evaluation of the EM convergence on synthetic data . . . 78
4.3.2 Evoked and induced effects characterization in MEG data . . . 81
4.3.3 Impact of model hyperparameter . . . 85
4.3.4 Experiments on Cam-CAN dataset . . . 87
4.3.5 Usual M/EEG data analysis . . . 91
4.4 Transcending limits with discretised parametric kernels . . . 95

5 FaDIn: Fast Discretized Inference for Hawkes Processes with General Parametric Kernels **97**
5.1 Mathematical formulation . . . 99
5.1.1 FaDIn . . . 100
5.1.2 Impact of the discretization . . . 103
5.2 Numerical experiments . . . 109
5.2.1 Consistency of Discretization . . . 109
5.2.2 Statistical and computational efficiency of FaDIn . . . 112
5.2.3 Sensitivity analysis regarding the parameter W . . . 116
5.3 Application to MEG data . . . 117
5.4 Discussion . . . 119

III Advancements in Convolutional Dictionary Learning for Large-Scale M/EEG Data Analysis: Stochastic Approaches and Population Studies **121**

6 Stochastic Windowing and Robust Convolutional Dictionary Learning for M/EEG Data **125**
6.1 Introduction . . . 126
6.2 Contextualizing the current work . . . 128

6.3 Inline outlier detection . . . 129
6.4 Stochastic windowing CDL . . . 132
6.4.1 Approximate sparse coding . . . 133
6.4.2 Stochastic sub-windowing . . . 134
6.4.3 Stochastic line search . . . 137
6.5 Experiments . . . 139
6.5.1 Data simulation . . . 139
6.5.2 Dictionary evaluation . . . 142
6.5.3 Experimental paradigm . . . 144
6.5.4 Results . . . 145
6.6 Conclusion . . . 156

7 Using Population CDL to Detect Task-Related Neuromagnetic Transients and Ageing Trends in a Large Open-Access Dataset 161
7.1 Methods . . . 166
7.1.1 Participants and experimental paradigm . . . 166
7.1.2 Data acquisition and processing . . . 167
7.1.3 Convolutional Dictionary Learning (CDL) . . . 167
7.1.4 Atom clustering . . . 169
7.1.5 Selection of task-related clusters . . . 178
7.1.6 Representative Atom Generation . . . 180
7.1.7 Demographic analysis . . . 181
7.1.8 Supplementary analysis . . . 183
7.2 Results . . . 183
7.3 Conclusion . . . 195

Conclusion and perspectives 197

Bibliography 199

Appendix 223
A Pioneers of modern neuroscience . . . 223
B Adaptation of FISTA for CDL's inner problem . . . 225
C FaDIn – Additional experiments . . . 229
C.1 Comparison of FaDIn with the negative log-likelihood loss . . . 229
C.2 Qualitative Comparison with a non-parametric approach . . . 229

List of Figures

1.1.1 Composition of the Central and Peripheral Nervous Systems. . . . 11
1.1.2 Structure and composition of a neuron. . . . 12
1.1.3 Synapse structure and function. . . . 13
1.1.4 Composition of the brain. . . . 14
1.1.5 Post-synaptic potentials in a neuron. . . . 16
1.1.6 Approximate plot of a typical action potential. . . . 17
1.1.8 Recordings of the electromagnetic field. . . . 18
1.1.7 An EEG recording setup. . . . 18
1.1.9 Person undergoing a MEG. . . . 20
1.1.10 Raw signals over some M/EEG sensors. . . . 22
1.2.1 Time-frequency analysis of a signal. . . . 25
1.2.2 Time-frequency plane for epoched signals. . . . 28
1.2.3 The first five components of an ICA 30
1.2.4 Evoked signals following an auditory stimulus. . . . 31

2.3.1 Decomposition of a noiseless univariate signal as a convolution. . . 42
2.4.1 Spacial and temporal representation of 3 atoms learned by CDL. . 47

3.2.1 Example of simple Poisson processes 52
3.3.1 A realisation of a multivariate Hawkes process. . . . 58

4.1.1 Convolutional dictionary learning applied to a univariate signal . . 69
4.1.2 Schematic operation of the CDL on MEG signals. . . . 70
4.3.1 True and estimated intensity functions on synthetic data. . . . 79
4.3.2 Mean and std of the relative infinite norm on synthetic data. . . . 80
4.3.3 Mean and 95 % CI of computation time for one EM algorithm. . . 81
4.3.4 DriPP results on MNE *sample* dataset. . . . 83
4.3.5 DriPP results on MNE *somato* dataset. . . . 85
4.3.6 DriPP results on 3 μ-wave atoms from MNE *somato* dataset. . . . 86
4.3.7 Influence of the threshold on the obtained results on MNE *sample* dataset. . . . 88
4.3.8 Influence of the kernel truncation upper bound on the obtained results on MNE *sample* dataset 89
4.3.9 DriPP results on Cam-CAN dataset, subject CC620264 91
4.3.10 DriPP results on Cam-CAN dataset, subject CC520597 92

4.3.11 DriPP results on Cam-CAN dataset, subject CC723395 93
4.3.12 ICA components of atoms in fig. 4.3.4 and their cosine similarity. . 94
4.3.13 Evoked signals following a visual stimulus on MNE *sample* dataset. 95

5.1.1 Median ℓ_2 norm between continuous and discrete EM algorithm . 109
5.2.1 Median ℓ_2 norm for estimated parameters with FaDIn and EM. . . 110
5.2.2 Median ℓ_2 norm for each parameter with continuous and discrete EM. 111
5.2.3 Median ℓ_2 norm of each parameters learned with continuous and discrete EM. 112
5.2.4 Influence of the discretization of FaDIn's results for raised cosine and exponential kernels. 113
5.2.5 Error on parameters for the Truncated Gaussian kernel as a function of T and Δ. 113
5.2.6 Error on parameters for the Raised Cosine kernel as a function of T and Δ. 114
5.2.7 Error on parameters for the Truncated Exponential kernel as a function of T and Δ. 115
5.2.8 FaDIn's statistical and computational efficiency benchmark. . . . 116
5.2.9 Influence of the kernel support size W of FaDIn's results. 117
5.3.1 FaDIn results on MNE *sample* dataset 118
5.3.2 FaDIn results on MNE *somato* dataset 119

6.3.1 Raw signal X, reconstruction error, threshold and learned outlier mask on subject a02 (minute 56) of Physionet Apnea-ECG data set. Detection method is based on modified z-score (MAD), with $\alpha = 3.5$. The method correctly identifies outliers blocks. 130
6.4.1 Evolution of the upper bound through epochs for cosine annealing. 139
6.5.1 Shapes considered for dictionary simulation. 140
6.5.2 True dictionary in experiments on synthetic data. 141
6.5.3 First 2000 timepoints of one trial (2 channels) of synthetic data, corrupted with outliers. Final data is $X \in \mathbb{R}^{10\times2\times5000}$ 141
6.5.4 WinCDL efficiency on simulated and real data 146
6.5.5 On uncorrupted synthetic data, effect of the learned atom size (left) and the chosen window length (right) to the final dictionary recovery score, for 20 repetitions. 147
6.5.6 Median recovery score evolution for different outliers detection methods on synthetic data (20 repetitions). 148
6.5.7 Recall 149
6.5.8 Precision 149
6.5.9 On currupted synthetic data, evolution of recall (left) and precision (right) of outliers detection. Note that by construction, when no detection method is applied, precision and recall are null 149
6.5.10 Descriptive statistics on Physionet Apnea-ECG dataset 150

6.5.11 Loss evolution for different outliers detection method, on 10 good trials of subject a02 of dataset Physionet Apnea-ECG. 151
6.5.12 Learned atoms with and without outliers detection method, on 10 bad trials of subject a02 of dataset Physionet Apnea-ECG. 152
6.5.13 First 200 s of raw data from *rodent striatum* dataset. 153
6.5.14 Loss evolution for multiple methods, on both clean and dirty segments of *rodent striatum* dataset. 154
6.5.15 Learned atoms on empirical time-series with strong artifacts . . . 155
6.5.16 30 atoms learned by WinCDL from a MEG recording. 158
6.5.17 30 atoms learned by `AlphaCSC` from a MEG recording. 159

7.1.1 Workflow diagram. 168
7.1.2 Representation of subject CC121428's 20 extracted atoms. 172
7.1.3 Representation of subject CC723395's 20 extracted atoms. 173
7.1.4 Correlation matrices of participants who were excluded or not from the dataset. 174
7.1.5 Histograms of the number of groups per participant with varying correlation coefficient. 175
7.1.6 Number of clusters and number of top clusters identified with varying R value thresholds. 178
7.1.7 Top clusters' representative atoms, by varying cutoff. 179
7.1.8 Cluster sets statistical description. 184
7.2.1 The seven task-related clusters identified in this work. 186
7.2.2 The dipole fits for each of the seven representative atoms. 187
7.2.3 Time-frequency representations for each of the 7 task-related atom clusters. 188
7.2.4 The age distribution of participants in LPostC_beta and LT_alpha clusters compared to the dataset's one. 189
7.2.5 Results of linear and quadratic regression of summed activation and burst rate with age during pre-movement and post-movement. 191
7.2.6 Mean and std of the distribution of activation values for each atom as a function of age. 192
7.2.7 Results of linear and quadratic regression of several burst characteristics with age. 193
A.1 Golgi's illustration of a dog's olfactory bulb. 224
C.1 Statistical and computational efficiency of FaDIn with the Log-Likelihood loss. 230
C.2 Comparison between our approach FaDIn and non-parametric approach. 232
C.3 Comparison between our approach FaDIn and non-parametric approach for a Truncated Gaussian kernel. 233
C.4 Comparison between our approach FaDIn and non-parametric approach for a Truncated Exponential kernel. 233

List of Tables

1.1 Natural brain frequencies bands. 26

4.1 Statistics of each MNE dataset 82
4.2 Statistical univariate Student t-test on MNE *sample* dataset 95

6.1 Global computation time, in seconds, on *rodent striatum* dataset. . 153

7.1 Summary table of excluded subjects. 176
7.2 Summary of the attributes of each cluster. 187
7.3 p-values and RMSE values for age-related regression. 190
7.4 Summary of the age-related effects for each cluster 194

Notations

$\mathbb{N}$	Set of integers
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^n$	Set of real valued vectors of dimension n
$\mathbb{R}^{n\times m}$	Set of real valued matrices of dimension $n \times m$
$[\![a\,,b]\!]$	Set of integers between a and b included
$\mathbb{1}_{\{E\}}$	0-1 indicator function of a set E
$\#E$	Cardinal of set E
0_E	Null element of set E
A^{-1}	Inverse of $A \in \mathbb{R}^{n\times m}$
$A^\top$	Transpose of $A \in \mathbb{R}^{n\times m}$
$\ker A$	Kernel space of $A \in \mathbb{R}^{n\times m}$
$*$	Convolution product
$E^\perp$	Orthogonal to the vector space E
$\langle u, v\rangle_E$	Inner product between $u \in E$ and $v \in E$
$\|\cdot\|_0$	Number of non-zero coordinates (ℓ_0 pseudo-norm)
$\|\cdot\|_p$	Euclidean ℓ_p norm, $p \geq 1$
$\|\cdot\|_\infty$	Euclidean ℓ_∞ norm
$\|\cdot\|_F$	Frobenius norm
$\mathrm{dom}(f)$	Definition domain of function f
∂_f	Sub-gradient of f
∇f	Gradient of f
sgn	Sign operator
prox_f	Proximal operator of f
ST	Soft-Thresholding operator
proj_E	Projection operator on set E
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution of mean μ and covariance matrix Σ

Introduction

Neuroscience, an interdisciplinary domain at the confluence of biology, psychology, and mathematics, is dedicated to unraveling the functioning of the brain and nervous system. Central to this endeavor is the analysis of neuronal activities, captured through sophisticated modalities such as magnetoencephalography (MEG) and electroencephalography (EEG). These two non-invasive techniques, leveraging the physical principles of magnetic and electric field measurements, respectively, offer a window into the brain's dynamic processes. In concrete terms, a few dozen or hundreds of sensors are placed around a subject's head, and the machine records the intensity of the electric or magnetic field at high temporal resolution, of the order of 1000 Hz, for durations ranging from a few minutes to several hours. Characterized by this high temporal resolution, M/EEG data, are represented as multivariate time-series signals – one per sensor –, encapsulating the intricate patterns of neuronal activity. In parallel, an equally critical component of our research involves the meticulous logging of external events. These events range from controlled stimuli and participant actions in experimental setups to clinical interventions, such as drug injections during surgical procedures. The integration of these logs with M/EEG data provides a comprehensive framework for understanding brain activity in diverse contexts.

MEG and EEG have been instrumental in advancing our comprehension of various cognitive and neurological processes. From these signals, we can extract a wealth of information including, but not limited to, neural oscillations, event-related potentials, and network connectivity patterns. These insights have profound implications in both clinical and research settings, offering pathways to novel therapeutic strategies and deeper understanding of the brain's functionality. The foundational processing and classical analysis methods applied to M/EEG data – including filtering and time-frequency analysis, independent component analysis, and segmentation and epoching – are crucial for accurately interpreting these complex signals and extracting meaningful information, thereby facilitating the exploration of the brain's intricate dynamics.

In this thesis, we confront two principal challenges.

1. **Developing an end-to-end framework for temporal dependency analysis.** Our first challenge involves leveraging Convolutional Dictionary Learning (CDL) decomposition to transform M/EEG signals into a stream of events. Our aim is to establish a framework capable of learning the temporal dependencies between recurrent neural patterns – called "atoms" – and external stimuli, rethinking how we model the dynamic between neural activity and external influences.

2. **Extending CDL decomposition across populations.** The second challenge is to scale and refine the CDL decomposition process for applicability across a broader population. This extension is two-fold. Firstly, we aim to enhance the speed and robustness of the CDL process, making it resilient to a variety of outliers. These outliers can arise from multiple sources, ranging from endogenous factors like sensor failures to exogenous factors such as subject movements. Secondly, our goal is to develop a novel aggregation methodology for synthesizing subject-specific CDL results. This approach will enable us to uncover population-level effects, offering new insights into the generalizability and variability of neural patterns across different individuals.

Developing an end-to-end framework for temporal dependency analysis

Thus, the first challenge is to develop an end-to-end framework aiming to elucidate the intricate relationships between external stimuli and neural responses with minimal data processing and expert intervention. At the core of this framework is the analysis of neural time-series data, rich in prototypical signal waveforms known as "atoms". These shift-invariant atoms, essential in clinical and cognitive research, are extracted to understand the timing and occurrence of neural events. Traditional methods like epoch averaging – *i.e.,* averaging of time-bound segments "centered" on a particular event – often fail to capture the nuances of these time-locked responses due to minor temporal deviations.

In response, this thesis leverages Convolutional Dictionary Learning (CDL), specifically tailored to the physical principles underlying electrophysiological signals, as described by Maxwell's equations. CDL offers an efficient, unsupervised approach for pattern extraction in electrophysiological signals. The model represents data as sparse linear combinations of convolutions between dictionary atoms

and shift-invariant codes, presenting a novel event-based representation of temporal dynamics. These atoms are defined by both spatial and temporal characteristics and may correspond to various physiological activities – *e.g.,* heartbeats or eye-blinks – or neural responses to external stimuli such as auditory or visual cues. Temporal point processes (TPP) provide a statistical framework ideally suited for modeling these discrete event activations. Historically used in neuroscience for modeling single-cell recordings and neural spike trains, these processes, however, have not directly addressed the interaction between deterministic stimuli onsets and stochastic atom activations.

To bridge this gap, the thesis introduces two models that contribute to a unified analytical framework for M/EEG data: Driven Temporal Point Processes (DriPP) and Fast Discretized Inference (FaDIn). DriPP extends CDL's capabilities by linking event occurrences with specific experimental conditions or tasks, using a novel statistical model that connects point process intensity functions to stimulation events. This approach is underpinned by an efficient expectation-maximization (EM) algorithm, showing promising results in uncovering both evoked and induced neural responses. FaDIn, on the other hand, tackles the inherent challenges in TPP inference, particularly with self-exciting Hawkes processes, by introducing a fast gradient-based solver. This method significantly enhances the accuracy and efficiency in modeling stimuli-induced patterns in brain signals, notably improving latency estimations.

Extending CDL decomposition across populations

The second challenge is dedicated to evolving CDL from a subject-wise application to a robust, population-level protocol. The ultimate goal is to establish a CDL-based methodology capable of generating a common dictionary of neural patterns applicable to entire populations. The current use of CDL is primarily limited to individual subjects, with computational constraints posed by existing algorithms, such as those in the `Python` package `AlphaCSC`, acting as significant bottlenecks for population studies. Additionally, experimental recordings often contain artifacts originating from various sources, including sensor failures (endogenous factors) or subject movements (exogenous factors). These outliers, if not identified and removed, can compromise the integrity of the analysis. Furthermore, the spatial variability of neural patterns, influenced by individual brain morphologies, presents a challenge in transferring learned atoms from one subject to another.

To address these challenges, the first significant contribution is the development of Stochastic Windowing and Robust Convolutional Dictionary Learning. This approach is designed to tackle the computational limitations encountered in

analyzing extensive time-series data and to manage the variable quality of measurements. By implementing stochastic windowing combined with GPU computation and `PyTorch` 's automatic differentiation, the process becomes more computationally efficient. Additionally, an in-learning outlier detection mechanism is integrated, enhancing the robustness of CDL against data anomalies.

The second key development involves applying CDL at the population level. Utilizing a data-driven approach on a large open-access dataset (Cam-CAN), coupled with a specialized aggregation algorithm, this work uncovers the intricate relationship between task performance and aging in the spatiotemporal characteristics of neuromagnetic transient bursts. This population-level analysis reveals age-related trends in activation levels of specific burst types, offering novel insights into the evolution of human brain activity across the lifespan.

Organization of the Thesis

This thesis is structured into three distinct parts, each focusing on a different aspect of M/EEG signal analysis and modeling, providing a comprehensive exploration of the field.

Part I, titled "General Background", lays the foundational knowledge necessary for the subsequent sections. It begins with chapter 1, which introduces the basic principles of neuroscience and the nature of neurophysiological signals. Chapter 2, delves into the technical aspects of dictionary learning, a key method used in this research, then chapter 3 provides a detailed examination of temporal point processes, setting the stage for their application in later chapters.

Part II, "Temporal Modeling and Inference in M/EEG Signals: A Point Process Approach", presents the two developed models that are at the core of the thesis: "DriPP", in Chapter 4, introduces a novel approach to modeling M/EEG signals using *driven point processes*; while "FaDIn" in chapter 5, further develops these methods, focusing on *fast discretized inference* techniques for complex temporal models.

Part III, "Advancements in Convolutional Dictionary Learning for Large-Scale M/EEG Data Analysis: Stochastic Approaches and Population Studies", shifts the focus to convolutional dictionary learning and its applications in large-scale data analysis. Chapter 6, "Stochastic Windowing CDL", explores innovative approaches to dictionary learning in the context of M/EEG data, and chapter 7 concludes the thesis by applying an CDL aggregation method to a large dataset to uncover patterns related to specific tasks and aging trends.

Part I

General Background

The first part of this thesis manuscript is dedicated to furnishing the reader with the essential general context and knowledge base required for a comprehensive understanding of the research conducted herein. This part is multifaceted in its intention. Not only does it aim to elucidate the complex interplay of neurophysiological signals and mathematical concepts that underpin the subsequent contributions, but it also serves as a reference tool addressed to a broad audience, including future doctoral students in neuroscience applications with magneto- and electroencephalography (M/EEG) signals and seasoned scholars at the intersection of neuroscience, mathematics, and machine learning.

This part begins with a detailed overview of the background on neuroscience and neurophysiological signals in chapter 1. It covers the fundamental biology of neuroscience, emphasizing the brain's role as the central organ of the neural system and the origin of neurophysiological signals. Further, it delves into the intricacies of M/EEG data pre-processing and classical analysis in neuroscience, including topics such as filtering and time-frequency analysis, independent component analysis, and segmentation and epoching. This chapter is particularly useful for a general knowledge in neuroscience as well as having a better grasp of experiments performed on real data in chapters 4 to 7.

The mathematical background in chapters 2 and 3 is equally rigorous. The former is a comprehensive exploration of dictionary learning that encompasses motivation and sparse representation, mathematical formulation and optimization problems, convolutional dictionary learning and its various applications. This chapter also includes a targeted examination of how convolutional dictionary learning finds application in neuroscience. Chapter 3 is a thorough look at temporal point processes, including definitions, characteristics, Hawkes processes, and goodness of fit analysis. The decomposition of M/EEG signals using CDL is used throughout this thesis, especially in part II as it is used as input of the developed point processes models when applied to real data.

This part has been diligently crafted to serve as a primer for those commencing a doctoral journey in neuroscience, particularly those dealing with M/EEG signals but lacking specific background or initial training in this domain. It also facilitates the learned reader to easily navigate the contributions, referring back to this section as needed to clarify points.

Chapter 1

Background on Neuroscience and Neurophysiological Signals

Contents

1.1 Biology of neuroscience **10**
1.1.1 The brain, the central organ of the neural system . . . 10
1.1.2 Origin of the neurophysiological signals 15
1.2 Pre-processing and classical analysis in neurosciences **21**
1.2.1 Filtering and time-frequency analysis 24
1.2.2 Independent Component Analysis 27
1.2.3 Segmentation and epoching 29

NEUROSCIENCE designates the scientific study of the structure, function, and development of the brain and the nervous system, which can be traced back to the Hellenistic world, with significant advancements in the late 19th and early 20th century (see section A). This vast field covers everything from the molecular level to the level of organs such as the brain, spinal cord, and peripheral nerves. This complex, multidisciplinary science involves various sub-disciplines – in particular neuroinformatics and computational neuroscience –, each of which brings its own specific knowledge and techniques to improve our understanding of the nervous system. Neuroscience is not only essential for understanding the brain's normal functions but also for identifying, treating, and preventing neurological disorders and injuries. Conditions such as Alzheimer's and Parkinson's diseases, schizophrenia, depression, and stroke can be better understood with a deeper knowledge of how the neural system operates.

This chapter aims to provide the reader with the requisite insights into the biological and technical aspects of the field. The first section begins by exploring the brain, the central organ of the neural system, offering an in-depth look into its structure and function (p. 10). It further delves into the biological origin of neurophysiological signals (p. 15), uncovering the mechanisms that give rise to the intricate signals that form the basis of neural communication.

Moving from the biological to the analytical, the second section (p. 21) provides an essential guide to the methods employed in the transformation and interpretation of magneto- and electroencephalography (M/EEG) data. This includes a comprehensive review of filtering and time-frequency analysis (p. 24), independent component analysis (p. 27), and segmentation and epoching (p. 29). These techniques are vital in translating complex neural signals into usable data, contributing to the advancements in treating and understanding neurological disorders and conditions.

This chapter serves as a bridge, connecting neuroscience's broad and multidisciplinary nature to the specific biological and analytical tools that enable researchers to unravel the mysteries of the nervous system. Whether for the investigation of normal brain functions or the pursuit of solutions to debilitating neurological diseases, the foundational knowledge provided in this chapter equips the reader with a nuanced perspective on the science of the neural system.

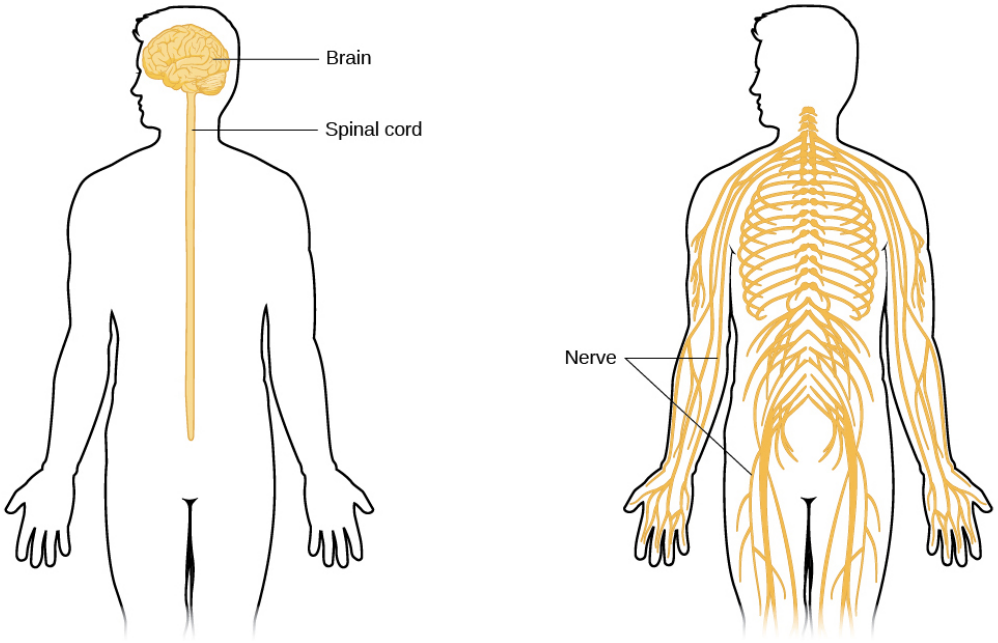
1.1 Biology of neuroscience

1.1.1 The brain, the central organ of the neural system

In this section, we delve into the foundational elements of the neural system, with a particular focus on the brain as its central organ. We explore the architecture of the nervous system, the role of neurons as its basic units, and the biological intricacies of the brain itself.

The nervous system, the network of neurons

The nervous system, a complex web of neural connections, serves as the body's control hub and communication network. This intricate system is composed of the brain, spinal cord, and a vast network of billions of neurons that pervade the body. Its key roles include organizing and coordinating bodily functions, as well as mediating interactions with the external environment. As established by Thomas



(a) Central Nervous System (CNS)

(b) Peripheral Nervous System (PNS)

Figure 1.1.1: Composition of the CNS (a) and the PNS (b). [Pelz]

Willis (1621-1675) very early on in his *Cerebri Anatome* (1664), this expansive network of neurons is fundamentally divided into two subsystems: the Central Nervous System (CNS) and the Peripheral Nervous System (PNS) [Dehaene, 2021], as presented in fig. 1.1.1.

The CNS, comprising the brain and spinal cord, is the primary control center of the body, tasked with system regulation, information processing, and memory formation. The PNS, on the other hand, consists of all the neurons linking the CNS to the rest of the body, responsible for transmitting signals to and from the brain or spinal cord. This division of the nervous system effectively facilitates communication between the CNS and the body's periphery.

Further, the PNS itself can be split into two components: the Autonomic Nervous System (ANS) and the Somatic Nervous System (SNS). The ANS governs the body's unconscious actions including regulating functions such as heart rate, respiration, blinking, and digestion. Conversely, the SNS handles voluntary control of the body, facilitating conscious and deliberate actions. Whenever a specific action, like raising a hand or pushing a button, is planned and executed, the SNS transmits signals from the brain to our muscles, instructing them on the necessary movements. Thus, the nervous system as a whole plays an integral role in the body's function, coordinating everything from basic physiological processes to complex voluntary actions.

The neuron, the basic unit of the neural system

A neuron, also known as a nerve cell, is an electrically excitable cell that serves as the fundamental unit of the nervous system. Characterized by their ability to transmit bioelectrical signals known as nerve impulses or *action potential*, neurons carry communication within the body, allowing for both primary and complex functions ranging from physiological regulation to cognition and movement. Physiologically, neurons exhibit two fundamental properties: excitability, which is the capacity to respond to stimuli and convert them into nerve impulses; and conductivity, which denotes the ability to transmit these impulses across the length of the neuron. These impulses propagate at remarkable speeds, capable of traveling through a nerve cell at velocities of up to 118 meters per second.

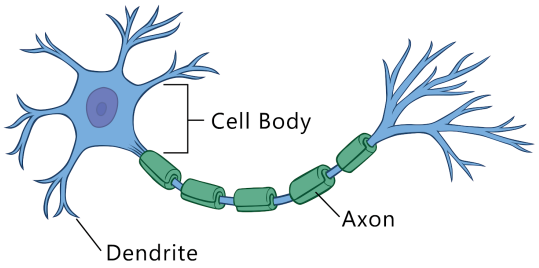


Figure 1.1.2: Structure and composition of a neuron. [Pollock, 2023]

Structurally, every neuron comprises three principal parts: the cell body (soma), dendrites, and an axon (see fig. 1.1.2). The cell body houses the neuron's nucleus and other vital organelles, serving as the metabolic center of the neuron. Dendrites, tree-like appendages extending from the cell body, receive signals from other neurons or the external environment and deliver them to the cell body. The axon, typically a singular elongated projection attached to the cell body, carries signals away from the cell body and towards other neurons, muscle cells, or glands. In humans, axons can be several centimeters, with some up to a dozen [Dehaene, 2021]. The dendrites and axons are called nerve fibers.

Contrary to the contiguous appearance of the nervous system, no two neurons directly touch each other. Instead, they communicate across a tiny space known as the synaptic gap or synapse, as shown in fig. 1.1.3. Indeed, the arrival of the *action potential*[1] at the axon terminals triggers a transformation of the signal from electrical to chemical in nature, which is essential for the signal to pass through the synaptic gap and into the next neuron. Precisely, within the axon terminals are vesicles loaded with neurotransmitters, which are chemical messengers that convey signals across the synapse. Some common examples of neurotransmitters are dopamine, serotonin, and epinephrine. Protecting these essential communication channels, a fatty layer of insulation known as the *myelin sheath* envelops the dendrites and the axon. The myelin sheath ensures the integrity of the nerve

[1] A detailed discussion of the *action potential* and its underlying mechanisms will be presented later in section 1.1.2.

signals by preventing them from interfering with one another, while also enhancing the speed of impulse propagation along the axon.

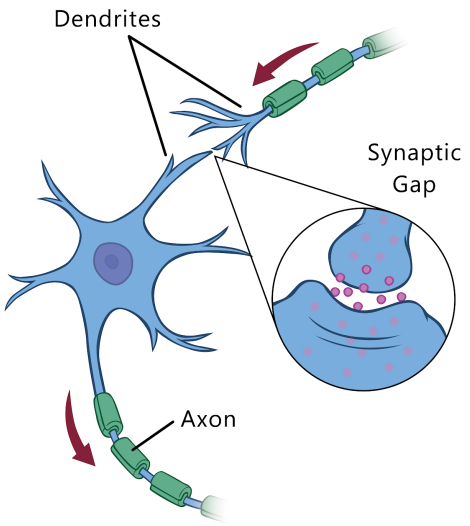


Figure 1.1.3: Synapse structure and function. [Pollock, 2023]

Despite the basic uniformity in the structure of neurons, there exist variations optimized for their specific functions, leading to the classification of neurons into three primary types: sensory neurons, interneurons, and motor neurons. Sensory neurons, also known as *afferent neurons*, carry signals from sensory organs such as the skin, tongue, ears, eyes, and nose to the brain and spinal cord, thereby enabling sensory perception. Interneurons, residing entirely within the central nervous system, form a link between sensory and motor neurons, often integrating and interpreting incoming sensory information and formulating an appropriate response. Motor neurons, or *efferent neurons*, then convey signals from the brain or spinal cord to the muscles and glands, facilitating voluntary and involuntary actions.

Contrary to most cells in the body, neurons possess a limited capacity for regeneration. An individual is born with over 100 billion neurons that grow and form connections as the individual develops. However, once an individual reaches adulthood, the total number of nerve cells begins to decrease due to natural cell death. This unique trait underscores the importance of maintaining neuronal health for lifelong function and well-being.

Biology of the brain

The human brain, an intricate organ, serves as the body's command center, controlling every action, whether it is an overt gesture like picking up an object or an unconscious action such as breathing or blinking. It forms a crucial part of the central nervous system, connecting with the rest of the body through the spinal cord and its extending nerves. Residing within the protective confines of the skull, the brain is shielded by three layers of membranes known as the meninges. Moreover, a fluid known as cerebrospinal fluid is present to buffer the brain and prevent its violent movement within the skull. At birth, a baby's brain boasts billions of neurons or nerve cells, and all the major cortical connection clusters are already in place. They are established on a mainly genetic basis. However, their termi-

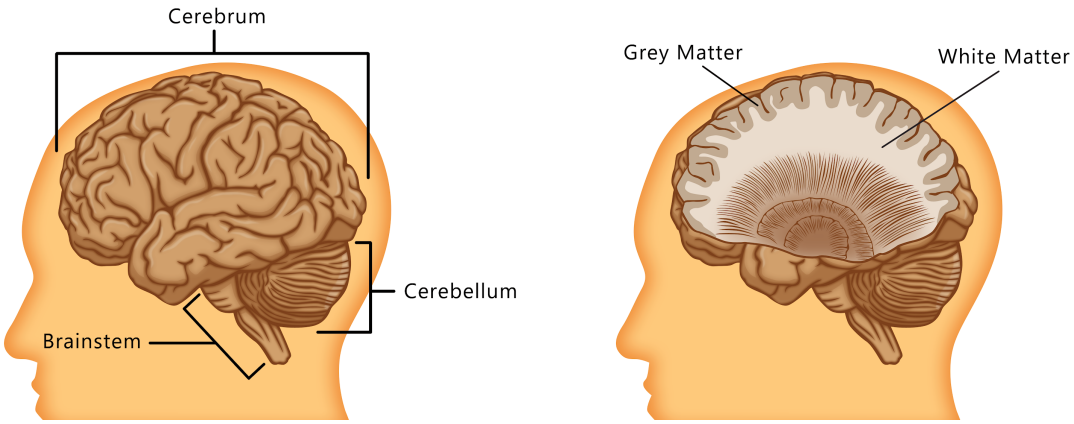


Figure 1.1.4: Composition of the brain. [Pollock, 2023]

nation has not yet been established with certainty. Axons search for their target for a long time and their wiring depends on experience. Synapse after synapse, learning shapes our nerve endings [Dehaene, 2021]. Indeed, as the child grows and matures, connections between these neurons continue to form, strengthening and proliferating with learning and memory formation. This intricate, dynamic network embodies the brain's incredible *plasticity*.

When we talk about brain plasticity, we mean it literally: neurons move, their axonal endings grow and shrink. Numerous buds (dendritic spines) are constantly appearing and disappearing on their dendritic trees, ready to receive new synapses. These organelles are constantly on the move: their size, their very presence, can change on the scale of a few tens of minutes, depending on the learning process [Dehaene, 2021]. Neuronal activity selectively modulates the strength and stability of synapses. A synapse whose two neurons, presynaptic and postsynaptic, activate together increases in strength. Its size increases to accommodate more receptor molecules. Conversely, if a synapse is not useful, its effectiveness decreases and it may retract completely [Dehaene, 2021].

The brain is not just made up of neurons: at least 50 % of cells are non-neuronal cells that do not produce electrical impulses called *glial cells*, or *neuroglia*. Neuroglia are also present in the spinal cord as well as in the Peripheral Nervous System. They are involved in all sorts of functions that support neurons, for example, recycling neurotransmitters released at the synapse, or acting as a link between synaptic activity and cerebral blood flow, ensuring neurons optimal functionality. For instance, it is the glial cells, the astrocytes, which detect that a region of the brain is working and which, in a few seconds, dilate the neighboring capillary arteries in order to supply the extra oxygen and glucose that the tissue needs – which generates the signal that we detect in functional MRI [Fields et al., 2014].

The human brain, home to approximately 86 billion neurons, predominantly comprises gray and white matter (fig. 1.1.4). Grey matter constitutes 40 % of the brain and includes neuron cell bodies, dendrites, and unmyelinated axons that appear gray. It forms the cortex – in Latin, the bark of the brain –, a thin surface a few millimetres thick that covers our two hemispheres hosting a myriad of neuronal cell bodies and synapses: 1 cubic millimeter of cortex contains around 50 000 neurons [Dehaene, 2021]. Meanwhile, white matter constitutes 60 % of the brain. It is situated in the brain's deeper regions, comprising bundles of myelinated axons that appear white, and serves as a conduit for connecting different brain regions.

The brain is organized into three major parts: the brainstem, the cerebellum, and the cerebrum. The brainstem, forming the bridge between the spinal cord and the brain's higher regions, houses control centers for automatic body functions such as breathing, swallowing, blinking, and vomiting. Situated above the brainstem, the cerebellum is integral to controlling muscle movements and maintaining posture and balance. The cerebrum, the brain's most prominent part, is divided into the left and right hemispheres. It governs higher cognitive processes, including memory formation, interpretation of signals from the senses, and emotional responses.

1.1.2 Origin of the neurophysiological signals

Neurophysiological signals refer to the magnetic field and the electrical currents generated by the biological processes within neurons. As previously described, the communication mechanism in neurons is an electrochemical process characterized by changes in the electric potential across the neuron's membrane, known as *action potentials*. Action potentials in neurons are also known as "nerve impulses" or "spikes", and the temporal sequence of action potentials generated by a neuron is called its "spike train". A neuron that emits an action potential, or nerve impulse, is often said to "fire". These action potentials are the basis of the electrical signals that can be measured on the scalp, the electroencephalogram (EEG), or on the skin, the electromyogram (EMG). The recording methods of such signals will be elaborated upon in this section.

In neuroscience, we are particularly interested in the electrical activity of the neurons and especially in the emission of post-synaptic potentials (PSP). PSP are changes in the membrane potential – the difference in electrical potential between the outside and inside of a cell – of the postsynaptic neuron consecutive to the binding of neurotransmitters to the postsynaptic cell's receptors, reflecting the exchange of ions between pre- and postsynaptic neurons (see fig. 1.1.5). There are two primary types of postsynaptic potentials, excitatory and inhibitory. Excitatory Postsynaptic Potentials (EPSPs) occur when the binding of neurotransmit-

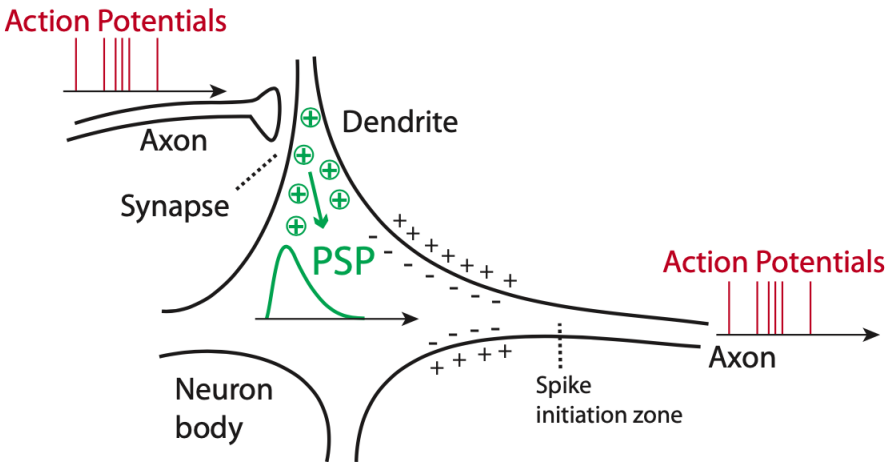


Figure 1.1.5: Post-synaptic potentials in a neuron, consecutive to ion exchange. [Gramfort]

ters to the postsynaptic receptors opens positive ion channels, allowing positively charged ions to enter the neuron, resulting into a net gain of positive charge across the membrane. This process depolarizes the neuron, making it more likely to fire an action potential. Inhibitory Postsynaptic Potentials (IPSPs) occur when the binding of neurotransmitters to the postsynaptic receptors opens channels for negatively charged ions or positively charged ions to exit. This hyperpolarizes the neuron, making it less likely to fire an action potential.

In a single patch of postsynaptic membrane, multiple EPSPs can likely occur. EPSPs, when occuring close in time, have an additive effect, which means that the sum of all the individual EPSPs will result in a combined effect. Greater membrane depolarization takes effect when there are larger EPSPs created. The larger the EPSPs become, the more it reaches the limit of firing an action potential. Indeed, action potentials are "all-or-nothing" events – contrary to PSPs that are said to be "graded potentials" –: when enough depolarization accumulates to bring the membrane potential up to a certain level, called the threshold, the action potential will fire [Kandel et al., 2000, chap. 9]. Note that human dendrites contain almost 30 000 synapses, and the activation of just 135 of them is enough to generate an activation potential [Dehaene, 2021]. If the membrane potential does not reach this threshold, the action potential will not fire. Action potential occurs thus within a neuron when it transmits electrical impulses. During this signal transmission, the membrane potential of the neuron – specifically the axon – fluctuates with rapid rises and falls.

Electro- and Magneto- encephalography Just as the moving charges on a wire – the ions in the axon – induce an electromagnetic field, a group of activated

neurons in the gray matter would form a current generator that produces such a field. More precisely the M/EEG signal measured is mostly emitted by pyramidal cells in the cortex, due to their unique topology and placement relative to the scalp. Indeed, pyramid cells, named for their characteristic pyramid-like shape, are a prominent type of excitatory neuron found in the cerebral cortex and play a pivotal role in cortical information processing.

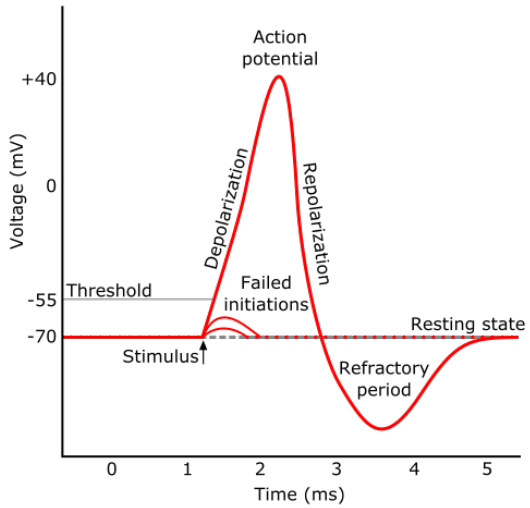
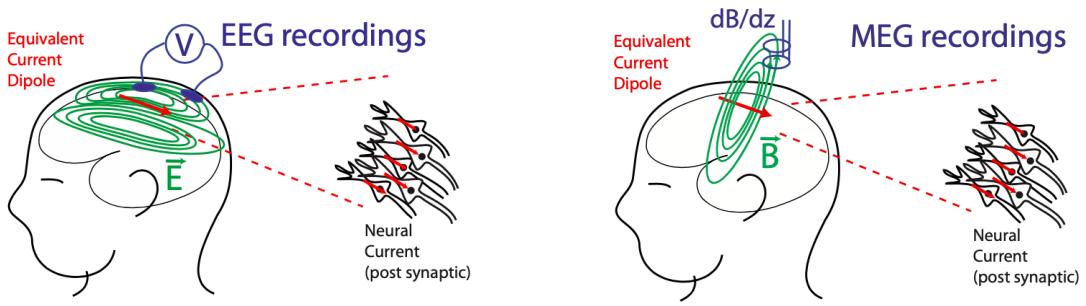


Figure 1.1.6: Approximate plot of a typical action potential and its different phases as the action potential passes a point on the cell membrane. [Wikipedia, 2023a]

These neurons exhibit a unique structure, with a single apical dendrite extending towards the cortical surface and multiple basal dendrites radiating from the cell body. They constitute approximately 80 % of the neurons of the cortex – the proportion varies with cortical regions –, each of them receiving around 10 000 synapses [Dehaene, 2021]. Their primary neurotransmitter, glutamate, facilitates excitatory communication within intricate neural networks that span various cortical layers and regions [Spruston, 2008, Yuste, 2015]. Pyramidal cells are instrumental in diverse cognitive functions, and their dysfunction has been implicated in neurological disorders such as Alzheimer's disease and schizophrenia [Selkoe, 2002, Konopaske et al., 2008]. The study of these cells has significantly contributed to understanding the complex connectivity and functional dynamics of the human brain [Markram et al., 2015].

Various techniques exist to record these signals, providing invaluable insights into brain function. Electroencephalography (EEG) is a technique that captures differences in electric potential in the brain using electrodes placed directly on the scalp. Magnetoencephalography (MEG), on the other hand, measures the magnetic flux density outside the head. For a measurable M/EEG signal, a large number of neurons, approximately 50 000, need to be simultaneously active. Both methods record synchronized neural activity at a very high temporal resolution, about the millisecond – sampling is often between 250 Hz to 1000 Hz –, and have the advantage of being non-invasive, unlike electrocorticography (ECoG) that uses electrodes surgically placed directly on the exposed surface of the brain. Such neurophysiological data also typically include a spatial dimension, as multiple sensors



(a) EEG recordings (b) MEG recordings

Figure 1.1.8: Recordings of the electrical field (a) and the magnetic field (b). [Gramfort]

are placed across the scalp, and different sensors will detect different signals depending on their location.

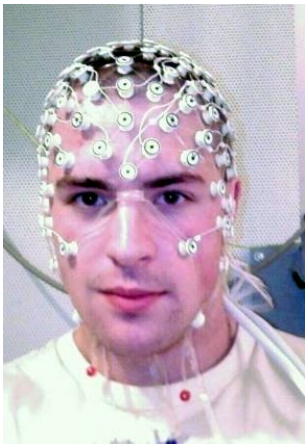


Figure 1.1.7: An EEG recording setup. [Wikipedia, 2023b]

This high temporal resolution is what makes MEG and EEG attractive for the functional study of the brain. Despite their poor spatial resolution, with only a few hundred spatial data points acquired simultaneously – about 300 to 400 sensors for MEG [Velmurugan et al., 2014] and up to 256 electrodes for EEG [Soufineyestani et al., 2020] –, these methods can localize neural activity with appropriate models and methods[2]. This is what is called the *inverse problem*, whose objective is to determine the current generators that produced the M/EEG measurements, as opposed to the *forward problem* whose objective is to predict the M/EEG surface signal to current dipoles in the brain [Pascual-Marqui, 1999, Galka et al., 2004, Grech et al., 2008]. While both EEG and MEG can provide information about brain activity, their sensitivity to different types of signals and spatial resolution differ. EEG, for instance, is often employed in sleep analysis due to its high temporal resolution, while MEG, capable of detecting signals from deeper brain structures, can provide more accurate source localization.

A critical limitation to recognize in the use of EEG as a neural imaging modality is its inherent constraints on spatial localization of brain activity. Given the external positioning of electrodes on the scalp, compounded by the brain tissue's highly conductive nature and surrounding cerebrospinal fluid, EEG signals often reflect a diffusion of neural activity across multiple regions. Essentially, an electrode situated externally can potentially capture signals emanating from anywhere within the neural structure. This becomes particularly complex when considering mul-

[2] Inria: MEG/EEG vs. other functional brain imaging modalities

tiple brain regions' simultaneous or rapidly successive activation. Consequently, any signal intercepted by an external electrode is usually an amalgam of activities from distinct, active neural sites capable of producing a signal potent and coherent enough to transverse the impedance presented by the skull [Newman et al., 2023]. From a mathematical standpoint, the inverse problem is ill-posed, offering an infinite array of potential solutions. Despite this, source localization is not entirely unfeasible. It has been demonstrated to be effective, mainly when neural activity is concentrated in well-defined regions, such as the primary visual or auditory cortices. However, a prevailing practice in EEG research remains the study of signals as they are captured on the scalp surface. Years of empirical studies have furnished robust evidence that correlates specific EEG signatures with particular cognitive functions or tasks. Hence, one of the main applications of EEG remains in identifying the presence, or assessing the magnitude, of these known signals in relation to specific cognitive activities or task conditions [Newman et al., 2023].

Other recording techniques In addition to EEG and MEG, other techniques, such as functional Magnetic Resonance Imaging (fMRI) and single-unit recording, are widely used in neuroscience research. fMRI leverages the blood-oxygen-level-dependent (BOLD) contrast to *indirectly* measure neural activity in the brain. The technique capitalizes on the hemodynamic response, a process where localized neural activity triggers a cascade of metabolic events leading to an increase in cerebral blood flow and a subsequent change in the ratio of oxygenated to deoxygenated hemoglobin [Ogawa et al., 1990]. By sensitizing the MRI signal to these blood oxygenation changes, BOLD fMRI provides a non-invasive window into the dynamic interplay of neural activation, metabolism, and vascular response. fMRI provide a very good spatial resolution but a rather poor temporal one[3] – of the order of a second for fMRI. On the other hand, single-unit recording involves placing a microelectrode directly onto a single neuron to record its electrical activity, offering precise measurement of individual neuron firing. Contrary to M/EEG and fMRI, it is an invasive recording technique. It provides insights into the basic physiological properties of neurons and neuronal networks.

Paradigm of a M/EEG experiment In an EEG experiment, a professionally trained nurse meticulously positions approximately 250 electrodes on the subject's scalp. These electrodes come equipped with a range of sensors, including some designed for particular localizations and electrooculography (EOG) sensors for eye movement tracking. These are also electrodes, the same as EEG electrodes, but intentionally placed close to the eyes specifically to monitor for blinks and eye movements, as shown in fig. 1.1.7. They are typically placed above and below one

[3]Other methods such as positron emission tomography (PET) and single-photon emission computed tomography (SPECT) have similar characteristic.

eye (to monitor blinks and vertical eye movements, as well as on the temples of the head laterally to the eyes (to monitor horizontal eye movements) [Newman et al., 2023]. MEG experiments, on the other hand, make use of a complex device comprising a vast array of around 300 to 400 ultra-low magnetic field sensors, cooled in liquid helium to $-269\,^{\circ}\mathrm{C}$, methodically situated above the subject's head to record neural activity with high precision [Dehaene, 2021]. Concurrent EEG recordings are often part of MEG experiments, offering supplementary information that bolsters the overall data quality. The duration of these recordings can fluctuate significantly based on the experimental design, with active experiments lasting mere minutes and more extensive analyses, such as sleep stage investigation, spanning several hours.

Given that the brain's magnetic signals measure in the femtotesla range, it is essential to shield the experiment from external magnetic disturbances, including those from Earth's magnetic field. To ensure data integrity, rooms used for M/EEG experiments are constructed with materials like aluminum and mu-metal, which effectively reduce high-frequency and low-frequency noise, respectively. This careful choice of materials creates a controlled environment for accurate detection and analysis of subtle neural signals – we are able to reconstruct the activity of the brain to the scale of a millisecond –, thereby improving the reliability of the data collected and subsequent findings.

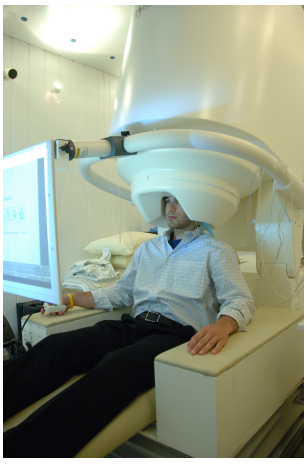


Figure 1.1.9: Person undergoing a MEG. [Wikipedia, 2023c]

During a typical task-based experiment, subjects are presented with external stimuli like auditory or visual signals, somatosensory inputs, or other pertinent cues. Occasionally, subjects are assigned active tasks, such as pressing a button in response to specific cues. The precise timing of these "events" is carefully recorded in tandem with the M/EEG signals to link them to brain activity patterns.

Additional sensors are often used to enhance data accuracy and reliability, including heartbeat monitors to track cardiac activity and sensors positioned near the eyes to capture eye movements (EOG, as previously mentioned), which can produce artifacts in the recordings[4]. These supplemental sensors are integral to

[4]The organization of neurons within the retina constitutes a specialized arrangement, giving rise to an electrical dipole characterized by distinct positive and negative poles. This configuration aligns precisely with the underlying principles of EEG detection, where such dipoles can be readily identified and monitored. Under stable conditions, where the eye's position remains fixed, the dipole generated by the retina exerts a constant effect, leaving the EEG recordings unaffected due to the unchanging nature of the dipole. However, the dynamics alter with changes in the eye's orientation, corresponding to different spatial locations. As the dipole adjusts with

the post-processing phase, aiding in the removal of artifacts from the recorded signals and leading to more precise and clean data. Indeed, in neuroscience, particularly within EEG studies, an artifact is any signal interference not originated from the primary source of interest: the brain's neuronal activity. These interferences, grouped as environmental, instrumentation, or biological, add noise to the data. Environmental artifacts can emerge from sources like AC power line oscillations or electromagnetic field noise from surrounding objects. Instrumentation artifacts come from the experimental setup itself, including electromagnetic interference from stimuli presentation or sensor malfunctions. Biological artifacts comprise disturbances from physiological actions such as heart electrical activity, eye movements, and muscle contractions during actions like swallowing. Even brain signals related to activities not central to the study can be considered artifacts. For reliable results, it is crucial to accurately identify and eliminate these artifacts, ensuring the interpreted data truly represent brain activity [Newman et al., 2023].

As shown in fig. 1.1.10, the final outcome of these experiments is multivariate *continuous* time series data that encapsulate brain activity over time, possibly including timestamps related to the events (stimuli or subject's actions). However, the inherent noise in the recorded data due to the nature of neural signals and external influences necessitates the use of advanced data analysis techniques to reveal meaningful patterns and provide insightful conclusions.

In sum, understanding how neurophysiological signals are produced and propagated in the human brain and the methodologies employed to record these signals is central to neuroscience. Such knowledge not only broadens our understanding of the human brain's workings but also paves the way for advances in diagnosing and treating neurological conditions. It thereby motivates the development of computational tools for learning such signals from data.

1.2 Pre-processing and classical analysis in neurosciences

Data pre-processing is an indispensable step when dealing with real-world signals obtained from magnetoencephalography (MEG) and electroencephalography

the movement of the eye, this shift is captured by the scalp electrodes used in EEG. Specifically, lateral eye movements, either to the left or right, induce variations in the electrical potential at the frontal electrodes. This results in an asymmetric response, with an increase in potential on one side of the head and a corresponding decrease on the other. The directionality of this asymmetry is contingent on the trajectory of the eye's movement, reflecting the complex interplay between ocular mechanics and neural electrical activity [Newman et al., 2023].

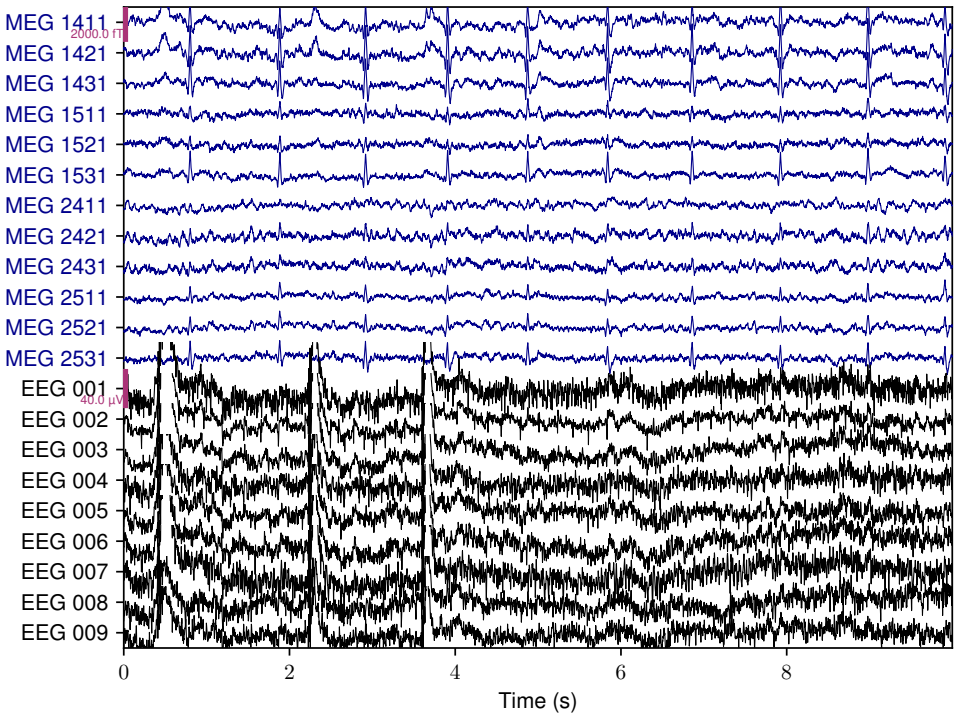


Figure 1.1.10: Raw signals over some M/EEG sensors, on MNE *sample* dataset. The most common way of viewing M/EEG data is in the time domain, with time plotted on the x-axis, and electrical potential (voltage) or magnetic field (tesla) on the y-axis. The observed signal can be represented as a matrix $X \in \mathbb{R}^{P \times T}$, with P sensors over T timestamps.

(EEG) experiments. Due to the inherent nature of the acquired signals and the fallibility of the machines used for measurement, these real-world signals often contain not only the desired neural information but also various forms of noise or artifacts. Therefore, a meticulous pre-processing routine, complemented by manual inspection rooted in domain knowledge, is necessary to discern artifacts and differentiate between evoked and induced responses. This section briefly presents some classic data pre-processing procedures and standard methods for analyzing M/EEG data. The aim is to provide a better understanding of the upstream processing carried out during real data experiments presented in the rest of the manuscript. Indeed, in the subsequent sections of this thesis, particularly in part II, encompassing chapter 4 and chapter 5, we introduce a comprehensive framework for temporal analysis in M/EEG data. This framework aims to simplify the pre-processing steps by utilizing only filtered raw data as its input. We thus start this section by giving an overview of usual pre-processing, before detailing each step.

General pre-processing As mentioned, M/EEG data pre-processing is a crucial step in the analysis pipeline, aiming to mitigate the effects of various artifacts and noise that are inevitably present in the raw data. Initially, the data undergoes a band-pass filtering process to focus on the frequency bands of interest, which largely depend on the specifics of the study [Widmann et al., 2015]. In a nutshell, this step entails applying mathematical transformations to remove high-frequency noise via low-pass filters and eliminate low-frequency drifts and slow trends using high-pass filters. Additionally, M/EEG signals can often be exposed to strong power line noise, typically a consistent interference at 50 Hz or 60 Hz depending on the geographic location. A commonly used method to remove line noise is the *notch filter*, but it comes with the risk of potentially severe signal distortions [Leske and Dalal, 2019]. This filter selectively attenuates a narrow frequency band centered around the power line frequency, thus effectively reducing its influence on the recorded signals without affecting the overall data integrity. As a general rule, filter cut-off frequencies must respect the *Nyquist theorem* to prevent aliasing [Shannon, 1949].

Whenever the acquired data contains defective or missing channels, *i.e.,* sensors, methods such as interpolation are used to deal with these data irregularities and ensure the overall integrity of the data set. The data might then be subjected to further steps, such as downsampling, to synchronize the sampling rates between different recordings and reduce the computational load during analysis; epoching, where continuous data is cut into shorter segments, typically time-locked to an event of interest; or ICA, to separate the data into statistically independent components and manually removed artifacts. Lastly, a procedure for dealing with remaining artifacts after epoching is often implemented, such as artifact rejection or correction, before the data is ready for analysis.

1.2.1 Filtering and time-frequency analysis

Time-frequency analysis is an essential step in the data pre-processing pipeline in neuroscience, particularly in the context of M/EEG, and allows for an in-depth examination of how the frequency content of signals changes over time. Leveraging mathematical transformations such as the Short-Time Fourier Transform (STFT) – a Fourier-related transform –, time-domain signals can be decomposed into the time-frequency space, presenting a dual view of the data. This method unveils the intricate spectral dynamics of M/EEG signals, offering insights into the brain's oscillatory activities and their variations across time. Thus, it offers a more detailed perspective of brain activity beyond the constraints of studies limited to either time or frequency domains.

Transition from time to frequency domain The *Time-Domain Representation*, where signals are displayed according to their evolution over time, is the most frequently used signal representation, as shown in fig. 1.1.10. However, these time-domain signals can be transformed into the *Frequency-Domain* using the Fourier Transform (FT). The principle behind FT is that any time-varying signal can be decomposed into a set of sine waves, each characterized by specific frequencies and amplitudes. Even complex EEG waveforms, where amplitude and frequency are variable, can be accurately described through this combination of sine waves. FT estimates the amplitudes of a broad range of sine waves, which can then be plotted in the frequency domain, with each frequency's *power* – or *amplitude* – represented along the y-axis, as represented in fig. 1.2.1.

However, this conversion into the frequency domain comes with a loss of temporal information. In the frequency domain, power at each frequency indicates the average power across the entire duration of the input data. There is no need to use the entire time range of the data to compute a FT and generate data in the frequency domain; however, it is pretty intuitive that, at a minimum, the signal should be long enough for at least one cycle of a sine wave to occur. In practice, 2 to 3 cycles would be preferable [Newman et al., 2023]. Therefore, the duration of the available data determines the lowest frequency that can be estimated, unveiling an inverse relationship between precision in time and frequency domains. Hence, an optimal balance between time and frequency precision is required for accurate analysis.

Relevance to neuroscience A common hypothesis in neuroscience is that brain activity involves periodic oscillations — oscillating sine wave — and aperiodic signals — "one-off" peaks and troughs. Generally, the time domain is often best suited for viewing and analyzing aperiodic signals, while the frequency domain

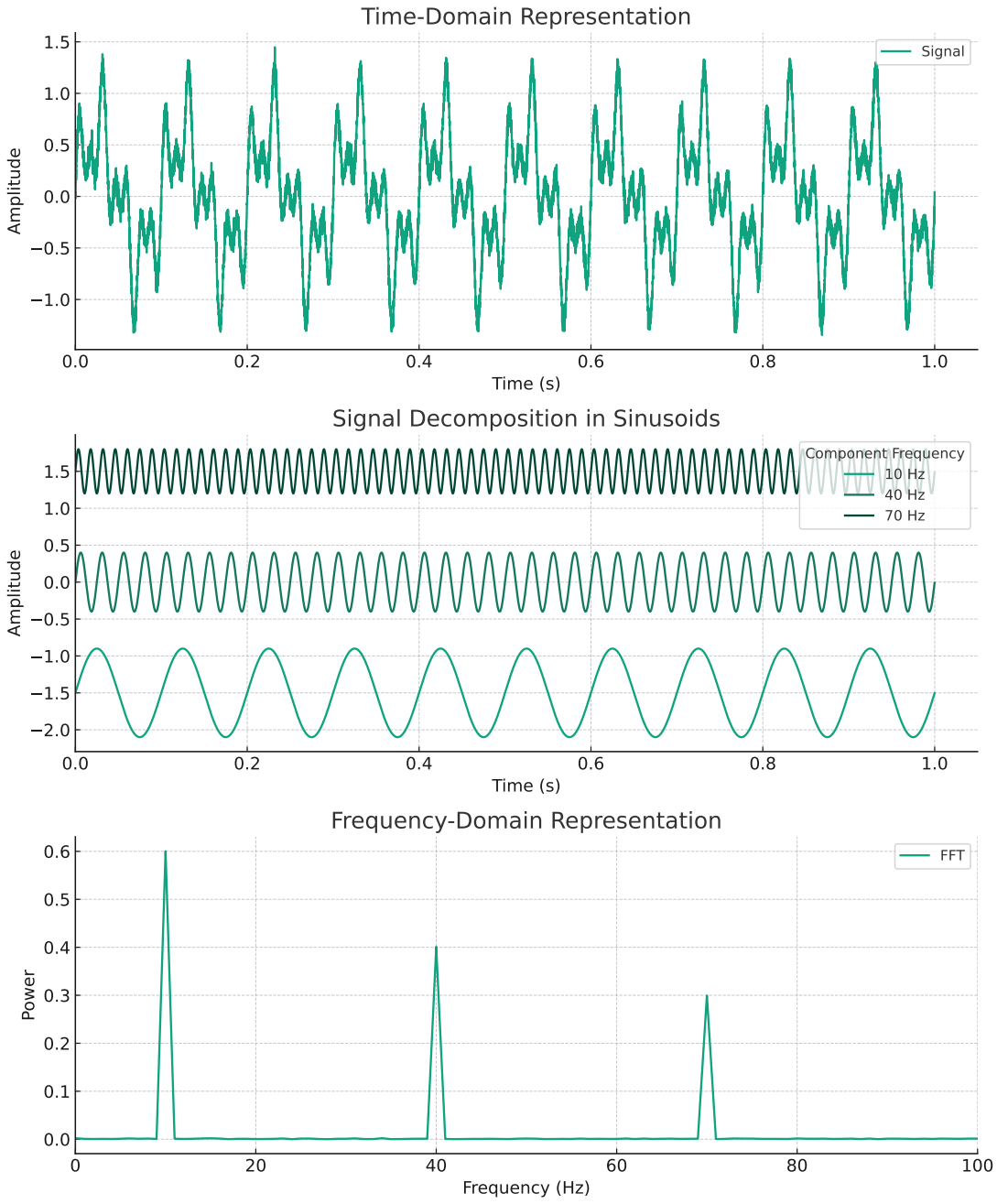


Figure 1.2.1: Time-frequency analysis of a signal. The top panel shows the original signal in the time domain. The middle panel depicts the signal decomposed into its sinusoidal components, with frequencies at 10 Hz, 40 Hz, and 70 Hz, respectively. The bottom panel presents the Fourier Transform (FT) of the signal, revealing its frequency content.

proves ideal for examining periodic signals. Different ranges of frequencies, called *frequency bands*, are often associated with various tasks and signals between different brain regions, revealing different aspects of brain activity.

For example, delta waves are typically observed during sleep, while alpha waves over occipital electrodes often appear when individuals close their eyes, thus blocking visual stimulation. Similarly, mu waves, while having a similar frequency band to alpha, are associated with motor activity and are typically focal over the motor cortex. However, these associations are not exclusive, and an increase in power in a specific frequency band could reflect various neurocognitive processes. Therefore, the context in which the EEG data were recorded is vital for accurate interpretation of these frequency bands, as well as their scalp localization where they are the most intense.

Table 1.1: Natural brain frequencies bands.

Band	**Frequency Range** (Hz)
Delta	<3
Theta	4–7
Alpha	8–14
Mu	8–12
Beta	15–30
Gamma	>30

Noise in EEG data and filtering The EEG signals that are most relevant to neuroscience research typically fall within the 1 Hz to 30 Hz frequency range [Newman et al., 2023]. However, EEG data inherently comes with noise from various sources, which show up as oscillating frequencies. While low-frequency noise may originate from head movements or electrode wires and appear as slow drifts in the signals, high-frequency noise might arise due to electromagnetic interference or muscle contractions, especially facial and neck ones. The frequencies of these noise sources may overlap with the crucial 1 Hz to 30 Hz EEG frequency band, thereby necessitating noise reduction to minimize the impact of noise on the signals of interest.

Reducing the signal's power at the frequencies above and below the range of experimental interest is called *filtering*. Filtering is typically performed at two stages: during data recording — *online filtering*, where a low pass filter by the amplifier attenuates high frequencies while passing lower frequencies through —, and during pre-processing. This approach is critical in digital recording to prevent *aliasing*, a phenomenon that causes high-frequency noise to appear as low-frequency artifacts in the data and occurs when a high-frequency signal is sampled at a much lower frequency.

The highest frequency that can be accurately recorded at a given sampling rate, known as the *Nyquist frequency* [Shannon, 1949], is defined as either $1/2$

or 1/3 of the sampling rate. In practice, the 1/3 rule is safer and used in most real-world situations where noise is unpredictable. In Event-Related Potentials (ERP) analysis, a high-pass (low frequency) cutoff of 0.1 Hz[5] and a low-pass (high frequency) cutoff of 30 Hz are usually applied. This is called a *bandpass filter*, as it preserves a "band" of frequencies between the high-pass and low-pass cutoffs. This setting strikes a balance between reducing artifacts outside the range of human EEG signals of interest and avoiding the induction of new artifacts. It should be noted that the high-pass cutoff must be much lower (by a factor of about 10) than the lowest frequency of interest, whereas the low-pass cutoff can be closer to the highest one.

Time-frequency visualization Time-domain and frequency-domain analyses offer valuable insights, but they cannot capture the full complexity of these signals, which often exhibit time-varying spectral characteristics [Cohen, 2014]. To address this issue, time-frequency analysis techniques, such as the Short-Time Fourier Transform (STFT) or wavelet transforms, are employed. These techniques provide a two-dimensional representation – known as a spectrogram, or scalogram in the case of wavelet – that reveals how the power of different frequency components in the signal evolves over time. In essence, these techniques extend the Fast Fourier Transform (FFT) to sliding windows of the signal, allowing for a dynamic view of the signal's spectral content [Mallat, 1999]. This is particularly useful in neuroscience, where transient, time-limited neural events – such as sensory responses, cognitive processing events, or epileptic spikes – may occur [Tallon-Baudry and Bertrand, 1999, Pfurtscheller and Da Silva, 1999]. These events can be better understood by observing their spectral content and how it changes over time. Thus, the time-frequency representation – as presented in fig. 1.2.2 – offers a powerful tool for the analysis and interpretation of M/EEG data, enabling researchers to capture the rich, time-varying spectral dynamics of neural activity.

1.2.2 Independent Component Analysis

ICA [Comon, 1994] forms a critical part of the pre-processing pipeline by separating the recorded signal into independent components. Each of these components

[5]The low frequency cutoff at 0.1 Hz is a standard value used in ERP analysis to reduce the influence of slow drifts and baseline shifts in the signal. By applying this high-pass filter, the signal is transformed so that it has a zero-mean, effectively removing the Direct Current (DC) component (frequency 0) which corresponds to the mean value of the signal. Indeed, the value of the frequency 0 component can be mathematically expressed as follow: $X(0) = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$, where N is the number of points in the signal, and $x[n]$ is the value of the signal at time n. This pre-processing step enhances the detection of event-related oscillatory components by minimizing potential distortions caused by low-frequency trends.

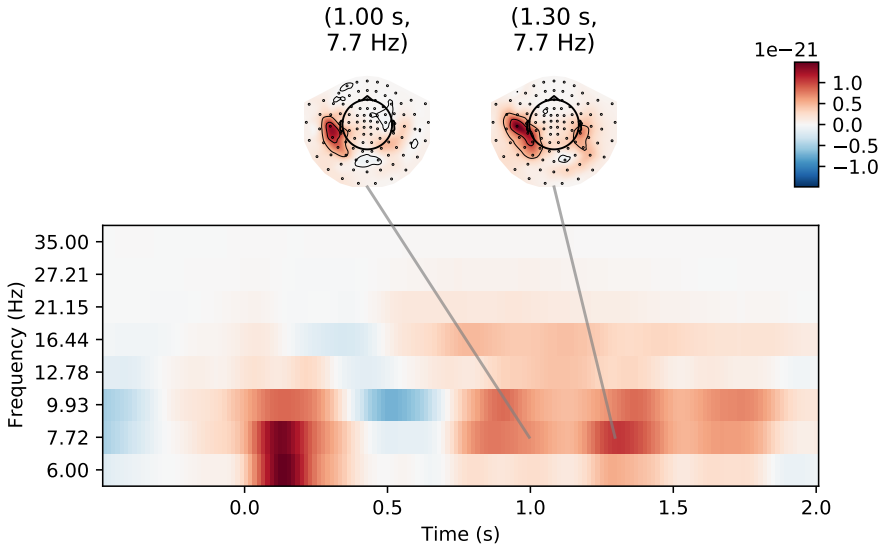


Figure 1.2.2: Time-frequency plane for epoched signals following a somatosensory stimulus (cue at time = 0), on MNE *somato* dataset, with overviews at 1 and 1.3 seconds and 7.7 Hz. Baseline correction applied from time point -0.5 until time point zero.

represents a full time course for a source signal. By manually identifying artifact components, it is possible to separate noise from the signal [Winkler et al., 2015].

More precisely, ICA is an algorithm for blind source separation. This task aims to extract independent source signals from a set of recordings where these signals are mixed together in unknown ratios. In the context of M/EEG analysis, this can be thought of as having several "microphones" (sensor channels) simultaneously recording many "instruments" (source signals like blinks, heartbeats, muscular activity, and brain activity) that are mixed together. ICA effectively separates these source signals based on their distinctive spatio-temporal properties.

Likewise, in EEG, the use of multiple electrodes is necessary for the successful application of ICA. The maximum number of ICA components is equivalent to the number of electrodes, though, in practice, the number of independent sources is typically less than the number of electrodes. ICA has found extensive use in EEG for identifying and removing artifacts, particularly ocular artifacts (blinks and eye movements) and muscle artifacts. Once these "noise" components are identified, they can be removed from the data without affecting the other components. Thus, the effects of artifacts can be eliminated from the data while preserving the EEG signals. ICA is particularly effective at capturing features in the data that account for the most variance, such as blinks and eye movements, which are larger than

EEG and therefore contain significant variance. Furthermore, low-frequency drift in the data explains substantial amounts of variance, making ICA most effective on data with more low-frequency power removed. Each independent component (IC) identified by ICA is a signal varying over time, with a "weighting" at each channel indicating the presence of that IC in the channel. These components can be visualized as scalp topography maps, showing where on the scalp the IC is most significant, as shown in fig. 1.2.3. Some ICs may be identified as ocular artifacts as they weigh most heavily around the front of the head.

While ICA is often used for its ability to reconstruct the original signals without identified artifacts, it comes with two key limitations. The first is the need for domain-specific knowledge to accurately identify artifacts. The second concerns the reconstruction process after removing some components. This step inevitably leads to an information loss across all channels, given that artifacts are generally shared across all sensors.

Mathematical formulation In the context of Independent Component Analysis (ICA), the observed data is considered to be a linear mixture of some unknown source signals. Let $X \in \mathbb{R}^{P \times T}$ be the observed data matrix, where P is the number of observed variables (*e.g.,* sensors in M/EEG) and T is the number of time points. The ICA model can be written as: $X = AS$, where $A \in \mathbb{R}^{P \times N}$ is the mixing matrix, N being the number of independent components, and $S \in \mathbb{R}^{N \times T}$ is the source matrix (independent components). The goal of ICA is to find an unmixing matrix $W \in \mathbb{R}^{N \times P}$ that transforms the observed data into signals that are as statistically independent from each other as possible: $S = WX$.

Once the independent components have been computed, some components can be visually identified and manually selected for removal, *e.g.,* those related to noise or artifacts. The reduced set of independent components is denoted as $S' \in \mathbb{R}^{N' \times T}$, where $N' \leq N$ is the number of remaining components after selection. The cleaned data $X' \in \mathbb{R}^{P \times T}$ can then be reconstructed by multiplying the selected components with the corresponding rows of the inverse of the unmixing matrix $W' \in \mathbb{R}^{N' \times P}$ and then adding back the mean $\mu \in \mathbb{R}^{P}$ of the original data: $X' = W'S' + \mu$.

1.2.3 Segmentation and epoching

Segmentation – or epoching – is another vital process in data pre-processing. An *epoch* is defined as a specific segment of time-bound data "centered" on a particular event. This event could range from a stimulus presentation to a specific behavioral response. The alignment of epochs to these specific events facilitates event-related analysis. The average of these epochs, often referred to as the *evoked* response,

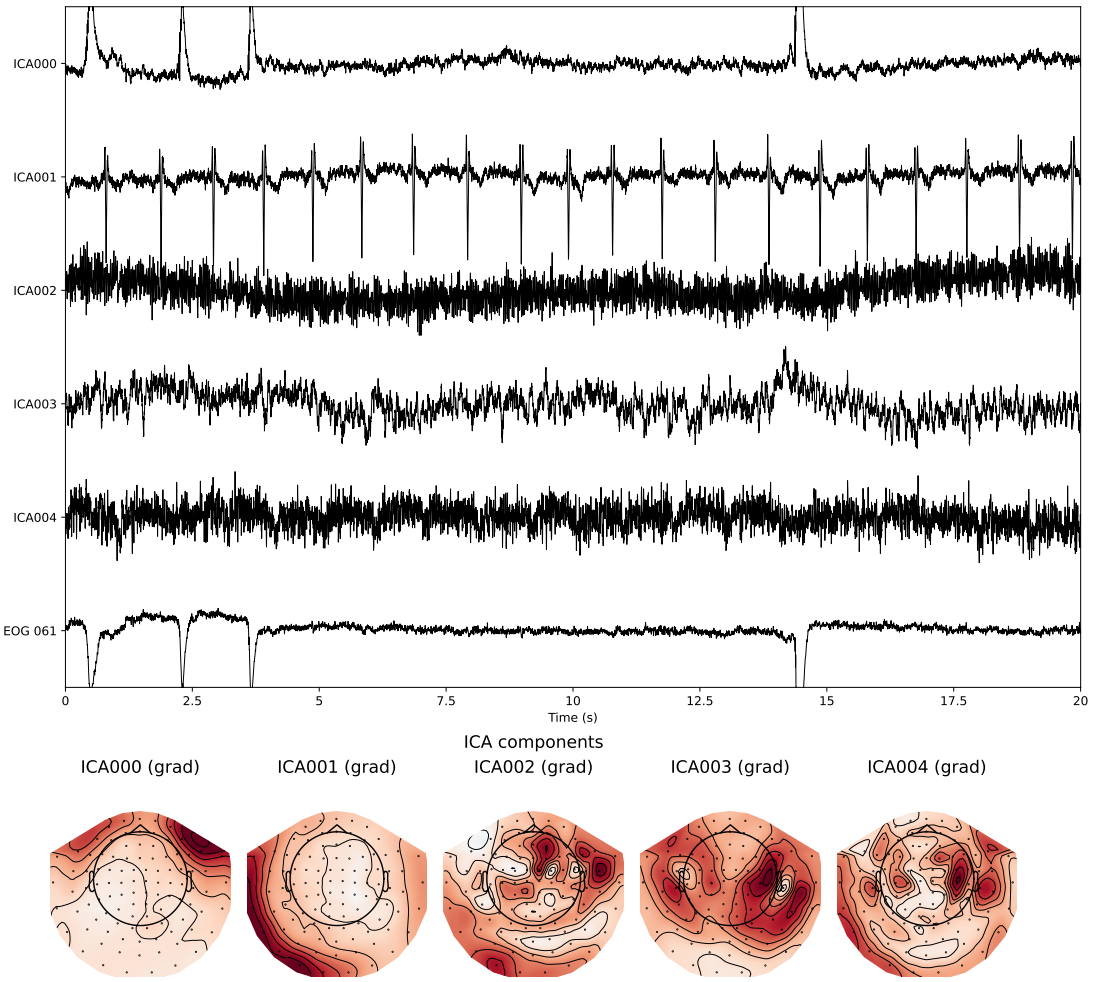


Figure 1.2.3: The first five independent components (over 15 fitted) alongside electrooculography (EOG) signal (top) and their corresponding topomaps (bottom), on MNE *sample* dataset. `ICA000` can be associated with the eye-blink artifact, whereas `ICA001` can be associated with the heartbeat artifact (this inference is drawn from the activation of sensors on the bottom-left of its topomap, which aligns with the heart's position relative to the head).

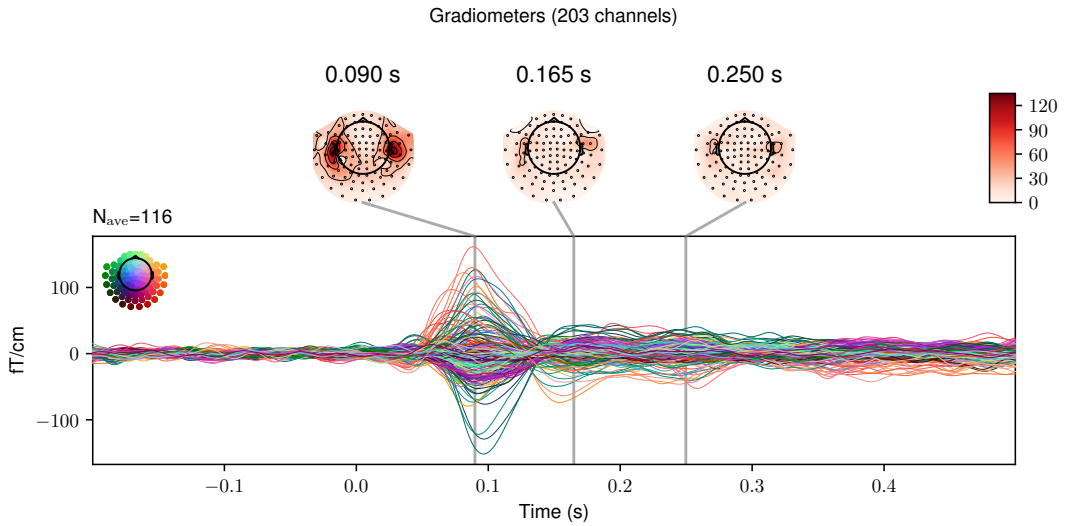


Figure 1.2.4: Evoked signals following an auditory stimulus (cue at time = 0), on MNE *sample* dataset. Baseline correction applied from beginning of the data until time point zero.

is frequently used for subsequent analysis. fig. 1.2.4 presents an example of such evoked signals on MNE *sample* dataset, obtained by averaging per channel, *i.e.,* per sensor, the segments "centered" – in fact 0.2 s before and 0.5 s after – around every auditory stimuli, left and right ones combined. A baseline correction has been applied before the stimulus, *i.e.,* the averaged value of the signal during this interval is calculated and then subtracted from the entire epoch, including both the pre-stimulus and post-stimulus periods. This effectively sets the average value of the signal during the baseline period to zero, allowing for a more accurate comparison of activity across different epochs and subjects. This correction serves to remove any slow drifts or shifts in the signal that are unrelated to the event of interest, ensuring that any observed changes in the signal after the event are not confounded by pre-existing trends or differences in baseline activity. By using a consistent baseline period across all epochs, researchers can more confidently attribute any observed differences in post-stimulus activity to the effects of the event itself, rather than to unrelated variations in the underlying signal. The choice of baseline period can depend on the specific research question and experimental design. Other studies may choose different baseline periods, depending on the timing of the events and the nature of the expected neural responses. Here, one can observe a strong response captured by sensors close to ears around 0.1 s after the stimuli.

Having discussed the intricacies of M/EEG data pre-processing and classical analysis techniques in neuroscience, we now turn our attention to more advanced

methods for information extraction. Specifically, we will explore how Convolutional Dictionary Learning (CDL) can be employed to analyze the cleaned and pre-processed M/EEG recordings.

Chapter 2

Background on Dictionary Learning

Contents

2.1 Motivation and sparse representation **34**
2.1.1 Sparsity and the Lasso 35
2.1.2 Lasso and its optimization 36
2.2 Dictionary Learning: mathematical formulation and optimization problem **38**
2.3 Convolutional Dictionary Learning **41**
2.4 Convolutional Dictionary Learning in neuroscience **44**
2.4.1 Rank-1 constraint 46

In the realm of data analysis and signal processing, developing efficient representations of data presents a key challenge. This quest becomes particularly complex when dealing with natural signals, such as images, audio, or magneto- and electroencephalography (M/EEG) recordings, as they are rarely sparse in their raw form. However, a closer examination reveals that these signals often possess an underlying structure that can allow for a more compact and meaningful representation. Emerging from the field of signal processing in the late 20th century, Dictionary Learning (DL) – a technique that stands at the intersection of sparsity and adaptability – has established itself as a cornerstone in this pursuit. As a powerful framework, DL offers the potential to create such representations, learning adaptable and efficient structures through the principle of *sparsity* [Aharon et al., 2006, Mairal et al., 2009].

Sparsity refers to the idea that signals, although living in high-dimensional spaces, can be represented succinctly as a linear combination of a few basic elements called *atoms*. These atoms are part of an overcomplete *dictionary*, a care-

fully chosen set that captures the inherent structure of the data [Olshausen and Field, 1997, Elad, 2010]. The sparser the representation, the better the dictionary [Le Magoarou and Gribonval, 2015, Le Magoarou et al., 2015]. For example, a M/EEG signal might consist of distinct chunks corresponding to heartbeats, eye-blinks, or visual stimulations. When viewed through the right "basis" or set of atoms, this signal becomes sparse and can be more easily understood and processed.

The application of this principle is not without challenges. First, one must identify the right family of atoms or a transformation that enables a sparse representation. This can be achieved either through expert knowledge or by learning directly from the data, a process known as *sparse coding* [Elad, 2010]. Second, the linear combination of atoms that constitutes each measurement must be recovered. While some analytical transforms like wavelets for images or Gaborlets for audio signals provide satisfactory results, the complexity and variability of signals often necessitate a more adaptive approach [Mallat, 1999].

This is where Dictionary Learning shines. Unlike traditional bases such as canonical bases or wavelets, DL learns the atoms *from* the data, leading to more adaptable and efficient representations [Aharon et al., 2006]. This adaptability is especially valuable in the analysis of natural images or biomedical applications like magnetoencephalography (MEG) and electrocardiograms (ECG), where specific patterns and structures are key to understanding the underlying phenomena [Cole and Voytek, 2017, Dupré la Tour et al., 2018, Xiang et al., 2018].

The field of Dictionary Learning has evolved to include various methods of decomposition, including the Lasso-based approach. This sparse representation paradigm offers significant benefits in terms of computational efficiency, interpretability, and even denoising, where it assists in removing noise not sparse in the same domain as the signal [Elad and Aharon, 2006].

2.1 Motivation and sparse representation

The mathematical nature of the inverse problems in Dictionary Learning can be articulated as follows. Let $x \in \mathbb{R}^n$ represent a real-valued signal of dimension n, and $y \in \mathbb{R}^m$ a real-valued measurement of dimension m. The forward model, denoted by $\mathcal{A}$, enables the computation of y from x, incorporating additive noise $b \in \mathbb{R}^m$ to model the degradation of the signal through the measurement device, leading to:

$$y = \mathcal{A}(x) + b \ . \tag{2.1.1}$$

When $\mathcal{A}$ is a linear operator[1], this relationship is expressed as:

$$y = Ax + b, \quad A \in \mathbb{R}^{m \times n} \ . \tag{2.1.2}$$

In the noiseless case, *i.e.,* $b = 0$, the inverse problem simplifies to matrix inversion. If $n = m$ and A is invertible, a unique and well-defined solution exists. However, complexities arise when these conditions are not met. A particularly pertinent scenario, especially in the context of M/EEG where the dimension of observation is smaller than the dimension of the signal, *i.e.,* $m < n$, is referred to as an *under-determined* problem. A common approach might involve seeking the Ordinary Least Square (OLS) solution, expressed as solving the following problem:

$$\min_{x \in \mathbb{R}^n} \|y - Ax\|_2^2 \ , \tag{2.1.3}$$

where $\|\cdot\|_2$ denotes the Euclidean ℓ_2 norm.

The Ordinary Least Square (OLS) solution, characterized as the orthogonal projection of y onto $(\ker A)^\perp$, leads to an ill-posed problem with infinite solutions. This complexity is compounded by the fact that noise cannot realistically be considered zero, especially in the case of Gaussian noise where the OLS solution often falls short. This insufficiency underlines the necessity of incorporating additional information or priors to construct a viable model. A strategic response to this issue is the introduction of regularization, where a specific function $\mathcal{R}$, reflecting the properties of the desired signal, is integrated into the model. The optimization problem we would like to solve becomes:

$$\min_{x \in \mathbb{R}^n} \|y - Ax\|_2^2 + \lambda \mathcal{R}(x) \ . \tag{2.1.4}$$

For example, Ridge – or Tikhonov regularization – [Ito and Jin, 2014], where $\mathcal{R}(x) \coloneqq \|x\|_2^2$, creates a strictly convex problem with a unique solution. Such regularization techniques, adaptable through hyperparameters, offer a pathway to penalize undesirable characteristics. The evolution of these strategies converges on the influential concept of *sparsity*.

2.1.1 Sparsity and the Lasso

Sparsity, where most entries of a vector or matrix are zero, is a crucial concept that facilitates the recovery process in signal processing, as it reduces the need for

[1]Though constraining $\mathcal{A}$ to be linear may appear restrictive, it is a realistic assumption in many applications, including imaging or neuroimaging.

extensive measurements. This idea is central to compressed sensing theory [Foucart and Rauhut, 2013] and has spawned a variety of optimization techniques. Within the context of under-determined scenarios, sparsity allows for the construction of effective penalization. One approach is to employ sparsity-inducing regularization $\mathcal{R}$, such as the ℓ_0 pseudo-norm, to constrain the number of non-zero coefficients in a signal [Elad, 2010]. This pseudo-norm is defined as follows:

$$\|x\|_0 = \#\{i,\ x_i \neq 0\} \quad , \tag{2.1.5}$$

where # denotes the cardinal operator. The set of indexes of non-zero coordinates is called the *support*. Then, the optimization problem becomes:

$$\min_{x \in \mathbb{R}^n} \|y - Ax\|_2^2 \quad \text{s.t.} \ \|x\|_0 \leq k_o \ , \tag{2.1.6}$$

where k_0 is a pre-determined upper bound on the number of non-zero coefficients in x. Though this leads to a non-convex and NP-hard problem, various heuristics and algorithms, like matching pursuit [Mallat and Zhang, 1994], have been developed to tackle it [Elad, 2010].

An alternative strategy involves convex relaxation that promotes sparsity, using ℓ_p norms, resulting in a convex optimization problem. ℓ_p norms are defined as follows:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} , \quad p \geq 1 \ . \tag{2.1.7}$$

A well-known example of this approach is the Least Absolute Shrinkage and Selection Operator (Lasso; Tibshirani 1996) problem, where the cost function is convex and bounded, leading to a convex set of solutions, using ℓ_1 norm as regularization:

$$\min_{x \in \mathbb{R}^n} \|y - Ax\|_2^2 + \lambda \|x\|_1 \ . \tag{2.1.8}$$

This convexity ensures that the solution set contains either a unique solution or an infinite number of solutions, providing a more tractable pathway to exploit sparsity.

2.1.2 Lasso and its optimization

The Lasso optimization problem expressed in eq. (2.1.8) can be expressed as a more general one:

$$\min_{x \in \mathbb{R}^n} L(x) + \lambda \mathcal{R}(x) \ , \tag{2.1.9}$$

where L measure the goodness of fit and is supposed to be differentiable, $\mathcal{L}$-smooth and convex; $\mathcal{R}$ penalizes the complexity and is supposed to be convex and "easy to

optimize"; and where $\lambda > 0$ is the regularization parameter that controls tradeoff between fit and complexity. A solution x^* of this problem must verify the following optimality condition:

$$0 \in \partial \left(L\left(x^*\right) + \lambda \mathcal{R}\left(x^*\right) \right) \tag{2.1.10}$$

$$\Leftrightarrow 0 \in \nabla L\left(x^*\right) + \lambda \partial \mathcal{R}\left(x^*\right) \tag{2.1.11}$$

$$\Leftrightarrow -\nabla L\left(x^*\right) \in \lambda \partial \mathcal{R}\left(x^*\right) \ , \tag{2.1.12}$$

where ∇f denotes, when its exists, the full gradient of f and ∂f denotes the sub-gradient of f, defined as follows, for $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ convex:

$$\partial f(x) \coloneqq \{g \in \mathbb{R}^n,\ f(y) \geq f(x) + \langle g, y - x \rangle, \forall y \in \mathrm{dom}(f)\} \ . \tag{2.1.13}$$

Applied to the Lasso problem – when $\mathcal{R}\left(\cdot\right) = \left\|\cdot\right\|_1$ – , it gives that:

$$-A^\top \left(Ax^* - y\right) \in \partial \left\|x^*\right\|_1 \ . \tag{2.1.14}$$

As the ℓ_1 norm of x is given by $\left\|x\right\|_1 = \sum_{i=1}^n |x_i|$, the subgradient of the ℓ_1 norm at a point x can be described as a set of vectors, and its calculation depends on the components of x. For each component x_i, the partial subgradient with respect to that component is:

$$\partial |x_i| = \begin{cases} 1 & \text{if } x_i > 0 \\ -1 & \text{if } x_i < 0 \\ [-1, 1] & \text{if } x_i = 0 \end{cases} \tag{2.1.15}$$

Thus, the subgradient of $\left\|x\right\|_1$ is the set of all vectors that can be formed by selecting a value for each component according to the above rules:

$$\partial \left\|x\right\|_1 = \left(\partial |x_1|, \ldots, \partial |x_n|\right) \ . \tag{2.1.16}$$

If all components of x are non-zero, the subgradient is a single vector, and it has the same sign as the corresponding components of x. If some components are zero, the subgradient is a set of vectors, reflecting the ambiguity in the derivative at those points.

Using the definition of smoothness of L, we have that, $\forall x, y \in \mathbb{R}^n$:

$$L(x) + \lambda \mathcal{R}(x) \leq L(y) + \langle \nabla L(y), x - y \rangle + \frac{\mathcal{L}}{2} \left\|x - y\right\|_2^2 + \lambda \mathcal{R}(x) \ . \tag{2.1.17}$$

By setting $y \coloneqq x^{(t)}$, minimizing the right-hand side of eq. (2.1.17) in x suggests a iterative procedure, known as the *Proximal Gradient Descent*:

$$x^{(t+1)} = \mathrm{prox}_{\frac{\lambda}{\mathcal{L}}\mathcal{R}} \left(x^{(t)} - \frac{1}{\mathcal{L}} \nabla L\left(x^{(t)}\right) \right) \ , \tag{2.1.18}$$

where prox denotes the proximal operator. It is defined as follows:

$$\text{prox}_{\gamma\mathcal{R}}(y) = \arg\min_x \frac{1}{2}\|x-y\|_2^2 + \gamma\mathcal{R}(x) \ . \tag{2.1.19}$$

The proximal operator for the ℓ_1 norm regularization, as in the Lasso problem, is known as the *soft-thresholding* operator. It is given by the following expression:

$$\text{ST}_\lambda(x) \coloneqq \text{prox}_{\lambda\|\cdot\|_1}(x) = \text{sign}(x)\max\left(|x|-\lambda, 0\right) \tag{2.1.20}$$

where the operations are applied element-wise. The soft-thresholding operator shrinks the absolute value of each component of the input vector by λ and retains the sign. Introduced in Daubechies et al. [2004], the Iterative Shrinkage Thresholding Algorithm (ISTA) produces the following procedure:

$$x^{(t+1)} = \text{ST}_{\tau\lambda}\left(x^{(t)} - \tau A^\top(Ax^{(t)} - y)\right) \ , \tag{2.1.21}$$

where $\tau > 0$ is the step size. ISTA has the advantage of simplicity and the ability to handle non-smooth regularization terms like the ℓ_1-norm. It converges to a solution of the Lasso problem under suitable conditions on the step size τ – for $0 < \tau < \frac{2}{L}$, L being the highest eigenvalue of $A^\top A$ – and other parameters. This algorithm is a foundational building block in sparse recovery and has inspired several variations and extensions, including the faster FISTA (Fast ISTA; Beck and Teboulle 2009) algorithm, that is an adaptation of gradient descent with momentum, described in algorithm 1. Other algorithms to solve the Lasso are presented in Hastie et al. [2015].

Algorithm 1: FISTA with constant stepsize [Beck and Teboulle, 2009]

Input $\text{prox}_{\frac{\lambda\mathcal{R}}{\mathcal{L}}}(\cdot)$, $\nabla L(\cdot)$, $\lambda > 0$ and $\mathcal{L}$;
Set $x^{(1)} = 0_{\mathbb{R}^n}$, $w^{(1)} = 0_{\mathbb{R}^n}$, $\beta^{(1)} = 1$;
for $t = 1, \ldots, T$ **do**
    $x^{(t+1)} = \text{prox}_{\frac{\lambda\mathcal{R}}{\mathcal{L}}}\left(w^{(t)} - \frac{1}{\mathcal{L}}\nabla L\left(w^{(t)}\right)\right)$;
    $\beta^{(t+1)} = \frac{1+\sqrt{1+4\left(\beta^{(t)}\right)^2}}{2}$;
    $w^{(t+1)} = x^{(t+1)} + \frac{\beta^{(t)}-1}{\beta^{(t+1)}}\left(x^{(t+1)} - x^{(t)}\right)$;
end
return $x^{(T+1)}$

2.2 Dictionary Learning: mathematical formulation and optimization problem

Dictionary Learning (DL) is typically formulated as an optimization problem with the objective of finding a dictionary that yields the sparsest representation of

the given data, often subject to a reconstruction error constraint [Aharon et al., 2006]. Thus, the mathematical formulation of the DL problem can be expressed as follows. Given a set of training signals $X \in \mathbb{R}^{N\times T}$, the objective is to find a dictionary $D = [\mathbf{d}_1\mathbf{d}_2\dots\mathbf{d}_K] \in \mathbb{R}^{N\times K}$, with $\mathbf{d}_k \in \mathbb{R}^N$ and a sparse code matrix $Z = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_K] \in \mathbb{R}^{K\times T}$ that minimize the following cost function:

$$\min_{D\in\mathcal{C}, Z\in\mathbb{R}^{K\times T}} F(D, Z; X) \coloneqq \frac{1}{2}\|X - DZ\|_F^2 + \lambda \sum_{k=1}^{K} \|\mathbf{z}_k\|_1 \quad , \tag{2.2.1}$$

where $\mathcal{C}$ is a convex set of constraints for D. $\|.\|_F$ denotes the Frobenius norm, defined as follows:

$$\|X\|_F \coloneqq \sqrt{\sum_{n=1}^{N}\sum_{t=1}^{T} |X_{n,t}|^2} \quad . \tag{2.2.2}$$

The reconstructed signal $\hat{X} \coloneqq \hat{D}\hat{Z}$ is invariant to several transformations of $\left(\hat{D}, \hat{Z}\right) \in \left(\mathcal{C} \times \mathbb{R}^{K\times T}\right)$, where $\hat{D}, \hat{Z}$ denote the arg min solutions of eq. (2.2.1).

Permutation invariant

Let $P \in \mathbb{R}^{K\times K}$ be a permutation matrix, *i.e.,* the identity matrix I_K with permuted columns, and $(D, Z) \in \left(\mathcal{C} \times \mathbb{R}^{K\times T}\right)$. Then $\hat{X} = \hat{D}\hat{Z} = \hat{D}P^\top P\hat{Z}$. Therefore, $\left(\hat{D}, \hat{Z}\right)$ and $\left(\hat{D}P^\top, P\hat{Z}\right)$ are equivalent representations of the signal, and $F\left(\hat{D}, \hat{Z}\right) = F\left(\hat{D}P^\top, P\hat{Z}\right)$.

Sign invariant

Let $S \in \mathbb{R}^{K\times K}$ be a sign change matrix, *i.e.,* a diagonal matrix with coefficients 1 and -1, and $(D, Z) \in \left(\mathcal{C} \times \mathbb{R}^{K\times T}\right)$. Then $\hat{X} = \hat{D}\hat{Z} = \hat{D}SS\hat{Z}$. Therefore, $\left(\hat{D}, \hat{Z}\right)$ and $\left(\hat{D}S, S\hat{Z}\right)$ are equivalent representations of the signal, and $F\left(\hat{D}, \hat{Z}\right) = F\left(\hat{D}S, S\hat{Z}\right)$.

Scale invariant

Let $\alpha > 0$, and $(D, Z) \in \left(\mathcal{C} \times \mathbb{R}^{K\times T}\right)$. Then $\hat{X} = \hat{D}\hat{Z} = \alpha\hat{D}\frac{1}{\alpha}\hat{Z}$. Therefore, $\left(\hat{D}, \hat{Z}\right)$ and $\left(\alpha\hat{D}, \frac{1}{\alpha}\hat{Z}\right)$ are equivalent representations of the signal. However, if

$\alpha > 1$, then $F\left(\hat{D}, \hat{Z}\right) > F\left(\alpha\hat{D}, \frac{1}{\alpha}\hat{Z}\right)$ and the optimization problem in Equation 2.2.1 will prefer solutions with smaller ℓ_1 norm, and thus α as big as possible. To alleviate this issue, the constraint set $\mathcal{C}$ generally includes a normalization criterion, where each atom of D is constrained to belong to a ball of fixed radius. Thus, the optimization problem becomes:

$$\min_{D\in\mathbb{R}^{N\times K}, Z\in\mathbb{R}^{K\times T}} \frac{1}{2}\|X - DZ\|_F^2 + \lambda \sum_{k=1}^{K} \|\mathbf{z}_k\|_1 \quad \text{s.t. } \forall k, \|\mathbf{d}_k\|_2 \leq 1 \ , \tag{2.2.3}$$

where constraint on the dictionary atoms $\mathbf{d}_k$ ensures that they have unit norm, which avoids trivial solutions as mentioned.

All this shows that the optimization problem in Equation 2.2.1 is highly non-convex. Indeed, for each solution (D, Z), there are at least $2^K K! - 1$ other equivalent solutions given by permutations and change of sign of coordinates of Z and columns of D.

The optimization problem is typically solved using an iterative approach, where two steps are alternated: *sparse coding*, where for a fixed dictionary D, the sparse code Z is updated; and *dictionary update*, where for a fixed sparse code Z, the dictionary D is updated. Despite the non-convex nature of the global problem, various efficient and effective algorithms have been proposed to solve each of these steps, including Orthogonal Matching Pursuit (OMP) for sparse coding and K-SVD for dictionary update [Aharon et al., 2006], as well as online dictionary learning [Mairal et al., 2009] and Proximal Alternating Linearized Minimization (PALM; Bolte et al. 2013).

The most simple example is the Method of Optimal Direction (MOD; Engan et al. 1999). It consists of minimizing F over Z with a sparse coding algorithm like FISTA (*cf.* algorithm 1), and then performing a projected gradient descent over D, with a sequence $\left(\tau^{(t)}\right)_{1\leq t\leq T}$ of step sizes:

$$Z^{(t+1)} = \arg\min_{Z\in\mathbb{R}^{K\times T}} \frac{1}{2}\|X - DZ\|_F^2 + \lambda \sum_{k=1}^{K} \|\mathbf{z}_k\|_1 \tag{2.2.4}$$

$$D^{(t+1)} = \text{proj}_{\mathcal{C}}\left(D^{(t)} - \tau^{(t)}\left(D^{(t)}Z^{(t+1)} - X\right)Z^{(t+1)^\top}\right) \tag{2.2.5}$$

The process of initializing the dictionary is pivotal in the success of alternating minimization algorithms, particularly due to the non-convex nature of the problem. Agarwal et al. [2016] give theoretical convergence guarantees regarding the "basin of attraction" of the true solution and establish that alternating minimization succeeds in its recovery when a dictionary is initialized with a limited error,

inversely proportional to the sparsity level. Common strategies for this initialization include employing random values or utilizing segments – "chunks" – of signals as starting points. Moreover, the selection of the hyperparameter λ significantly influences the algorithm's performance. Determining an optimal value for λ can be especially challenging in unsupervised scenarios, where there may be limited guidance or constraints to inform this choice.

2.3 Convolutional Dictionary Learning

Traditional Dictionary Learning is not well suited to handle the local structure and shift-invariance that are common in images and other high-dimensional signals. This led to the development of convolutional dictionary learning (CDL) which extends traditional dictionary learning by introducing the convolution operation and provides an even more efficient representation for signals that have a temporal or spatial structure. In CDL, the data is modeled as a sparse linear combination of convolutions between dictionary atoms and shift-invariant codes. This paradigm allows signals to be represented as convolutions of sparse activation maps with dictionary filters, capturing local and shift-invariant structures in the data [Papyan et al., 2017, Zeiler et al., 2010]. These convolutional dictionaries have the advantage of being able to represent signals with fewer parameters due to shared filter usage, making them more efficient and scalable for large datasets [Bristow et al., 2013].

The concept of CDL was first introduced by Grosse et al. [2007] in the context of audio classification. This approach was further expanded and refined in the subsequent years, with significant contributions from researchers such as Mairal et al. [2009] and Sulam et al. [2018], among others. CDL has since found applications in a wide range of areas, including image and video processing, audio signal processing, and neuroscience.

The mathematical formulation of CDL is similar to that of dictionary learning, but with the inner product replaced by convolution, and with appropriate modifications to the constraints and regularization term. Here, and for what follows, the adopted formulation and notations are adapted from Moreau et al. [2018] and Dupré la Tour et al. [2018], that follows Grosse et al. [2007]. We thus denote the value of signals at time $t \in [\![0\,, T-1]\!]$ into brackets, *i.e.,* for $\mathbf{x} \in \mathbb{R}^T$, $\mathbf{x}[t] \in \mathbb{R}$ and for $\mathbf{X} \in \mathbb{R}^{P\times T}$, $X[t] \in \mathbb{R}^P$. Note that, $\forall t \notin [0\,, T-1]\,, \mathbf{x}[t] = 0$ and $X[t] = 0_{\mathbb{R}^P}$. The convolution of two signals $\mathbf{z} \in \mathbb{R}^{T-L+1}$ and $\mathbf{d} \in \mathbb{R}^L$ is denoted by $\mathbf{z} * \mathbf{d} \in \mathbb{R}^T$ and is defined by:

$$\forall t \in [\![0\,, T-1]\!]\,, \quad (\mathbf{z} * \mathbf{d})\,[t] \coloneqq \sum_{\tau=0}^{L-1} \mathbf{z}[t-\tau]\mathbf{d}[\tau] \ . \tag{2.3.1}$$

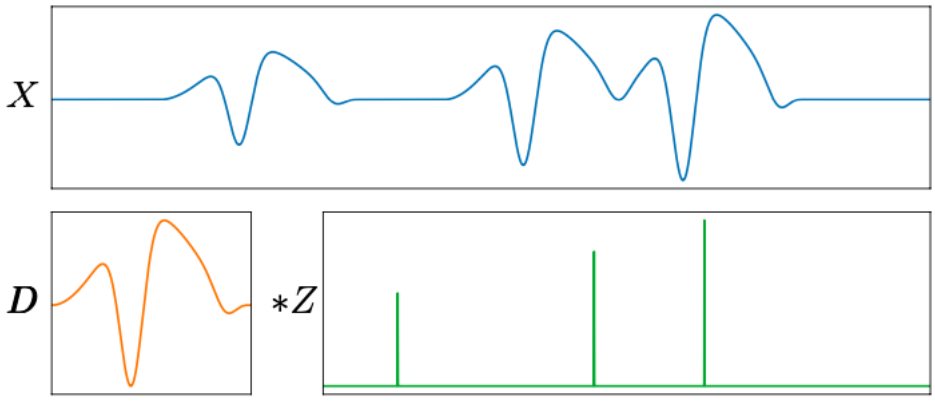


Figure 2.3.1: Decomposition of a noiseless univariate signal X (blue) as the convolution $Z * \boldsymbol{D}$ between a temporal pattern $\boldsymbol{D}$ (orange) and a sparse activation signal Z (green). [Moreau and Gramfort, 2019]

Thus, for N measures of a univariate signal $X = \left\{\mathbf{x}^n \in \mathbb{R}^T, n = 1, \ldots, N\right\}$, the goal is to recover a dictionary of K atoms $D = \left\{\mathbf{d}_k \in \mathbb{R}^L, k = 1, \ldots, K\right\}$ and, for each measurement n, its associated sparse codes $Z^n = \left\{\mathbf{z}_k^n \in \mathbb{R}^{\widetilde{T}}, k = 1, \ldots, K\right\}$, by solving the following optimization problem:

$$\min_{D \in \mathcal{C}, \mathbf{z}_k^n \in \mathbb{R}^{\widetilde{T}}} F(D, Z; X) \coloneqq \sum_{n=1}^{N} \left(\frac{1}{2} \left\| \mathbf{x}^n - \sum_{k=1}^{K} \mathbf{z}_k^n * \mathbf{d}_k \right\|_2^2 + \lambda \sum_{k=1}^{K} \|\mathbf{z}_k^n\|_1 \right) \quad , \tag{2.3.2}$$

where $\mathcal{C} = \left\{D \in \mathbb{R}^{K \times L},\ \|\mathbf{d}_k\|_2 \leq 1, k = 1, \ldots, K\right\}$ is the set of dictionaries composed of filters with unit norm, and where $\mathbf{z}_k^n * \mathbf{d}_k \in \mathbb{R}^T$, with $\widetilde{T} \coloneqq T - L + 1$. With a slight abuse of notation, Z represents the set of all sparse codes $Z^n, n = 1, \ldots, N$.

Convolutional Dictionary learning can be written as a bi-level optimization problem to minimize the cost function with respect to the dictionary only, as mentioned in Mairal et al. [2009], by solving

$$\begin{aligned} &\min_{D \in \mathcal{C}} \ G(D; X) \coloneqq F(D, Z^*(D); X) \\ &\text{with } Z^*(D) \coloneqq \arg\min_{Z} F(D, Z; X) \ . \end{aligned} \tag{2.3.3}$$

Computing the sparse codes $Z^*(D)$ is often referred to as the *inner problem*, while the global minimization is the *outer problem*.

Conventional approaches to dictionary learning address this bi-convex optimization challenge through the *Alternating Minimization* (AM) technique [Mairal et al., 2009]. The procedure iteratively refines two distinct components. Initially, the cost function F is minimized over Z while keeping the dictionary D fixed. This

is commonly achieved using sparse coding algorithms such as (F)ISTA [Daubechies et al., 2004, Beck and Teboulle, 2009], coordinate descent [Wu et al., 2008], or ADMM [Boyd et al., 2011]. Subsequently, the second stage either employs the gradient $\nabla_1 F(D, Z; X)$ – where ∇_1 specifies the gradient with respect with the first variable in F – to execute one or more steps of projected gradient descent for dictionary refinement, or it directly identifies the optimal D by solving a least squares problem, which typically involves the computation of a pseudo-inverse.

As mentioned, the main advantage of CDL over traditional dictionary learning is its ability to capture spatial and temporal dependencies in data, making it particularly effective for image and audio signals. However, CDL involves a higher computational cost due to the convolution operation $z * d$, especially when the size of the signal gets large. Empirically, CDL has been shown to outperform traditional dictionary learning in tasks such as image denoising and audio signal separation [Grosse et al., 2007, Sulam et al., 2018].

Initialization strategies and optimization techniques for efficient learning As for classical Dictionary Learning, initialization plays a crucial role in the performance and convergence of the dictionary learning algorithms. There are two main strategies for initialization. The first is random initialization, where the initial dictionary is populated with random values typically drawn from a Gaussian distribution. The second strategy is the "chunk" strategy, where the initial dictionary is composed of segments ("chunks") of the original signal.

Additionally, an important aspect of the optimization process is resampling. Resampling comes into play when an atom in the dictionary is underutilized, *i.e.,* not used, or only used once (which is particularly likely if the atom was initialized using the chunk strategy, as it might be represented at least once in the signal). In such cases, the ineffective atom is discarded and replaced with a new one. The new atom is generated based on the initialization strategy: it could be another random atom or a new chunk of the original signal. This resampling process ensures that all atoms in the dictionary are effectively contributing to the signal representation, thereby enhancing the overall efficiency and performance of the dictionary learning algorithm.

Trade-offs between dictionary size, sparsity, computational time, and reconstruction quality Choosing the appropriate dictionary size, degree of sparsity, and computational resources involves significant trade-offs. Larger dictionaries and higher levels of sparsity can lead to better reconstruction quality but at the expense of increased computational time. An essential aspect of this trade-off is the selection of the hyperparameter λ, which regulates the balance between

reconstruction fidelity and sparsity. Finding an optimal value for λ is particularly challenging, especially in unsupervised scenarios where there is limited guidance. One effective method to address this challenge is through cross-validation, where different values of λ can be systematically evaluated to determine the one that offers the best compromise between accuracy, sparsity, and computational efficiency. This approach is crucial in settings where the optimal balance between dictionary size, sparsity, computational time, and reconstruction quality must be achieved for the specific requirements of the application [Aharon et al., 2006]. Therefore, the choice of λ through cross-validation becomes a pivotal factor in tuning the algorithm for desired performance.

2.4 Convolutional Dictionary Learning in neuroscience

Dictionary learning, and in particular convolutional dictionary learning, has shown promising results in the field of neuroscience. The nature of neural data, especially electroencephalogram (EEG) and magnetoencephalogram (MEG) data, is such that it exhibits temporal structures, making it a good candidate for CDL. For example, in the analysis of M/EEG data, the goal is often to identify the underlying sources in the brain that gave rise to the recorded signals. Since these sources often have a temporal structure, CDL can provide a more efficient representation of the data, making it easier to identify the sources [Dupré la Tour et al., 2018].

The data recorded from one subject via M/EEG is complex, as presented in fig. 1.1.10. Most observed MEG signal $X \in \mathbb{R}^{P \times T}$ – with P sensors also called *channels*, and T timestamps – contains heavy noise bursts and have low signal-to-noise ratio [Jas et al., 2017]. Thus, we cannot work directly with this result, we have to pre-process it in some way. Neural time-series data contain a wide variety of prototypical signal waveforms – atoms – that are of significant importance in clinical and cognitive research. In order to analyze such data, one of the goals is hence to extract such "shift-invariant" atoms[2], as events can happen at any instant [Jas et al., 2017]. While alpha waves (8 Hz to 14 Hz, *cf.* table 1.1) are known to closely resemble short sinusoids, and thus are revealed by Fourier analysis or wavelet transforms, there is an evolving debate that electromagnetic neural signals are composed of more complex waveforms that cannot be analyzed by linear filters and traditional signal representations [Cole and Voytek, 2017, Dupré la Tour et al.,

[2]In the context of CDL and the decomposition of neural signals, the time-invariant property implies that the spatio-temporal atoms (or patterns) learned from the signal are consistent across different time points. This means that the same patterns can be observed at various time shifts within the signal, and these patterns are not dependent on the specific time instance.

2018]. Such patterns in electrophysiological signals can be extracted efficiently in an unsupervised way using CDL [Barthélemy et al., 2013, Dupré la Tour et al., 2018].

This approach provides a unsupervised data-driven way to uncover meaningful and interpretable features in the data, paving the way for new insights into neural processes. For brain signals, specific techniques have been developed based on *convolutional sparse coding* (CSC) [Jas et al., 2017, Dupré la Tour et al., 2018, Moreau and Gramfort, 2019]. This method aims at finding a dictionary of atoms and some associated activation vectors, in order to recover the original signal X by doing a convolution between the dictionary of atoms $\mathbf{D}$ and their sparse activation vectors Z, as shown in Figure 2.3.1. Similarly as eq. (2.3.1), the convolution between $\mathbf{z} \in \mathbb{R}^{T-L+1}$ and $D \in \mathbb{R}^{P \times L}$ is denoted by $\mathbf{z} * D \in \mathbb{R}^{P \times T}$ and obtained by convolving every row of D by $\mathbf{z}$, *i.e.,* it is defined by:

$$\forall t \in [\![0\,, T-1]\!]\,, \quad (\mathbf{z} * D)\,[t] \coloneqq \sum_{\tau=0}^{L-1} \langle \mathbf{z}[t-\tau], D[\tau] \rangle \quad . \tag{2.4.1}$$

We now are in the case of multivariate signals, thus the cost function from eq. (2.3.2) is adapted, and the optimization problem is as follows,

$$\begin{aligned} \min_{D_k \in \mathbb{R}^{P \times L}, \mathbf{z}_k^n \in \mathbb{R}^{\tilde{T}}} \sum_{n=1}^{N} \left(\frac{1}{2} \left\| X^n - \sum_{k=1}^{K} \mathbf{z}_k^n * D_k \right\|_F^2 + \lambda \sum_{k=1}^{K} \|\mathbf{z}_k^n\|_1 \right) \\ \text{s.t.} \quad \|D_k\|_F^2 \leq 1 \text{ and } \mathbf{z}_k^n \geq 0_{\mathbb{R}^{\tilde{T}}} \quad , \end{aligned} \tag{2.4.2}$$

where $\mathbf{X} = \left\{ X^n \in \mathbb{R}^{P \times T}, n = 1, \ldots, N \right\}$ are N observed multivariate signals of length T that are recorded over P channels (mapping to space locations), $\lambda > 0$ is the regularization parameter, $\mathbf{D} = \left\{ D_k \in \mathbb{R}^{P \times L}, k = 1, \ldots, K \right\}$ are the K spatio-temporal atoms constituting the dictionary, $\mathbf{Z} = \left\{ Z^n \in \mathbb{R}^{K \times \widetilde{T}}, n = 1, \ldots, N \right\}$ the set of all sparse codes, with $Z^n = \left\{ \mathbf{z}_k^n \in \mathbb{R}^{\widetilde{T}}, k = 1, \ldots, K \right\}$ the K sparse signals of activations associated with X^n, with $\widetilde{T} \coloneqq T - L + 1$. The element-wise constraint on $\mathbf{z}_k^n$ comes from the fact that it is assumed that its entries are positive, which means that the temporal patterns are present each time with the same polarity [Dupré la Tour et al., 2018].

This problem is bi-convex and solved with alternate minimization: first, given K fixed atoms D_k and a regularization parameter $\lambda > 0$, retrieve the NK activation signals z_k^n associated to the signals X^n, *e.g.,* by locally greedy coordinate descent (LGCD); then, given NK fixed activation signals $\mathbf{z}_k^n$, update the K spatial patterns D_k, and so forth.

2.4.1 Rank-1 constraint

The incorporation of a rank-1 constraint into multivariate CDL is motivated by the physical properties governing electrophysiological signals such as EEG and MEG. Rooted in Maxwell's equations, the physical model dictates that each sensor instantaneously receives a linear transformation of every source emanating from the brain. Importantly, this transformation is characterized by a constant topographic map; that is, a signal originating from the same neural source but captured at disparate time points will be projected across the sensor array via the same linear transformation. As electromagnetic waves traverse the brain at the speed of light, all sensors register the same waveform simultaneously, albeit with varying intensity levels due to factors like tissue conductivity and sensor-source distance. To model these nuances effectively, Dupré la Tour et al. [2018] advocate for the use of multivariate CSC with a rank-1 constraint. This constraint ensures that each learned atom possesses distinct spatial and temporal patterns, as depicted in Figure 2.4.1. Mathematically, the rank-1 constraint formalizes each atom $D_k \in \mathbb{R}^{P \times L}$ as the outer product $\mathbf{u}_k \mathbf{v}_k^\top$, where $\mathbf{u}_k \in \mathbb{R}^P$ encapsulates the spatial activations across channels and $\mathbf{v}_k \in \mathbb{R}^L$ embodies the temporal evolution. This constraint not only adheres to the underlying physical model but also significantly constrains the solution space to more realistic and interpretable dictionary atoms.

Thus, the $\|D_k\|_F^2 \leq 1$ constraint in (2.4.2) is now replaced by $\|\mathbf{u}_k\|_2^2 \leq 1$ and $\|\mathbf{v}_k\|_2^2 \leq 1$, making the problem tri-convex:

$$\begin{aligned} \min_{\mathbf{u}_k \in \mathbb{R}^P, \mathbf{v}_k \in \mathbb{R}^L, \mathbf{z}_k^n \in \mathbb{R}^{\tilde{T}}} & \sum_{n=1}^{N} \left(\frac{1}{2} \left\| X^n - \sum_{k=1}^{K} \mathbf{z}_k^n * \left(\mathbf{u}_k \mathbf{v}_k^\top \right) \right\|_F^2 + \lambda \sum_{k=1}^{K} \|\mathbf{z}_k^n\|_1 \right) \\ \text{s.t.} \quad & \|\mathbf{u}_k\|_2^2 \leq 1, \|\mathbf{v}_k\|_2^2 \leq 1 \text{ and } \mathbf{z}_k^n \geq 0_{R^{\tilde{T}}} \ . \end{aligned} \tag{2.4.3}$$

Despite these successes, there are still many challenges and open questions in the application of DL and CDL to neuroscience data. One challenge is the high dimensionality and complexity of the data, which makes the dictionary learning problem more difficult to solve. Another challenge is the interpretability of the learned dictionaries. While DL provides a data-driven and adaptive basis for representing the data, it does not provide an explicit model of the data generation process, which can be a disadvantage in certain applications [Mairal et al., 2014]. Furthermore, the choice of dictionary size, the level of sparsity, and the balance between representation accuracy and computational cost are all critical yet tricky issues to address.

In conclusion, Dictionary Learning and its extensions such as Convolutional Dictionary Learning provide a powerful framework for representing data in a sparse and efficient manner. While there are still many challenges to overcome and open

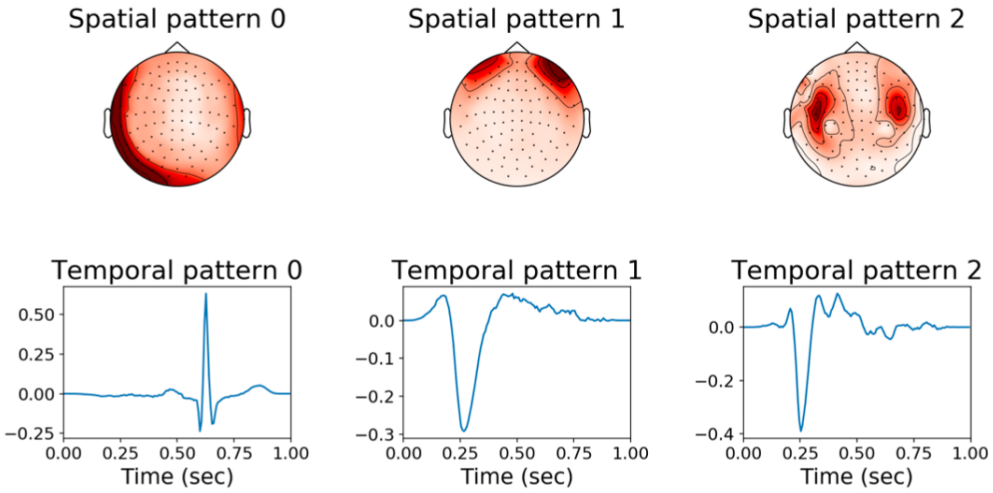


Figure 2.4.1: Spacial (up) and temporal (down) representation of three atoms obtained by dictionary learning. (alphacsc)

questions to answer, the potential of these methods in various applications, including neuroscience, is vast and promising. As we continue to develop more advanced algorithms and gain access to larger and more complex datasets, we can expect to see even more impressive results from these methods in the future.

Chapter 3

Background on Temporal Point Processes

Contents

3.1 Definitions **50**
3.2 Temporal point processes **51**
3.2.1 Poisson process and likelihood function 54
3.3 Hawkes processes **55**
3.4 Goodness of fit **59**

The study of temporal point processes has evolved significantly since its inception, serving as a rich intersection between statistical physics, stochastic processes, and more recently, machine learning and finance. Originating from the foundational work on point processes by Cox and Moran in the mid-20th century [Moran, 1953, Cox, 1955], the field underwent a pivotal transition to incorporate time-dependent phenomena, notably advanced by Daley and Vere-Jones [Daley and Vere-Jones, 2003, 2007]. This shift opened new vistas for applications ranging from seismology to financial markets. A landmark in this development was the introduction of Hawkes processes by Alan G. Hawkes in 1971 [Hawkes, 1971], which enabled the modeling of complex, self-exciting systems, finding immediate applications in epidemiology and later in neural spike train analysis. As computational capabilities expanded, so did methodological advancements, including the development of efficient algorithms for parameter estimation, such as the Expectation-Maximization (EM) algorithm and Maximum Likelihood Estimation (MLE) [Lewis and Shedler, 1979]. The 21st century witnessed a resurgence of temporal point processes in machine learning, tackling challenges in anomaly detection, event prediction, and recommendation systems [Zhou et al., 2013a].

In this chapter, we will give a short introduction on point processes, particularly on temporal point processes, with a focus on Hawkes processes. Further details can be found in Daley and Vere-Jones [2003, 2007]. We use the notation from Achab [2017], as described in section 3.1.

3.1 Definitions

A point process is a type of mathematical model used to describe patterns formed by points randomly distributed in a space. This space, denoted as S, is a locally compact metric space equipped with its Borel σ-algebra $\mathscr{B}$. Essentially, S is a space where points can be placed, and the Borel σ-algebra is a mathematical framework that allows us to measure and analyze these points. The term "locally finite counting measures on S" refers to a way of counting points in S such that in any bounded region, there are only a finite number of points. This concept is crucial for ensuring that the point process is well-defined and manageable. We denote this set of counting measures as X_S. Let $\mathcal{N}_S$ the smallest σ-algebra on X_S such that all point counts $f_B : X_s \to \mathbb{N}, \omega \mapsto \#(\omega \cap B)$ are measurable for B relatively compact in $\mathscr{B}$, where $\#A$ denotes the cardinality of the set A. This means that we can count the number of points in any reasonably sized (relatively compact) region in a consistent and well-defined manner. A point process on S is a measurable map ξ from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to the measurable space $(X_S, \mathcal{N}_S)$.

In simpler terms, one can think of a point process as a way to randomly scatter a number of points in a space, like stars in the night sky or raindrops falling on a surface, where the exact pattern of these points is governed by the rules of probability.

Every realization of a point process ξ can be written as $\xi = \sum_{i=1}^{n} \delta_{X_i}$, where δ denotes the Dirac measure, n is an integer-valued random variable, and each X_i is a randomly located point within the space S. A point process can be equivalently represented by a counting process $N(B) \coloneqq \int_B \xi(x)\,\mathrm{d}x$. In simple terms, this counting process $N(B)$ counts the number of events in each Borel subset $B \in \mathscr{B}$. The mean measure M of a point process ξ is a measure on S that assigns to every $B \in \mathscr{B}$ the expected number of event of ξ in B, *i.e.,*

$$\forall B \in \mathscr{B}, \quad M(B) \coloneqq \mathbb{E}[N(B)] \quad .$$

3.2 Temporal point processes

A temporal point process is a stochastic, or random, process composed of a time series of binary events that occur in continuous time [Paninski, 2019]. However, unlike time series that model events occurring at a fixed rate or interval, temporal point processes can study multiple time scales at once [Bompaire, 2019], as they do not assume a fixed time interval between events.

In this particular case, S is the time interval $[0\,, T)$, equipped with the Borel σ-field of the real line $\mathscr{B}(\mathbb{R})$. Here, a realization of a point process is simply a set of time points $t_i \in S$: $\xi = \sum_{i=1}^{n} \delta_{t_i}$. With a slight abuse of notation, we associate to the set of distinct random timestamps $\xi = \{t_1, \dots, t_n\}$ occurring before T, the counting process $N_t = \sum_{t_i \in \xi} \mathbb{1}_{\{t_i \leq t\}}$, which is then simply the number of points in the time interval $[0\,, t]$. This counting process is a random process that evolves over time by jumps of size 1. Studying temporal point processes consists in analyzing when these jumps occur. The *conditional intensity* function $\lambda\left(t|\mathscr{F}_t\right)$ is the usual way to characterize temporal point processes, where the present depends only on the past. It is defined as the expected infinitesimal rate at which events are expected to occur after t given the information $\mathscr{F}_t$ available up to – but not including – time t, *i.e.,* the history of the counting process N_t prior to t. Namely,

$$\lambda\left(t|\mathscr{F}_t\right) = \lim_{\mathrm{d}t \to 0} \frac{\mathbb{P}\left(N_{t+\mathrm{d}t} - N_t = 1|\mathscr{F}_t\right)}{\mathrm{d}t} \quad , \tag{3.2.1}$$

where $\mathscr{F}_t = \{t_i,\ t_i < t, i = 1, \dots, n\}$ is the natural filtration of the process. The conditional intensity function is sometimes denoted $\lambda^*(t)$.

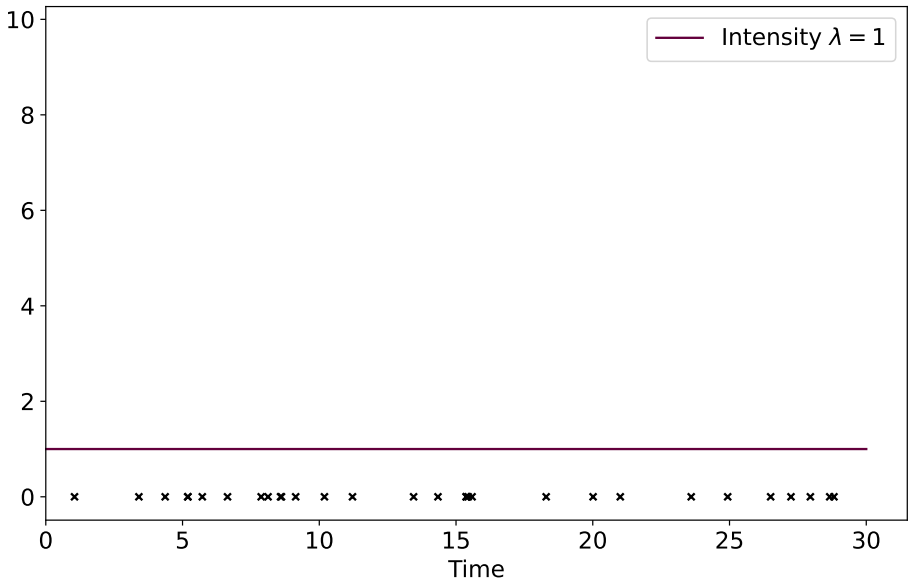
As the quantity $\mathrm{d}N_t \coloneqq N_{t+\mathrm{d}t} - N_t \in \{0, 1\}$ can only increase by one event at each $\mathrm{d}t$, it readily follows that $\mathbb{P}\left(\mathrm{d}N_t = 1|\mathscr{F}_t\right) = \lambda^*(t)\mathrm{d}t$ and

$$\mathbb{E}\left[\mathrm{d}N_t|\mathscr{F}_t\right] = 1 \times \mathbb{P}\left(\mathrm{d}N_t = 1|\mathscr{F}_t\right) + 0 \times \mathbb{P}\left(\mathrm{d}N_t = 0|\mathscr{F}_t\right) = \lambda^*(t)\mathrm{d}t \ . \tag{3.2.2}$$

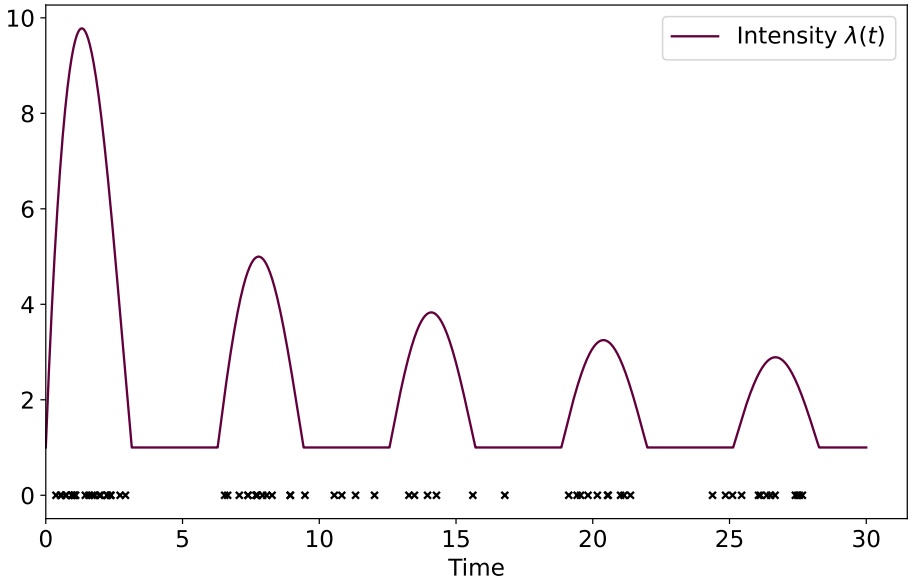
Hence, we can also think of the conditional intensity function $\lambda^*(t)$ as an instantaneous rate of events per time of unit.

The **homogeneous Poisson process** is the simplest temporal point process, which assumes that the events arrive at a constant rate, which corresponds to a constant intensity function $\lambda\left(t|\mathscr{F}_t\right) = \lambda^*(t) = \lambda > 0$, as shown in fig. 3.2.1a. In other words, it describes a phenomenon with no memory and a constant probability of occurrence, in which $N_{t+\Delta t} - N_t$ follows a Poisson distribution of parameter Δt for any $\Delta t > 0$. For this process, $\forall B \in \mathscr{B}(\mathbb{R})$, $M(B) = \lambda\,|B|$, where $|\cdot|$ denotes the Lebesgue measure on $(S, \mathscr{B}(\mathbb{R}))$.

The **inhomogeneous Poisson process** is a more general process, for which the conditional intensity function is not constant as it depends on t but *not* on the



(a) Homogeneous Poisson process



(b) Inhomogeneous Poisson process

Figure 3.2.1: Comparison of event occurrence in one-dimensional Poisson Processes. The red line represents the intensity function $\lambda(t)$ for two Poisson processes, demonstrating, in the case of the inhomogeneous one, the variability of event rates over time. The black crosses denote the actual events, each positioned to indicate the time at which the event occurred. The density of crosses reflects the probability of events occurring at different times, according to the intensity function. The underlying objective in point process inference is to uncover the latent intensity function from these observed events.

history, *i.e.,* $\lambda\left(t|\mathscr{F}_t\right) = \lambda^*(t) = \lambda(t)$, as depicted in fig. 3.2.1b. For this process, $M\left(B\right) = \int_B \lambda(x)\,\mathrm{d}x$, for all $B \in \mathscr{B}\left(\mathbb{R}\right)$.

Let us denote $f^*(t) = f\left(t|\mathscr{F}_t\right)$ the conditional probability density function of the inter-event time, *i.e.,* the probability that the next event will occur during the interval $[t\,, t + \mathrm{d}t)$ conditioned on the history $\mathscr{F}_t$. Let us also denote

$$F^*(t) = F\left(t|\mathscr{F}_t\right) = \mathbb{P}\left(t_n \leq t_{n+1} \leq t|\mathscr{F}_t\right) = \int_{t_n}^{t} f^*(\tau)\,\mathrm{d}\tau \quad , \tag{3.2.3}$$

the conditional cumulative density function, *i.e.,* the probability that the next event t_{n+1} will occur before time t conditioned on the history $\mathscr{F}_t$, where here t_n is the last event in $\mathscr{F}_t$, *i.e.,* the last event before time t, and t_{n+1} the random next one. Finally, we denote $S^*(t) = 1 - F^*(t) = \mathbb{P}\left(t_{n+1} \geq t|\mathscr{F}_t\right)$ the complementary cumulative distribution, also called the survival function, *i.e.,* the probability that the next event will not occur before time t conditioned on the history $\mathscr{F}_t$ [De et al., 2019].

Now,

$$\begin{aligned}
\lambda^*(t) &= \lim_{\mathrm{d}t \to 0} \frac{\mathbb{P}\left(t \leq t_{n+1} \leq t + \mathrm{d}t|t_{n+1} > t\right)}{\mathrm{d}t} \\
&= \lim_{\mathrm{d}t \to 0} \frac{1}{\mathrm{d}t} \frac{\mathbb{P}\left(t \leq t_{n+1} \leq t + \mathrm{d}t\right)}{\mathbb{P}\left(t_{n+1} > t\right)} \\
&= \lim_{\mathrm{d}t \to 0} \left(\frac{1}{\mathrm{d}t} \frac{f^*(t)\mathrm{d}t}{S^*(t)} + o(1) \right) \\
&= \frac{f^*(t)}{S^*(t)} \\
&= -\frac{1}{S^*(t)} \frac{\mathrm{d}S^*(t)}{\mathrm{d}t} \\
&= -\frac{\mathrm{d}\log S^*(t)}{\mathrm{d}t}
\end{aligned} \tag{3.2.4}$$

By integrating the left and right-hand sides in the above equation, we have that

$$\int_{t_n}^{t} \lambda^*(\tau)\,\mathrm{d}\tau = \int_{t_n}^{t} -\frac{\mathrm{d}\log S^*(\tau)}{\mathrm{d}\tau}\,\mathrm{d}\tau = -\log S^*(t) + \underbrace{\log S^*(t_n)}_{=0} \quad , \tag{3.2.5}$$

and thus,

$$S^*(t) = \exp\left(-\int_{t_n}^{t} \lambda^*(\tau)\,\mathrm{d}\tau \right) \quad . \tag{3.2.6}$$

Finally, we get that

$$f^*(t) = \lambda^*(t) \exp\left(-\int_{t_n}^{t} \lambda^*(\tau)\,\mathrm{d}\tau \right) \quad . \tag{3.2.7}$$

3.2.1 Poisson process and likelihood function

This section is taken and adapted from [Daley and Vere-Jones, 2003, chap. 2, p. 19-23] and aims to give a more in-depth presentation of the Poisson processes, with a focus on the computation of the likelihood function, as it is crucial information for what follows in the manuscript.

The stationary Poisson process – what we previously called the homogeneous Poisson process, *cf.* fig. 3.2.1a – on the line is completely defined by the following equation, in which we use $N(a_i, b_i]$ to denote the number of events of the process falling in the half-open interval $(a_i, b_i]$ with $a_i < b_i \leq a_i + 1$:

$$\mathbb{P}\left(N(a_i, b_i] = n_i, i = 1, \ldots, k\right) = \prod_{i=1}^{k} \frac{[\lambda(b_i - a_i)]^{n_i}}{n_i!} e^{-\lambda(b_i - a_i)} \quad . \tag{3.2.8}$$

This definition embodies three important features: i) the number of points in each finite interval $(a_i, b_i]$ has a Poisson distribution of parameter λ; ii) the numbers of points in disjoint intervals are independent random variables; and iii) the distributions are stationary, *i.e.,* they depend only on the lengths $b_i - a_i$ of the intervals.

The likelihood of a finite realization of a Poisson process may be defined as the probability of obtaining the given number of observations in the observation period, times the joint conditional density for the positions of those observations, given their number.

Suppose that there are N observations on $(0, T]$ at time points $t_1, \ldots, t_N$. From 3.2.8, we can write down immediately the probability of obtaining single events in $(t_i - \Delta, t_i]$ and no points on the remaining part of $(0, T]$. Let A and B be respectively those events, namely,

$$A = \{N(t_i - \Delta, t_i] = 1, i = 1, \ldots, N\} \quad , \tag{3.2.9}$$

and

$$B = \{N(0, t_1 - \Delta] = 0, N(t_N, T] = 0, N(t_i, t_{i+1} - \Delta] = 0, i = 1, \ldots, N-1\}$$

$$\begin{aligned}\mathbb{P}\left(A \cap B\right) &= \left(\prod_{i=1}^{N} \lambda\Delta e^{-\lambda\Delta}\right) \times e^{-\lambda(t_1-\Delta)} \times e^{-\lambda(T-t_N)} \times \prod_{i=1}^{N-1} e^{-\lambda(t_{i+1}-\Delta-t_i)} \\ &= \left(\prod_{i=1}^{N} \lambda\Delta\right) \times e^{-\lambda\Delta N} \times e^{-\lambda(T-N\Delta)} \\ &= e^{-\lambda T} \prod_{i=1}^{N} \lambda\Delta \\ &= \lambda^N \Delta^N e^{-\lambda T} \ .\end{aligned}$$

Dividing by Δ^N and letting $\Delta \to 0$, to obtain the density, we find as the required likelihood function:

$$L_{(0\,,T]}\left(N; t_1, \ldots, t_n\right) = \lambda^N e^{-\lambda T} \ . \tag{3.2.10}$$

We can extend this result to a Poisson process with time-varying rate $\lambda(t)$, commonly called the *nonhomogeneous* or *inhomogeneous* Poisson process, *cf.* fig. 3.2.1b. The process can be defined exactly as in 3.2.8, with the quantities $\lambda\left(a_i\,, b_i\right] = \int_{a_i}^{b_i} \lambda \,\mathrm{d}x$ replaced wherever they occur by quantities

$$\Lambda\left(a_i\,, b_i\right] = \int_{a_i}^{b_i} \lambda(x)\,\mathrm{d}x \ , \tag{3.2.11}$$

called the *compensator* of the point process. Thus, the joint distributions are still Poisson, and the independence property still holds. The likelihood function takes the more general form

$$\begin{aligned}L_{(0\,,T]}\left(N; t_1, \ldots, t_n\right) &= e^{-\Lambda(0\,,T]} \prod_{i=1}^{N} \lambda(t_i) \\ &= \exp\left(-\int_0^T \lambda(t)\,\mathrm{d}t + \sum_{i=1}^{N} \log\lambda(t_i)\right) \\ &= \exp\left(-\int_0^T \lambda(t)\,\mathrm{d}t + \int_0^T \log\lambda(t) N\left(\mathrm{d}t\right)\right) \ .\end{aligned} \tag{3.2.12}$$

Note that this result could also be obtained by using the Equation (3.2.7).

3.3 Hawkes processes

In this section, we give the main definitions and properties of Hawkes processes and multivariate Hawkes processes, and set the notations that will be used in

the rest of the manuscript. Hawkes processes [Hawkes, 1971, Hawkes and Oakes, 1974] are temporal point processes in which the intensity depends on the process history with an excitation mechanism. They can be understood as the equivalent of auto-regressive time series models (AR; Box et al. 2015) but in continuous time. This allows to study cross causality that might occur in one or several events series [Bompaire, 2019].

A Hawkes process is defined by a history dependent intensity λ defined as follows,

$$\lambda\left(t|\mathscr{F}_t\right) = \psi\left(\mu + \int_{-\infty}^{t} \phi(t-s)\,\mathrm{d}N_s\right) \quad , \tag{3.3.1}$$

where

$$\int_{-\infty}^{t} \phi(t-s)\,\mathrm{d}N_s = \sum_{i,t_i \leq t} \phi(t-t_i) \quad . \tag{3.3.2}$$

The parameter $\mu \geq 0$ is referred as the *baseline intensity* – or *background intensity* – and it corresponds to the exogenous intensity of the considered events. The function $\phi(\cdot) : \mathbb{R}^+ \to \mathbb{R}$ is called the *kernel function*, or the transfer function [Chen et al., 2017], and quantifies over time and in magnitude the influence of past events. Note that the occurrence of each event t_i makes the intensity vary by a certain amount, determined by the kernel, making the intensity history dependent and a stochastic process by itself [De et al., 2019].

If the *link function* ψ on the right-hand side of Equation 3.3.1 is non-linear, then $\lambda(t)$ is the intensity of a non-linear Hawkes process [Brémaud and Massoulié, 1996]. Various forms of non-linear link functions exist, each with their own specific applications. For instance, $\psi(\cdot) = \exp(\cdot)$ is often used in financial markets to capture exponential increases in trading activities. The logistic function $\psi(\cdot) = \frac{1}{1+\exp(-\cdot)}$ is commonly applied in social network analysis, particularly in modeling the virality of information spread, as it accounts for saturation effects. In seismology, a power-law function $\psi(\cdot) = (\cdot)^\alpha$, $\alpha \neq 1$, is employed to model the heavy-tailed distribution of aftershock occurrences over time. More recently, neural network-based link functions have been introduced in computational neuroscience to capture the complex, nonlinear dependencies in neural spike trains. Each of these non-linear forms offers specific advantages in capturing the dynamics of the system being modeled, thereby extending the applicability of Hawkes processes to a wider range of phenomena.

In contrast to non-linear Hawkes processes, a linear link function ψ offers a simplified yet flexible model. Specifically, if $\psi(\cdot) = a(\cdot) + b$, where a and b are constants, the Hawkes process retains its linearity but allows for a scaling and shifting of the intensity function. This can be particularly beneficial in systems where the impact of historical events on future occurrences is scaled by a factor a or shifted by b. For instance, in queueing systems, the linear scaling factor a can model the

rate at which incoming tasks trigger additional tasks in the queue [Daley and Vere-Jones, 2007]. In epidemiology, the constant b can account for external factors such as vaccination rates that uniformly shift the intensity of disease spread [Becker, 1977]. Thus, a linear ψ provides a useful compromise between model complexity and interpretability, allowing for adjustments to the intensity function without introducing non-linearity [Hawkes, 1971].

In what follows, we restrict ourselves to the simplest case where ψ is the identity function.

Multivariate Hawkes process

We can extend the univariate Hawkes process to model the interactions of $K \geq 1$ temporal point processes, called *nodes*.

Namely, it models timestamps $\left\{t_k^{(i)}\right\}_{k\geq 1}$ of nodes $i = 1, \ldots, K$ associated with a multivariate counting process $N_t = \left[N_t^{(1)}, \ldots, N_t^{(K)}\right]$. Note that for all nodes $i = 1, \ldots, K$, we still have that all of its timestamps $t_k^{(i)}$ occur in the time interval $[0\,, T]$. The excitation dynamic between the nodes is encompassed by the auto-regressive structure of the conditional intensity. For component $N_t^{(i)}$ it writes:

$$\lambda_i\left(t|\mathscr{F}_t\right) = \mu_i + \sum_{j=1}^{K} \int_{-\infty}^{t} \phi_{i,j}(t-s)\,\mathrm{d}N_s^{(j)} \quad , \tag{3.3.3}$$

where $\phi_{i,j}(t)$ quantifies the excitation rate of an event of type j on the arrival rate of events of type i after a time lag t. In general it is assumed that each kernel is causal and positive, meaning that Hawkes processes can only account for mutual excitation effects since the occurrence of some event can only increase the future arrival intensity of other events. If the kernels are integrable, each entry of the $K \times K$ matrix $(\Phi)_{i,j} = \int_0^T \phi_{i,j}(t)\,\mathrm{d}t$ denotes the expected number of events of type i directly triggered by an event of type j.

Kernels parametrisation

The kernel function in a Hawkes process plays an indispensable role, determining how past events exert influence on future occurrences. The choice of kernel is often predicated on the specific dynamics of the system under examination. In this section, we explore a variety of kernel parametrizations, elucidating their applications, scholarly references, and inherent limitations.

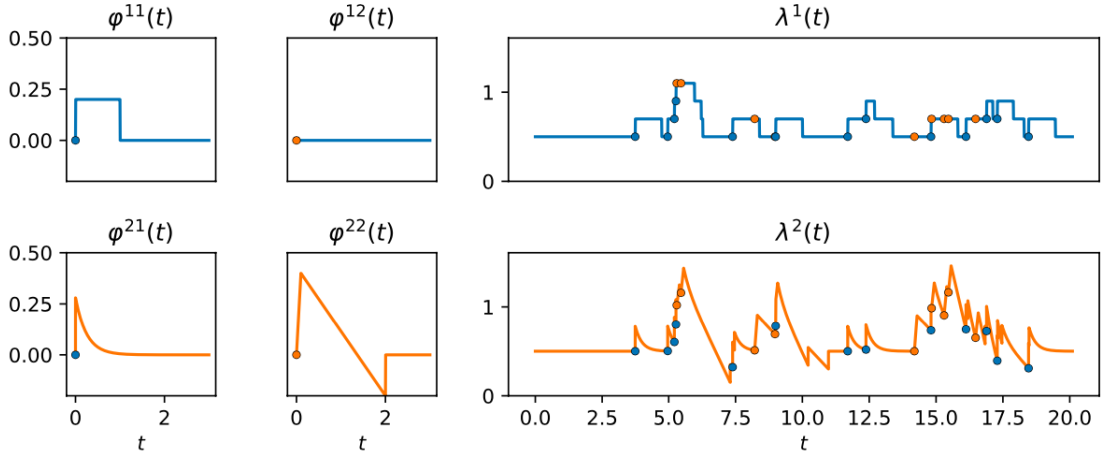


Figure 3.3.1: A realisation of a 2 nodes multivariate Hawkes process using `Tick` package [Bacry et al., 2017b]. The four excitation kernels are shown on the left-hand side. The intensities are displayed on the right-hand side (against time, up to time 20), where events are represented by colored dots (blue corresponding to node 1 and orange to node 2). [Bompaire, 2019]

Exponential kernel The main parametric model is the so-called *exponential kernel*, in which the kernels have the following form:

$$\phi_{i,j}(t) = \alpha_{i,j}\beta \exp(-\beta t), \quad \alpha_{i,j} > 0, \beta > 0 \ . \tag{3.3.4}$$

In this model the integral matrix $\Phi = (\alpha_{i,j})_{1 \leq i,j \leq K}$ and $\beta > 0$ is a memory parameter. A more general approach is the *sum of exponentials kernels* [Lemonnier and Vayatis, 2014], namely

$$\phi_{i,j}(t) = \sum_{u=1}^{U} \alpha_{i,j}^{(u)} \beta^{(u)} \exp\left(-\beta^{(u)} t\right), \quad \alpha_{i,j}^{(u)} > 0, \beta^{(u)} > 0 \ . \tag{3.3.5}$$

Exponential parametrization is predominantly employed in financial markets for the modeling of high-frequency data, as it has the capacity to model instant response to event with a rapid return to baseline. One significant limitation is the kernel's assumption of exponential decay in the influence of past events, which may not be suitable for capturing long-range dependencies

Gaussian kernel Similarly, we can define the *gaussian kernel* and the *sum of gaussians kernels*:

$$\phi_{i,j}(t) = \sum_{u=1}^{U} \frac{1}{\sqrt{2\pi\sigma_{i,j}^{(u)}}} \exp\left(-\frac{\left(t - \mu_{i,j}^{(u)}\right)^2}{2{\sigma_{i,j}^{(u)}}^2}\right), \quad \sigma_{i,j}^{(u)} > 0 \ . \tag{3.3.6}$$

This kernel is commonly used in natural language processing and seismology. Ogata [1999] utilized this kernel for modeling aftershocks following earthquakes. The kernel's tendency to smooth out abrupt changes in event dynamics serves as a limitation.

Power-law kernel The Power-Law kernel can be expressed as:

$$\phi_{i,j}(t) = t^{-\alpha_{i,j}} \quad . \tag{3.3.7}$$

It is a prevalent choice in network analysis and studies within the social sciences. Crane and Sornette [2008] applied this kernel in the modeling of social systems. The main drawbacks are its computational complexity and the necessity for large datasets for accurate parameter estimation.

Piecewise-constant kernel A piecewise-constant kernel can be defined as:

$$\phi_{i,j}(t) = \alpha_{i,j}, \quad t \in [t_{n-1}\,, t_n) \quad . \tag{3.3.8}$$

This form of kernel is especially useful in epidemiological models that feature diseases spreading in distinct phases. One of the risks is overfitting if the number of intervals is not correctly specified.

In `Python`, the `Tick` package allows us to easily manipulate Hawkes process with exponential and gaussian kernels [Bompaire, 2019, Bacry et al., 2017b]. Other kernel functions are presented in Mehrdad and Zhu [2014].

3.4 Goodness of fit

We call goodness-of-fit a function telling how well a statistical model fits a set of observations [Bompaire, 2019]. It has roots in classical statistics, tracing back to the foundational works of Pearson [1900] and Fisher [1922].

Negative log-likelihood The notion of negative log-likelihood (NLL) finds its origins in the maximum likelihood estimation (MLE) framework proposed by Ronald A. Fisher in the early 20th century. For temporal point processes, the concept was prominently highlighted in the seminal works of Daley and Vere-Jones [2003] and Ogata [1988], obtained from (3.2.12) and given by:

$$\mathcal{L}\left(\lambda, \mathscr{F}_T\right) = -\log L\left(\lambda, \mathcal{F}_T\right) = \int_0^T \lambda\left(s|\mathscr{F}_s\right)\, \mathrm{d}s - \sum_{k=1}^{N_T} \log \lambda\left(t_k, \mathscr{F}_{t_k}\right) \quad , \tag{3.4.1}$$

where $\mathscr{F}_T = \{t_1, \ldots, t_n\}$ is the full history of the process and N_T is the total number of events that have occurred in $[0\,, T]$.

The minimization of NLL essentially aligns with the maximization of the likelihood of the observed events under the model parameters θ. It has been extensively used in fields such as seismology, telecommunications, and more recently, neuroscience, particularly in the modeling of neuronal spike trains [Brown et al., 2003]. Under some assumptions, the maximum likelihood estimator obtained by minimizing this error is consistent, asymptotically normal, and asymptotically efficient [Bompaire, 2019, Ogata, 1978].

Least squares error We now focus on the least squares loss inspired from empirical risk minimization (ERM). Assuming a class of parametric kernel parametrized by η, the objective is to find parameters that minimize the following loss function (see *e.g.,* eq. (I.2) in Bompaire, 2019, Chapter 1; Reynaud-Bouret and Rivoirard [2010], Hansen et al. [2015], Bacry et al. [2020]):

$$\mathcal{L}\left(\theta, \mathscr{F}_T\right) = \frac{1}{N_T} \sum_{i=1}^{p} \left(\int_0^T \lambda_i\left(s \middle| \mathscr{F}_s^i\right)^2 \, \mathrm{d}s - 2 \sum_{t_n^i \in \mathscr{F}_T^i} \lambda_i\left(t_n^i \middle| \mathscr{F}_{t_n^i}^i\right) \right), \tag{3.4.2}$$

where $\mathscr{F}_T \coloneqq \{\mathscr{F}_T^1, \ldots, \mathscr{F}_T^p\}$ is the set of all considered timestamps across all processes, $N_T \coloneqq \sum_{i=1}^p N_T^i$ is the total number of timestamps, and where $\theta \coloneqq (\boldsymbol{\mu}, \boldsymbol{\eta})$, where bold version of parameter denotes the associated vector, *e.g.,* $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$. Interestingly, when used with an exponential kernel, this loss benefits from some precomputations of complexity $\mathcal{O}\left(N_T\right)$, making the subsequent iterative optimization procedure independent of N_T [Bompaire, 2019]. This computational ease is the main advantage of the loss $\mathcal{L}$ over the log-likelihood function. However, when using a general parametric kernel, these precomputations require $\mathcal{O}\left(\left(N_T\right)^2\right)$ operations, killing the computational benefit of the ℓ_2 loss $\mathcal{L}$ over the log-likelihood. It is worth noting that this loss differs from the quadratic error minimized between the counting processes and the integral of the intensity function, as used in Wang et al. [2016], Eichler et al. [2017] and Xu et al. [2018].

Part II

Temporal Modeling and Inference in M/EEG Signals: A Point Process Approach

Traditionally, several methodologies have been employed to unearth hidden patterns from high-dimensional time-series data produced by M/EEG, ranging from statistical models to machine learning algorithms. However, the need for more advanced techniques to model these time-series data's intricacies has never been more evident. This chapter amalgamates two crucial contributions in this direction: the *Driven Temporal Point Processes (DriPP)* and the *Fast Discretized Inference for Hawkes Processes with General Parametric Kernels (FaDIn)*. These two methods contribute to a more nuanced understanding of event-related neuronal activity by exploiting the power of point processes.

Driven Temporal Point Processes (DriPP) extend the capabilities of Convolutional Dictionary Learning, an unsupervised learning technique that has been successfully applied to extract temporal patterns in M/EEG data. DriPP takes this a step further by modeling how these event occurrences are influenced by specific tasks or experimental conditions. Using a novel statistical point process model, DriPP links the intensity function of the point process to stimulation events, thus providing a framework to study how cognitive or sensorial stimulation modulates neural events. An efficient expectation-maximization (EM) algorithm has been developed to estimate the parameters of this model, which has shown promising results in revealing both evoked and induced event-related neural responses.

Fast Discretized Inference (FaDIn), on the other hand, focuses on the inherent challenges in Temporal Point Process (TPP) inference, particularly in the context of Hawkes processes. Traditional Hawkes processes often employ exponential or non-parametric kernels, which have limitations regarding data requirements or their ability to model latencies effectively. FaDIn addresses these issues by introducing a fast ℓ_2 gradient-based solver for TPPs using general parametric kernels with finite support. The method offers a more accurate and computationally efficient approach to model stimuli-induced patterns in brain signals, significantly improving the estimation of pattern latency.

Both DriPP and FaDIn contribute to a unified framework for the analysis of M/EEG data, offering robust and efficient methods to identify and interpret temporal patterns modulated by external stimuli. By integrating these methods, this second part aims to provide a comprehensive view of advanced point process models in analyzing neural time-series data. It underlines the statistical and computational advancements in point process modeling and highlights their applicability and effectiveness in neuroscience research.

Chapter 4

DriPP: Driven Point Processes to Model Stimuli Induced Patterns in M/EEG Signals

Contents

4.1 Mathematical formulation **68**
4.2 Parameters inference with an EM-based algorithm **70**
4.3 Experiments **78**
4.3.1 Evaluation of the EM convergence on synthetic data . 78
4.3.2 Evoked and induced effects characterization in MEG data 81
4.3.3 Impact of model hyperparameter 85
4.3.4 Experiments on Cam-CAN dataset 87
4.3.5 Usual M/EEG data analysis 91
4.4 Transcending limits with discretised parametric kernels **95**

The content of this chapter was published in:

Cédric Allain, Alexandre Gramfort, and Thomas Moreau. DriPP: Driven point processes to model stimuli induced patterns in M/EEG signals. *International Conference on Learning Representations*, 2022

Statistical analysis of human neural recordings is at the core of modern neuroscience research. Thanks to non-invasive recording technologies such as elec-

troencephalography (EEG) and magnetoencephalography (MEG), or invasive techniques such as electrocorticography (ECoG) and stereotactic EEG (sEEG), the ambition is to obtain a detailed quantitative description of neural signals at the millisecond timescale when human subjects perform different cognitive tasks [Baillet, 2017]. During neuroscience experiments, human subjects are exposed to several external stimuli, and we are interested in knowing how these stimuli influence neural activity.

After pre-processing steps, such as filtering or Independent Component Analysis (ICA; Winkler et al. 2015) to remove artifacts, common techniques rely on epoch averaging – to highlight evoked responses – or time-frequency analysis to quantify power changes in certain frequency bands [Cohen, 2014] – for induced responses (*cf.* section 1.2 for more details on this matter). While these approaches have led to numerous neuroscience findings, it has also been criticized. Indeed, averaging tends to blur out the responses due to small jitters in time-locked responses, and the Fourier analysis of different frequency bands tends to neglect the harmonic structure of the signal, leading to the so-called "Fourier fallacy" [Jasper, 1948, Jones, 2016]. In so doing, one may conclude to a spurious correlation between components that have actually the same origin. Moreover, artifact removal using ICA requires a tedious step of selecting the correct components.

Driven by these drawbacks, a recent trend of work aims to go beyond these classical tools by isolating prototypical waveforms related to the stimuli in the signal [Cole and Voytek, 2017, Dupré la Tour et al., 2018, Donoghue et al., 2020]. The core idea consists in decomposing neural signals as combinations of time-invariant patterns, which typically correspond to transient bursts of neural activity [Sherman et al., 2016], or artifacts such as eye blinks or heartbeats. In machine learning, various unsupervised algorithms have been historically proposed to efficiently identify patterns and their locations from multivariate temporal signals or images [Lewicki and Sejnowski, 1999, Jost et al., 2006, Heide et al., 2015, Bristow et al., 2013, Wohlberg, 2016b], with applications such as audio classification [Grosse et al., 2007] or image inpainting [Wohlberg, 2016a]. For neural signals in particular, several methods have been proposed to tackle this task, such as the sliding window matching (SWM; Gips et al. 2017), the learning of recurrent waveforms [Brockmeier and Príncipe, 2016], adaptive waveform learning (AWL; Hitziger et al. 2017) or convolutional dictionary learning (CDL; Jas et al. 2017, Dupré la Tour et al. 2018). Equipped with such algorithms, the multivariate neural signals are then represented by a set of spatio-temporal patterns, called *atoms*, with their respective onsets, called *activations*. Out of all these methods, CDL has emerged as a convenient and efficient tool to extract patterns, in particular due to its ability to easily include physical priors for the patterns to recover. For example, for M/EEG data, Dupré la Tour et al. [2018] have proposed a CDL method which extracts atoms that appertain to electrical dipoles in the brain by imposing a rank-1

structure. While these methods output characteristic patterns and an event-based representation of the temporal dynamics, it is often tedious and requires a certain domain knowledge to quantify how stimuli affect the atoms' activations. Knowing such effects allows determining whether an atom is triggered by a specific type of stimulus, and if so, to quantify by how much, and with what latency. See section 2.4 for more details on this matter.

As activations are random signals that consist of discrete events, a natural statistical framework is the one of temporal point processes (PP). PP have received a surge of interest in machine learning [Bompaire, 2019, Shchur et al., 2020, Mei et al., 2020] with diverse applications in fields such as healthcare [Lasko, 2014, Lian et al., 2015] or modelling of communities on social networks [Long et al., 2015]. In neuroscience, PP have also been studied in the past, in particular to model single cell recordings and neural spike trains [Truccolo et al., 2005, Okatan et al., 2005, Kim et al., 2011, Rad and Paninski, 2011], sometimes coupled with spatial statistics [Pillow et al., 2008] or network models [Galves and Löcherbach, 2015]. However, existing models do not directly address our question, namely, the characterization of the influence of a deterministic PP – the stimuli onsets – on a stochastic one – the neural activations derived from M/EEG recordings.

This work proposes a novel method – called driven point process (DriPP) – to model the activation probability for CDL. This method is inspired from Hawkes processes (HP; Hawkes 1971), and models the intensity function of a stochastic process conditioned on the realization of a set of PP, called *drivers*, parametrized using truncated Gaussian kernels to better model latency effects in neural responses. The resulting process can capture the surge of activations associated to external events, thus providing a direct statistical characterization of how much a stimulus impacts the neural response, as well as the mean and standard deviation of the response's latency. We derive an efficient expectation-maximization (EM) based inference algorithm and show on synthetic data that it reliably estimates the model parameters, even in the context of M/EEG experiments with tens to hundreds of events at most. Finally, the evaluation of DriPP on the output of CDL for standard MEG datasets shows that it reveals neural responses linked to stimuli that can be mapped precisely both in time and in brain space. Our methodology offers a unified approach to decide if some waveforms extracted with CDL are unrelated to a cognitive task, such as artifacts or spontaneous brain activity, or if they are provoked by a stimulus – no matter if they are "evoked" or "induced" as more commonly described in the neuroscience literature [Tallon-Baudry et al., 1996]. While these different effects are commonly extracted using different analysis pipelines, DriPP simply reveals them as stimuli-induced neural responses using a single unified method, that does not require any manual tuning or selection.

4.1 Mathematical formulation

Recall from chapter 3 that the conditional intensity function $\lambda : \mathbb{R}^+ \to \mathbb{R}^+$ of a temporal point process (TPP) of events $\{t_i\}$ is defined as follows:

$$\lambda\left(t|\mathscr{F}_t\right) \coloneqq \lim_{\mathrm{d}t\to 0} \frac{\mathbb{P}\left(N_{t+\mathrm{d}t} - N_t = 1|\mathscr{F}_t\right)}{\mathrm{d}t} \quad , \tag{4.1.1}$$

where $N_t \coloneqq \sum_{i\geq 1} \mathbb{1}_{\{t_i \leq t\}}$ is the counting process associated to the process and where $\mathscr{F}_t \coloneqq \{t_i,\ t_i < t\}$. This function corresponds to the expected infinitesimal rate at which events are occurring at time t given the arrival times of past events prior to t [Daley and Vere-Jones, 2003].

The proposed model DriPP is adapted from the Hawkes process (HP; Hawkes 1971), as the occurrence of a past event in the driver increases the likelihood of occurrence of activation events in the near future. However, here we suppose that the stochastic point process in our model of neural activations does not have the self-excitatory behavior characteristic of HP. Instead, the sources of activation in the DriPP model are either the drivers or some spontaneous background activity, but not its own previous activations. More specifically, in DriPP, the intensity function at time t between a stochastic process k – whose set of events is denoted $\mathcal{A}_k$ – and a non-empty set of drivers $\mathcal{P}$ – whose events are denoted $\mathcal{T}_p \coloneqq \{t_1^{(p)}, \ldots, t_{n_p}^{(p)}\}, p \in \mathcal{P}$ – is composed of a baseline intensity $\mu_k \geq 0$ and triggering kernels $\kappa_{k,p} : \mathbb{R}^+ \to \mathbb{R}$:

$$\lambda_{k,\mathcal{P}}(t) = \mu_k + \sum_{p\in\mathcal{P}} \sum_{i, t_i^{(p)} \leq t} \alpha_{k,p} \kappa_{k,p}\left(t - t_i^{(p)}\right) \quad , \tag{4.1.2}$$

where $\alpha_{k,p} \geq 0$ is a coefficient which controls the relative importance of the driver p on the occurrence of events on the stochastic process k. Note that when the driver processes are known, the intensity function is deterministic, and thus corresponds to the intensity of an inhomogeneous Poisson process [Daley and Vere-Jones, 2003]. The coefficient $\alpha_{k,p}$ is set to be non-negative so that we only model excitatory effects, as events on the driver only increase the likelihood of occurrence of new events on the stochastic process. Inhibition effects are assumed non-existent. Figure 4.1.1 illustrates how events $\mathcal{T}_p$ on the driver influence the intensity function after a short latency period.

A critical parametrization of this model is the choice of the triggering kernels $\kappa_{k,p}$. To best model the latency, we use a parametric truncated normal distribution of mean $m_{k,p} \in \mathbb{R}$ and standard deviation $\sigma_{k,p} > 0$, with support $[a\,, b] \subset \mathbb{R}^+$, $b > a$. Namely,

$$\kappa_{k,p}(x) \coloneqq \kappa\left(x; m_{k,p}, \sigma_{k,p}, a, b\right) = \frac{1}{\sigma_{k,p}} \frac{\phi\left(\frac{x-m_{k,p}}{\sigma_{k,p}}\right)}{\Phi\left(\frac{b-m_{k,p}}{\sigma_{k,p}}\right) - \Phi\left(\frac{a-m_{k,p}}{\sigma_{k,p}}\right)} \mathbb{1}_{\{a\leq x\leq b\}} \quad , \tag{4.1.3}$$

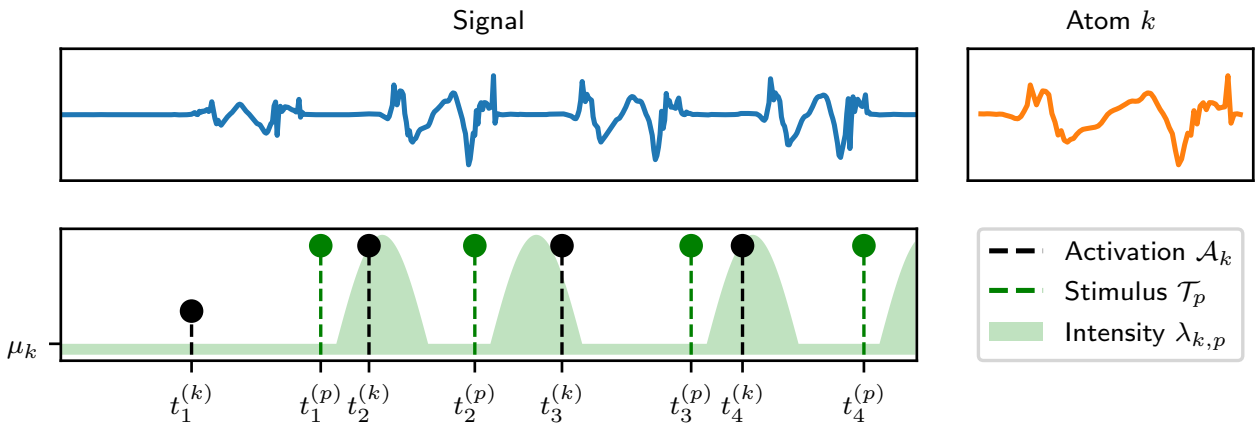


Figure 4.1.1: **Top:** Convolutional dictionary learning (CDL) applied to a univariate signal (blue) decomposes it as the convolution of a temporal pattern (orange) and a sparse activation signal (black). **Bottom:** Intensity function $\lambda_{k,p}$ defined by its baseline μ_k and the stimulus events $\mathcal{T}_p$ (green). Intensity increases following stimulation events with a certain latency.

where here, ϕ (resp. Φ) denotes the probability density function (resp. cumulative distribution function) of the standard normal distribution. This parametrization differs from the usual exponential kernel usually considered in HP, that captures responses with low latency. Note that the truncation values $a, b \in \mathbb{R}^+$ are supposed independent of both the stochastic process and the drivers, hence they are similar for all kernel $p \in \mathcal{P}$. Indeed, in the context of this paper, those values delimit the time interval during which a neuronal response might occur following an external stimulus. In other words, the interval $[a\,, b]$ denotes the range of possible latency values. In the following, we denote by $T \coloneqq T^{(k)}$ the duration of the process k.

As previously mentioned, our research into point processes is motivated by the need to unravel the complex relationships between external stimuli and neuronal responses in the context of M/EEG data. The Convolutional Dictionary Learning framework (CDL; *cf.* section 2.4) has provided us with a means to extract meaningful atoms from raw signals. Our endeavor is to use point processes to analyze these temporal, sparse activations, thereby paving the way for a more nuanced understanding of how the brain processes stimuli. Figure 4.1.2 offers a visual depiction of this decomposition of raw signals into atoms and their respective activations.

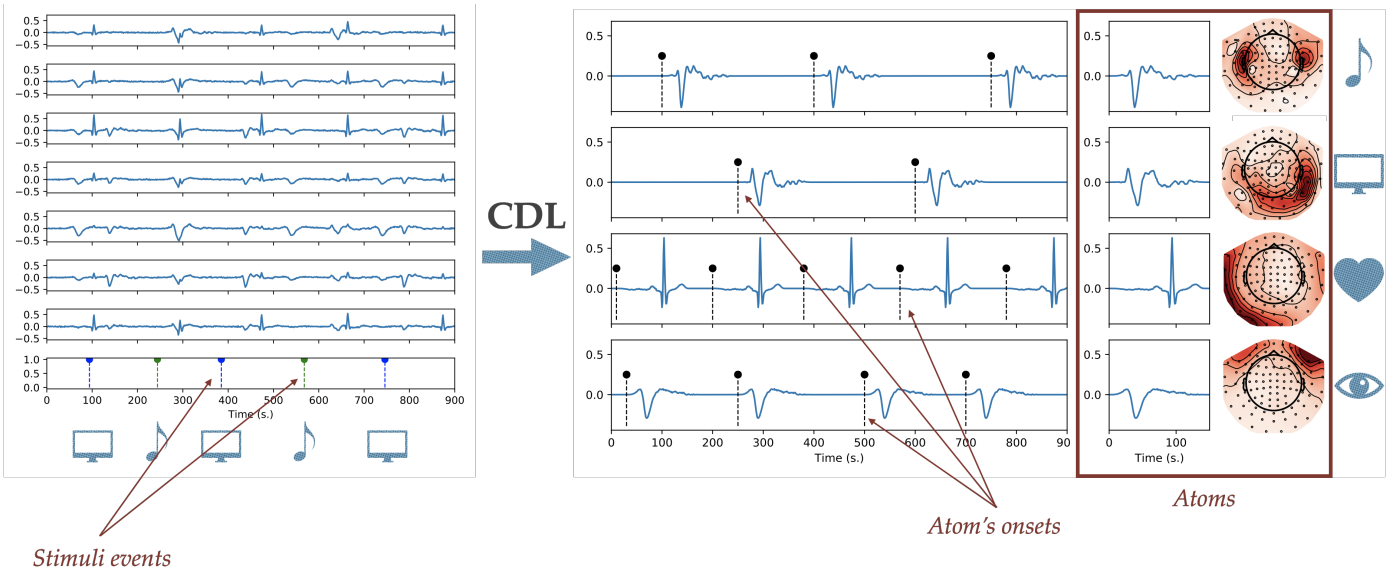


Figure 4.1.2: Schematic operation of the CDL on MEG signals. Raw MEG signals alongside timestamps of external stimuli of type visual and auditory (left). CDL output composed of a set of spatio-temporal atoms alongside their respective onsets (right). One may claim to associate each atom to a physical phenomenom, *i.e.,* heartbeat or eye blink artifact, auditory or visual neural response.

4.2 Parameters inference with an EM-based algorithm

We propose to infer the model parameters $\Theta_{k,\mathcal{P}} = (\mu_k, \boldsymbol{\alpha}_{k,\mathcal{P}}, \boldsymbol{m}_{k,\mathcal{P}}, \boldsymbol{\sigma}_{k,\mathcal{P}})$, where we denote in bold the vector version of the parameter, *i.e.,* $\boldsymbol{x}_{k,\mathcal{P}} = (x_{k,p})_{p\in\mathcal{P}}$, via maximum-likelihood using an EM-based algorithm [Lewis and Mohler, 2011, Xu et al., 2016]. The pseudocode of the algorithm is presented in Algorithm 2. The expectation-maximization (EM) algorithm [Dempster et al., 1977] is an iterative algorithm that allows to find the maximum likelihood estimates (MLE) of parameters in a probabilistic model when the latter depends on non-observable latent variables.

First, we derive the negative log-likelihood of the model.

PROPOSITION 4.2.1. *Given an intensity function modelled as follows,*

$$\lambda_{k,\mathcal{P}}(t) = \mu_k + \sum_{p\in\mathcal{P}} \sum_{i,t_i^{(p)}\leq t} \alpha_{k,p}\kappa_{k,p}\left(t - t_i^{(p)}\right) \ ,$$

with $\mu_k \geq 0$, $\alpha_{k,p} \geq 0$ and where $\kappa_{k,p}(\cdot)$ denotes the truncated normal distribution of mean $m_{k,p} \in \mathbb{R}$ and standard deviation $\sigma_{k,p} > 0$, with support $[a\,,b] \subset \mathbb{R}^+$, $b > a$, the negative log-likelihood for the model's parameters $\Theta_{k,\mathcal{P}} = (\mu_k, \boldsymbol{\alpha}_{k,\mathcal{P}}, \boldsymbol{m}_{k,\mathcal{P}}, \boldsymbol{\sigma}_{k,\mathcal{P}})$ is:

$$\mathcal{L}_{k,\mathcal{P}}\left(\Theta_{k,\mathcal{P}}\right) = \mu_k T + \sum_{p\in\mathcal{P}} \alpha_{k,p} n_p - \sum_{t\in\mathcal{A}_k} \log\left(\mu_k + \sum_{p\in\mathcal{P}} \sum_{i,t_i^{(p)}\leq t} \alpha_{k,p}\kappa_{k,p}\left(t - t_i^{(p)}\right)\right) \ . \tag{4.2.1}$$

Proof 4.2.1

From (3.4.1), the negative log-likelihood for our model's parameters $\Theta_{k,\mathcal{P}}$ is as follows:

$$\mathcal{L}_{k,\mathcal{P}}\left(\Theta_{k,\mathcal{P}}\right) = \int_0^T \lambda_{k,\mathcal{P}}(t)\mathrm{d}t - \sum_{t\in\mathcal{A}_k} \log \lambda_{k,\mathcal{P}}(t) \ . \tag{4.2.2}$$

Now, we show that

$$\int_0^T \lambda_{k,p}(t)\mathrm{d}t = \mu_k T + \sum_{p\in\mathcal{P}} \alpha_{k,p} n_p$$

Without loss of generality, we can first assume that $\forall p \in \mathcal{P}, t_{n_p}^{(p)} + b \leq T$, i.e., *the signal ends after every possible neurological response driven by a stimulus, or equivalently, $T \geq \max_{p=1,\dots,P} t_{n_p}^{(p)} + b$. Hence,*

$$\forall p \in \mathcal{P}, \forall i = 1, \dots, n_p, \int_0^T \kappa_{k,p}\left(t - t_i^{(p)}\right) \mathrm{d}t = 1 \ .$$

Thus, we have that

$$\begin{aligned}
\int_0^T \lambda_{k,p}(t)\,\mathrm{d}t &= \int_0^T \left(\mu_k + \sum_{p\in\mathcal{P}} \sum_{i,t_i^{(p)}<t} \alpha_{k,p}\kappa_{k,p}\left(t - t_i^{(p)}\right)\right) \mathrm{d}t \\
&= \mu_k T + \sum_{p\in\mathcal{P}} \alpha_{k,p} \left(\sum_{t_i^{(p)}\in\mathcal{T}_p} \int_0^T \kappa_{k,p}\left(t - t_i^{(p)}\right) \mathrm{d}t\right) \\
&= \mu_k T + \sum_{p\in\mathcal{P}} \alpha_{k,p} n_p \ .
\end{aligned}$$

Finally, we have that the negative log-likelihood writes

$$\mathcal{L}_{k,\mathcal{P}}(\Theta_{k,\mathcal{P}}) = \mu_k T + \sum_{p \in \mathcal{P}} \alpha_{k,p} n_p - \sum_{t \in \mathcal{A}_k} \log \lambda_{k,\mathcal{P}}(t) \ .$$

Substituting $\lambda_{k,\mathcal{P}}(t)$ by its full expression using (4.1.2) concludes the proof.

Expectation step

For a given estimate, the E-step aims at computing the events' assignation, *i.e.,* the probability that an event comes from either the kernel or the baseline intensity. At iteration n, let $P_k^{(n)}(t) \coloneqq P_k^{(n)}(t; p, k)$ be the probability that the activation at time $t \in [0\,, T]$ has been triggered by the baseline intensity of the stochastic process k, and $P_p^{(n)}(t) \coloneqq P_p^{(n)}(t; p, k)$ be the probability that the activation at time t has been triggered by the driver p. By the definition of our intensity model (4.1.2), we have:

$$P_k^{(n)}(t) = \frac{\mu_k^{(n)}}{\lambda_{k,\mathcal{P}}^{(n)}(t)} \qquad \text{and} \qquad \forall p \in \mathcal{P}, P_p^{(n)}(t) = \frac{\alpha_{k,p}^{(n)} \sum_{i, t_i^{(p)} \leq t} \kappa_{k,p}^{(n)} \left(t - t_i^{(p)}\right)}{\lambda_{k,\mathcal{P}}^{(n)}(t)} \ , \tag{4.2.3}$$

where $\theta^{(n)}$ denotes the value of the parameter θ at step n of the algorithm, and similarly, if f is a function of parameter θ, $f^{(n)}(x; \theta) \coloneqq f\left(x; \theta^{(n)}\right)$. Note that,

$$\forall t \in [0\,, T]\,, P_k^{(n)}(t) + \sum_{p \in \mathcal{P}} P_p^{(n)}(t) = 1 \ .$$

Maximization step

Once this assignation has been computed, one needs to update the parameters of the model using MLE. To obtain the update equations, we fix the probabilities $P_k^{(n)}$ and $P_p^{(n)}$, and cancel the negative log-likelihood derivatives with respect to each parameter.

For given values of probabilities $P_k^{(n)}(t)$ and $P_p^{(n)}(t)$, we here derive succinctly the update for parameters μ and $\boldsymbol{\alpha}$:

$$\begin{aligned}
&\frac{\partial \mathcal{L}_{k,\mathcal{P}}}{\partial \mu_k^{(n)}}\left(\Theta_{k,\mathcal{P}}^{(n)}\right) = 0 \\
\Leftrightarrow\ & T - \sum_{t\in\mathcal{A}_k} \frac{1}{\lambda_{k,\mathcal{P}}^{(n)}(t)} = 0 \\
\Leftrightarrow\ & T - \sum_{t\in\mathcal{A}_k} \frac{P_k^{(n)}(t)}{\mu_k^{(n)}} = 0 \\
\Leftrightarrow\ & \mu_k^{(n+1)} = \frac{1}{T}\sum_{t\in\mathcal{A}_k} P_k^{(n)}(t)
\end{aligned} \tag{4.2.4}$$

$$\frac{\partial \mathcal{L}_{k,\mathcal{P}}}{\partial \alpha_{k,p}^{(n)}}\left(\Theta_{k,\mathcal{P}}^{(n)}\right) = 0 \Leftrightarrow n_p - \sum_{t\in\mathcal{A}_k} \frac{P_p^{(n)}(t)}{\alpha_{k,p}^{(n)}} = 0 \Leftrightarrow \alpha_{k,p}^{(n+1)} = \frac{1}{n_p}\sum_{t\in\mathcal{A}_k} P_p^{(n)}(t) \tag{4.2.5}$$

Note that by definition of $P_p^{(n)}$, $\alpha_{k,p}^{(n+1)}$ maintains the same sign as its initialization $\alpha_{k,p}^{(0)}$. This property stems from the inherent characteristics of the algorithm and remains consistent throughout its iterations. These two updates amount to maximizing the probabilities that the events assigned to the driver or the baseline stay assigned to the same generation process.

Then, we give the update equations for $\boldsymbol{m}$ and $\boldsymbol{\sigma}$, which corresponds to parametric estimates of each truncated Gaussian kernel parameter with events assigned to the kernel. First, let us rewrite the kernel function to have it under a form that simplifies further computations.

$$\begin{aligned}
\kappa(x; m, \sigma, a, b) &= \frac{1}{\sigma}\frac{\phi\left(\frac{x-m}{\sigma}\right)}{\Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right)}\mathbb{1}_{\{a\le x\le b\}} \\
&= \frac{\exp\left(-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}\right)}{C\left(m, \sigma, a, b\right)}\mathbb{1}_{\{a\le x\le b\}}
\end{aligned}$$

where

$$\begin{aligned}
C\left(m, \sigma, a, b\right) &\coloneqq \int_a^b \exp\left(-\frac{1}{2}\frac{(u-m)^2}{\sigma^2}\right)\mathrm{d}u \\
&= \sigma\sqrt{2\pi}\left(\Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right)\right)
\end{aligned}$$

and similarly, we denote by subscripts the partials derivatives:

$$\begin{aligned} C_m(m,\sigma,a,b) &\coloneqq \frac{\partial}{\partial m} C(m,\sigma,a,b) \\ &= \int_a^b \frac{u-m}{\sigma^2} \exp\left(-\frac{1}{2}\frac{(u-m)^2}{\sigma^2}\right) \mathrm{d}u \\ &= \left[-\exp\left(-\frac{1}{2}\frac{(u-m)^2}{\sigma^2}\right)\right]_a^b \\ &= \exp\left(-\frac{1}{2}\frac{(a-m)^2}{\sigma^2}\right) - \exp\left(-\frac{1}{2}\frac{(b-m)^2}{\sigma^2}\right) \ , \end{aligned}$$

and

$$\begin{aligned} C_\sigma(m,\sigma,a,b) \coloneqq& \frac{\partial}{\partial \sigma} C(m,\sigma,a,b) \\ =& \int_a^b \frac{(u-m)^2}{\sigma^3} \exp\left(-\frac{1}{2}\frac{(u-m)^2}{\sigma^2}\right) \mathrm{d}u \\ =& \left[-\frac{u-m}{\sigma}\exp\left(-\frac{(u-m)^2}{2\sigma^2}\right)\right]_a^b + \frac{1}{\sigma}\int_a^b \exp\left(-\frac{1}{2}\frac{(u-m)^2}{\sigma^2}\right)\mathrm{d}u \\ =& \frac{a-m}{\sigma}\exp\left(-\frac{(a-m)^2}{2\sigma^2}\right) - \frac{b-m}{\sigma}\exp\left(-\frac{(b-m)^2}{2\sigma^2}\right) \\ &+ \frac{1}{\sigma} C(m,\sigma,a,b) \ . \end{aligned}$$

Hence, we can precompute the kernel's derivatives with respect to m and σ:

$$\frac{\partial}{\partial m}\kappa(x;m,\sigma,a,b) = \left(\frac{x-m}{\sigma^2} - \frac{C_m(m,\sigma,a,b)}{C(m,\sigma,a,b)}\right)\kappa(x;m,\sigma,a,b) \ , \tag{4.2.6}$$

and

$$\frac{\partial}{\partial \sigma}\kappa(x;m,\sigma,a,b) = \left(\frac{(x-m)^2}{\sigma^3} - \frac{C_\sigma(m,\sigma,a,b)}{C(m,\sigma,a,b)}\right)\kappa(x;m,\sigma,a,b) \ . \tag{4.2.7}$$

Update equation for $m_{k,p}$

$$\begin{aligned}\frac{\partial \mathcal{L}_{k,\mathcal{P}}}{\partial m_{k,p}}\left(\Theta_{k,\mathcal{P}}\right) &= -\sum_{t\in\mathcal{A}_k}\sum_{i,t_i^{(p)}\leq t}\left(\frac{t-t_i^{(p)}-m_{k,p}}{\sigma_{k,p}^2}-\frac{C_m\left(m_{k,p},\sigma_{k,p},a,b\right)}{C\left(m_{k,p},\sigma_{k,p},a,b\right)}\right)\frac{\alpha_{k,p}\kappa_{k,p}\left(t-t_i^{(p)}\right)}{\lambda_{k,\mathcal{P}}(t)}\\ &= \left(\frac{m_{k,p}}{\sigma_{k,p}^2}+\frac{C_m\left(m_{k,p},\sigma_{k,p},a,b\right)}{C\left(m_{k,p},\sigma_{k,p},a,b\right)}\right)\sum_{t\in\mathcal{A}_k}P_p(t)\\ &\quad -\frac{\alpha_{k,p}}{\sigma_{k,p}^2}\sum_{t\in\mathcal{A}_k}\sum_{i,t_i^{(p)}\leq t}\frac{\left(t-t_i^{(p)}\right)\kappa_{k,p}\left(t-t_i^{(p)}\right)}{\lambda_{k,\mathcal{P}}(t)}\end{aligned} \tag{4.2.8}$$

Hence, by canceling the previous derivative,

$$m_{k,p}^{(n+1)} = \frac{\alpha_{k,p}^{(n)}\sum_{t\in\mathcal{A}_k}\sum_{i,t_i^{(p)}\leq t}\frac{\left(t-t_i^{(p)}\right)\kappa_{k,p}^{(n)}\left(t-t_i^{(p)}\right)}{\lambda_{k,\mathcal{P}}^{(n)}(t)}}{\sum_{t\in\mathcal{A}_k}P_p^{(n)}(t)} - {\sigma_{k,p}^{(n)}}^2\frac{C_m\left(m_{k,p}^{(n)},\sigma_{k,p}^{(n)},a,b\right)}{C\left(m_{k,p}^{(n)},\sigma_{k,p}^{(n)},a,b\right)}\ , \tag{4.2.9}$$

where $C_m\left(m,\sigma,a,b\right) \coloneqq \frac{\partial C}{\partial m}\left(m,\sigma,a,b\right)$, as previously defined.

Update equation for $\sigma_{k,p}$

$$\begin{aligned}\frac{\partial \mathcal{L}_{k,\mathcal{P}}}{\partial \sigma_{k,p}}\left(\Theta_{k,\mathcal{P}}\right) &= -\sum_{t\in\mathcal{A}_k}\sum_{i,t_i^{(p)}\leq t}\left(\frac{\left(t-t_i^{(p)}-m_{k,p}\right)^2}{\sigma_{k,p}^3}-\frac{C_\sigma\left(m_{k,p},\sigma_{k,p},a,b\right)}{C\left(m_{k,p},\sigma_{k,p},a,b\right)}\right)\frac{\alpha_{k,p}\kappa_{k,p}\left(t-t_i^{(p)}\right)}{\lambda_{k,\mathcal{P}}(t)}\\ &= \frac{C_\sigma\left(m_{k,p},\sigma_{k,p},a,b\right)}{C\left(m_{k,p},\sigma_{k,p},a,b\right)}\sum_{t\in\mathcal{A}_k}P_p(t)\\ &\quad -\frac{\alpha_{k,p}}{\sigma_{k,p}^3}\sum_{t\in\mathcal{A}_k}\sum_{i,t_i^{(p)}\leq t}\frac{\left(t-t_i^{(p)}(t)-m_{k,p}\right)^2\kappa_{k,p}\left(t-t_i^{(p)}\right)}{\lambda_{k,\mathcal{P}}(t)}\end{aligned} \tag{4.2.10}$$

Hence, by canceling the previous derivative,

$$\sigma_{k,p}^{(n+1)} = \left(\frac{C\left(m_{k,p}^{(n)}, \sigma_{k,p}^{(n)}, a, b\right)}{C_\sigma\left(m_{k,p}^{(n)}, \sigma_{k,p}^{(n)}, a, b\right)} \frac{\alpha_{k,p}^{(n)} \sum_{t \in \mathcal{A}_k} \sum_{i, t_i^{(p)} \leq t} \frac{\left(t - t_i^{(p)}(t) - m_{k,p}^{(n)}\right)^2 \kappa_{k,p}^{(n)}\left(t - t_i^{(p)}\right)}{\lambda_{k,\mathcal{P}}^{(n)}(t)}}{\sum_{t \in \mathcal{A}_k} P_p^{(n)}(t)} \right)^{1/3}, \tag{4.2.11}$$

where $C_\sigma(m, \sigma, a, b) \coloneqq \frac{\partial C}{\partial \sigma}(m, \sigma, a, b)$, as previously defined.

Finally, to ensure that the σ coefficient stays strictly positive in order to avoid computational errors, we add a projection step onto $[\varepsilon, +\infty)$, with $\varepsilon > 0$ that is pre-determined:

$$\sigma_{k,p}^{(n+1)} = \mathrm{proj}_{[\varepsilon, +\infty)}\left(\sigma_{k,p}^{(n+1)}\right) . \tag{4.2.12}$$

In practice, we set ε such that we avoid the overfitting that can occur when the kernel's mass is too concentrated. Note that once the initial values of the parameters are determined, the EM algorithm is entirely deterministic.

Algorithm 2: EM-based algorithm

input : $\mathcal{A}_k, \mathcal{T}_p, a, b, T, N$
output: The estimated values for parameters $\mu, \boldsymbol{\alpha}, \boldsymbol{m}$ and $\boldsymbol{\sigma}$

```
Initialize μ^(0), α^(0), m^(0), σ^(0)        // Eq (4.2.14), (4.2.15) and (4.2.16)
for i = 0, ..., N − 1 do
    if α^(i) = 0_{ℝ^{#𝒫}} then
        μ^(i+1) = μ^(MLE)                                   // Eq (4.2.13)
        break
    end
    Define λ^(i)                                            // Eq (4.1.2)
    Compute μ^(i+1), α^(i+1), m^(i+1), σ^(i+1)   // Eq (4.2.4), (4.2.5), (4.2.9), (4.2.12)
end
return μ^(i+1), α^(i+1), m^(i+1), σ^(i+1)
```

Also, when the estimate of parameter m is too far from the kernel's support $[a, b]$, we are in a pathological case where EM is diverging due to indeterminacy between setting $\alpha = 0$ and pushing m to infinity due to the discrete nature of our events. Thus, we consider that the stochastic process is not linked to the considered driver, and fall back to the MLE estimator defined in (4.2.13). The algorithm is therefore stopped and we set $\boldsymbol{\alpha} = \mathbf{0}_{\mathbb{R}^{\#\mathcal{P}}}$.

It is worth noting that if $\forall p \in \mathcal{P}, \alpha_{k,p} = 0$, then the intensity is reduced to its baseline, thus the negative log-likelihood is $\mathcal{L}_{k,p}(\Theta_{k,p}) = \mu_k T - \#\mathcal{A}_k \log \mu_k$, where $\#\mathcal{A}$ denotes the cardinality of the set $\mathcal{A}$. Thus, we can terminate the EM

algorithm by directly computing the MLE for μ_k, namely:

$$\mu_k^{(\mathrm{MLE})} = \frac{\#\mathcal{A}_k}{T} \ , \tag{4.2.13}$$

that corresponds to the average number of event per timestamps, like the parameter λ in a homogeneous Poisson Process.

Initialization strategy

We propose an initialization strategy based on *moment matching* [Bowman and Shenton, 2004], where parameters are initialized based on their role in the model. It reads:

$$\mu_k^{(0)} = \frac{\#\mathcal{A}_k - \#\left(\bigcup_{p\in\mathcal{P}} \mathcal{D}_{k,p}\right)}{T - \boldsymbol{\lambda}\left(\bigcup_{p\in\mathcal{P}}\bigcup_{t'\in\mathcal{T}_p}[t'+a\,,t'+b]\right)} \tag{4.2.14}$$

$$\alpha_{k,p}^{(0)} = \frac{\#\mathcal{D}_{k,p}}{\boldsymbol{\lambda}\left(\bigcup_{t'\in\mathcal{T}_p}[t'+a\,,t'+b]\right)} - \mu_k^{(0)}, \quad \forall p \in \mathcal{P} \tag{4.2.15}$$

$$m_{k,p}^{(0)} = \frac{1}{\#\mathcal{D}_{k,p}} \sum_{d\in\mathcal{D}_{k,p}} d \quad \text{and} \quad \sigma_{k,p}^{(0)} = \sqrt{\frac{1}{\#\mathcal{D}_{k,p}} \sum_{d\in\mathcal{D}_{k,p}} \left|d - m_{k,p}^{(0)}\right|^2}, \quad \forall p \in \mathcal{P} \ , \tag{4.2.16}$$

where $\boldsymbol{\lambda}(\cdot)$ denotes the Lebesgue measure, and where

$$\mathcal{D}_{k,p} \coloneqq \left\{t - t_*^{(p)}(t),\ t \in \mathcal{A}_k\right\} \cap [a\,,b]$$

is the set of all empirical delays possibly linked to the driver p, with

$$t_*^{(p)}(t) \coloneqq \max\left\{t',\ t' \in \mathcal{T}_p, t' \leq t\right\}$$

denoting the timestamp of the last event on driver p that occurred before time t.

Here, the initial baseline intensity $\mu^{(0)}$ is set to the average number of process' events that occur outside any kernel support, *i.e.,* the events that are guaranteed to be exogenous or spontaneous. Similarly, the kernel intensity $\alpha^{(0)}$ is computed as the increase in the average number of activations over the kernel support, compared to $\mu^{(0)}$. The initial guess for $m^{(0)}$ and $\sigma^{(0)}$ are obtained with their parametric estimates, *i.e.,* the mean and standard deviation of delays, considering that all event on the kernel support are assigned to the considered driver.

4.3 Experiments

We evaluated our model on several experiments, using both synthetic and empirical MEG data. We used `Python` [Python Software Foundation, 2019] and its scientific libraries [Virtanen et al., 2020, Hunter, 2007, Harris et al., 2020]. We relied on `alphacsc` for CDL with rank-1 constraints on MEG [Dupré la Tour et al., 2018] and we used `MNE` [Gramfort et al., 2013] to load and manipulate the MEG datasets. Computations were run on CPU Intel(R) Xeon(R) E5-2699, with 44 physical cores.

4.3.1 Evaluation of the EM convergence on synthetic data

For a given number of drivers and a set of corresponding parameters Θ, we first generate the drivers' processes and then simulate the stochastic process for a predetermined duration T. Each driver's timestamps are simulated as follows: given an interstimuli interval (ISI), a set of $S = \left\lfloor \frac{T}{\text{ISI}} \right\rfloor$ equidistant timestamps is generated – where $\lfloor \cdot \rfloor$ denotes the floor function. Then P timestamps are uniformly sampled without replacement from this set. In all our experiments, we fixed the ISI to 1 s for the "wide" kernel, and to 1.4 s for the "sharp" one. Finally, a one-dimensional non-homogeneous Poisson process is simulated following Lewis' thinning algorithm (Algorithm 3; Lewis and Shedler 1979), given the predefined intensity function λ and the drivers' timestamps.

Figure 4.3.1 illustrates the intensity function recovery with two drivers considered together: the first one has a "wide" kernel with standard deviation $\sigma = 0.2\,\text{s}$, and the second one has a "sharp" kernel with $\sigma = 0.05\,\text{s}$. Both kernels have support $[0.03\,\text{s}, 0.8\,\text{s}]$ and mean $m = 0.4\,\text{s}$, the coefficients α are both set to 0.8 and the baseline intensity parameter μ to 0.8. We report 8 estimated intensities obtained from independent simulations of the processes – using $T = 10\,000\,\text{s}$ and $P/S = 0.6$ – that we plot over each one of the driver's kernel's support. The EM algorithm is run for 50 iterations using the moment matching initialization strategy described in section 4.1. Note that here, the randomness only comes from the data generation, as the EM algorithm uses a deterministic initialization. Figures demonstrate that the EM algorithm is able to successfully recover the parameters for both shapes of kernels.

To provide a quantitative evaluation of the parameters' recovery, we compute, for each driver $p \in \mathcal{P}$, the ℓ_∞ norm between the intensity λ^* computed with the true parameters Θ_p^* and the estimated intensity λ_p with parameters $\widehat{\Theta}_p$:

$$\left\| \lambda^* - \widehat{\lambda}_p \right\|_\infty := \max_{t \in [0,T]} \left| \mu^* + \alpha_p^* \kappa_p^*(t) - \hat{\mu} - \hat{\alpha}_p \hat{\kappa}_p(t) \right| \quad . \tag{4.3.1}$$

Algorithm 3: Lewis and Shedler [1979], p.7, Algorithm 1, One-dimensional nonhomogeneous Poisson process simulation

Data: $\lambda(t)$, T

initialize $n = m = 0$, $t_0 = s_0 = 0$, $\bar{\lambda} = \max_{t \in [0,T]} \lambda(t)$;

while $s_m \leq T$ **do**

  Draw $u \sim \mathcal{U}nif_{[0,1]}$;

  $w \leftarrow -\ln(u)/\bar{\lambda}$; `// so that` $w \sim \mathcal{E}xp\left(\bar{\lambda}\right)$

  $s_{m+1} \leftarrow s_m + w$;

  Draw $D \sim \mathcal{U}nif_{[0,1]}$;

  if $D \leq \lambda(s_{m+1})/\bar{\lambda}$ **then**

    $t_{n+1} \leftarrow s_{m+1}$;

    $n \leftarrow n + 1$;

  end

  $m \leftarrow m + 1$;

end

if $t_n \leq T$ **then**

  return $\{t_k\}_{k=1,2,\ldots,n}$;

else

  return $\{t_k\}_{k=1,2,\ldots,n-1}$;

end

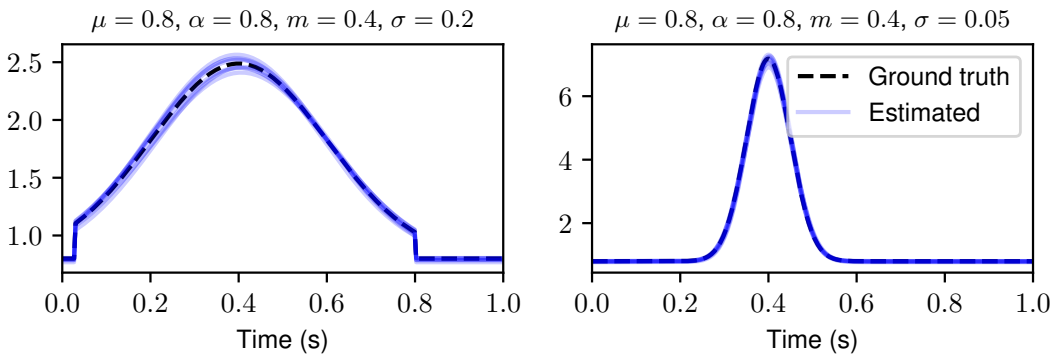


Figure 4.3.1: True and estimated intensity functions following a driving event at time zero for two different kernels, on synthetic data. **Left**: "wide" kernel with $\sigma = 0.2$. **Right**: "sharp" kernel with $\sigma = 0.05$. Parameters used are $T = 10000$, $P/S = 0.6$. On synthetic data, the EM algorithm successfully retrieves the true values of parameters, for both shapes of kernels.

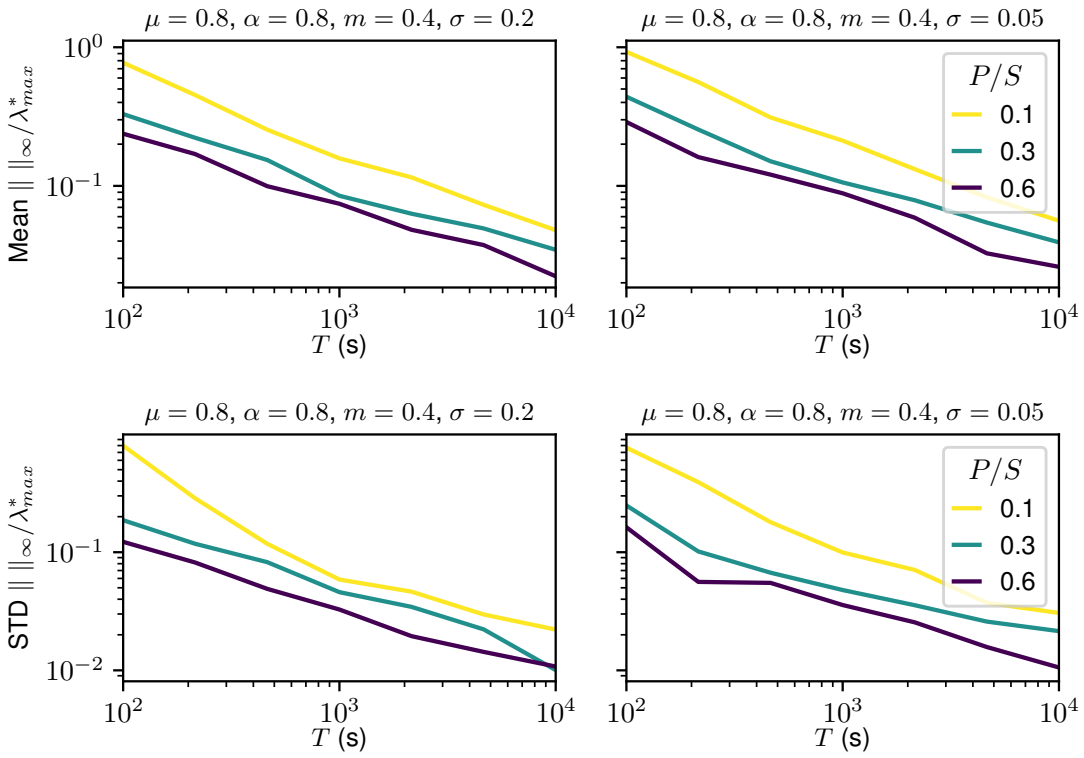


Figure 4.3.2: Mean (top) and standard deviation (bottom) of the relative infinite norm as a function of process duration T and the percentage of events kept P/S, for two kernel shapes on synthetic data: wide kernel (**left**) and sharp kernel (**right**). The accuracy of the EM estimates increases with longer and denser processes.

The rationale for using the ℓ_∞ norm is to ensure that errors during baseline and within the kernel support are given equal importance. Figure 4.3.2 presents the parameter recovery for the same scenario with varying P/S and T. To easily compare the EM performances on the two shapes of kernels, Figure 4.3.2 reports the mean and standard deviation of the relative ℓ_∞ norm – that is the ℓ_∞ divided by the maximum of the true intensity λ^* – computed for each of the driver over 30 repetitions with different realizations of the process. The results show that the more data are available, either due to a longer process duration (increase in T) or due to a higher event density (increase in P/S), the better are the parameter estimates. The convergence appears to be almost linear in these two cases. Moreover, the average computation time for an EM algorithm in Figure 4.3.2 took 18.16 s, showing the efficiency of our inference method. In addition, we report the scaling of the EM computation time as a function of T, presented in Figure 4.3.3. For each value of T, the mean computation time is computed over 90 experiments (3 values of P/S times 30 random seeds).

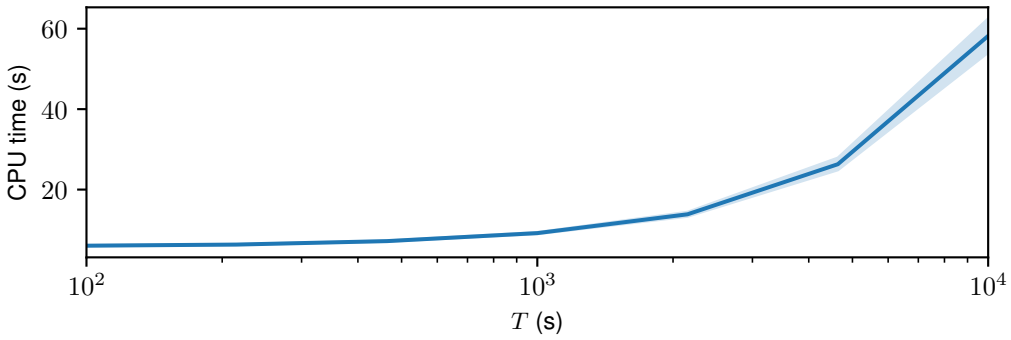


Figure 4.3.3: Mean and 95 % CI of computation time (in seconds) for one EM algorithm, as a function of the process duration T (in seconds). Results are obtained on synthetic data.

4.3.2 Evoked and induced effects characterization in MEG data

Datasets Experiments on MEG data were run on two datasets from `MNE` Python package [Gramfort et al., 2013, 2014]: the *sample* dataset and the somatosensory (*somato*) dataset[1]. These datasets were selected as they elicit two distinct types of event-related neural activations: evoked responses which are time locked to the onset of the driver process, and induced responses which exhibit random jitters. Complementary experiments were performed on the larger Cam-CAN dataset [Shafto et al., 2014][2]. Presentation of the dataset, data pre-processing and obtained results on 3 subjects are presented in subsection 4.3.4. The presented results are self-determined as they exhibit, for each subject, the atoms that have the higher ratio α/μ. For all studied datasets, full results are presented in supplementary materials.

The *sample* dataset contains M/EEG recordings of a human subject presented with audio and visual stimuli. In this experiment, checkerboard patterns are presented to the subject in the left and right visual field, interspersed by tones to the left or right ear. The experiment lasts about 4.6 min and approximately 70 stimuli per type are presented to the subject. The interval between the stimuli is on average of 750 ms, all types combined, with a minimum of 593 ms. Occasionally, a smiley face is presented at the center of the visual field. The subject was asked to press a button with the right index finger as soon as possible after the appearance of the face. In the following, we are only interested in the four main stimuli types: auditory left, auditory right, visual left, and visual right. For the *somato* dataset,

[1] Both available at https://mne.tools/stable/overview/datasets_index.html
[2] Available at https://www.cam-can.org/index.php?content=dataset

a human subject is scanned with MEG during 15 min, while 111 stimulations of his left median nerve were made. The minimum ISI is 7 s.

Experimental setting For both datasets, only the 204 gradiometer channels are analyzed. The signals are pre-processed using high-pass filtering at 2 Hz to remove slow drifts in the data, and are resampled to 150 Hz to limit the atom size in the CDL. CDL is computed using `alphacsc` [Dupré la Tour et al., 2018] with the `GreedyCDL` method. For the *sample* dataset, 40 atoms of duration 1 s each are extracted, and for the *somato* dataset, 20 atoms of duration 0.53 s are estimated. The extracted atoms' activations are binarized using a threshold of 6×10^{-11} (resp. 1×10^{-10}) for *sample* (resp. *somato*), and the times of the events are shifted to make them correspond to the peak amplitude time of the atom. Then, for every atom, the intensity function is estimated using the EM-based algorithm with 400 iterations and the moment matching initialization strategy. Kernels' truncation values are hyper-parameters for the EM and thus must be pre-determined. The upper truncation value b is chosen smaller than the minimum ISI. Here, we used in addition some previous domain knowledge to set coherent values for each dataset. Hence, for the *sample* (resp. *somato*) dataset, kernel support is fixed at $[0.03\,\mathrm{s}\,, 0.5\,\mathrm{s}]$ (resp. $[0\,\mathrm{s}\,, 2\,\mathrm{s}]$). See subsection 4.3.3 for an analysis on how these hyperparameters influence on the obtained results presented below.

Table 4.1 presents the main information related to real MEG datasets available with the `MNE` Python package [Gramfort et al., 2013, 2014].

Dataset	# Atoms	Duration of Atoms (s.)	# Atom's events	# Drivers	# Driver's events	Sequence length (min.)
sample	40	1	≈ 401.025	4	≈ 72.25	4.6
somatosensory	20	0.53	10408	1	111	15

Table 4.1: Statistics of each MNE dataset. $\approx N$ denotes that N is the average number.

Evoked responses in sample dataset Results on the *sample* dataset are presented in Figure 4.3.4. We plot the spatial and temporal representations of four selected atoms, as well as the estimated intensity functions related to the two types of stimuli: auditory (*blue*) and visual (*orange*). The first two atoms are specifically handpicked to exhibit the usual artifacts, and the last two are selected as they have the two bigger ratios α/μ for their respective learned intensity functions. Even though the intensity is learned with the two stimuli conjointly, we plot the two corresponding "intensities at the kernel" separately, *i.e.,* $\forall p \in \mathcal{P}, \forall t \in [0\,, 0.5]$, we plot $\lambda_{k,p}(t), k = 0, 1, 2, 6$.

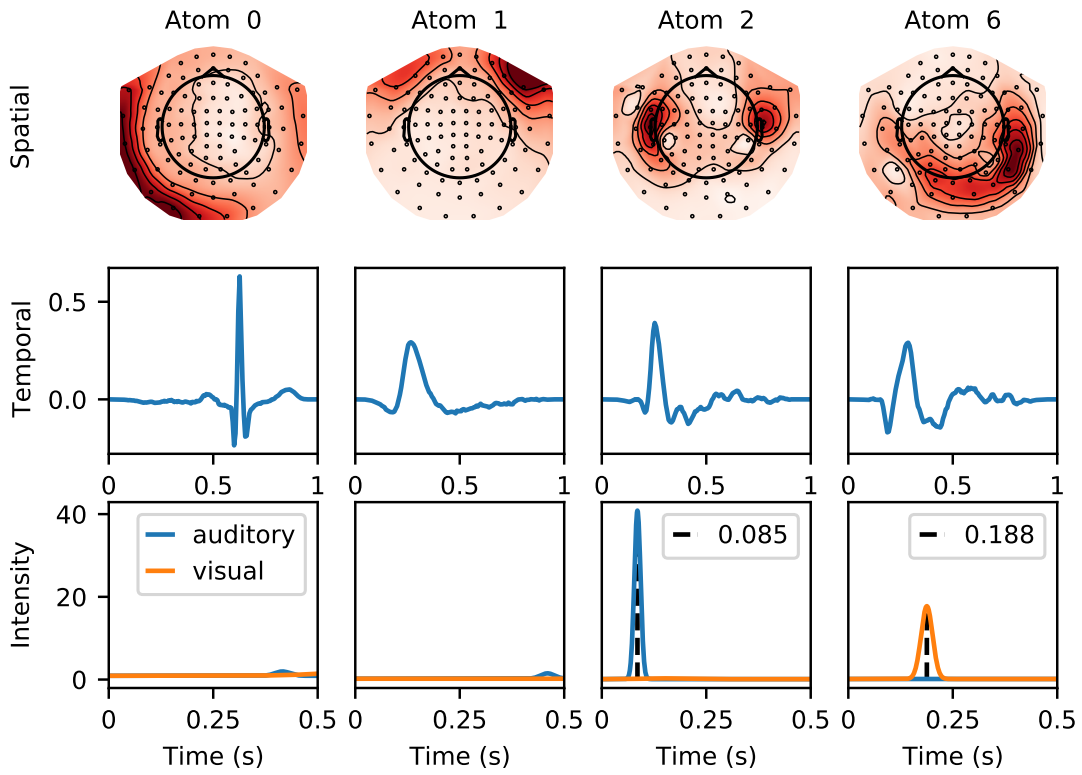


Figure 4.3.4: Spatial and temporal patterns of 4 atoms from *sample* dataset, and their respective estimated intensity functions following a stimulus (cue at time = 0 s), for auditory and visual stimuli. The heartbeat and eye-blink artifacts are not linked to any stimuli. An auditory stimulus will induce a neural response similar to atom 2, with a mean latency of 85 ms.

Spatial and temporal representations of atom 0 (resp. atom 1) indicate that it corresponds to the heartbeat (resp. the eye blink) artifact. These two atoms are thus expected not to be linked to any stimuli. This is confirmed by the shape of the intensities estimated with DriPP that is mostly flat, which indicates that the activation of these two atoms are independent of auditory and visual stimuli. Note that these two artifacts can also be recovered by an Independent Component Analysis (ICA), as shown in Figure 1.2.3. Indeed, the cosine similarity between the spatial maps of the eye blink (resp. the heartbeat) artifact extracted with CDL and its corresponding component in ICA analysis is 99.58 % (resp. 99.78 %), as presented in Figure 4.3.12. In contrast, by looking at the spatial and temporal patterns of atom 2 (resp. atom 6), it can be associated with an auditory (resp. visual) evoked response. Given the spatial topography of atom 2, we conclude to a bilateral auditory response and the peak transient temporal pattern suggests an evoked response that is confirmed by the estimated intensity function that contains a narrow peak around 85 ms post-stimulus. This is the M100 response – here the auditory one – well known in the MEG literature (its equivalent in EEG is the N100) [Näätänen and Picton, 1987]. The M100 is indeed a peak observed in the evoked response between 80 and 120 milliseconds after the onset of a stimulus in an adult population. Regarding atom 6, topography is right lateralized in the occipital region, suggesting a visual evoked response. This is confirmed by the intensity function estimated that reports a relationship between this atom and the visual stimuli. Here also, the intensity peak is narrow, which is characteristic of an evoked response. This reflects a right lateralized response along the right ventral visual stream in this subject. This may be connected to the P200, a peak of the electrical potential between 150 and 275 ms after a visual onset. Moreover, the intensities estimated with DriPP for the unrelated tasks are completely flat. We have $\alpha = 0$, which indicates that atoms' activations are exogenous or spontaneous relatively to unrelated stimuli. For comparison, we present in subsection 4.3.5 similar results obtained with dedicated MEG data analysis tools, such as evoked responses and time-frequency plots.

Induced response in somato dataset Results on the *somato* dataset are presented in Figure 4.3.5. Similar to the results on *sample*, spatial and temporal patterns of 3 handpicked atoms are plotted alongside the intensity functions obtained with DriPP. Thanks to their spatial and temporal patterns, and with some domain knowledge, it is possible to categorize these 3 atoms: atom 2 corresponds to a μ-wave located in the secondary somatosensory region (S2), atom 7 corresponds to an α-wave originating in the occipital visual areas, whereas atom 0 corresponds to the eye-blink artifact. As α-waves are spontaneous brain activity, they are not phase-locked to the stimuli. It is thus expected that atom 7 is not linked to the task, as confirmed by its estimated intensity function where $\alpha = 0$. For atom 2 – that corresponds to a μ-wave –, its respective intensity is nonflat with a broad peak close to 1 s, which characterizes an induced response. Moreover, similar to

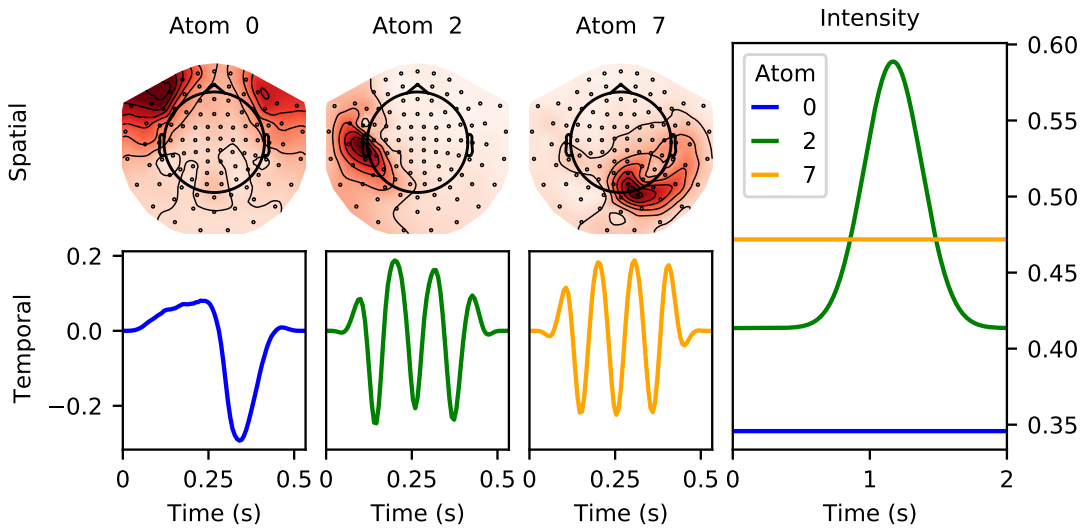


Figure 4.3.5: Spatial and temporal patterns of 3 atoms from *somato* dataset, and their respective estimated intensity functions following a somatosensory stimulus (cue at time = 0 s). The eye-blink artifact (atom 0) is not linked to the stimulus, and neither is the α-wave (atom 7). A somatosensory stimulus will induce a neural response similar to atom 2, with a mean latency of 1 s.

results on the *sample* dataset, we recover the eye-blink artifact that also has a flat intensity function. This allows us to be confident in the interpretation of the obtained results. Figure 4.3.6 shows 3 atoms that correspond all to a μ-wave located in the secondary somatosensory region (S2), with three different shapes of kernels in their estimated intensity functions. They have an estimated intensity similar to atom 2, *i.e.,* non-flat with a broad peak close to 1 s. The usual time/frequency analysis reported in Figure 1.2.2 exhibits the induced response of the μ-wave.

4.3.3 Impact of model hyperparameter

In this section, we dwell on the analysis of how setting hyperparameter values may impact the obtained results, with the aim of determining whether it is possible to set these parameters using a general rule of thumb, without degrading previous results. More specifically, we will look at the impact of two hyperparameters on the results we obtained on the MNE sample dataset: the threshold value – applied to the atoms' activation values to binarized them, currently set at 6×10^{-11} –, and the kernel support, currently set at $[0.03\,\text{s}, 0.5\,\text{s}]$ using previous and domain knowledge. To do so, we conducted two experiments on *sample* where each varies one hyperparameter. We plot the intensity function learned by DriPP for the same four atoms – namely, the two artifacts (heartbeat and eye blink) and the auditory

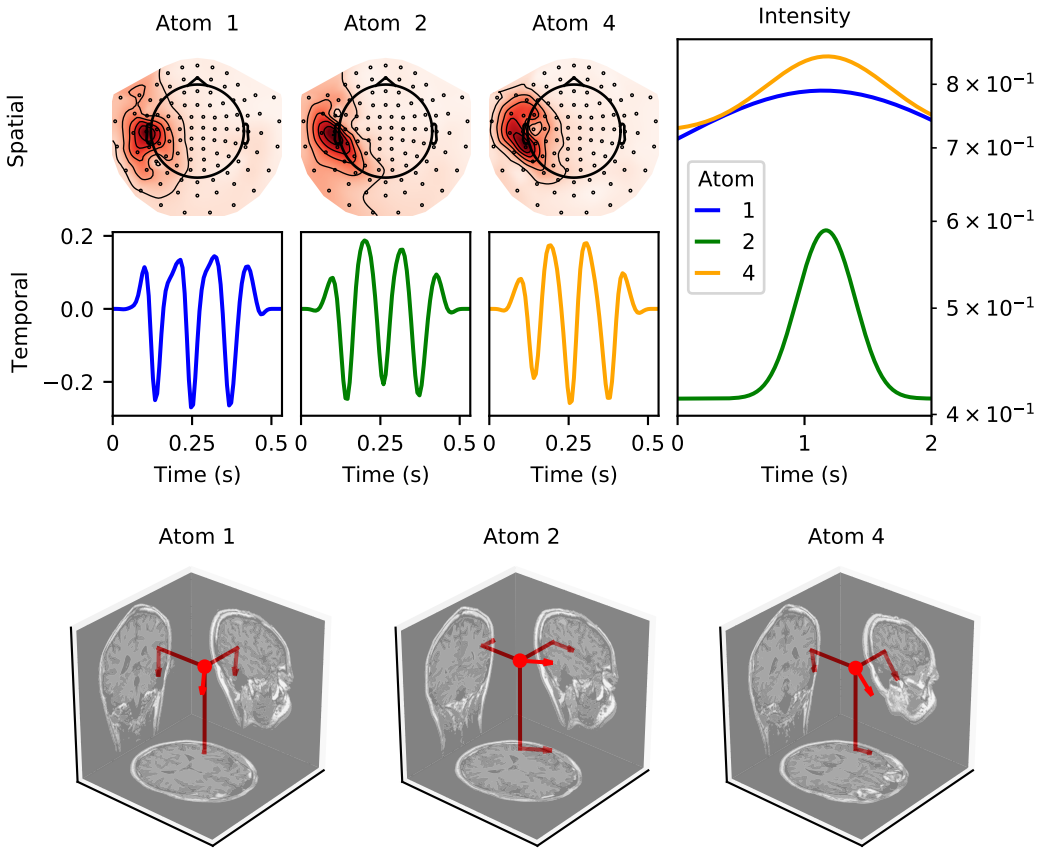


Figure 4.3.6: Spatial and temporal patterns of 3 μ-wave atoms from *somato* dataset, alongside with their respective estimated intensity functions. Below, in order to provide further information, are the corresponding brain locations for each atom obtained with dipole fitting.

and visual responses –, separately for the two stimuli (auditory and visual), similarly to Figure 4.3.4. We report its "true" value – the one presented in Figure 4.3.4 – in black déshed lines.

For the first experiment, presented in Figure 4.3.7, we varied the threshold, expressed as a percentile, between 0 and 80. A threshold of 20 means that we only keep activations whose values are above the 20 % percentile computed over all strictly positive activations, *i.e.,* the smaller the threshold is, the more activations are kept. The value of the threshold used in all other experiments is the 60 % percentile. One can observe that for the two artifacts (atoms 0 and 1), when the threshold gets smaller, the learned intensity functions get flatter, indicating that the stimulus has no influence on the atom activation. However, for the two others atoms, the effect of a smaller threshold is the opposite, as the intensity functions have a higher peak, indicating a bigger value of the α parameter, and thus strengthening the link stimulus-atom. Thus, the threshold value could be set to a small percentile and therefore computed without manual intervention, without degrading the current results.

For the second experiment, we now focus on the kernel truncation values a and b. Results are presented in Figure 4.3.8. We set $a = 0$, as we did for somato and Cam-CAN datasets, and we vary b between 0.5 – the current value – and 10, a large value compared to the ISI. One can observe that for $b = 0.5$ or $b = 1$, the results are either unchanged (atoms 1, 2, and 6) or better (atom 0, as the intensity function is totally flat for this artifact), indicating that setting $a = 0$ and b close to the average ISI of 0.750 s does not hinder the results. However, when b is too large, the results degrade quickly, up to the point where all learned intensities are flat lines, indicating that our model does not find any link between the stimuli and the atoms. This is due to the fact that this hyperparameter is of great importance in the initialization step, as the greater b is, the more atom's activations are considered being on a kernel support. Thus, setting the upper truncation value to a value close to the average ISI seems to give reliable results.

4.3.4 Experiments on Cam-CAN dataset

The Cam-CAN dataset contains data of M/EEG recordings of 643 human subjects submitted to audio and visual stimuli. In this experiment, 120 bimodal audio/visual trials and eight unimodal trials – included to discourage strategic responding to one modality (four visual only and four auditory only) – are presented to each subject. For each bimodal trial, participants see two checkerboards presented to the left and right of a central fixation (34 milliseconds duration) and simultaneously hear a 300 milliseconds binaural tone at one of three frequencies (300 Hz, 600 Hz, or 1200 Hz, equal numbers of trials pseudorandomly ordered). For

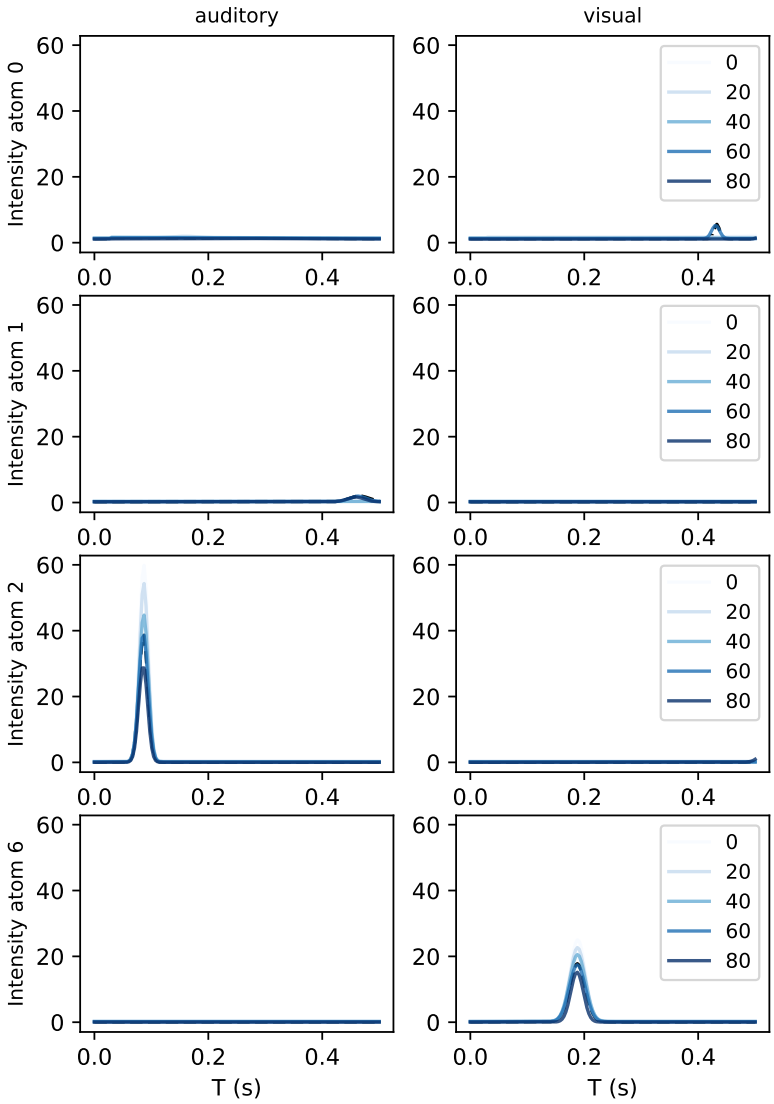


Figure 4.3.7: Influence of the threshold (expressed as a percentile) on the obtained results on MNE *sample* dataset, for the 4 main atoms, for auditory and visual stimulus. The value of the threshold presented in Figure 4.3.4 is $\tau = 60\,\%$. The threshold value has limited impact on obtained results, and thus could be determined with a general rule of thumb, as a percentile over all activations values.

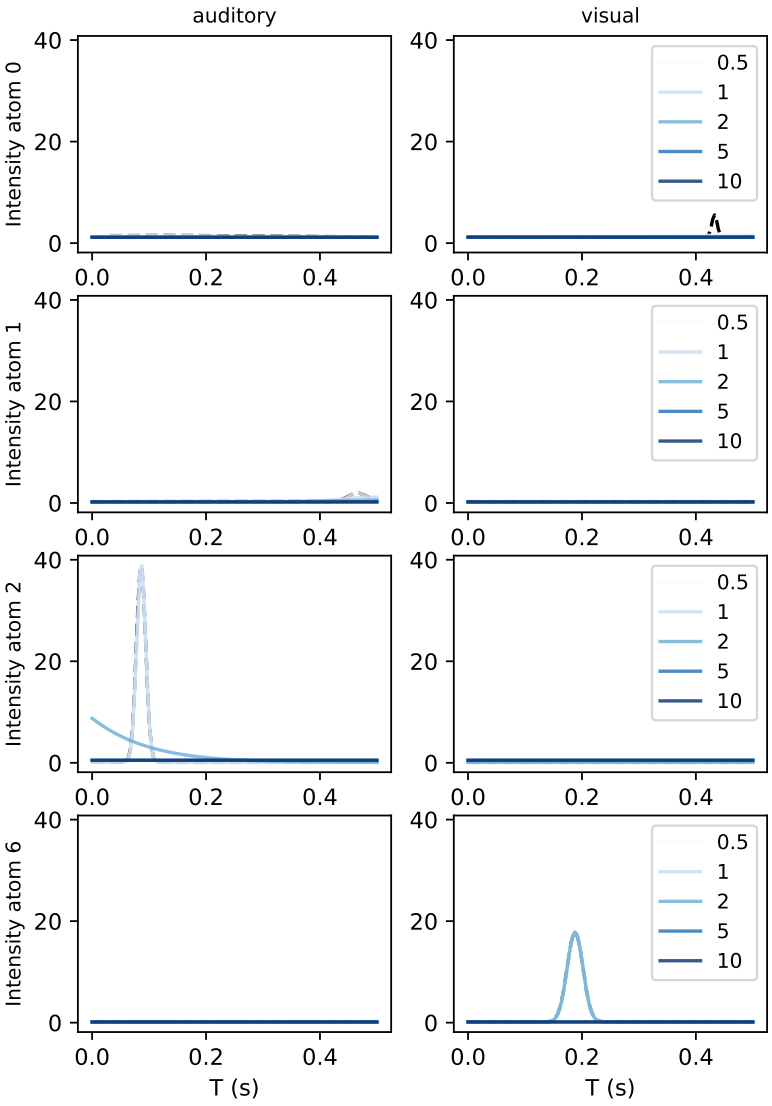


Figure 4.3.8: Influence of the kernel truncation upper bound b (with $a = 0$) on the obtained results on MNE *sample* dataset, for the 4 main atoms, for auditory and visual stimulus. By setting b too high, all intensity functions are completely flat, indicating a total loss of information.

unimodal trials, participants either only hear a tone or see the checkerboards. For each trial, participants respond by pressing a button with their right index finger if they hear or see any stimuli [Shafto et al., 2014]. For each subject, the experiment lasts less than 4 min.

The signals are pre-processed using low-pass filtering at 125 Hz to remove slow drifts in the data, and are resampled to 150 Hz to limit the atom size in the CDL. CDL is computed using `alphacsc` [Dupré la Tour et al., 2018] with the `GreedyCDL` method. Twenty atoms of duration 0.5 s are extracted from the signals. The extracted atoms'activations are binarized and, similarly as previous experiments, the events time are shifted to make them correspond to the peak amplitude time in the atom. Then, for every atom, the intensity function is estimated using the EM-based algorithm with 200 iterations and the moment matching initialization strategy. Kernels' truncation values are set at $[0\,\text{s}, 0.9\,\text{s}]$. Two drivers per atom are considered. The first driver contains timestamps of all bimodal trials (all frequencies combined) with in addition the 4 auditory unimodal stimuli (denoted `audivis_catch0`). The second driver is similar to the first one, but instead of the auditory unimodal stimuli, it contains the 4 visual unimodal stimuli (denoted `audivis_catch1`).

This experiment is performed for 3 subjects, for each one we plot the 5 atoms that have the highest ratio α/μ, so that we automatically exhibit atoms that are highly linked to the presented stimuli. Similarly as results presented in section 4.3, we plot the spatial and temporal representation of each atom, as well as the two learned intensity functions that we plot "at kernel". Results are presented on Figure 4.3.9, Figure 4.3.10 and Figure 4.3.11.

First, we can observe that, as for experiments on *sample* and *somato* datasets, we recover the heartbeat artifact (atom 0 in Figure 4.3.10) as well as the eye-blink artifact (atoms 0 in Figure 4.3.9 and in Figure 4.3.11). Because of the paradigm of the experiment, and in particular the instructions given to the subjects – that is, to press a button after every stimulus, the majority of them being visual –, it is not surprising if the eye blink artifact is slightly linked to the stimuli. Indeed, it is often observed in such experiments – where there is no designated time to blink – that the subject "allows" themself to blink after the visual stimulus.

As the majority of the presented stimuli are a combination of visual and auditory, the CDL model struggles to separate the two corresponding neural responses. Hence, that is why it can be observed that most of the first atoms reported exhibit a mixture between auditory and visual response in their topographies (first row). However, one can observe that for some such atoms, DriPP learned intensity is able to indicate what is the main contributing stimulus in the apparition of the atom. For instance, for atom 10 in Figure 4.3.9, the auditory stimulus is the main

responsible stimulus of the presence of this atom, despite this latter presenting a mixture of both neural responses. A similar analysis can be made for atom 1 in Figure 4.3.10, but this time for the auditory stimulus.

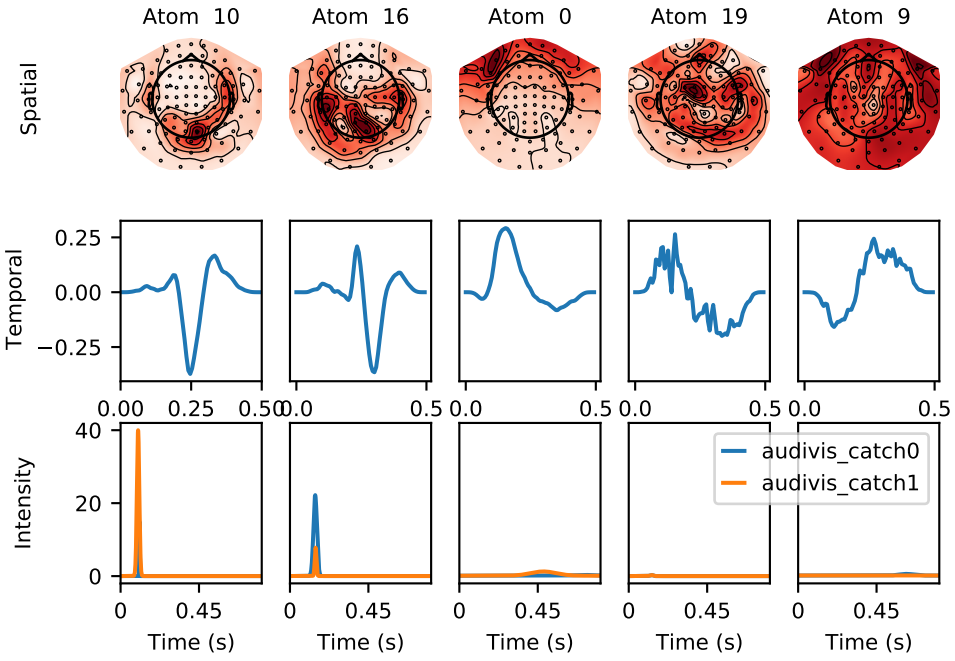


Figure 4.3.9: Spatial and temporal patterns of the 5 atoms from Cam-CAN dataset subject CC620264, a 76.33-year-old female, and their respective estimated intensity functions following an audiovisual stimulus (cue at time = 0 s) (`audivis_catch0` represents timestamps for bimodal trials and 4 auditory stimuli, while `audivis_catch1` represents the same but with visual stimuli instead of auditory). Atoms are ordered by their bigger ratio α/μ.

4.3.5 Usual M/EEG data analysis

We present in this final section some results obtained using usual M/EEG data analysis, such as Independent Component Analysis (ICA), epoch averaging, or time/frequency analysis. See section 1.2 for technical details on the usual methods for analysing neurophysiological data. First, on MNE *sample* dataset, we proceed to an ICA to manually identify usual artifacts. To do so, similarly as the CDL pre-processing, the raw signals are filtered (high-pass filter at 2 Hz), and 40 independent components are fitted. The two components 0 and 1, that we manually identify as corresponding to the eye blink and heartbeat artifacts respectively, are presented in fig. 1.2.3. In fig. 4.3.12, we associate for each of the CDL atoms presented in fig. 4.3.4 the ICA component that has the maximum cosine similarity.

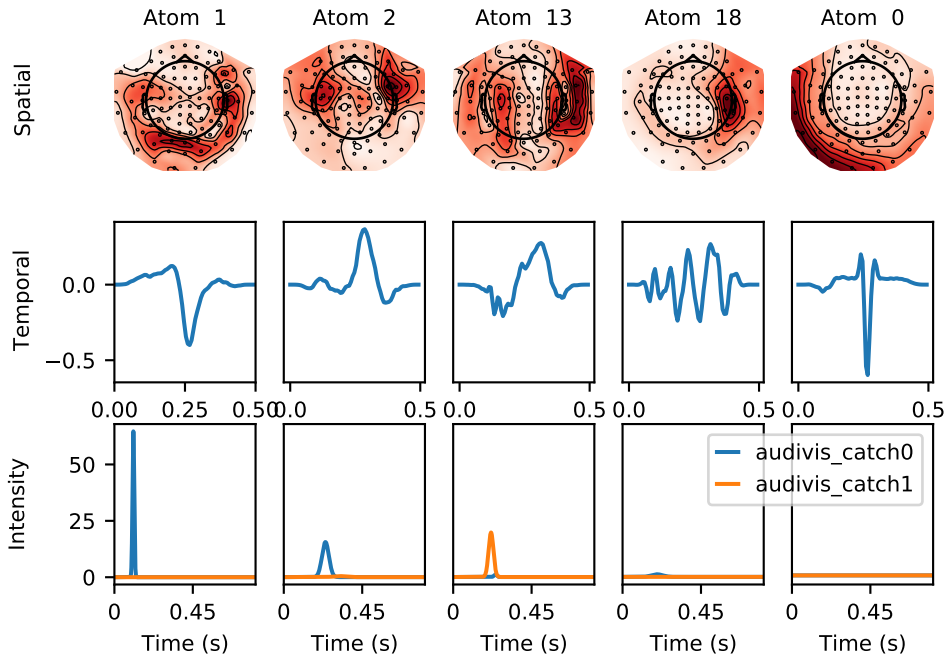


Figure 4.3.10: Spatial and temporal patterns of the 5 atoms from Cam-CAN dataset subject CC520597, a 64.25-year-old male, and their respective estimated intensity functions following an audiovisual stimulus (cue at time = 0 s) (`audivis_catch0` represents timestamps for bimodal trials and 4 auditory stimuli, while `audivis_catch1` represents the same but with visual stimuli instead of auditory). Atoms are ordered by their bigger ratio α/μ.

One can observe that the artifact atoms and components are highly similar, suggesting that CDL and ICA have equal performance on the artifact detection. Note that this high similarity is only based on the spatial pattern. Indeed, ICA does not provide temporal waveforms for the atoms as well as their temporal onsets, contrarily to CDL.

However, for the auditory and visual response, the result is different. For the auditory one (atom 1), there is not really an ICA equivalent, as it is most correlated with the eye-blink ICA component. Regarding the visual atom (atom 6), there is an ICA component that presents a high similarity. While the two related components correspond to neural sources on the occipital cortex, the atom 6 obtained with CSC is more right lateralized, suggesting a source in the right ventral visual stream. Note that unlike ICA, which recover full time courses for each source, CDL also provides the onset of the patterns, which we later use for automated identification of event related components. Finally, this demonstrates that CDL is a strong competitor to ICA for artifact identification, while simultaneously enabling to reveal evoked or induced neural responses in an automated way.

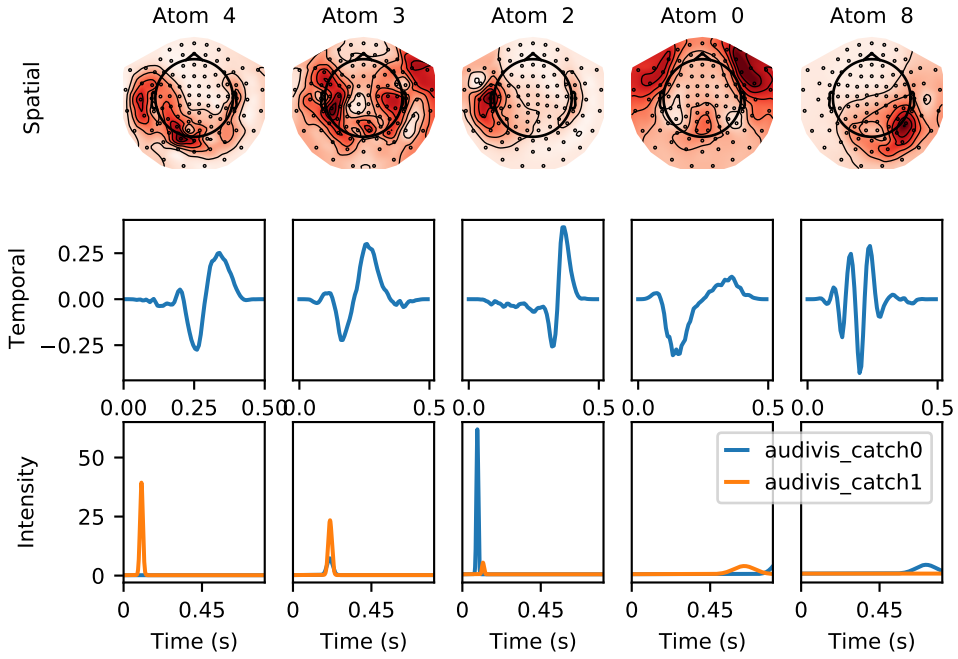


Figure 4.3.11: Spatial and temporal patterns of the 5 atoms from Cam-CAN dataset subject CC723395, a 86.08-year-old female, and their respective estimated intensity functions following an audiovisual stimulus (cue at time = 0 s) (`audivis_catch0` represents timestamps for bimodal trials and 4 auditory stimuli, while `audivis_catch1` represents the same but with visual stimuli instead of auditory). Atoms are ordered by their bigger ratio α/μ.

As mentioned, the ICA is commonly used to remove manually identified artifacts to reconstruct the original signals free of those artifacts. However, there are two drawbacks of this method, the first one being that some domain-related knowledge is needed in order to correctly identify the artifacts. The second drawbacks happen when the signal is reconstructed after the removal of certain components. Indeed, such a reconstruction will lead to a loss of information across all channels, as the artifacts are shared by all the sensors. Thanks to the Convolutional dictionary learning (CDL) that extracts the different artifacts directly from the raw data, our method does not suffer from these drawbacks.

Still on MNE *sample* dataset, we compute from raw data the epoch average following an auditory stimulus (stimuli in the left ear and in the right ear are combined) and plot on fig. 1.2.4 the obtained evoked signals. fig. 4.3.13 is similar but for the visual stimuli (again, stimuli on the left visual field and on the right visual field are combined).

On *somato* dataset, in order to exhibit the induced response to the somatosensory stimulus on the left median nerve of the subject, we perform a time/frequency

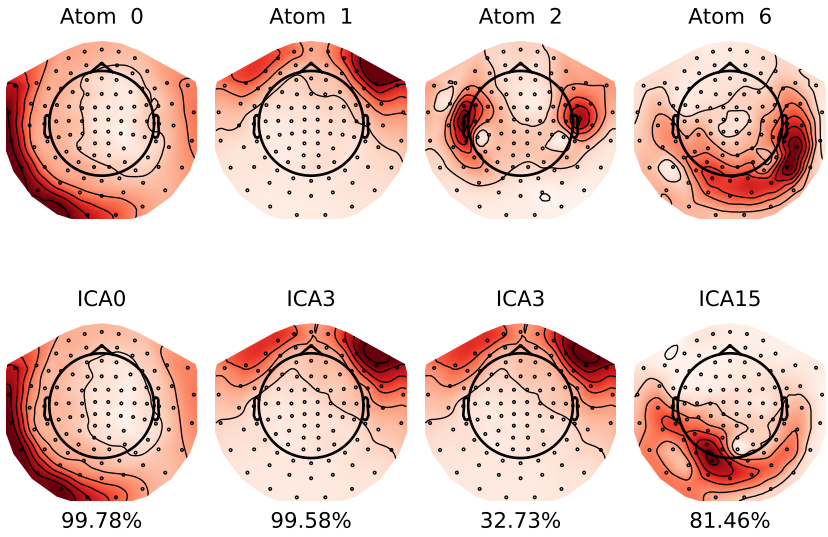


Figure 4.3.12: Spatial representation of the four atoms presented in fig. 4.3.4 (top) alongside their ICA components with the maximum cosine similarity (value indicated at the bottom). Atoms and ICA are computed on MNE *sample* dataset.

analysis, presented on fig. 1.2.2. In order to perform this analysis, a complex process including the use of Morlet wavelets is performed.

Finally, note that these methods that are commonly used in the M/EEG data analysis comprise a thoughtfully data pre-processing as well as a manual intervention requiring domain knowledge to identify both artifacts and evoked and induced responses. As the proposed method in this paper is composed of a unified pipeline, it is a significant gain. Indeed, DriPP is able to automatically isolate artifacts without prior removal of artifacts using ICA, as well as capture the diversity of latencies corresponding to induced responses.

Regarding the statistical significance of the link between a stimulus and a neural pattern, one can consider performing a statistical test, where one wishes to reject the null hypothesis of independence. We have performed a statistical test using a student t-test to check if the mean activation probability on $[a, b]$ segments is the same as the mean estimated on the baseline $[0, T] \setminus \bigcup_{t_i} [t_i + a, t_i + b]$. We present the results of the performed test in table 4.2. We can clearly see that both artifacts (heartbeat and eye-blink) are not linked to stimuli while neural responses are linked to related stimuli (p-values are very small). We would like to stress that while this test is interesting, it is much more dependent on the selection of the support interval $[a, b]$ than the proposed method. Moreover, this test does not

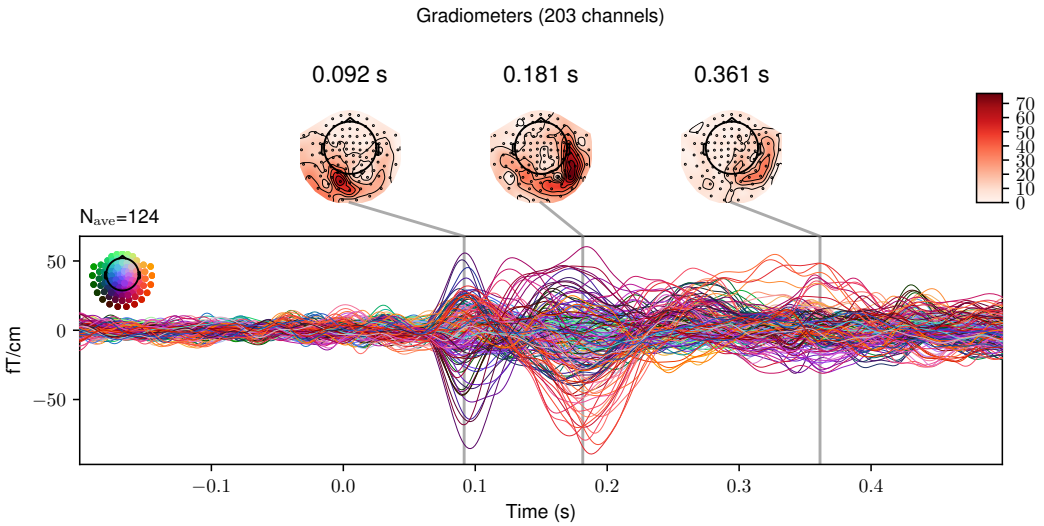


Figure 4.3.13: Evoked signals following a visual stimulus (cue at time = 0) on MNE *sample* dataset. Baseline correction applied from beginning of the data until time point zero.

atom id	atom type	p-value auditory	p-value visual
0	heartbeat	2.25×10^{-1}	1.19×10^{-1}
1	eye-blink	5.85×10^{-1}	8.48×10^{-1}
2	auditory	2.31×10^{-97}	6.31×10^{-1}
6	visual	7.78×10^{-1}	5.12×10^{-50}

Table 4.2: Statistical univariate Student t-test on MNE *sample* dataset, H_0: independence between an atom and the stimulus (auditory or visual). The atom ids correspond to the ones presented in fig. 4.3.4.

allow to assess the latency of the responses, and whether the atom is an induced response or an evoked one.

4.4 Transcending limits with discretised parametric kernels

This work proposed a point process (PP) based approach specially designed to model how external stimuli can influence the occurrences of recurring spatio-temporal patterns, called *atoms*, extracted from M/EEG recordings using convolutional dictionary learning (CDL). The key advantage of the developed method

is that by estimating few parameters (one baseline parameter and 3 parameters per considered driver), it provides a direct statistical characterization of when and how each stimulus is responsible for the occurrences of neural responses. Importantly, it can achieve this with relatively limited data which is well adapted to MEG/EEG experiments that last only a few minutes, hence leading to tens or hundreds of events at most. This work proposed an EM algorithm derived for a novel kernel function: the truncated Gaussian, which differs from the usual parametrization in PP models that capture immediate responses, *e.g.,* with exponential kernels. As opposed to competing methods that can involve manual selection of task-related neural sources, DriPP offers a unified approach to extract waveforms and automatically select the ones that are likely to be triggered by the considered stimuli. Note however that DriPP has been developed based on a point process framework, which is event-based. When working with continuous stimuli, other techniques must be considered, *e.g.,* spatio-temporal response functions (STRF; Drennan and Lalor 2019).

This chapter also unveils inherent limitations, most notably in the realms of kernel function selection and computational efficiency. The choice of kernel functions has proven to be a double-edged sword; while non-parametric kernels offer flexibility, they require a considerably larger dataset for robust statistical inference. Parametric kernels like the exponential kernel, although computationally less demanding, impose their own set of constraints, such as a fixed decay parameter, which may not be universally applicable across different experimental setups. The next chapter aims to address these limitations by introducing a novel inference method known as FaDIn. This approach will bring improvements in kernel selection flexibility and computational scalability. Specifically, FaDIn employs a discretization strategy that enhances computational efficiency without sacrificing statistical robustness, thereby offering a more versatile and efficient framework for TPP modeling in neuroscience.

Chapter 5

FaDIn: Fast Discretized Inference for Hawkes Processes with General Parametric Kernels

Contents

5.1 Mathematical formulation **99**

5.1.1 FaDIn 100

5.1.2 Impact of the discretization 103

5.2 Numerical experiments **109**

5.2.1 Consistency of Discretization 109

5.2.2 Statistical and computational efficiency of FaDIn . . . 112

5.2.3 Sensitivity analysis regarding the parameter W 116

5.3 Application to MEG data **117**

5.4 Discussion **119**

The content of this chapter was carried out in collaboration with Guillaume Staerman and was published in:

Guillaume Staerman, Cédric Allain, Alexandre Gramfort, and Thomas Moreau. FaDIn: Fast Discretized Inference for Hawkes Processes with General Parametric Kernels. *International Conference on Machine Learning*, 2023

As we saw, the statistical framework of Temporal Point Processes (TPPs; see *e.g.,* Daley and Vere-Jones 2003) is well adapted for modeling event-based data, as it offers a principled way to predict the rate of events as a function of time and the previous events' history. Multivariate Hawkes processes (MHP; Hawkes 1971) are likely the most popular, as they can model interactions between each univariate process, as well as self-excitation behavior, *i.e.,* a past event will increase the probability of having another event in the future on the same process.

A key feature of MHP modeling is the choice of kernels that can be either non-parametric or parametric. In the non-parametric setting, kernel functions are approximated by histograms [Lewis and Mohler, 2011, Lemonnier and Vayatis, 2014], by a linear combination of pre-defined functions [Zhou et al., 2013a, Xu et al., 2016] or, alternatively, by functions lying in a Reproducing kernel Hilbert space (RKHS; Yang et al. 2017). In addition to the frequentist approach, many Bayesian approaches, such as Gibbs sampling [Ishwaran and James, 2001] or (stochastic) variational inference [Hoffman et al., 2013], have been adapted to MHP in particular to fit non-parametric kernels. Bayesian methods also rely on the modeling of the kernel by histograms [Donnet et al., 2020] or by a linear combination of pre-defined functions [Linderman and Adams, 2015]. These approaches are designed whether in continuous-time [Rasmussen, 2013, Zhang et al., 2018, Donnet et al., 2020, Sulem et al., 2021] or in discrete-time [Mohler et al., 2013, Linderman and Adams, 2015, Zhang et al., 2018, Browning et al., 2022]. These functions allow great flexibility for the shape of the kernel, yet this comes at the risk of poor estimation of it when only a small amount of data is available [Xu et al., 2017].

Another approach is to consider parametrized kernels to estimate the intensity function. Although it can introduce a potential bias by assuming a particular kernel shape, this approach has several benefits. First, it reduces inference burden, as the parameter, say η, is typically lower dimensional than non-parametric kernels. Moreover, for kernels satisfying the Markov property [Bacry et al., 2015], computing the conditional intensity function is linear in the total number of timestamps/events. The most popular kernel belonging to this family is the exponential kernel [Ogata, 1981]. It is defined by $\eta = (\alpha, \gamma) \mapsto \alpha\gamma \exp(-\gamma t)$, where α and γ are the scaling and the decay parameters, respectively [Veen and Schoenberg, 2008, Zhou et al., 2013b]. However, as pointed out by Lemonnier and Vayatis [2014], the maximum likelihood estimator for MHP with exponential kernels is efficient only if the decay γ is fixed. Thus, only the scaling parameter α is usually inferred. This implies that the hyperparameter γ must be chosen in advance, usually using a grid search, a random search, or Bayesian optimization. This leads to a computational burden when the dimension of the MHP is high. The second option is to define a γ decay parameter common to all kernels, which results in a loss of expressiveness of the model. In both cases, the relevance of the exponential kernel relies on the choice of the decay parameter, which may not be adapted to the data [Hall

and Willett, 2016]. For more general parametric kernels which do not verify the Markov property, the inference procedure with both MLE or ℓ_2 loss scales poorly as they have quadratic computational scaling with the number of events, making their use limited in practice (see *e.g.,* Bompaire, 2019, Chapter 1). Recently, neural network-based MHP estimation has been introduced, offering, with sufficient data, relevant models at the cost of high computational cost [Mei and Eisner, 2017, Shchur et al., 2019, Pan et al., 2021]. These limitations for parametric and non-parametric kernels prevent their usage in some applications, as pointed out by Carreira [2021] in finance or in neuroscience (*cf.* chapter 4) The latter application is one of the main motivations for this work.

This work proposes a new inference method – named FaDIn – to estimate any parametric kernels for Hawkes processes. Our approach is based on two key features. First, we use finite-support kernels and a discretization applied to the empirical risk minimization (ERM; Reynaud-Bouret and Rivoirard 2010, Hansen et al. 2015, Bacry et al. 2020) inspired least-squares loss. Second, we propose to employ some precomputations that significantly reduce the computational cost. We then show, empirically and theoretically, that the implicit bias induced by the discretization procedure is negligible compared to the statistical error. Further, we highlight the efficiency of FaDIn in computation and statistical estimation over the non-parametric approach. Finally, we demonstrate the benefit of using a general kernel with MEG data. The flexibility of FaDIn allows us to model neural response to external stimuli with a much better-adapted kernel than the existing method derived in chapter 4.

5.1 Mathematical formulation

Recall from chapter 3 that given p sets of timestamps $\mathscr{F}_T^i = \{t_n^i,\ t_n^i \in [0, T]\}_{n=1}^{N_T^i}$, $i = 1, \dots, p$, each process of a Multivariate Hawkes processes (MHP; Hawkes, 1971) is described by the following intensity function:

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{p} \int_0^t \phi_{ij}(t-s)\ \mathrm{d}N_s^j \ , \tag{5.1.1}$$

where μ_i is the baseline parameter, $[N_t^1, \dots, N_t^p]$ the associated multivariate counting process and $\phi_{ij} : \mathbb{R}_+ \to \mathbb{R}_+$ the excitation function – called *kernel* – representing the influence of j-th process' past events onto i-th process' future events.

In this chapter, we will focus on the least squares loss inspired from empirical risk minimization (ERM; eq. (3.4.2)). We denote $\mathscr{F}_T \coloneqq \{\mathscr{F}_T^1, \dots, \mathscr{F}_T^p\}$ the set of all considered timestamps across all processes, $N_T \coloneqq \sum_{i=1}^p N_T^i$ the total number of timestamps, and where $\theta \coloneqq (\boldsymbol{\mu}, \boldsymbol{\eta})$.

5.1.1 FaDIn

The approach we propose in this work fills the need for general parametric kernels in many applications. We provide a computationally and statistically efficient solver – coined FaDIn – that works with many parametric kernels using gradient-based algorithms. Precisely, it relies on three key ideas: (i) the use of parametric finite-support kernels, (ii) a discretization of the time interval $[0, T]$, and (iii) precomputations allowing an efficient optimization procedure detailed below.

Finite support kernels

A core bottleneck for MLE or ℓ_2 estimation of parametric kernels is the need to compute the intensity function for all events. For general kernels, the intensity function usually requires $\mathcal{O}\left((N_T)^2\right)$ operations, which makes it intractable for long-time-length processes. To make this computation more efficient, we consider finite support kernels. Using a finite support kernel amounts to setting a limit in time for the influence of a past event on the intensity, *i.e.*, $\forall t \notin [0\,, W]\,, \phi_{ij}(t) = 0$, where $W \ll T$ denotes the length of the kernel's support. This assumption matches applications where an event cannot have influence far in the future, such as in neuroscience (as in chapter 4 with DriPP and in Krumin et al. 2010, Eichler et al. 2017), genetics Reynaud-Bouret and Schbath [2010] or high-frequency trading [Bacry et al., 2015, Carreira, 2021]. The intensity function eq. (5.1.1) can then be reformulated as a convolution between the kernel ϕ_{ij} and the sum of Dirac functions $z_i \coloneqq \sum_{t_n^i \in \mathcal{F}_T^i} \delta_{t_n^i}$ located at the event occurrences t_n^i:

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{p} \left(\phi_{ij} * z_j\right)(t), \quad t \in [0\,, T] \quad . \tag{5.1.2}$$

where

$$\left(\phi_{ij} * z_j\right)(t) \coloneqq \int_{-\infty}^{+\infty} \phi_{ij}(t-s) z_j(s)\,\mathrm{d}s = \int_{t-W}^{t} \phi_{ij}(t-s) z_j(s)\,\mathrm{d}s \quad ,$$

as ϕ_{ij} is finite support. Thus, the intensity can be computed efficiently with this formula. Indeed, only events in the interval $[t - W\,, t]$ need to be considered. See section 5.2.3 for more details.

Discretization

To make these computations even more efficient, we propose to rely on discretized processes. Most Hawkes processes estimation procedures involve a continuous

paradigm to minimize (3.4.2) or its log-likelihood counterpart. Discretization has been investigated so far for non-parametric kernels [Kirchner, 2016, Kirchner and Bercher, 2018, Kurisu, 2016]. The discretization of a TPP consists in projecting each event t_n^i on a regular grid $\mathcal{G} = \{0, \Delta, 2\Delta, \dots, G\Delta\}$, where $G = \left\lfloor \frac{T}{\Delta} \right\rfloor$. We refer to Δ as the stepsize of the discretization[1]. Here $\lfloor \cdot \rfloor$ denotes the floor function. Let $\widetilde{\mathscr{F}_T^i}$ be the set of projected timestamps of $\mathscr{F}_T^i$ on the grid $\mathcal{G}$, with the projection done to the nearest grid point. The intensity function of the i-th process of our discretized MHP is defined as:

$$\begin{aligned}\tilde{\lambda}_i[s] &= \mu_i + \sum_{j=1}^{p} \sum_{\tilde{t}_m^j \in \widetilde{\mathscr{F}}_{s\Delta}^j} \phi_{ij}(s\Delta - \tilde{t}_m^j) \\ &= \mu_i + \sum_{j=1}^{p} \underbrace{\sum_{\tau=0}^{L} \phi_{ij}^{\Delta}[\tau] z_j[s-\tau]}_{(z_j * \phi_{ij}^{\Delta})[s]}, \quad s \in [\![0\,, G]\!]\,, \end{aligned} \tag{5.1.3}$$

where $L \coloneqq \left\lfloor \frac{W}{\Delta} \right\rfloor$ denotes the number of points on the discretized support, where $\phi_{ij}^{\Delta}[\tau] = \phi_{ij}(\tau\Delta)$ is the kernel value on the grid and where $\forall s \in [\![0\,, G]\!]$, $z_i[s] \coloneqq \#\left\{|t_n^i - s\Delta| \leq \frac{\Delta}{2}\right\}$ denotes the number of events projected on the grid timestamp s. Note that for $s \notin [\![0\,, G]\!]$, $z_i[s] = 0$. From now and throughout the rest of the chapter, we denote $\phi_{ij} : \mathbb{R}_+ \to \mathbb{R}_+$ as a function while $\phi_{ij}^{\Delta} \in \mathbb{R}_+^{L+1}$ represents the associated discrete vector, with $\phi_{ij}^{\Delta}[t]$ being the t-th element of that vector. Compared to the continuous formulation, the intensity function can be computed more efficiently as one can rely on discrete convolutions, whose worst-case complexity scales as $\mathcal{O}(N_T L)$. It can also be further accelerated using Fast Fourier Transform when N_T is large. Another benefit of the discretization is that for kernels whose values are costly to compute, at most L values need to be calculated. This can have a strong computational impact when $N_T \gg L$ as all values can be precomputed and stored.

While discretization improves the computational efficiency, it also introduces a bias in the computation of the intensity function and, thus potentially, in estimating the kernel parameters. The impact of the discretization on the estimation is considered in section 5.1.2 and section 5.2.1. Note that this bias is similar to the one incurred by quantizing the kernel as histograms for non-parametric estimators.

Loss and precomputations

FaDIn aims at minimizing the discretized ℓ_2 loss, which approximates the integral on the left part of (3.4.2) by a sum on the grid $\mathcal{G}$ after projecting timestamps of

[1] In practice, we would take $\Delta < 1$.

$\mathscr{F}_T$ on it. It boils down to optimizing the following loss $\mathcal{L}_\mathcal{G}$ defined as:

$$\mathcal{L}_\mathcal{G}\left(\theta, \widetilde{\mathscr{F}_T}\right) \coloneqq \frac{1}{N_T}\sum_{i=1}^{p}\left(\Delta\sum_{s=0}^{G}\left(\tilde{\lambda}_i[s]\right)^2 - 2\sum_{\tilde{t}_n^i \in \widetilde{\mathscr{F}}_T^i}\tilde{\lambda}_i\left[\frac{\tilde{t}_n^i}{\Delta}\right]\right) \quad , \tag{5.1.4}$$

where the Δ factor comes from estimating the integral in eq. (3.4.2) by the method of triangles. To find the parameters of the intensity function θ, FaDIn minimizes $\mathcal{L}_\mathcal{G}$ using a first-order gradient-based algorithm. The computational bottleneck of the proposed algorithm is thus the computation of the gradient $\nabla\mathcal{L}_\mathcal{G}$ regarding parameters θ. Using the discretized finite-support kernel, this gradient can be computed using convolution, giving the same computational complexity as the computation of the intensity function $\mathcal{O}(N_T L)$. However, gradient computation can still be too expensive for long processes with many events to get reasonable inference times.

Using the least squares error of the process expressed in eq. (5.1.4), one can further reduce the complexity of computing the gradient by precomputing some constants $\Phi_j(\tau; G)$, $\Psi_{j,k}(\tau, \tau'; G)$ and $\Phi_j\left(\tau; \widetilde{\mathscr{F}_T^i}\right)$ that do not depend on the parameter θ. Indeed, by developing and rearranging the terms in eq. (5.1.4), one obtains:

$$\begin{aligned} N_T\ \mathcal{L}_\mathcal{G}\left(\theta, \widetilde{\mathscr{F}_T}\right) = (T+\Delta)\sum_{i=1}^{p}\mu_i^2 + 2\Delta\sum_{i=1}^{p}\mu_i\sum_{j=1}^{p}\sum_{\tau=0}^{L}\phi_{ij}^\Delta[\tau]\underbrace{\left(\sum_{s=0}^{G} z_j[s-\tau]\right)}_{=:\Phi_j(\tau;G)} \\ + \Delta\sum_{i,j,k}\sum_{\tau=0}^{L}\sum_{\tau'=0}^{L}\phi_{ij}^\Delta[\tau]\phi_{ik}^\Delta[\tau']\underbrace{\left(\sum_{s=0}^{G} z_j[s-\tau]\ z_k\left[s-\tau'\right]\right)}_{=:\Psi_{j,k}(\tau,\tau';G)} \\ - 2\left(\sum_{i=1}^{p} N_T^i\mu_i + \sum_{i,j}\sum_{\tau=0}^{L}\phi_{ij}^\Delta[\tau]\underbrace{\left(\sum_{\tilde{t}_n^i\in\widetilde{\mathscr{F}}_T^i} z_j\left[\frac{\tilde{t}_n^i}{\Delta} - \tau\right]\right)}_{=:\Phi_j\left(\tau;\widetilde{\mathscr{F}}_T^i\right)}\right) \quad , \end{aligned}$$

where $(T+\Delta)$ comes from the fact that $(G+1)\Delta = T+\Delta$.

The term $\Psi_{j,k}(\tau, \tau'; G)$ dominates the computational cost of our precomputations. It requires $\mathcal{O}(G)$ operations for each tuples (τ, τ') and (j, k). Thus, it has a total complexity of $\mathcal{O}(p^2L^2G)$ and is the bottleneck of the precomputation phase. For any $m \in \{1, \dots, p\}$, the gradient of the loss w.r.t. the baseline parameter is

given by:

$$N_T \frac{\partial \mathcal{L}_G}{\partial \mu_m} = 2\left(T+\Delta\right)\mu_m + 2\Delta \sum_{j=1}^{p}\sum_{\tau=1}^{L} \phi_{mj}^{\Delta}[\tau]\Phi_j\left(\tau; G\right) - 2N_T^m \quad .$$

For any tuple $(m,l) \in \{1,\dots,p\}^2$, the gradient w.r.t. η_{ml} that parametrizes by ϕ_{ml} is:

$$\begin{aligned} N_T \frac{\partial \mathcal{L}_G}{\partial \eta_{ml}} = {} & 2\Delta\mu_m \sum_{\tau=0}^{L} \frac{\partial \phi_{ml}^{\Delta}[\tau]}{\partial \eta_{ml}}\ \Phi_l\left(\tau; G\right) \\ & + 2\Delta \sum_{k=1}^{p}\sum_{\tau=0}^{L}\sum_{\tau'=0}^{L} \frac{\partial \phi_{ml}^{\Delta}[\tau]}{\partial \eta_{ml}} \phi_{mk}^{\Delta}[\tau']\Psi_{l,k}\left(\tau,\tau'; G\right) \\ & - 2\sum_{\tau=0}^{L} \frac{\partial \phi_{ml}^{\Delta}[\tau]}{\partial \eta_{m,l}}\ \Phi_l\left(\tau; \widetilde{\mathscr{F}}_T^m\right). \end{aligned}$$

Gradients of kernel parameters dominate the computational cost of gradients. The complexity is of $\mathcal{O}\left(pL^2\right)$ for each kernel parameter, leading to a total complexity of $\mathcal{O}\left(p^3L^2\right)$ and is independent of the number of events N_T. Thus, a trade-off can be made between the precision of the method and its computational efficiency when varying the size of the kernel's support or the discretization.

Remark The primary motivation for the ℓ_2 loss is the presence of terms that can be precomputed in contrast to the log-likelihood [Reynaud-Bouret and Rivoirard, 2010, Reynaud-Bouret et al., 2014, Bacry et al., 2020]. A comparison is performed in section C.1.

Optimization The inference is then conducted using gradient descent for the ℓ_2 loss $\mathcal{L}_{\mathcal{G}}$. FaDIn thus allows for very general parametric kernels, as exact gradients for each parameter involved in the kernels can be derived efficiently as long as the kernel is differentiable and has finite support. Gradient-based optimization algorithms can, therefore, be used without limitation, in contrast with the EM algorithm, which requires a close-form solution to zero the gradient, which is difficult for many kernels. A critical remark is that the problem is generally non-convex and may converge to a local minimum.

5.1.2 Impact of the discretization

While discretization allows for efficient computations, it also introduces a perturbation in the loss value. In this section, we quantify the impact of this perturbation

on the parameter estimation when Δ goes to 0. Through this section, we observe a process $\mathscr{F}_T$ whose intensity function is given by the parametric form $\lambda(\cdot;\theta^*)$. Note that if the process $\mathscr{F}_T$'s intensity is not in the parametric family $\lambda(\cdot;\theta)$, θ^* is defined as the best approximation of its intensity function in the ℓ_2 sense. The goal of the inference process is thus to recover the parameters θ^*.

When working with the discrete process $\widetilde{\mathscr{F}_T}$, the events t_n^i of the original process are replaced with a projection on a grid $\tilde{t}_n^i := t_n^i + \delta_n^i$. Here, δ_n^i is uniformly distributed on $[-\Delta/2, \Delta/2]$. We consider the discrete FaDIn estimator $\widehat{\theta}_\Delta$ defined as $\widehat{\theta}_\Delta := \arg\min_\theta \mathcal{L}_\mathcal{G}(\theta)$. We can upper-bound the error incurred by $\widehat{\theta}_\Delta$ by the decomposition:

$$\left\|\widehat{\theta}_\Delta - \theta^*\right\|_2 \leq \underbrace{\left\|\widehat{\theta}_c - \theta^*\right\|_2}_{(*)} + \underbrace{\left\|\widehat{\theta}_\Delta - \widehat{\theta}_c\right\|_2}_{(**)} , \tag{5.1.5}$$

where $\widehat{\theta}_c := \arg\min_\theta \mathcal{L}(\theta)$ is the reference estimator for θ^* based on the standard ℓ_2 estimator for *continuous* point processes. This decomposition involves the statistical error $(*)$ and the bias error $(**)$ induced by the discretization. The statistical term measures how far the parameters obtained by minimizing the ℓ_2 continuous loss having access to a finite amount of data are from the true ones. In contrast, the term $(**)$ represents the discretization bias induced by minimizing the discrete loss (eq. (5.1.4)) instead of the continuous one (eq. (3.4.2)). In the following proposition, we focus on the discretization error $(**)$, which is related to the computational trade-off offered by our method and not on the statistical error of the continuous ℓ_2 estimator $(*)$. Our work showcases that this disregarded estimator can be efficiently computed, and we hope it will promote research to describe its asymptotic behavior. We now study the perturbation of the loss due to discretization.

PROPOSITION 5.1.1. *Let $\mathscr{F}_T$ and $\widetilde{\mathscr{F}_T}$ be respectively a MHP process and its discretized version on a grid $\mathcal{G}$ with stepsize Δ. Assume that the intensity function of $\mathscr{F}_T$ possesses continuously differentiable finite support kernels on $[0, W]$. Thus, assuming $\Delta < \min_{t_n^i, t_m^j \in \mathscr{F}_T} |t_n^i - t_m^j|$, for any $i, j \in [\![1, p]\!]$, it holds:*

$$\widetilde{\lambda}_i[s] = \lambda_i(s\Delta) - \sum_{j=1}^{p} \sum_{t_m^j \in \mathscr{F}_{s\Delta}^j} \delta_m^j \frac{\partial \phi_{ij}}{\partial t}\left(s\Delta - t_m^j; \theta\right) + \mathcal{O}(\Delta^2),$$

and

$$\mathcal{L}_{\mathcal{G}}(\theta, \widetilde{\mathscr{F}_T}) = \frac{2}{N_T} \sum_{i,j=1}^{p} \sum_{t_n^i \in \mathscr{F}_T^i} \sum_{t_m^j \in \mathscr{F}_T^j} \left(\delta_m^j - \delta_n^i\right) \frac{\partial \phi_{ij}}{\partial t}\left(t_n^i - t_m^j; \theta\right) + \mathcal{L}(\theta) + \Delta h(\theta) + \mathcal{O}\left(\Delta^2\right) \quad .$$

Proof 5.1.1

Recall that by definition,

$$\lambda_i(s\Delta) = \mu_i + \sum_{j=1}^{p} \sum_{t_m^j \in \mathscr{F}_{s\Delta}^j} \phi_{ij}\left(s\Delta - t_m^j\right) \quad ,$$

and

$$\begin{aligned} \tilde{\lambda}_i[s] &= \mu_i + \sum_{j=1}^{p} \sum_{\tilde{t}_m^j \in \widetilde{\mathscr{F}}_{s\Delta}^j} \phi_{ij}\left(s\Delta - \tilde{t}_m^j\right) \\ &= \mu_i + \sum_{j=1}^{p} \sum_{t_m^j \in \mathscr{F}_{s\Delta}^j} \phi_{ij}\left(s\Delta - t_m^j - \delta_m^j\right) \quad , \end{aligned} \tag{5.1.6}$$

where eq. (5.1.6) is a consequence of hypothesis $\Delta < \min_{t_n^i, t_m^j \in \mathscr{F}_T} |t_n^i - t_m^j|$ which ensures that no event collapses on the same bin of the grid, i.e., *each $\tilde{t}_m^j$ corresponds to a unique t_m^j, and that $\#\widetilde{\mathscr{F}}_{s\Delta}^j = \#\mathscr{F}_{s\Delta}^j$.*

Note that this hypothesis also implies that the intensity function is smooth for all points on the grid $\mathcal{G}$. Applying the first-order Taylor expansion to the kernels ϕ_{ij} in $s\Delta - t_m^j$ and bounding the perturbation δ_n^i by Δ yields the first result of the proposition.

For the perturbation of the loss $\mathcal{L}_\mathcal{G}$, we have:

$$\begin{aligned}\mathcal{L}_\mathcal{G}\left(\theta, \widetilde{\mathscr{F}_T}\right) &= \frac{1}{N_T}\sum_{i=1}^{p}\left(\Delta\sum_{s=0}^{G}\left(\tilde{\lambda}_i[s]\right)^2 - 2\sum_{\tilde{t}_n^i\in\widetilde{\mathscr{F}}_T^i}\tilde{\lambda}_i\left[\frac{\tilde{t}_n^i}{\Delta}\right]\right) + \mathcal{L}(\theta,\mathscr{F}_T) - \mathcal{L}(\theta,\mathscr{F}_T)\\ &= \mathcal{L}(\theta,\mathscr{F}_T)\\ &\quad + \frac{1}{N_T}\sum_{i=1}^{p}\left(\underbrace{\Delta\sum_{s=0}^{G}\tilde{\lambda}_i[s]^2 - \int_0^T\lambda_i(t)^2\mathrm{d}t}_{(*)} \underbrace{-2\sum_{t_n^i\in\mathscr{F}_T^i}\left(\tilde{\lambda}_i\left[\frac{\tilde{t}_n^i}{\Delta}\right] - \lambda_i\left(t_n^i\right)\right)}_{(**)}\right).\end{aligned}$$

The first term $(*)$ is the error of a Riemann approximation of the integral. Theorem 1.2 in Tasaki [2009] shows that asymptotically with $\Delta\to 0$,

$$\Delta\sum_{s=0}^{G}\tilde{\lambda}_i[s]^2 - \int_0^T\lambda_i(t)^2\,\mathrm{d}t = \Delta.h_i\left(\theta\right) + \mathcal{O}\left(\Delta^2\right)\quad, \tag{5.1.7}$$

where $h_i(\theta) \coloneqq \frac{1}{2}\left(\int_0^T\left|\lambda_i\left(t;\theta\right)\frac{\partial\lambda_i}{\partial t}\left(t;\theta\right)\right|^{1/2}\mathrm{d}t\right)^2$ and we denote $h(\theta) = \frac{1}{N_T}\sum_{i=1}^{p}h_i(\theta)$.

For the second term $(**)$, we re-use the expression from eq. (5.1.6) but use a Taylor expansion in $t_n^i - t_m^j$. The perturbation becomes $\delta_m^j - \delta_n^i$,

$$\sum_{t_n^i\in\mathscr{F}_T^i}\left(\tilde{\lambda}_i\left[\frac{\tilde{t}_n^i}{\Delta}\right] - \lambda_i\left(t_n^i\right)\right) = \sum_{t_n^i\in\mathscr{F}_T^i}\left(\delta_n^i - \delta_m^j\right)\frac{\partial\phi_{ij}}{\partial t}\left(t_n^i - t_m^j;\theta\right) + \mathcal{O}\left(\Delta^2\right). \tag{5.1.8}$$

Summing eq. (5.1.7) and eq. (5.1.8) concludes the proof.

The first result is a direct application of the Taylor expansion of the intensity for the kernels. For the loss, the first perturbation term $\Delta.h(\theta)$ comes from approximating the integral with a finite Euler sum [Tasaki, 2009] while the second one derives from the perturbation of the intensity. This proposition shows that, as the discretization step Δ goes to 0, the perturbed intensity and ℓ_2 loss are good estimates of their continuous counterpart. We now quantify the discretization error $(**)$ as Δ goes to 0.

PROPOSITION 5.1.2. *We consider the same assumption as in proposition 5.1.1. Then, if the estimators* $\widehat{\theta}_c = \arg\min_\theta \mathcal{L}(\theta)$ *and* $\widehat{\theta}_\Delta = \arg\min_\theta \mathcal{L}_\mathcal{G}(\theta)$ *are uniquely defined,* $\widehat{\theta}_\Delta$ *converges to* $\widehat{\theta}_c$ *as* $\Delta \to 0$*. Moreover, if* $\mathcal{L}$ *is* C^2 *and its hessian* $\nabla^2\mathcal{L}\left(\widehat{\theta}_c\right)$ *is positive definite with* $\varepsilon^2 > 0$ *its smallest eigenvalue, then:*

$$\left\|\widehat{\theta}_\Delta - \widehat{\theta}_c\right\|_2 \leq \frac{\Delta}{\varepsilon} g\left(\widehat{\theta}_\Delta\right) \quad ,$$

with $g\left(\widehat{\theta}_\Delta\right) = \mathcal{O}(1)$.

Proof 5.1.2

We consider the two estimators $\widehat{\theta}_\Delta = \arg\min_\theta \mathcal{L}_\mathcal{G}(\theta)$ *and* $\widehat{\theta}_c = \arg\min_\theta \mathcal{L}(\theta)$. *With the loss approximation from proposition 5.1.1, we have a pointwise convergence of* $\mathcal{L}_\mathcal{G}(\theta)$ *towards* $\mathcal{L}(\theta)$ *for all* $\theta \in \Theta$ *as* Δ *goes to 0. By continuity of* $\mathcal{L}_\mathcal{G}$*, we have that the limit of* $\widehat{\theta}_\Delta$ *when* Δ *goes to 0 exists and is equal to* $\widehat{\theta}_c$*. This proves that the discretized estimator converges to the continuous one as* Δ *decreases.*

We now characterize its asymptotic speed of convergence. The KKT conditions impose that:

$$\nabla\mathcal{L}_\mathcal{G}\left(\widehat{\theta}_\Delta\right) = 0 \qquad \text{and} \qquad \nabla\mathcal{L}\left(\widehat{\theta}_c\right) = 0. \tag{5.1.9}$$

Using the approximation from proposition 5.1.1, one gets in the limit of small Δ*:*

$$\begin{aligned}\nabla\mathcal{L}_\mathcal{G}\left(\widehat{\theta}_\Delta\right) = \nabla\mathcal{L}\left(\widehat{\theta}_\Delta\right) + \Delta.\frac{\partial h}{\partial\theta}\left(\widehat{\theta}_\Delta\right) + \mathcal{O}\left(\Delta^2\right) \\ + \frac{2}{N_T}\sum_{i,j=1}^{p}\sum_{t_n^i\in\mathscr{T}_T^i}\sum_{t_m^j\in\mathscr{T}_T^j}\left(\delta_m^j - \delta_n^i\right)\frac{\partial^2\phi_{ij}}{\partial t\partial\theta}\left(t_n^i - t_m^j;\widehat{\theta}_\Delta\right).\end{aligned}$$

Combining this with eq. (5.1.9), we get:

$$\begin{aligned}\nabla\mathcal{L}\left(\widehat{\theta}_\Delta\right) = -\Delta.\frac{\partial h}{\partial\theta}\left(\widehat{\theta}_\Delta\right) \\ + \frac{2}{N_T}\sum_{i,j=1}^{p}\sum_{t_n^i\in\mathscr{T}_T^i}\sum_{t_m^j\in\mathscr{T}_T^j}\left(\delta_n^i - \delta_m^j\right)\frac{\partial^2\phi_{ij}}{\partial t\partial\theta}\left(t_n^i - t_m^j;\widehat{\theta}_\Delta\right) + \mathcal{O}\left(\Delta^2\right),\end{aligned}$$

and as $\nabla\mathcal{L}\left(\widehat{\theta}_c\right) = 0$ *and* $\delta \in \left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right]$,

$$
\begin{aligned}
\left\|\nabla\mathcal{L}\left(\widehat{\theta}_\Delta\right) - \nabla\mathcal{L}\left(\widehat{\theta}_c\right)\right\|_2 = \Bigg\| & -\Delta.\frac{\partial h}{\partial\theta}\left(\widehat{\theta}_\Delta\right) \\
& + \frac{2}{N_T}\sum_{i,j=1}^{p}\sum_{t_n^i\in\mathscr{F}_T^i}\sum_{t_m^j\in\mathscr{F}_T^j}\left(\delta_n^i-\delta_m^j\right)\frac{\partial^2\phi_{ij}}{\partial t\partial\theta}\left(t_n^i-t_m^j;\widehat{\theta}_\Delta\right)\Bigg\|_2 \\
& + \mathcal{O}\left(\Delta^2\right) \\
\leq \Delta & \left\|\frac{\partial h}{\partial\theta}\left(\widehat{\theta}_\Delta\right) + \frac{2}{N_T}\sum_{i,j=1}^{p}\sum_{t_n^i\in\mathscr{F}_T^i}\sum_{t_m^j\in\mathscr{F}_T^j}\frac{\partial^2\phi_{ij}}{\partial t\partial\theta}\left(t_n^i-t_m^j;\widehat{\theta}_\Delta\right)\right\|_2 \\
& + \mathcal{O}\left(\Delta^2\right) \\
\leq \Delta & g\left(\widehat{\theta}_\Delta\right) \quad ,
\end{aligned}
$$

where

$$
g(\theta) \coloneqq \left\|\frac{\partial h}{\partial\theta}\left(\widehat{\theta}_\Delta\right) + \frac{2}{N_T}\sum_{i,j=1}^{p}\sum_{t_n^i\in\mathscr{F}_T^i}\sum_{t_m^j\in\mathscr{F}_T^j}\frac{\partial^2\phi_{ij}}{\partial t\partial\theta}\left(t_n^i-t_m^j;\widehat{\theta}_\Delta\right)\right\|_2 + \mathcal{O}\left(\Delta\right) \quad .
$$

This function is a $\mathcal{O}(1)$*. Using the hypothesis that the hessian* $\nabla^2\mathcal{L}\left(\widehat{\theta}_c\right)$ *exists and is positive definite with smallest eigenvalue* $\varepsilon^2 > 0$ *(*i.e., $\mathcal{L}$ *is locally strongly convex at point* $\widehat{\theta}_c$*), we have:*

$$
\begin{aligned}
\varepsilon^2\left\|\widehat{\theta}_\Delta - \widehat{\theta}_c\right\|_2^2 &\leq \left\|\nabla\mathcal{L}\left(\widehat{\theta}_\Delta\right) - \nabla\mathcal{L}\left(\widehat{\theta}_c\right)\right\|_2^2 \\
\text{i.e.,} \qquad \left\|\widehat{\theta}_\Delta - \widehat{\theta}_c\right\|_2 &\leq \frac{\Delta}{\varepsilon}g\left(\widehat{\theta}_\Delta\right).
\end{aligned}
$$

This concludes the proof.

This proposition shows that asymptotically on Δ, the estimator $\widehat{\theta}_\Delta$ is equivalent to $\widehat{\theta}_c$. It also shows that the discrete estimator converges to the continuous one at the same speed as Δ decreases. This is confirmed experimentally by results shown in fig. 5.1.1. Thus, one would need to select Δ so that the discretization error is small compared to the statistical one. Notice that assumptions from proposition 5.1.2 are not too restrictive. Indeed, they require the existence of a unique minimizer of $\mathcal{L}$, $\mathcal{L}_\mathcal{G}$ and $\mathcal{L}$. Moreover, if $\mathcal{L}$ is $\mathcal{C}^2$ in $\widehat{\theta}_c$, the previous hypothesis also implies the strong local convexity at this point.

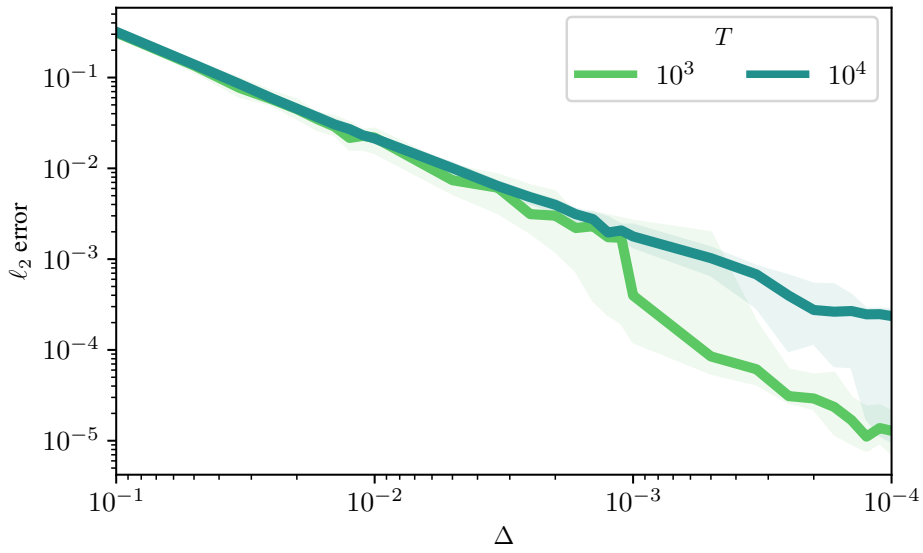


Figure 5.1.1: Median and interquartile error bar of the ℓ_2 norm between the parameters estimated computed with EM algorithm, continuously and discretely, w.r.t. the stepsize Δ. This figure confirms the results from proposition 5.1.1; that is, that the convergence of $\widehat{\theta}_\Delta$ towards $\widehat{\theta}_c$ is linear with respect to Δ.

5.2 Numerical experiments

We present various synthetic data experiments to support the advantages of the proposed approach. To begin, we investigate the bias induced by the discretization in section 5.2.1. Afterwards, the statistical and computational efficiency of FaDIn is highlighted through a benchmark with popular non-parametric approaches section 5.2.2. Sensitivity analysis regarding the parameter W and additional non-parametric comparisons are also provided, respectively in section 5.2.3 and section C.

5.2.1 Consistency of Discretization

In order to study the estimation bias due to discretization, we run two experiments and report the results in fig. 5.2.1. The general paradigm is a one-dimensional TPP with intensity parametrized as in eq. (5.1.1) with a Truncated Gaussian kernel of mean $m \in \mathbb{R}$ and standard deviation $\sigma > 0$, with fixed support $[0\,, W] \subset \mathbb{R}^+$, $W > 0$. It corresponds to $\phi(x) = \alpha\kappa(x), \alpha \geq 0$ with

$$\kappa(x) \coloneqq \kappa\left(x; m, \sigma, W\right) = \frac{1}{\sigma} \frac{f\left(\frac{x-m}{\sigma}\right)}{F\left(\frac{W-m}{\sigma}\right) - F\left(\frac{-m}{\sigma}\right)} \mathbb{1}_{\{0 \leq x \leq W\}},$$

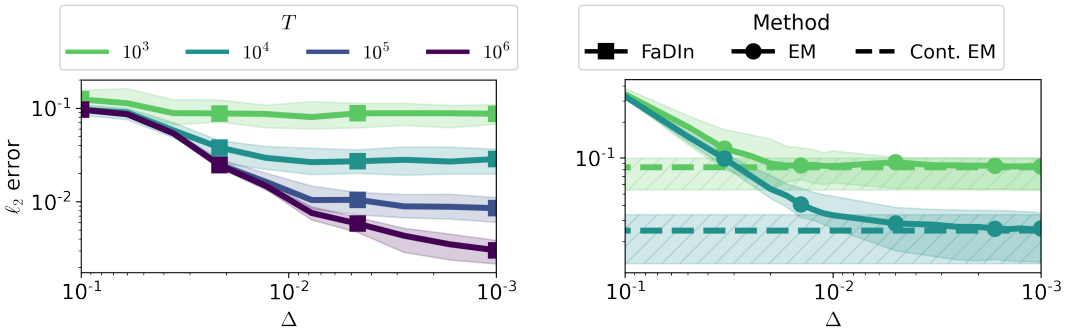


Figure 5.2.1: Median and interquartile error bar of the ℓ_2 norm between true parameters and parameter estimates computed with FaDIn (left) and with EM algorithm (right), continuously and discretely, w.r.t. the stepsize of the grid Δ.

where f (resp. F) is the probability density function (resp. cumulative distribution function) of the standard normal distribution. Hence, the parameters to estimate are μ and $\eta = (\alpha, m, \sigma)$.

In both experiments, for multiple process length T, the discrete estimates are computed for varying grid stepsize Δ, from 10^{-1} to 10^{-3}. The parameter W is set to 1. The ℓ_2 norm of the difference between estimates and the true parameter values – the ones used for data simulation – is computed and reported. We first computed the parameter estimates with our FaDIn method for $T \in \{10^3, 10^4, 10^5, 10^6\}$, for 100 simulations each time. Second, since we wish to separate discretization bias from statistical bias, we compute the estimates with an EM algorithm, both continuously and discretely, and that for 50 random data simulations.

One can observe that the ℓ_2 errors between discrete estimates and true parameters tend towards zero as T increases. For T fixed, one can see plateaus starting for stepsize values that are not particularly small, indicating that the discretization bias is limited. The second experiment with the EM algorithm shows that when plateau is reached, it corresponds to some statistical error. In other words, even for a reasonably coarse stepsize, the bias induced by the discretization is slight compared to the statistical error.

Discretization on EM estimates (DriPP)

Figure 5.2.2 presents the detailed results – *i.e.,* parameter-wise – of the experiment shown in Figure 5.2.1 (right). In this experiment, we are interested in the context of Driven Point Process (DriPP; chapter 4) with an exogenous homogeneous PP. The simulation parameter of the latter is set to 0.5, meaning that on average, 1 event occurs every 2 seconds on the driver.

Figure 5.2.3 presents the results of the same experiment with Poisson parameter set to 0.1 which represents roughly five times less events.

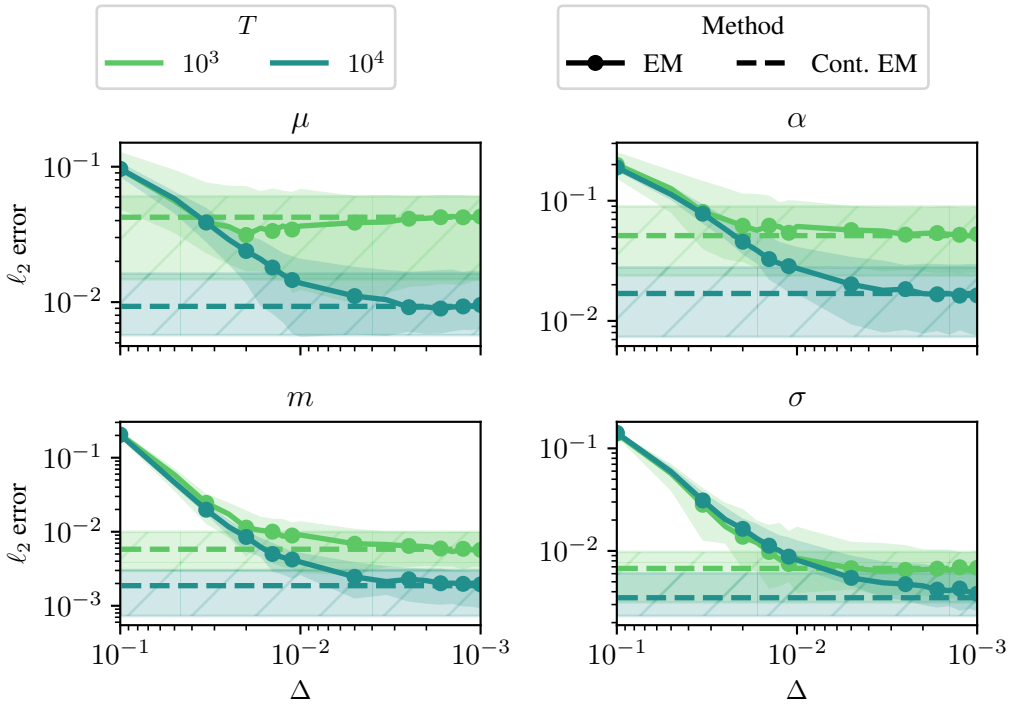


Figure 5.2.2: Median and interquartile error bar of the ℓ_2 norm between true parameters and parameter estimates computed with EM algorithm, continuously and discretely, w.r.t. the stepsize Δ.

Discretization effect on FaDIn estimates

This section presents additional results.We reproduce the experiments as above with FaDIn and two other kernels: Raised Cosine and Truncated Exponential. The Raised Cosine kernel is defined by:

$$\phi(x) = \alpha \left[1 + \cos\left(\frac{x-u}{\sigma}\pi - \pi\right)\right] \mathbb{1}_{\{x \in [u\,, u+2\sigma]\}} \ . \tag{5.2.1}$$

The parameters to estimate are μ, α, u and σ. The Truncated Exponential kernel of decay parameter $\gamma \in \mathbb{R}_+$, with fixed support $[a\,, b] \subset \mathbb{R}^+$, $b > a$ is defined as $\phi(x) = \alpha\kappa(x), \alpha \geq 0$ with

$$\kappa(x) \coloneqq \kappa\left(x; \gamma, a, b\right) = \frac{h\left(x\right)}{H\left(b\right) - H\left(a\right)} \mathbb{1}_{\{a \leq x \leq b\}}, \tag{5.2.2}$$

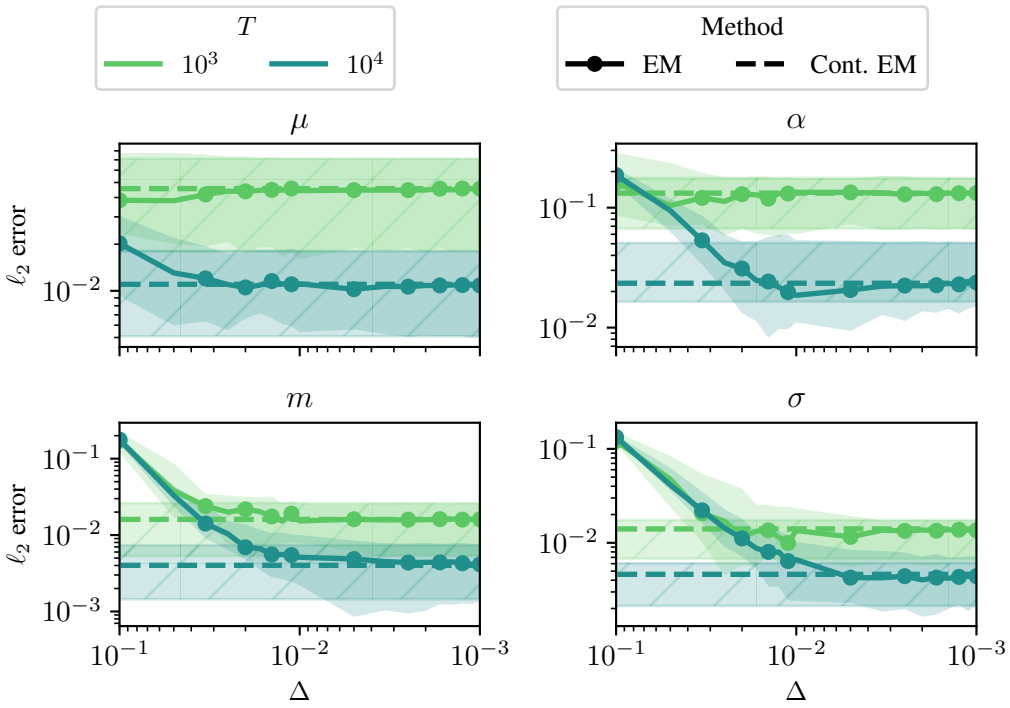


Figure 5.2.3: Median and interquartile error bar of the ℓ_2 norm between true parameters and parameter estimates computed with EM algorithm, continuously and discretely, w.r.t. the stepsize Δ.

where here h (resp. H) is the probability density function of parameter γ (resp. cumulative distribution function) of the exponential distribution. The parameters to estimate are μ, α and γ.

Estimation results (median and 20-80% quantiles) are displayed in Figure 5.2.4 and confirm the conclusion presented above about the consistency of the discretization for FaDIn. In addition, we display the quadratic error for each parameter separately in Figure 5.2.5 for the Truncated Gaussian, Figure 5.2.6 for the Raised Cosine and Figure 5.2.7 for the Truncated Exponential kernels.

5.2.2 Statistical and computational efficiency of FaDIn

We compare FaDIn with non-parametric and parametric methods by assessing approaches' statistical and computational efficiency. To learn the non-parametric kernel, we select various existing methods. The first benchmarked method uses histogram kernels and relies on the EM algorithm, provided in Zhou et al. [2013a] and implemented in the `tick` library [Bacry et al., 2017a]. The kernel is set with one basis function. The three other approaches involve a linear combination of

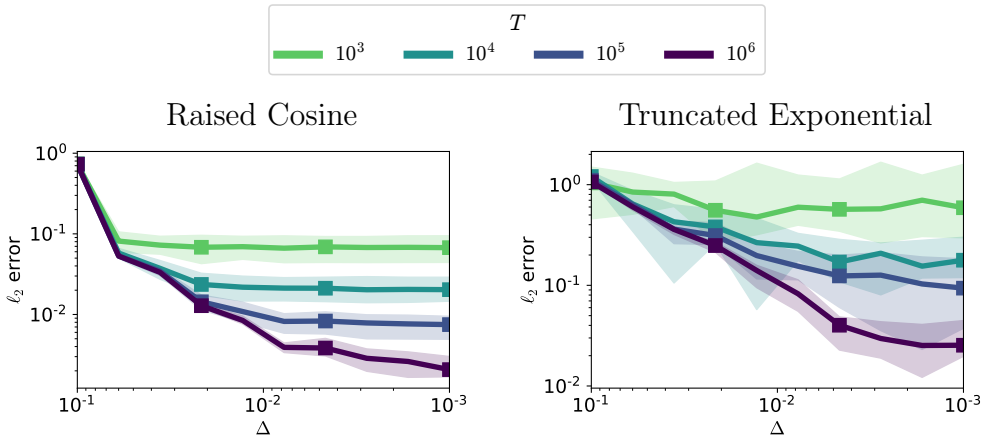


Figure 5.2.4: Comparison of the influence of the discretization on the parameter estimation of FaDIn for a Raised Cosine kernel (left) and an Exponential kernel (right) w.r.t. the stepsize of the grid Δ.

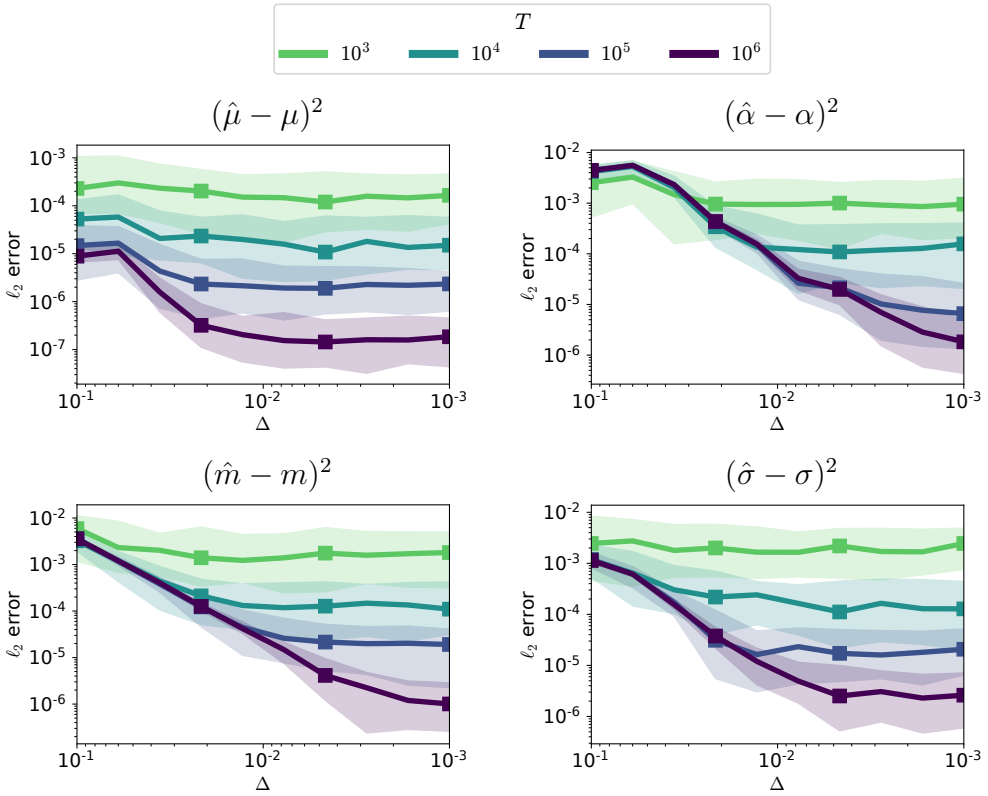


Figure 5.2.5: Error on parameters for the Truncated Gaussian kernel as a function of T and Δ.

pre-defined raised cosine functions as non-parametric kernels. The inference is

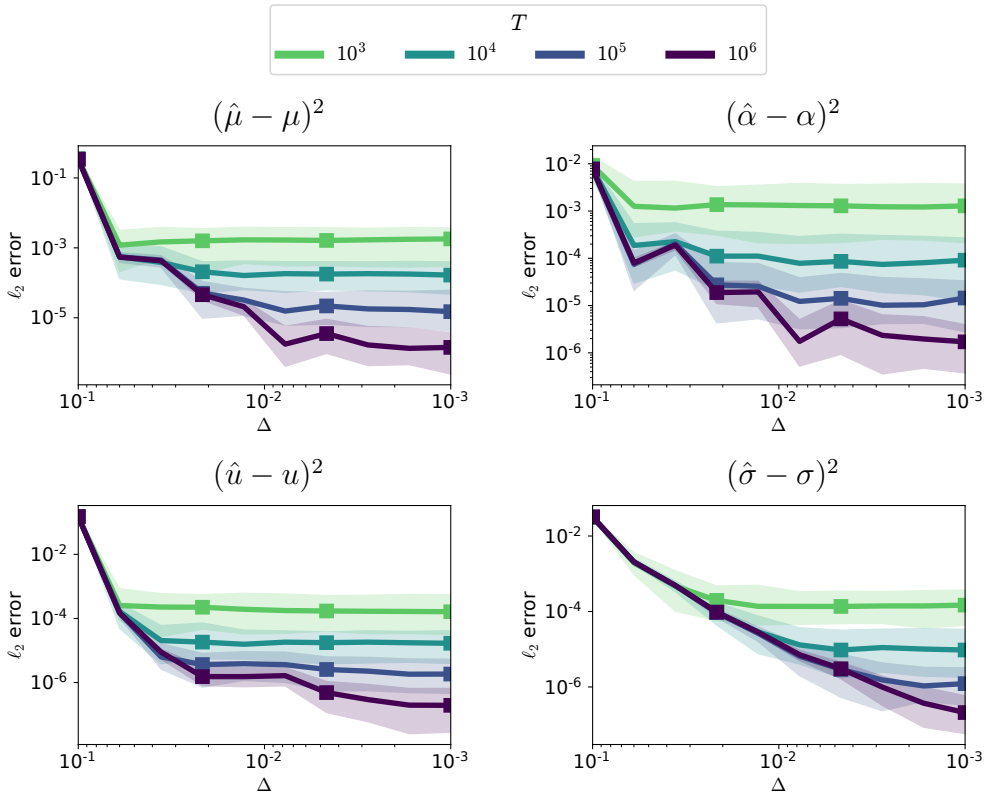


Figure 5.2.6: Error on parameters for the Raised Cosine kernel as a function of T and Δ.

made either by stochastic gradient descent algorithm (Non-param SGD; Linderman and Adams, 2014) or by Bayesian approaches such as Gibbs sampling (Gibbs) or Variational Inference (VB) from Linderman and Adams [2015]. These algorithms are implemented in the `pyhawkes` library[2]. In the following experiments, we set the number of basis to five for each method. The parametric approach we compare with is the Neural Hawkes Process (NeuralHawkes; Mei and Eisner, 2017) where authors represents the intensity function by a LSTM module. The latter is calculated on a GPU. The experiment is conducted as follows. We simulate a two-dimensional Hawkes process (repeated ten times) using the `tick` library with baseline $\boldsymbol{\mu} = [0.1, 0.2]$ and Raised Cosine kernels:

$$\phi_{i,j}(x) = \alpha_{i,j}\left[1 + \cos\left(\frac{x - u_{i,j}}{\sigma_{i,j}}\pi - \pi\right)\right], (i,j) \in \{1,2\}^2$$

on the support $[u_{i,j}, u_{i,j} + 2\sigma_{i,j}]$ and zero outside with parameters $\boldsymbol{\alpha} = \begin{bmatrix} 1.5 & 0.1 \\ 0.1 & 1.5 \end{bmatrix}$, $\mathbf{u} = \begin{bmatrix} 0.1 & 0.3 \\ 0.3 & 0.3 \end{bmatrix}$ and $\boldsymbol{\sigma} = \begin{bmatrix} 0.3 & 0.25 \\ 0.3 & 0.3 \end{bmatrix}$. Further, we infer the intensity function

[2] https://github.com/slinderman/pyhawkes

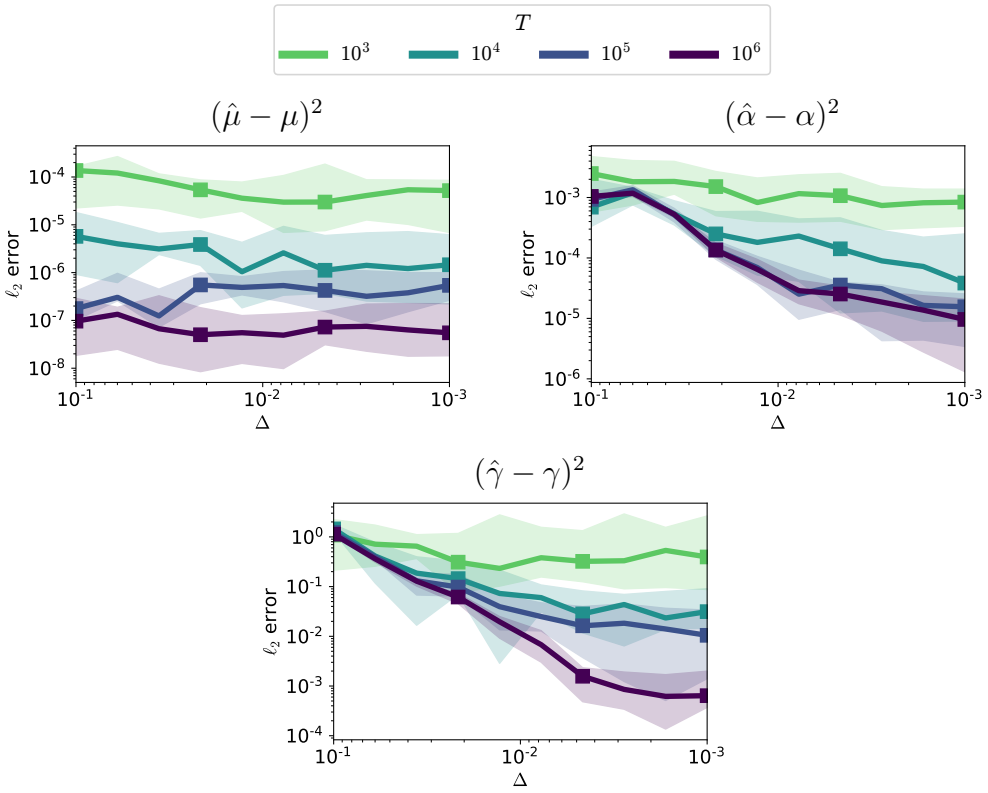


Figure 5.2.7: Error on parameters for the Truncated Exponential kernel as a function of T and Δ.

of these underlying Hawkes processes using FaDIn and the four previously mentioned methods setting $\Delta = 0.01$ for all these discrete approaches. The parameter W of FaDIn is set to 1. This experiment is repeated for varying values of $T \in \{10^3, 10^4, 10^5\}$. The averaged (over the ten runs) normalized ℓ_1 error on the intensity (evaluated on the same discrete grid), as well as the associated computation time, are reported in fig. 5.2.8. Due to the high computational times of NeuralHawkes, this approach is performed once and is not applied for $T = 10^5$.

From a statistical perspective, we can observe the advantages of FaDIn inference for varying T over the benchmarked methods. It is worth noting that this result is expected by a parametric approach when the used kernel belongs to the same family as the one with which events have been simulated. Also, only one (long) sequence of data has been used, explaining the poor statistical results of the Neural Hawkes, which is efficient on many repetitions of short sequences due to the massive amount of parameters to infer. From a computational perspective, FaDIn is very efficient compared to benchmarked approaches. Indeed, it scales very well with an increasing time T and then with a growing number of events. In

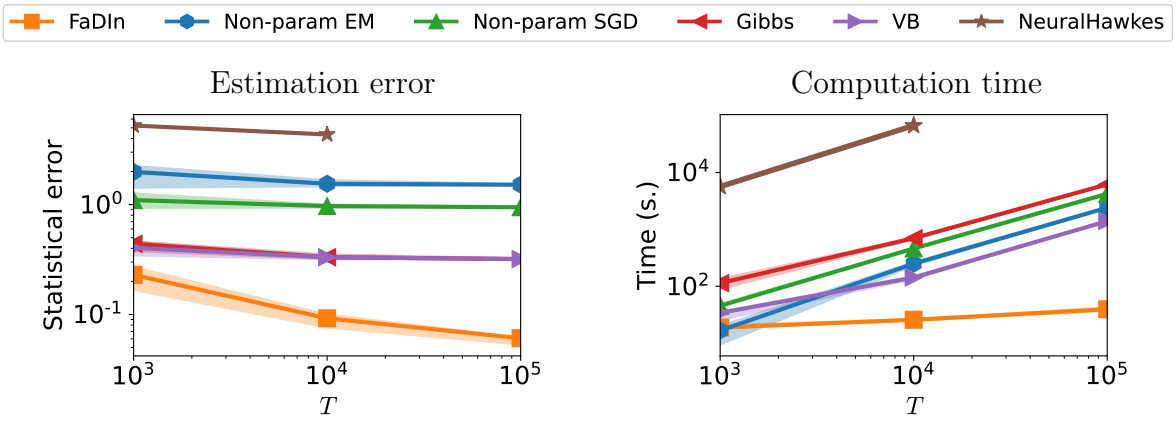


Figure 5.2.8: Comparison of the statistical and computational efficiency of FaDIn with five benchmarked methods. The averaged (over ten runs) statistical error on the intensity function (left) and the computational time (right) are computed regarding the time T (and thus the number of events).

contrast, other methods depend on the number of events and scale linearly with the time T.

5.2.3 Sensitivity analysis regarding the parameter W

To study the estimation bias induced by the finite support kernels, we conduct an experiment using FaDIn with a (Truncated) Exponential kernel. The general framework is a one-dimensional TPP with intensity parametrized as in eq. (5.1.1) with a Truncated Exponential kernel having a decay parameter γ, with fixed support $[0, W] \subset \mathbb{R}^+$, $W > 0$. It corresponds to $\phi(x) = \alpha\kappa(x), \alpha \geq 0$ with $\kappa(x)$ defined in eq. (5.2.2). In our present case, we fix $a = 0$ and $b = W$. Therefore, when $W \to \infty$, this Truncated Exponential kernel converges to the standard exponential kernel, *i.e.*, $t \mapsto \alpha\gamma \exp(-\gamma t)$. The parameters to estimate are μ and $\eta = (\alpha, \gamma)$.

The experiment is conducted as follows. We simulate events (10 repetitions) from a Hawkes process with baseline $\mu = 1.1$ and a standard Exponential kernel (non-truncated) with $\alpha = 0.8$, $\gamma = 0.5$ for varying $T \in \{10^3, 10^4, 10^5, 10^6\}$ using the `tick` Python library. FaDIn is then computed on each of these sets of events using a Truncated Exponential kernel of length $W \in [1, 100]$ and a stepsize $\Delta = 0.01$. The averaged (over ten runs) and the 25 % and 75 % quantiles statistical ℓ_2-error of parameters (left) and computational time (right) are displayed w.r.t. the support length W in fig. 5.2.9. On the one hand, one can observe that the ℓ_2-error converges to a plateau once $W > 10$, *i.e.*, the bias induced by the finite

support length is reduced. On the other hand, the computational time increase when W increases. Interestingly, for each T, the computational time is close when W is high enough (close to 100). Indeed, optimizing the loss becomes the bottleneck of FaDIn since the grid size $G = TL+1$ only intervenes in the precomputation part.

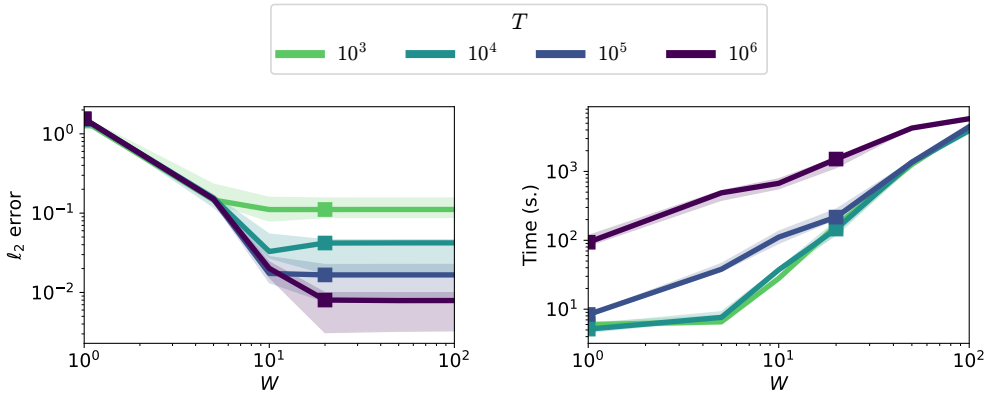


Figure 5.2.9: Comparison of the influence of the kernel support size W on the parameter estimation of FaDIn for a Truncated Exponential kernel. The averaged (over 10 runs) statistical ℓ_2-error (left) and computational time (right) are displayed w.r.t. the support length W.

5.3 Application to MEG data

Recall that the response's latency related to a stimulus has been identified as a Biomarker of ageing [Price et al., 2017] and many diseases such as epilepsy [Kannathal et al., 2005], Alzheimer's [Dauwels et al., 2010], Parkinson's [Tanaka et al., 2000] or multiple sclerosis [Gil et al., 1993]. Therefore, obtaining information on such a feature after auditory or visual stimuli is critical to characterize and eventually detect the presence of a specific disease for a given subject. FaDIn allows fitting a statistical model on this latency by inferring a model on the latency of these responses through Hawkes processes kernels. This approach characterizes the delays' distribution more finely compared to the latency estimates.

As in chapter 4, we use Convolutional Dictionary Learning (CDL; Jas et al. 2017) with rank-1 constraint [Dupré la Tour et al., 2018] to decompose raw signal into a set of spatio-temporal patterns, called *atoms*, with their respective onsets, called *activations* (*cf.* Figure 4.1.2). Experiments on MEG data were run on the same two datasets from the `MNE` Python package [Gramfort et al., 2013, 2014]: the *sample* dataset and the somatosensory (*somato*) dataset[3] (*cf.* section 4.3.2 for more

[3]Both available at https://mne.tools/stable/overview/datasets_index.html

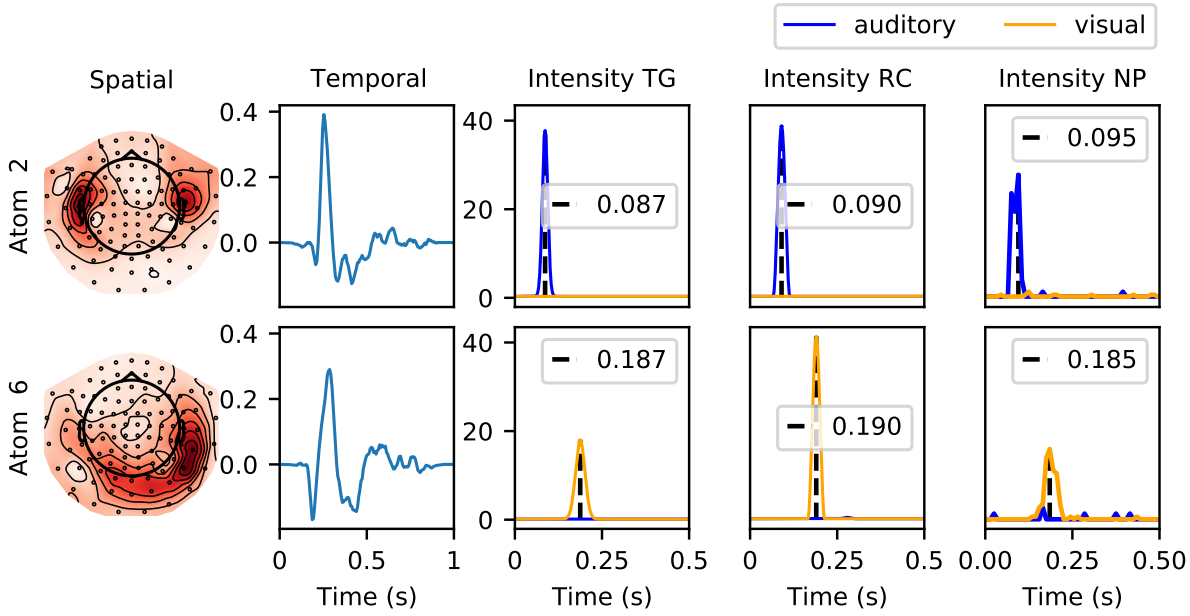


Figure 5.3.1: Spatial and temporal patterns of 2 atoms from MNE *sample* dataset, and their respective estimated intensity functions after a stimulus (cue at time = 0 s), for auditory and visual stimuli with non-parametric (NP), Truncated Gaussian (TG) and Raised Cosine (RC) kernels.

details regarding these datasets). Recall that the *sample* dataset contains M/EEG recordings of a human subject presented with audio and visual stimuli, while for the *somato* dataset, a human subject is scanned with MEG during 15 min, while 111 stimulations of his left median nerve were made.

We are interested in the paradigm of Driven Point Process (DriPP; chapter 4) and for every extracted atom, its intensity function related to the corresponding stimuli is estimated using a non-parametric kernel (NP) and two kernel parametrizations: Truncated Gaussian (TG) and Raised Cosine (RC). Results on the *sample* (resp. *somato*) dataset are presented in Figure 5.3.1 (resp. Figure 5.3.2), where only the kernel related to each type of stimulus is plotted, for the sake of clarity.

Results show that all three kernels agree on a peak latency around 90 ms for the auditory condition and 190 ms for the visual condition. Due to the limited number of events, one can observe that the non-parametric kernel estimated is less smooth, with spurious peaks later in the interval. Overall, these results on real MEG data demonstrate that our approach with a RC kernel parametrization can recover correct latency estimates even with the discretization of stepsize 0.02. Furthermore, the usage of RC allows us to have sharper peaks in intensity compared to TG, enforcing the link between the external stimulus and the atom's

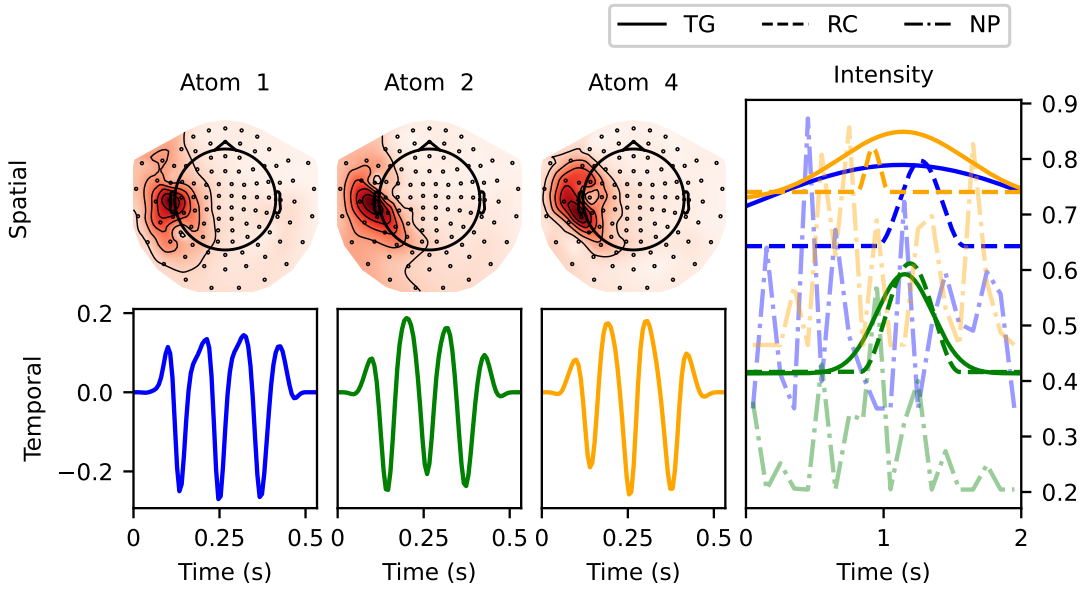


Figure 5.3.2: Spatial and temporal patterns of three μ-wave atoms from MNE *somato* dataset, and their respective estimated intensity functions following a stimulus (cue at time = 0 s), for somatosensory stimuli with non-parametric kernel (NP) and two parametrized kernels: Truncated Gaussian (TG) and Raised Cosine (RC).

activation. This difference mainly comes from the fact that RC does not need pre-determined support. This advantage is even more pronounced in the case of induced responses, such as in the *somato* dataset (see Figure 5.3.2), where the range of possible latency values is more difficult to determine beforehand.

5.4 Discussion

This work proposed an efficient approach to infer general parametric kernels for Multivariate Hawkes processes. Our method makes the use of parametric kernels computationally tractable, beyond exponential kernels. The development of FaDIn is based on three key features: (*i*) finite-support kernels, (*ii*) timeline discretization, and (*iii*) precomputations reducing the computational cost of the gradients. These allow for a computationally efficient gradient-based approach, improving state-of-the-art methods while providing flexible use of kernels well-fitted to the considered applications. Moreover, this work shows that the bias induced by the discretization is negligible, both theoretically and numerically. By allowing the use of a general parametric kernel in Hawkes processes, this contribution opens new possibilities for many applications. This is the case with M/EEG data, where estimating information about the rate and latency of occurrences of brain sig-

nal patterns is at the core of neuroscience questions. Therefore, FaDIn makes it possible to use a Raised Cosine kernel, allowing for efficient retrieval of these parameters.

Part III

Advancements in Convolutional Dictionary Learning for Large-Scale M/EEG Data Analysis: Stochastic Approaches and Population Studies

The analysis of magneto- and electroencephalography (M/EEG) data has been at the forefront of neuroscience research, offering a high-resolution temporal window into the intricate activities of the human brain. M/EEG data present specific challenges and opportunities: the data are high-dimensional, inherently noisy, and exhibit complex spatiotemporal patterns that can be both task-related and age-dependent. Convolutional Dictionary Learning (CDL) has emerged as a powerful tool to distill meaningful features from these intricate data sets. However, the application of CDL to M/EEG data is fraught with computational difficulties and has yet to be fully exploited for large-scale, population-level studies. This part tries to unify two critical advancements in this field.

The first contribution, *Stochastic Windowing and Robust Convolutional Dictionary Learning*, addresses the computational bottlenecks in applying CDL to large time-series data. By introducing a stochastic windowing technique and leveraging the computational capabilities of GPU, this work offers a scalable solution to CDL's high computational demand. The proposed method has been rigorously benchmarked against existing libraries and algorithms, showcasing its efficacy in handling large-scale problems, including M/EEG multivariate time series.

The second work, *Using Population CDL to Detect Task-Related Neuromagnetic Transients and Ageing Trends*, takes CDL to a population level. By applying a data-driven CDL approach to a large open-access dataset (Cam-CAN), this research reveals the complex interplay between task performance and ageing in the spatiotemporal characteristics of neuromagnetic transient bursts. The study shows age-related trends in activation levels of specific burst types for the first time, providing valuable insights into how human brain activity evolves across the lifespan.

Together, these contributions present a holistic view of the advancements in CDL for M/EEG data analysis. They not only solve the computational challenges but also extend the applicability of CDL to large-scale, population-based studies. This part aims to provide a comprehensive understanding of these advancements, offering methodologies that are both computationally efficient and capable of revealing complex neural dynamics at both individual and population scales.

Chapter 6

Stochastic Windowing and Robust Convolutional Dictionary Learning for M/EEG Data

Contents

6.1 Introduction **126**
6.2 Contextualizing the current work **128**
6.3 Inline outlier detection **129**
6.4 Stochastic windowing CDL **132**
6.4.1 Approximate sparse coding 133
6.4.2 Stochastic sub-windowing 134
6.4.3 Stochastic line search 137
6.5 Experiments **139**
6.5.1 Data simulation 139
6.5.2 Dictionary evaluation 142
6.5.3 Experimental paradigm 144
6.5.4 Results 145
6.6 Conclusion **156**

The content of this chapter was carried out in collaboration with Benoît Malezieux and was submitted to:

Cédric Allain, Benoît Malezieux, and Thomas Moreau. Fast and robust convolutional dictionary learning for large m/eeg data. *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. Under review

THIS chapter aims to delve into the advancements in the field of convolutional dictionary learning (CDL), with a particular focus on M/EEG data. An extension and enhancement of the `AlphaCSC` framework is presented.

Convolutional Dictionary Learning (CDL) consists of finding a sparse representation from noisy data and is a common way to encode data-driven knowledge on signals. Yet, it is computationally expensive on data-sets with large time series or images, and the ones that are encountered in neuroimaging applications often suffer from anomalies that make it hard to learn relevant patterns, in particular when dealing with M/EEG data.. To overcome the computational burden and the varying quality of measurements – real world data often suffer from outliers that make the training process less efficient –, we introduce a GPU-based implementation relying on stochastic windowing – with `PyTorch`'s automatic differentiation – and anomaly detection, and propose an analysis of its behavior on both simulated and real data.

6.1 Introduction

With cheap wearable devices, physiological signals are now routine to monitor patient conditions during hospital stays or everyday life. Monitoring produces a huge amount of data which needs automated processing to extract valuable insights. A common task to process these signals is to extract recurring physiological events such as QRS complex – *a.k.a.* heartbeats – in ECG [Luz et al., 2016], neural oscillations in the local field potential [Cole and Voytek, 2017], or brain responses in Magneto and Electroencephalography (M/EEG) datasets [Hämäläinen et al., 1993, Dupré la Tour et al., 2018]. Once extracted, the frequency, variability, and waveforms of these events can give important information about the processes that produced the signals.

Many different event detection algorithms for physiological signals have been developed in the literature. Traditional techniques rely on signal processing tools to detect prototypical features of the events, such as peak-detection [Pan and Tompkins, 1985] or wavelet-based approaches Martinez et al. [2004]. These methods require a lot of domain expertise and can be hard to fine-tune or adapt when working with new signals or events. More recently, deep learning-based approaches have been proposed to deal with detection tasks [Xiang et al., 2018, Craik et al., 2019], framing the problem as a classification of sub-windows based on annotated signals. While these approaches show very good performances on specific tasks, they need a large amount of labeled data for each specific signal or event and require heavy post-processing from individual sub-window prediction to time-localized prediction.

Another route to detect events is to consider unsupervised methods, where the events are identified through their repeating prototypical waveforms. This problem is typically seen as a factorization problem where the signal is assumed to be the product of a redundant basis of patterns or atoms – the dictionary – and of a sparse representation vector – the sparse codes. In particular, Convolutional Dictionary Learning (CDL; Grosse et al. 2007) enables the decomposition of signals into sparse combinations of localized spatio-temporal atoms, allowing the discovery of patterns associated with physiological activities. A benefit of this approach is that the algorithm learns the prototypical patterns associated with events directly from the data, removing the need for supervision. A typical example is given by the work of Dupré la Tour et al. [2018] who have successfully leveraged the physical properties of the signals to learn meaningful prototypical patterns through rank-1 constrained CDL.

However, despite the appeal of such a holistic method to detect events, its usage is still limited to very few studies. One core limitation is that for large signals, learning the waveforms is computationally expensive. Fast numerical solvers for Dictionary Learning on large data have been successful on a wide range of tasks [Mairal et al., 2009, Mensch et al., 2016]. While some efforts have also been made to develop fast CDL solvers [Wohlberg, 2016b, Dupré la Tour et al., 2018], these methods are still limited to small-scale studies as they are too computationally demanding for huge datasets. Using distributed optimization, Moreau and Gramfort [2020] proposed a method that scales to large datasets, however not reducing the computational burden. Moreover, CDL is also not robust to anomalies and artifacts often found in this kind of data due to potential sensor instabilities or external disturbances. Indeed, large artifacts tend to set astray the learning algorithm, resulting in senseless dictionaries that are not usable for physiological signal analysis. This lack of scalability and robustness limits the usability of CDL-based analysis for population studies.

Contributions. To unlock CDL for unsupervised event detection on population-level physiological signals, we propose to tackle these two challenges at once. First, we extend the formulation of CDL to the setting of robust regression, which allows to learn a model that is robust to outliers in the training distribution. Using the ability of the model to properly encode the signal as a proxy to detect outliers, we introduce a novel and robust CDL formulation that integrates the principles of the Least Trimmed Squares (LTS) method [Rousseeuw, 1984]. Next, we derive an efficient procedure based on approximate sparse coding and sub-windowing. By performing sparse coding on signal frames, we compute cheap stochastic gradient approximations that allow us to make fast progress on learning the dictionary. The implementation synergistically combines the principles of stochastic windowing with the computational prowess of `PyTorch` [Paszke et al., 2019] This approach also

couples very well with the robust regression scheme, allowing the effective removal of outliers during training. We demonstrate the efficiency of our procedure for learning convolutional dictionaries on large ECG and M/EEG datasets, in order to pave the way for applications of CDL to population-level datasets.

6.2 Contextualizing the current work

From chapter 2, recall that Convolutional Dictionary Learning can be written as a bi-level optimization problem to minimize the cost function with respect to the dictionary only, as mentioned by Mairal et al. [2009], by solving

$$\min_{\mathbf{D}\in\mathcal{C}} G(\mathbf{D};\mathbf{X}) = F\left(\mathbf{D}, \mathbf{Z}^*(\mathbf{D}); \mathbf{X}\right) \tag{6.2.1}$$

$$\text{with } \mathbf{Z}^*(\mathbf{D}) = \arg\min_{\mathbf{Z}} F(\mathbf{D}, \mathbf{Z}; \mathbf{X}) \ , \tag{6.2.2}$$

with $\mathcal{C} \coloneqq \left\{\mathbf{D} \in \mathbb{R}^{K\times P\times L},\ \|D_k\|_F^2 \leq 1, \forall k = 1, \dots, K\right\}$. Computing the sparse codes $\mathbf{Z}^*(\mathbf{D})$ is often referred to as the inner problem, while the global minimization is the outer problem.

In the following, we will focus on gradient descent on $\mathbf{D}$. Once $\mathbf{Z}^*(\mathbf{D})$ is known, Danskin [1967, Theorem 1] states that the gradient $\nabla G(\mathbf{D};\mathbf{X})$ is equal to $\nabla_{\mathbf{D}} F(\mathbf{D}, \mathbf{Z}^*(\mathbf{D}); \mathbf{X})$. Even though the inner problem is non-smooth, this result holds as long as the solution $\mathbf{Z}^*(\mathbf{D})$ is unique. Denoting by $\mathbf{D}^\top$ the adjoint operator of $\mathbf{D}$, we will assume that $\mathbf{D}^\top\mathbf{D}$ is invertible on the support of $\mathbf{Z}^*(\mathbf{D})$ in the following. This implies the uniqueness of $\mathbf{Z}^*(\mathbf{D})$.

While CDL users generally want to get both $\mathbf{D}$ and $\mathbf{Z}$, it is of interest to efficiently optimize eq. (2.3.3) in order to get a solution $\mathbf{D}^*$ as fast as possible, and then compute an optimal $\mathbf{Z}^*(\mathbf{D}^*)$. Using this strategy, one relies on cheap approximations of the sparse code $\mathbf{Z}^N\left(\mathbf{D}^*;\mathbf{X}\right) \approx \mathbf{Z}^*\left(\mathbf{D}^*;\mathbf{X}\right)$ to update the value of $\mathbf{D}$. Indeed, the tedious part of the calculation is generally the sparse coding step, especially for large signals for which convolutions are expensive. To do that, existing methods rely on approximate gradients, an approach popularized by the usage of unrolled algorithms and automatic differentiation [Mairal et al., 2010, Scetbon et al., 2021, Tolooshams and Ba, 2021, Malézieux et al., 2022] in the context of Dictionary Learning.

However, when dealing with physiological signals, the gradient estimation is often made harder by the presence of anomalies in the data. As a matter of fact, occasional sensor malfunctions and external disturbances can corrupt the measurements and impair the ability of traditional algorithms to estimate a good

descent direction. This results in poor-quality dictionaries and makes CDL hard to use on real-world neuroimaging data-sets.

In what follows, we derive a novel CDL formulation that is robust to the presence of anomalies in the data and propose cheap gradient approximation to efficiently learn robust dictionaries on large ECG and M/EEG signals.

6.3 Inline outlier detection

When the recording contains segments corrupted by large amplitude artifacts, the standard formulation of CDL usually encodes corrupted segments but fails to learn relevant atoms that compose the majority of the signal. Indeed, the ℓ_2 norm gives strong weight to parts of the signal with large variance, thus making classical algorithms converge to a dictionary which only succeeds in encoding artifacts. To cope with this issue, we propose to leverage the robust regression framework to design a robust formulation of CDL.

We consider a setting where a few segments in the signal are corrupted by large artifacts. The core idea behind our work is that there is no need for the dictionary to properly encode the corrupted segments. Therefore, the reconstruction errors for corrupted segments can be considered outliers for the distribution of reconstruction errors on all segments in the signal. Instead of minimizing the mean reconstruction error for each segment, we propose to minimize a robust estimate of the mean, less sensitive to outliers.

To develop our framework, let us first rewrite the original loss F for CDL as a minimization of the mean reconstruction error on segments, *i.e.,*

$$\begin{aligned} F(D, Z; X) &= \sum_{t=-L}^{T} \frac{1}{L} \boldsymbol{\varepsilon}_t \ , \\ \text{with } \boldsymbol{\varepsilon}_t &= \frac{1}{2} \left\| R\left[t : t+L\right] \right\|_2^2 + \lambda \sum_{k=1}^{K} \left\| z_k\left[t : t+L\right] \right\|_1 \end{aligned} \tag{6.3.1}$$

where $R[t] = (X - \sum_{k=1}^{K} Z(X; D) * \mathbf{d}_k)[t]$ for $t \in [1, T]$ and $R[t] = 0$ when $t \notin [1, T]$. Note that $X\left[a : b\right]$ denotes the restriction of X to the interval $[a, b]$. To get a robust estimate of the mean, we resort to the trimmed mean, which writes

$$\min_{D,Z} \widetilde{F}_\rho(D, Z; X) = \sum_{t=-L}^{T} \mathbb{1}_{\{\boldsymbol{\varepsilon}_t < \rho\}} \frac{1}{L} \boldsymbol{\varepsilon}_t \tag{6.3.2}$$

where ρ is a threshold that depends on the distribution of the errors $\{\boldsymbol{\varepsilon}_t\}_{t=-L}^{T}$, and which allows keeping only a clean fraction of the signal segments. This formulation

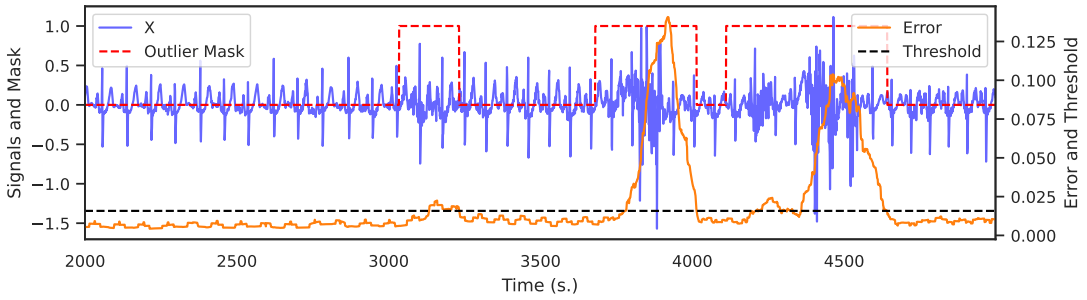


Figure 6.3.1: Raw signal X, reconstruction error, threshold and learned outlier mask on subject a02 (minute 56) of Physionet Apnea-ECG data set. Detection method is based on modified z-score (MAD), with $\alpha = 3.5$. The method correctly identifies outliers blocks.

corresponds to a Least Trimmed Square for the regression – a robust method to fit model parameters to data containing outliers. A critical design choice is the selection of the statistic ρ used as the threshold.

Outliers statistics and thresholds. From the empirical distribution of patch reconstruction errors $\{\boldsymbol{\varepsilon}_t\}_{t=-L}^{T}$, one needs to compute a threshold ρ which separates the outlier segments in X from the normal segments. The goal is to detect extreme points in the segments' error distribution, which are too large compared to the population of segment errors. Assuming a contamination level α, we propose several statistics from the outlier detection literature.

- **Method of quantiles**

$$\rho = Q_{\boldsymbol{\varepsilon},(1-\alpha)} \tag{6.3.3}$$

  where $Q_{\boldsymbol{\varepsilon},q}$ denotes the quantile of order q of the set $\boldsymbol{\varepsilon}$.

- **Method of interquartile range (IQR)**

$$\rho = Q_{\boldsymbol{\varepsilon},0.75} + \alpha \mathrm{IQR}_{\boldsymbol{\varepsilon}} \tag{6.3.4}$$

  where $\mathrm{IQR}_{\boldsymbol{\varepsilon}} \coloneqq Q_{\boldsymbol{\varepsilon},0.75} - Q_{\boldsymbol{\varepsilon},0.25}$ denotes the interquartile range of the set $\boldsymbol{\varepsilon}$.

- **Method of z-score** [Iglewicz and Hoaglin, 1993] Let us define the mean and standard deviation of the reconstruction errors, *i.e.*, $\mu_{\boldsymbol{\varepsilon}} = \frac{1}{NPW}\sum_{n,p,w=1}^{N,P,W} \boldsymbol{\varepsilon}_{n,p,w}$ and $\sigma_{\boldsymbol{\varepsilon}} = \sqrt{\frac{1}{NPW}\sum_{n,p,w=1}^{N,P,W} (\boldsymbol{\varepsilon}_{n,p,w} - \mu_{\boldsymbol{\varepsilon}})^2}$. The z-score is computed as follows:

$$\mathbf{Z}_{n,p,w} \coloneqq \frac{\boldsymbol{\varepsilon}_{n,p,w} - \mu_{\boldsymbol{\varepsilon}}}{\sigma_{\boldsymbol{\varepsilon}}} \ . \tag{6.3.5}$$

  Outliers are defined as observations such that $|\mathbf{Z}_{n,p,w}| > \alpha$, generally $\alpha = 2$ or 3. Hence,

$$\rho = \mu_{\boldsymbol{\varepsilon}} + \alpha\sigma_{\boldsymbol{\varepsilon}} \tag{6.3.6}$$

- **Method of modified z-score** [Iglewicz and Hoaglin, 1993] Let us define the mean absolute deviation (MAD) as follows:

$$\text{MAD} = \text{median}_{n,p,w}\left(|\boldsymbol{\varepsilon}_{n,p,w} - \text{Med}_{\boldsymbol{\varepsilon}}|\right) \quad , \tag{6.3.7}$$

where $\text{Med}_{\boldsymbol{\varepsilon}}$ denotes the median of the set $\boldsymbol{\varepsilon}$. The modified z-score is thus defined as follows:

$$\mathbf{M}_{n,p,w} = c\frac{\boldsymbol{\varepsilon}_{n,p,w} - \text{Med}_{\boldsymbol{\varepsilon}}}{\text{MAD}} \quad , \tag{6.3.8}$$

where $c = 0.6745$ is a constant needed as $\mathbb{E}\left[\text{MAD}\right] = 0.6745\sigma$ for large N [Iglewicz and Hoaglin, 1993].

Outliers are observations for which $|\mathbf{M}_{n,p,w}| > \alpha$, generally $\alpha = 3.5$. Hence,

$$\rho = \text{Med}_{\boldsymbol{\varepsilon}} + \alpha\frac{\text{MAD}}{c} \tag{6.3.9}$$

Note that these methods are initially bilateral, but only the upper bound is considered in this work, as the goal is to detect outliers with large reconstruction errors.

Optimization for the least trimmed square Due to the indicator in (6.3.2) whose threshold depends on the parameters Z and D, minimizing this loss directly would be computationally untractable and thus not adapted to large physiological signals. To decouple the estimation of $\boldsymbol{\varepsilon}_t$ and the computation of the threshold ρ, we consider the following bi-level formulation

$$\begin{aligned} &\min_{D} \widetilde{F}_{\rho(D)}(D, Z^*(D); X) \quad , \\ &s.t. \quad \rho(D) = \mathcal{S}(\boldsymbol{\varepsilon}(D, Z^*(D))) \quad , \\ &\text{and} \quad Z^*(D) = \arg\min_Z F(D, Z; X) \quad . \end{aligned} \tag{6.3.10}$$

where $\mathcal{S}$ is the statistic chosen from the previous paragraph and $\boldsymbol{\varepsilon}(D, Z)$ denotes the set of the segment reconstruction error for Z and D. Using this formulation, $Z^*(D; X)$ does not depend on the threshold ρ and can be computed directly using classical sparse coding algorithms. Then, using $Z^*(D; X)$, one can compute the observed $\boldsymbol{\varepsilon}_t$ and thus compute the statistics necessary to get ρ.

Computation of outliers and a robust gradient estimate. One issue with the proposed approach is that in CDL, if a sample $X[t]$ is corrupted, them all $Z[t-L:t]$ coefficient can be corrupted. Indeed, due to the convolutional nature of

the model, there is a delay effect between a non-zero coefficient in Z and its effect in X. To take this into account, we propose to modify slightly our formulation. In our original formulation, a mask such as $\mathbf{m} = \mathbb{1}_{\{\varepsilon > \rho\}} \in \{0, 1\}^{N \times T}$ is used to select part of the signal used to compute the loss. Here, each element indicates whether the corresponding timestamp in the signal is an outlier (1) or not (0). To accommodate for the delayed nature of the convolution, we modify this mask to consider as outliers samples before an outlier, that might be badly reconstruted to to their dependency on corrupted samples. For each identified outlier timestamp, we also mark each of the preceding $L-1$ timestamps in the same channel as outliers. This approach is inspired by the concept of the *opening* in mathematical morphology and image filtering [Bolon et al., 1995, p. 386]. It ensures that the temporal structure of the convolution process is taken into account when determining the spread of outlier effects across timestamps. Mathematically, it can be formulated as follows, by creating an "extended" mask[1] $\mathbf{m}'$ such that

$$\mathbf{m}'[n, i] = 1 \text{ if } \sum_{l=0}^{L-1} \mathbf{m}[n, i + l] \geq 1 \text{ else } 0 \ . \tag{6.3.11}$$

Finally, the "opposite" outlier mask is used as weights in the ℓ_2 loss which is used to learn $\mathbf{D}$ to select only valid samples. Figure 6.3.1 illustrates the outliers detection method on real ECG data, with the outliers mask computed with the reconstruction error and the threshold.

6.4 Stochastic windowing CDL

In the context of large population studies with electrophysiological data, it is not rare to deal with very large time series consisting of minutes-long recordings sampled at high frequency (from 100 Hz to 1000 Hz) from hundreds of subjects. Thus, we must adapt the learning process to make it usable for very long time series, *i.e.,* where $T \gg 1$, which is not accounted for in CDL literature. The complexity of the existing methods mostly comes from the need to compute full signal convolution, with a complexity of $\mathcal{O}(T \log T)$. To reduce the need for full signal convolution, we propose the Stochastic Windowing CDL algorithm (WinCDL). The core idea of WinCDL is to process randomly sampled sub-windows of the signal to get a stochastic estimate of the gradient. By processing subwindows of length $W \ll T$, we ensure that our algorithm is efficient as only small convolutions need to be performed. To make this algorithm even more efficient, we also leverage some empirical observations from unrolled DL's studies [Tolooshams and Ba, 2021, Malézieux et al., 2022], stating that only a small number of iterations

[1] Note that in practice, one can simply do a convolution between $\mathbf{m}$ and a unit uniform filter of size L with proper alignment.

is sufficient to get a good estimate of the gradient. Using a few sparse coding iterations on small windows allows us to reduce the computational costs associated with gradient estimation. Another issue that comes up is that the choice of step sizes is critical to the optimization process in dictionary learning, and SGD methods based on simple heuristics like rate decay are difficult to tune in this context. We stabilize our algorithm by using the stochastic line search procedure, proposed by Vaswani et al. [2019]. The use of a line search is possible due to the efficient computation of the loss with the approximate sparse codes $\mathbf{Z}^{(M)}(\mathbf{D})$ when M is taken sufficiently small. The algorithm is detailed in Algorithm 4, and we describe each step in detail below.

Algorithm 4: Pseudo-code for WinCDL

```
Set 𝒯 the number of iterations;
Set (U_{α,t})_{t∈ℕ} a sequence of maximal step sizes decreasing to 0;
Set 𝒞 a set of constraints for (u, v);
for t = 1, ..., 𝒯 do
    Sample index i and window X[i, i + W] in the dataset;
    Compute gradient ∇G_i(D^(t); X[i, i + W]) of D^(t) with approximate
     sparse coding and inline outlier detection;
    Compute best step size α_t with line search and starting point U_{α,t};
    D^(t+1) ← P_𝒞(D^(t) − α_t ∇G_i(D^(t); X[i, i + W]));
end
return D^(t+1)
```

6.4.1 Approximate sparse coding

As illustrated in Malézieux et al. [2022] and Tolooshams and Ba [2021], optimizing over $\mathbf{D}$ does not necessarily require precise sparse coding at each gradient descent step. Here, we build our algorithm from a fixed number of FISTA iterations in the same spirit as unrolling, but without leveraging back-propagation. Indeed, for a given dictionary $\mathbf{D}$ composed of K atoms D_k and a regularization parameter $\lambda > 0$, we would like to retrieve the NK activation signals $Z^n \in \mathbb{R}^{K \times \widetilde{T}}$ (with $\widetilde{T} := T - L + 1$) associated to the signals $X^n \in \mathbb{R}^{P \times T}$ by solving the following optimization problems: $\forall n = 1, \dots, N$,

$$(Z^n)^*(\mathbf{D}) := \underset{\mathbf{z}_k^n \in \mathbb{R}^{\widetilde{T}}}{\arg\min} \frac{1}{2} \left\| X^n - \sum_{k=1}^{K} \mathbf{z}_k^n * D_k \right\|_F^2 + \lambda \sum_{k=1}^{K} \|\mathbf{z}_k^n\|_1 \quad \text{s.t.} \quad \mathbf{z}_k^n \geq 0_{\mathbb{R}^{\widetilde{T}}} \ . \tag{6.4.1}$$

Instead of employing FISTA [Beck and Teboulle, 2009] to solve each of these convex problems in $\mathbf{z}_k^n$ to full convergence, we opt for a more computationally efficient approach by limiting the algorithm to a fixed number of iterations, M. This strategy provides a practical balance between computational efficiency and solution accuracy. However, it should be noted that this approach may not guarantee the complete resolution of the problem within these M iterations. Thus, $\mathbf{Z}^*(\mathbf{D})$ would be approximated by $\mathbf{Z}^{(M)}(\mathbf{D}) \coloneqq \left\{ (Z^n)^{(M)}(\mathbf{D}), n = 1, \dots, N \right\}$, the set of all solutions of eq. (6.4.1) after M iterations of FISTA. The adapted FISTA algorithm for CDL's inner problem is presented in algorithm 6, in section B.

6.4.2 Stochastic sub-windowing

To update the dictionary, the loss that we want to minimize is

$$G(\mathbf{D}; \mathbf{X}) = \sum_{t=-L}^{T} \mathbb{1}_{\{\boldsymbol{\varepsilon}_t < \rho\}} \left\| R\left[t : t + L\right] \right\|_2^2 \quad . \tag{6.4.2}$$

where R is computed as in (6.3.1). For datasets composed of large time series, such as in the case of M/EEG, the computational cost of FISTA increases dramatically with the length of the signal, as the convolutions become more expensive, and the sum over all windows becomes very expensive. Instead of processing all windows at each iteration, we propose to sample small chunks of data from the recordings when evaluating the loss. This stochastic approach allows for a dramatic reduction of the computational cost, by reducing the number of terms in the sum as well as reducing the cost of sparse coding. The procedure consists in choosing a random chunk of signal $\mathbf{X}\left[i - L : i + W\right]$ at each iteration, where W is a hyper-parameter corresponding to the length of the window and where i is an index in $[0, T - W]$. Then we compute the sparse code $Z_{n,k,i}$ for this chunk of data and get the gradient of

$$G_i(\mathbf{D}; \mathbf{X}) = \sum_{t=i}^{i+W} \mathbb{1}_{\{\boldsymbol{\varepsilon}_t < \rho\}} \left\| R\left[t : t + L\right] \right\|_2^2 \quad . \tag{6.4.3}$$

with respect to $\mathbf{D}$. The sparse codes $Z_{N,k,i}$ are further approximations of the original sparse code on the chunk due to border effects. To minimize their effect, we sample chunks of the signal, and not uniform segments from the loss. This extra approximation does not hinder the convergence of the algorithm in practice. As $\boldsymbol{\varepsilon}_t$ only depends on the reconstruction error on a local chunk, it can be computed directly from $Z_{N,k,i}$. This algorithm is amenable to a stochastic version of Alternating Minimization, with sub-windows of the full signal samples from the original signal X.

Difference between gradients

Now, we would like to demonstrate that the stochastic gradient of L_i is a good estimate of the true one of L. Let $\mathbf{x} \in \mathbb{R}^T$ a univariate signal. Let $\mathbf{d} \in \mathbb{R}^L$ a dictionary with a single atom. Let $L \leq W \leq T$ be the length of a window. Let $i \in [\![0\,, T-1]\!]$ be an index. We define,

$$S_{W,i} \coloneqq [i : i + W - 1] \tag{6.4.4}$$
$$S_{W,i,L} \coloneqq [i - L + 1 : i + W + L - 2] \tag{6.4.5}$$
$$\partial S_{W,i,L} \coloneqq [i - (L-1) : i] \cup [i + W - 1 : i + W - 1 + (L-1)] \tag{6.4.6}$$

Note that $S_{W,i} \cup \partial S_{W,i,L} = S_{W,i,L}$. We also define the restriction of Z onto $S_{W,i}$ (resp. $S_{W,i,L}$) by $Z_{S_{W,i}} \coloneqq Z\,[i : i + W - L] \in \mathbb{R}^{W-L+1}$ (resp. $Z_{S_{W,i,L}} \coloneqq Z\,[i - L + 1 : i + W - 1] \in \mathbb{R}^{W+L-1}$).

To avoid border effects in the gradient estimation, it is possible to compute the sparse codes over the extended window $S_{W,i,L}$ and to use only its restriction over $S_{W,i}$ where $L \in \mathbb{N}_*$ is the width of the buffer zone [Moreau and Gramfort, 2020]. In the following, we will make the assumption that the sparse code estimator obtained as explained above is equal to the original window $Z_{S_{W,i}}$.

Let us define the following loss function:

$$L_i(D) = \frac{1}{2} \left\| \mathbf{x}\,[i : i + W - 1] - \mathbf{z}_{S_{W,I,L}} * \mathbf{d} \right\|_2^2 \quad . \tag{6.4.7}$$

We define the estimated gradient as follows:

$$\hat{g}_{W,i} \coloneqq \nabla L_i(D) \quad , \tag{6.4.8}$$

and proposition 6.4.2 shows that this is an unbiased estimator of the true gradient g^*.

PROPOSITION 6.4.1. *For $\mathbf{x} \in \mathbb{R}^T$, $\mathbf{z} \in \mathbb{R}^{T-L+1}$ and $\mathbf{d} \in \mathbb{R}^L$, the gradient of the loss function $L(\mathbf{d}) = \frac{1}{2} \|\mathbf{x} - \mathbf{z} * \mathbf{d}\|_2^2$ is:*

$$\nabla L(\mathbf{d}) = -\mathbf{z}^{\curvearrowleft} * (\mathbf{x} - \mathbf{z} * \mathbf{d}) \in \mathbb{R}^L \quad , \tag{6.4.9}$$

where $\forall t \in \left[\!\!\left[0\,, \widetilde{T} - 1\right]\!\!\right], \mathbf{z}^{\curvearrowleft}[t] = \mathbf{z}\left[\widetilde{T} - 1 - t\right]$, with $\widetilde{T} \coloneqq T - L + 1$.

Proof 6.4.1

The loss function can be rewritten as:

$$L(\mathbf{d}) = \frac{1}{2}\sum_{t=0}^{T-1}\left(\mathbf{x}[t] - \sum_{\tau=0}^{L-1}\mathbf{z}[t-\tau]\mathbf{d}[\tau]\right)^2 \quad . \tag{6.4.10}$$

Taking the derivative with respect to $\mathbf{d}[l], \forall l \in [\![0\,, L-1]\!]$*, we have:*

$$\frac{\partial L(\mathbf{d})}{\partial \mathbf{d}[l]} = -\sum_{t=0}^{T-1}\left(\mathbf{x}[t] - \sum_{\tau=0}^{L-1}\mathbf{z}[t-\tau]\mathbf{d}[\tau]\right)\mathbf{z}[t-l] \tag{6.4.11}$$

$$= -\sum_{t=0}^{T-1}\left(\mathbf{x} - \mathbf{z} * \mathbf{d}\right)[t]\mathbf{z}[t-l] \tag{6.4.12}$$

$$= -\left(\mathbf{z}^{\curvearrowleft} * (\mathbf{x} - \mathbf{z} * \mathbf{d})\right)[l] \ , \tag{6.4.13}$$

hence the result.

PROPOSITION 6.4.2. *Under the assumption that for each window* i*, we have access to the correct value of* $Z_{S_{W,i,L}}$ *on the interval* $S_{W,i,L}$*, then*

$$\mathbb{E}_i\left[\hat{g}_{W,i}\right] = g^* \ .$$

Proof 6.4.2

The true gradient is

$$g^* = \mathbf{z}^{\curvearrowleft} * \mathbf{x} - \mathbf{z}^{\curvearrowleft} * \mathbf{z} * \mathbf{d} = \psi - \phi * \mathbf{d} \in \mathbb{R}^L \tag{6.4.14}$$

where $\forall s \in [1, L]$*,*

$$\psi[s] = \sum_{\tau=1}^{T-L+1} z[\tau]X[s+\tau-1] \tag{6.4.15}$$

$$\text{and } \phi[s] = \sum_{\tau=1}^{T-L+1} z[\tau]z[s+\tau-1] \tag{6.4.16}$$

The estimated gradient on window $S_{W,i} = [i : i+W-1]\,, i \in [\![1\,, T-W]\!]$ *is*

$$\hat{g}^{W,i} = Z^-_{S_{W,i}} * X_{S_{W,i}} - Z^-_{S_{W,i}} * Z_{S_{W,i}} * D \tag{6.4.17}$$

$$= \psi^{W,i} - \phi^{W,i} * D \in \mathbb{R}^L \tag{6.4.18}$$

where

$$\forall s \in [\![0\,, L-1]\!]\,, \psi^{W,i}[s] = \sum_{\tau=i}^{i+W} z[\tau+s]x[\tau] \tag{6.4.19}$$

$$\forall s \in [\![-L+1\,, L-1]\!]\,, \phi^{W,i}[s] = \sum_{\tau=i}^{i+W} z[\tau+s]z[\tau] \tag{6.4.20}$$

Here, we make the assumption that for each window i, we have access to the correct value of z on the interval $S_{W,i,L}$ (we add the border of size L $\partial S_{W,i,L}$).

For $s \in [0, L-1]$, we have that the value of $\psi[s]$ for the full signal is:

$$\psi[s] = \sum_{\tau=0}^{T-L+1} z[\tau+s]x[\tau] \tag{6.4.21}$$

$$= \sum_{i=-W+1}^{T-L+1} \frac{1}{W} \sum_{\tau=i}^{i+W} z[\tau+s]x[\tau] \tag{6.4.22}$$

$$= \frac{1}{W} \sum_{i=-W+1}^{T-L+1} \psi^{W,i}[s] \tag{6.4.23}$$

$$\tag{6.4.24}$$

where the second line derives from the fact that each coefficient is seen through W windows. Here, to avoid border effect, we consider we can take windows on the extended interval $[-W+1, T-L+1+W]$, for instance with zero padding.

Thus, we can see that

$$\mathbb{E}_i\left[\psi^{W,i}[s]\right] = \frac{1}{T-L+W} \sum_{i=-W+1}^{T-L+1} \psi^{W,i}[s] = \frac{W}{T-L+W}\psi[s] \tag{6.4.25}$$

Similarly, we can show that $\mathbb{E}_i\left[\phi^{W,i}[s]\right] = \frac{W}{T-L+W}\phi[s]$ for all $s \in [-L+1, L-1]$.

Thus, by linearity of the convolution and the expectation, we get:

$$\mathbb{E}_i\left[\hat{g}^{W,i}\right] = \mathbb{E}_i\left[\psi^{W,i}\right] - \mathbb{E}_i\left[\phi^{W,i}\right] * D = \psi - \phi * D = g^* \tag{6.4.26}$$

6.4.3 Stochastic line search

Standard stochastic gradient descent algorithms fail to produce satisfying results on our problem because of the difficulty to tune the step size. Thus, we leverage a method called *Stochastic line search*, introduced in Vaswani et al. [2019], that

extends the line search to the case where the gradient is stochastic. Indeed, line search algorithms are very helpful in situations where the step size of the gradient descent is hard to tune.

In usual gradient descent, the parameters $\theta^{(t)} = \left(D_1^{(t)}, D_K^{(t)}\right)$ are updated at each gradient step t with the help of $\nabla_\theta G\theta^{(t)})$ and a step size parameter $\alpha^{(t)}$, as follows

$$\theta^{(t+1)} = \theta^{(t)} - \alpha^{(t)} \nabla_\theta G\left(\theta^{(t)}\right) \quad . \tag{6.4.27}$$

The line search algorithm consists of finding the largest possible step size $\alpha^{(t)}$ starting from an upper bound $\bar{\alpha}$, by iterating over $n \in \mathbb{N}$ and evaluating the loss $G\left(\theta^{(t)} - \rho^n \bar{\alpha} \nabla_\theta G\left(\theta^{(t)}\right)\right)$ for $0 < \rho < 1$ until the loss satisfies a condition of the form

$$G\left(\theta^{(t)} - \rho^n \bar{\alpha} \nabla_\theta G\left(\theta^{(t)}\right)\right) < G\left(\theta^{(t)}\right) - c \ , \tag{6.4.28}$$

where c can either be constant or depend on the problem, or until $\rho^n < \epsilon$ where $\epsilon > 0$ is a stopping criterion. Then, set $\alpha^{(t)} \coloneqq \rho^{n_f} \bar{\alpha}$ where n_f is the final number of iterations.

The idea of the Stochastic line search is to extend this principle in a scenario where we only have access to an estimate of the true gradient. In this case, naively applying a line search with each sample of gradient would lead to a non converging sequence of $\left(\theta^{(t)}\right)_{t\in\mathbb{N}}$, because the algorithm restarts the process at the upper bound $\bar{\alpha}$ for each new window of signal. Instead, the Stochastic extension uses a decreasing sequence of upper bounds $\left(\bar{\alpha}^{(t)}\right)_{t\in\mathbb{N}}$ that should converge to 0, with a well-chosen heuristic. For instance, our implementation is based on a sequence obtained through cosine annealing, starting from an initial upper bound $\alpha_{\max}$ that is a hyper-parameter of the algorithm, as described by the following equation and presented in fig. 6.4.1:

$$\bar{\alpha}^{(t)} = \alpha_{\max} \times \frac{1}{2}\left(1 + \cos\left(\frac{\pi t}{T}\right)\right) \quad . \tag{6.4.29}$$

Thus, for each random sample i, once a gradient has been computed, the parameters are updated by evaluating $G_i\left(\theta^{(t)} - \rho^n \bar{\alpha}^{(t)} \nabla_\theta G_i\left(\theta^{(t)}\right)\right)$ for increasing values of $n \in \mathbb{N}$, until the stopping criterion is reached, and then $\alpha^{(t)} \coloneqq \rho^{n_f} \bar{\alpha}^{(t)}$.

Line search gradient descent is known to be time-consuming because it is necessary to compute the new loss for each potential choice of parameters. In the case of Dictionary Learning, this is usually a major issue, because the computation of the loss involves a sparse coding procedure. Our implementation replaces this expensive step by N iterations of proximal gradient descent that are fast to compute on GPU, which makes it possible to rely on a line search algorithm.

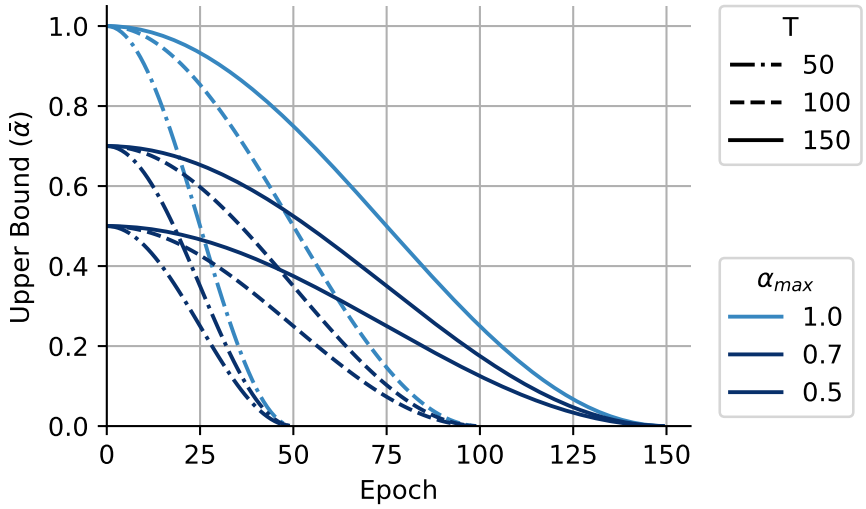


Figure 6.4.1: Evolution of the upper bound $\bar{\alpha}$ over epochs following a cosine annealing evolution, highlighting the impact of varying total number of epochs T and initial upper bounds $\alpha_{\max}$. Line styles in the legend correspond to different T values, while the color legend represents varying $\alpha_{\max}$ values.

6.5 Experiments

6.5.1 Data simulation

In the analysis of multi-dimensional time-series data, the synthesis of representative signals is vital for the evaluation and validation of signal processing algorithms. This subsection describes the data generation, encompassing the generation of a dictionary, sparse vector, signal, and optional contamination. The contamination process, although not mandatory, allows for further complexity in the generated data.

Dictionary generation The atoms dictionary $\mathbf{D} \in \mathbb{R}^{K \times P \times L}$ – K the number of atoms, P the number of channels, and L the atom's duration – can be simulated following different methods.

- **Random dictionary** The dictionary can be generated randomly, following either a Gaussian or uniform distribution, *i.e.,* each element of the random dictionary is defined as:

$$\mathbf{D}[k,p,l] \sim \begin{cases} \mathcal{N}(\mu, \sigma^2) & \text{if Gaussian} \\ \mathcal{U}nif_{[\text{lower},\text{upper}]} & \text{if Uniform} \end{cases} \tag{6.5.1}$$

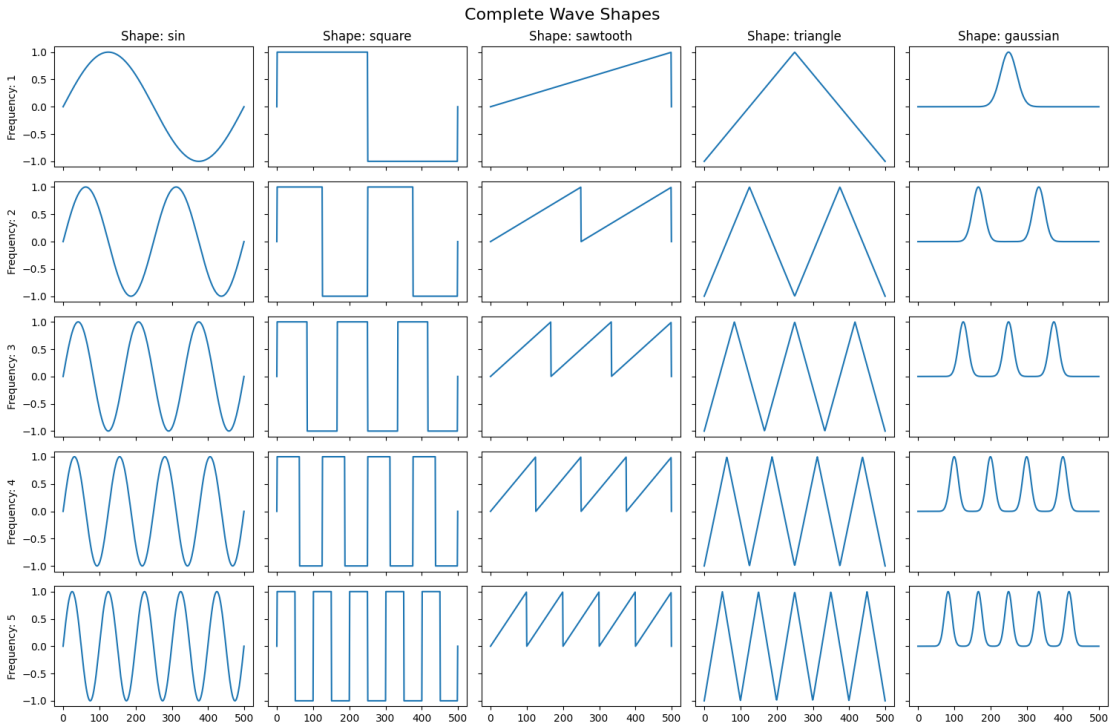


Figure 6.5.1: Shapes considered for dictionary simulation.

- **Wave dictionary** Alternatively, the dictionary can be composed of basic wave shapes referred to as atoms. These atoms can be sinusoidal, square, sawtooth, triangular, or Gaussian, with variations such as positive versions to ensure non-negative values, as presented in fig. 6.5.1. The frequency of the atoms' pattern is determined by an integer parameter, and the shapes are iteratively reused with incremented frequencies to ensure diversity. In the case of multiple channels (*i.e.,* $P \geq 2$), atoms are distributed across channels one at a time.

A Tukey window can be applied to ensure that each atom starts and ends at 0, thereby reducing artifacts.

Sparse vector generation The sparse vector $\mathbf{Z} \in \mathbb{R}^{N \times K \times \widetilde{T}}$ encapsulates the activation values and sparsity level, with $\widetilde{T} := T - L + 1$. The activation values can be either random (following Gaussian or uniform distribution) or set at a constant value. The sparsity level is specified as a ratio representing the number of non-zero values in each vector $\mathbf{z}_k^n \in \mathbb{R}^{\widetilde{T}}, n = 1, \ldots, N, k = 1, \ldots, K$. An absolute value function can be applied to obtain only positive activation values while maintaining the desired sparsity.

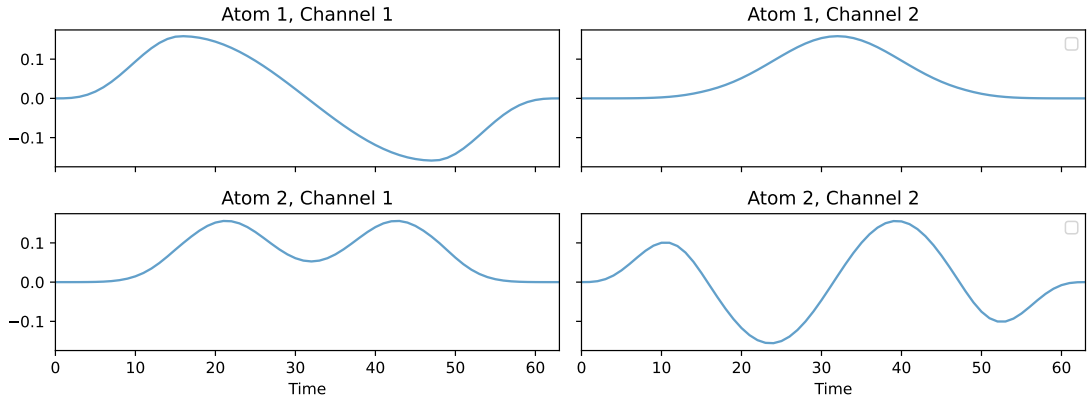


Figure 6.5.2: True dictionary in experiments on synthetic data.

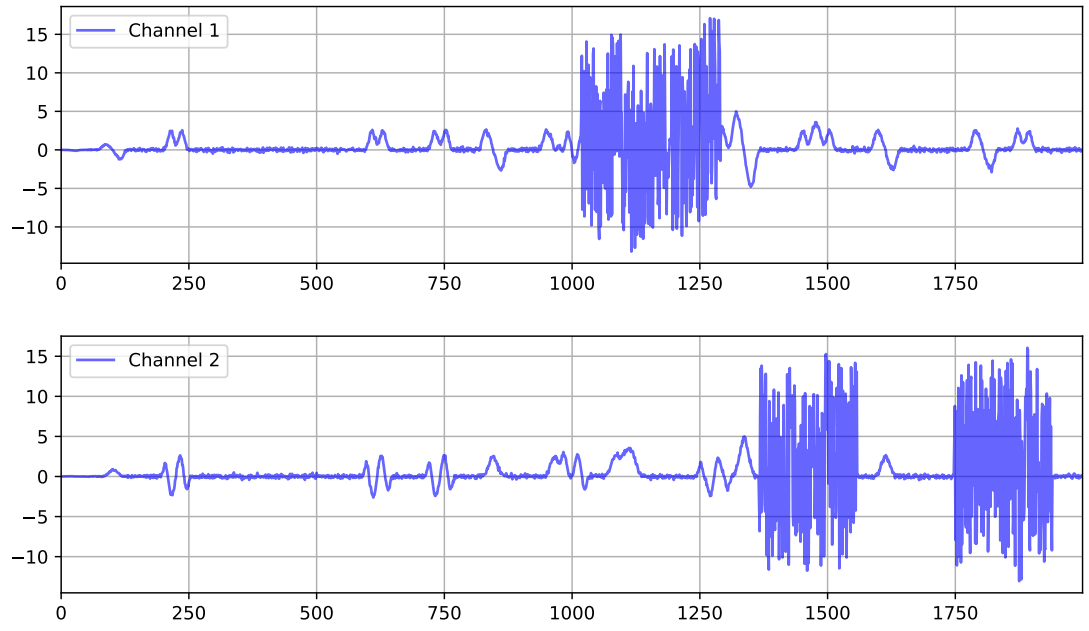


Figure 6.5.3: First 2000 timepoints of one trial (2 channels) of synthetic data, corrupted with outliers. Final data is $X \in \mathbb{R}^{10\times 2\times 5000}$

Signal construction The signal $\mathbf{X} \in \mathbb{R}^{N \times P \times T}$ is generated by computing the convolution between the dictionary and the sparse vector, with a possible random Gaussian noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$:

$$\mathbf{X}[n, p, t] = \sum_{k=1}^{K} (\mathbf{D}[k, p, :] * \mathbf{Z}[n, k, :]) + \boldsymbol{\varepsilon} \tag{6.5.2}$$

where $*$ denotes the convolution operation.

Contamination To mimic the presence of significant artifacts found in real-world data, a contamination term $\mathbf{X}_{\text{contam}}$ can be added to the signal. This involves convolving a randomly generated dictionary containing P atoms – one per channel – of length L_{contam} (preferably longer than the normal atom length, *e.g.,* $L_{\text{contam}} \geq 3L$) with a sparse vector, following a given method and sparsity level. In practice, the contamination signal is randomized again, *i.e.,* each non-null timestamp is redrawn so that it is not possible to "learn" the contamination atoms.

Simulation parameters If not stated otherwise, we used the following parameters in the follow-up experiments on synthetic data:

- dictionary of shape $(K, P, L) = (2, 2, 64)$, composed of "sin" and "gaussian" atoms shapes;
- the sparse vector is composed of constant activations of value 1, with a sparsity of $0.4\,\%$ and a length of $\widetilde{T} = 50000 - 64 + 1$, *i.e.,* roughly 200 activations per atom;
- if contaminated, we add a contamination signal of sparsity $6 \times 10^{-3}\,\%$, *i.e.,* on average, 30 contamination atoms per channels, each one of length $L_{\text{contam}} = 3L = 192$;
- the final signal is hence of duration 50 000 over 2 channels, with an additional gaussian noise of standard deviation $\sigma = 0.01$.

6.5.2 Dictionary evaluation

In our methodology, we evaluate the effectiveness of a learned dictionary, denoted as $\widehat{\mathbf{D}} \in \mathbb{R}^{K' \times P \times L'}$, by comparing it against a set of true dictionary patterns, represented as $\mathbf{D} \in \mathbb{R}^{K \times P \times L}$ and computing a "recovery score", using the convolutional cosine similarity following optimal assignment, as defined by Moreau and

Gramfort [2020]. The learned dictionary and the true patterns are structured as three-dimensional arrays, where dimensions correspond to the number of atoms, channels, and atoms' duration. The learned dictionary may differ from the true dictionary in terms of the number of atoms and the length of time atoms, typically featuring more atoms and extended durations.

The evaluation process involves a computational step known as multi-channel correlation. In this step, each atom of the learned dictionary is systematically compared with each pattern in the true dictionary. This comparison is carried out channel by channel, aggregating the results to capture the overall similarity between the dictionary atom and the pattern.

After performing these comparisons for all combinations of atoms and patterns, we create a matrix that represents the correlation strengths between each pair. To objectively assess the quality of the learned dictionary, we use an optimization technique called the Hungarian algorithm. This algorithm finds the best possible "matching" between the learned dictionary atoms and the true patterns, aiming to maximize the overall correlation.

The final score, which quantifies the performance of the learned dictionary, is derived by averaging the values of these optimal matchings. This score is scaled between 0 and 1, where 1 represents the best possible performance. A higher score indicates that the learned dictionary more accurately represents the true dictionary patterns, providing a measure of its quality and effectiveness in capturing the essential features of the data.

Mathematically, the recovery score between the dictionaries $\widehat{\mathbf{D}}$ and $\mathbf{D}$ can be expressed as follow:

$$\text{score} = \frac{1}{K} \sum_{i=1}^{K} C_{i,j^*(i)} \ , \tag{6.5.3}$$

where $j^*(i), i = 1, \dots, K$ denote the results of the linear sum assignment problem [Crouse, 2016][2] on correlation matrix $C \coloneqq \text{Corr}\left(\mathbf{D}, \widehat{\mathbf{D}}\right) \in \mathbb{R}^{K \times K'}$, with $\forall i \in [\![1\,, K]\!]$, $\forall j \in [\![1\,, K']\!]$,

$$C_{i,j} = \max_{l=1,\dots,L+L'-1} \text{Corr}_{\text{2D}}\left(D_i, \widehat{D}_j\right)[l] \in \mathbb{R} \ , \tag{6.5.4}$$

where $D_i \in \mathbb{R}^{P \times L}$ and $\widehat{D}_j \in \mathbb{R}^{P \times L'}$. The multivariate "2D" correlation between the two matrices D and $\widehat{D}$ is defined as follow:

$$\text{Corr}_{\text{2D}}\left(D, \widehat{D}\right) = \sum_{p=1}^{P} \text{Corr}_{\text{1D}}\left(d_p, \hat{d}_p\right) \in \mathbb{R}^{L+L'-1} \ , \tag{6.5.5}$$

[2]We use the `SciPy`'s implementation.

where $d_p \in \mathbb{R}^L$ and $\hat{d}_p \in \mathbb{R}^{L'}$. The 1D "*full*" correlation between the two vectors d and $\hat{d}$ is defined as follow, $\forall t \in [\![1\,, L + L' - 1]\!]$:

$$\mathrm{Corr}_{\mathrm{1D}}\left(d, \hat{d}\right)[t] = \left(d * \hat{d}\right)[t - T + 1] = \sum_{l=1}^{L} d[l]\hat{d}\,[l - t + T] \in \mathbb{R}\ , \tag{6.5.6}$$

where $T \coloneqq \max(L, L')$.

6.5.3 Experimental paradigm

In our study, we implemented a specific experimental paradigm to evaluate the performance of convolutional dictionary learning. This paradigm was carefully designed to balance computational efficiency with the need to capture sufficient data characteristics. The following parameters and strategies were employed.

- **Atom Size Determination** Our objective was to learn atoms that are slightly larger than those used in the simulation. To achieve this, we increased the size of the atoms by a factor of 1.5. Specifically, the size of each atom was determined using the formula $L' = \lfloor 1.5L \rfloor$, where L represents the original size of atoms in the simulation, and $\lfloor \cdot \rfloor$ denotes the floor function. This adjustment was made to capture more extensive features in the data.

- **Window Size Configuration** To capture a larger context around each atom for more comprehensive analysis, we set the window size to be ten times the size of an atom. This was calculated as $W = \lfloor 10L' \rfloor$. The larger window size allows for a broader view of data characteristics around each atom, providing a more extensive analysis context.

- **Batch Size Computation** Our goal was to efficiently process subsets of data. Each batch was designed to contain $p = 10\,\%$ of the data segments. We calculated the length of data loaders as $\mathcal{D} = \left\lfloor \frac{NT}{W} \right\rfloor$ and then set the batch size to $\lfloor p\mathcal{D} \rfloor$. This strategy ensured that each batch processed a manageable portion of data, promoting computational efficiency while ensuring thorough data coverage.

- **Epoch and Batch Processing** To ensure complete data coverage in each training epoch, we processed the data in batches, covering all available data within each epoch. The maximum number of batches per epoch was calculated is thus simply $\left\lfloor \frac{1}{p} \right\rfloor$. This approach guaranteed that the entire dataset was utilized for training in each epoch, enhancing the learning process.

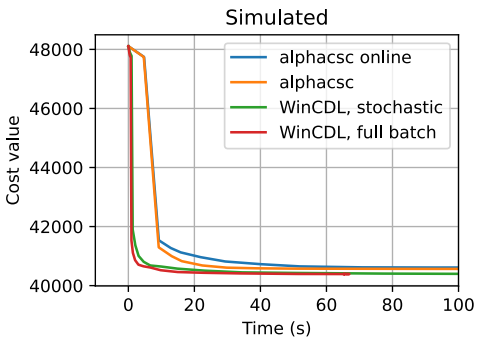
- **Number of Learned Atoms and dictionary initialization** We aimed to increase the potential for successful dictionary learning by initializing the model with a higher number of atoms than those explicitly present in the simulation. Specifically, we set the number of atoms as $K' = 2K$. This strategy was intended to enhance the model's capability to capture a broader and more diverse range of patterns, thereby increasing the likelihood of finding a set of atoms that effectively represents the underlying data and perfectly recovering the simulation dictionary. For the initial dictionary, we opted for a random generation strategy where the initial values were randomly chosen between 0 and 1, followed by norm-1 rescaling. This approach provided a diverse and stochastic starting point for the dictionary learning process. The combination of a higher initial number of atoms and random initialization allowed the model to explore a more extensive feature space than what was explicitly present in the data, potentially leading to more robust and versatile dictionary learning outcomes.

- **Optimization algorithm** We employed 50 iterations of FISTA for each batch. The experiment was configured to run for 100 epochs, implying a total of $100 \times \left\lfloor \frac{1}{p} \right\rfloor$ updates of the dictionary.

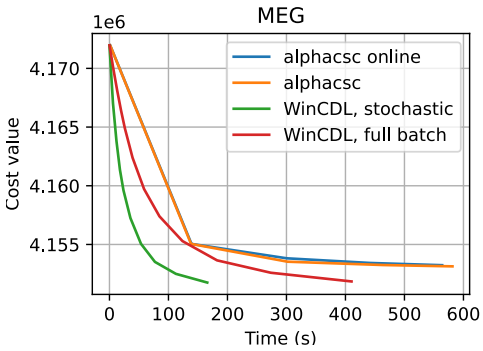
6.5.4 Results

On WinCDL efficiency In fig. 6.5.4, we provide a comparison of performance between WinCDL and two methods for rank-1 CDL implemented in the `Python` package `AlphaCSC`. The cost value is taken as $G(\mathbf{D}, \mathbf{X}) = F(\mathbf{D}, \mathbf{Z}^*(\mathbf{D}); \mathbf{X})$, where the unknown to optimize is the dictionary. Thus, we do not take into account the ability of the algorithm to compute D and the sparse codes Z at the same time, but we evaluate the dictionary D by computing F with the exact sparse codes $Z^*(D)$. The figure shows that the usage of line search and sub-windowing allows to speed up the Dictionary Learning process, on both simulated and real MEG data from the MNE *sample* dataset – previously introduced in section 4.3.2.

Parameter sensitivity of WinCDL To assess the sensitivity of WinCDL to hyperparameters, we first conduct experiments on synthetic data without contamination by outliers. We rigorously evaluate the impact of two critical hyperparameters on our algorithm's performance: the length of the learned atom and the window size relative to the true atom size. To assess the influence of the learned atom length, we varied it as a percentage of the true atom size, testing at 75 %, 100 %, 150 %, and 200 % across different true values (32, 64, and 128). This experiment was replicated 20 times for each setting. Remarkably, our findings in-



(a) Cost evolution on simulated data



(b) Cost evolution on real MEG data

Figure 6.5.4: Comparison of cost evolution during optimization between WinCDL and `AlphaCSC` methods.

dicate that the model achieves near-perfect recovery scores once the learned atom length is sufficiently long (ratio ≥ 1), highlighting that performance is robust to overestimation of the atom size.

In a parallel set of experiments, we focused on the effect of window size, setting it to 5, 10, and 15 times the true atom size, coupled with different batch sizes (25 %, 50 %, 75 %, and 100 % of the total available batches). The results consistently showed near-perfect recovery across all window size factors when the batch size was at least 50 % of the total batches, emphasizing the effectiveness of our algorithm to recover the dictionary with clean data.

These experiments – which results are presented in fig. 6.5.5 – collectively underscore the resilience of our model's performance to variations in key hyper-parameters. They reveal that once the learned atom size is reasonably estimated and sufficient data is utilized (batch factor ≥ 0.5), the model reliably achieves high recovery accuracy, demonstrating its robustness and effectiveness in diverse scenarios.

On outliers detection methods with synthetic data. We evaluated the performance of outlier detection methods on contaminated synthetic data. We examined the final recovery score over 20 repetitions and its evolution across iterations. Additionally, we investigated the score when no detection method is applied to such contaminated data. This comparison was intended to provide a clearer understanding of the relative effectiveness of each detection algorithm.

To generate corrupted synthetic data, we proceed as follows. Starting from a clean but noisy signal, we randomly select a given percentage of timestamps, that would be the first timestamps of a "outlier block". Each outlier block is composed of Gaussian noise of length $3L$ – *i.e.,* 3 times longer than the true atoms – and

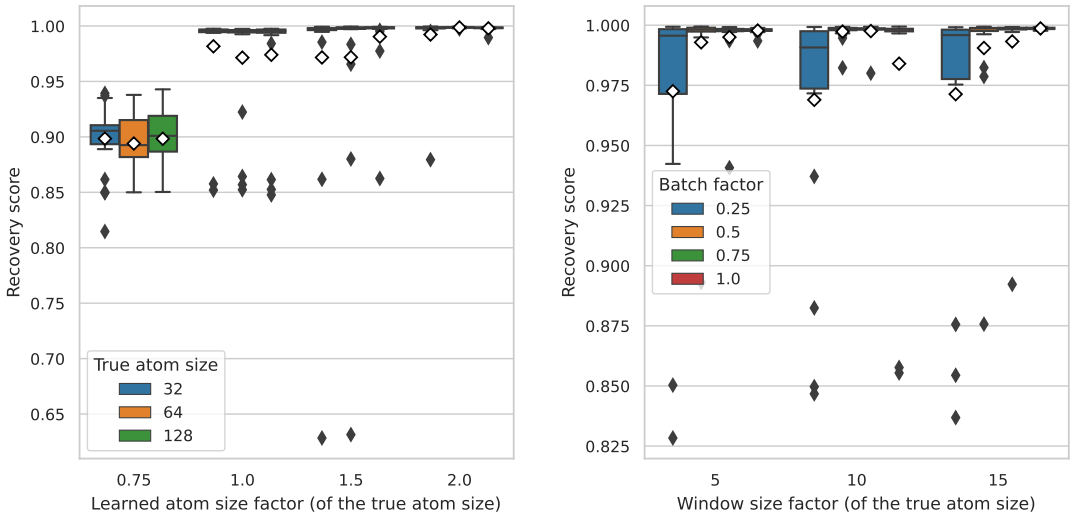


Figure 6.5.5: On uncorrupted synthetic data, effect of the learned atom size (left) and the chosen window length (right) to the final dictionary recovery score, for 20 repetitions.

their respective activation values activation is 50 times bigger than for the true atoms. Figure 6.5.3 show the first 2000 timepoints of one trial (2 channels) with outliers.

The contamination process results in approximately 10 % of the total timestamps being labeled as contaminated. This is due to the fact that each contamination activation corrupts L_{contam} timestamps. Consequently, we implemented the quantile detection method with three different α values: 5 %, 10 %, and 20 %.

The results are presented in fig. 6.5.6, based on 20 repetitions. It is observed that 4 method perform particularly well, namely MAD, IQR, z-score (with $\alpha = 1$) and the quantile (20 %) methods, with recovery score above 0.8. In contrast, when no detection method is applied, the final score is significantly lower, falling at 0.4. However, the sub-optimal performance of the other detection methods can be explained by the fact that they do not remove enough outliers in order to learn clean enough patterns.

Figure 6.5.9 presents additional metrics, namely recall and precision, computed between the true outlier mask (used to generate the corrupted synthetic data) and the learned one. In order to compute fair metrics, the learned mask is taken before the opening step.

We observed a notable inverse relationship between precision and recall across the different methods employed. Specifically, methods demonstrating higher recall, indicative of their ability to successfully identify a larger number of true outliers, concurrently exhibited lower precision. This lower precision reflects a higher rate

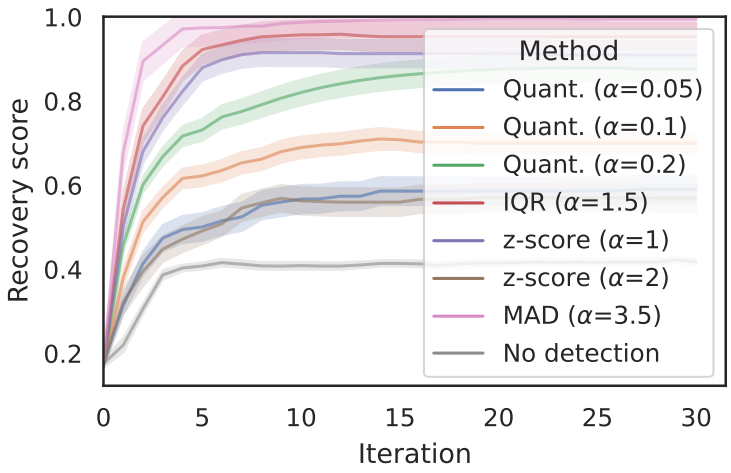


Figure 6.5.6: Median recovery score evolution for different outliers detection methods on synthetic data (20 repetitions).

of false positives, suggesting these methods may be overly permissive in classifying data points as outliers. Conversely, methods with elevated precision levels were adept at accurately pinpointing true outliers, yet this came at the expense of recall, implying that a considerable number of actual outliers were overlooked.

This dichotomy highlights a classic trade-off between precision and recall, common in classification tasks, where optimizing one measure inversely affects the other. The pattern observed in our study suggests that high-recall methods can potentially cast a wider net, but at the risk of catching non-outlier data, whereas high-precision methods are more conservative, prioritizing precision over coverage.

Evaluation of Outlier Detection Methods on ECG Data To assess the efficiency of outlier detection methods on real-world data, we utilized the Physionet Apnea-ECG dataset Penzel et al. [2000].

The ECG measures the electrical activity of the heart by using electrodes (the number depends on the test) that are connected to the skin, which detects small electrical changes due to depolarization and repolarization of heart muscles [Mostafa et al., 2019]. The Apnea-ECG Database (AED) [Penzel et al., 2000] is one of the most commonly used databases for ECG analysis. A total of 70 nighttime ECG recordings – with obstructive sleep apnoea (OSA) –, with one-minute annotations, were provided by Philipps University, Marburg, Germany and are freely available for download on the PhysioNet site[3] [Goldberger et al., 2000].

[3]Data available at: https://www.physionet.org/content/apnea-ecg/1.0.0/

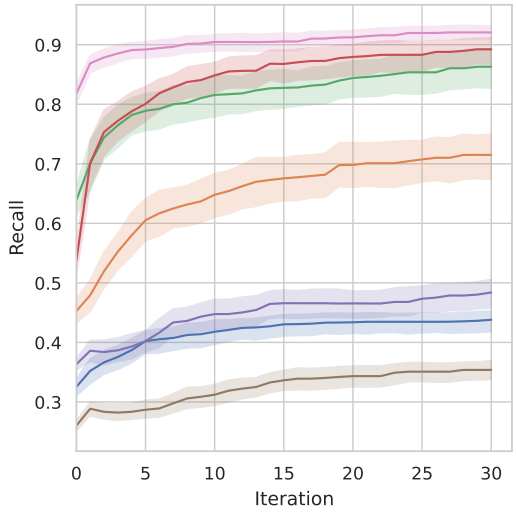


Figure 6.5.7: Recall

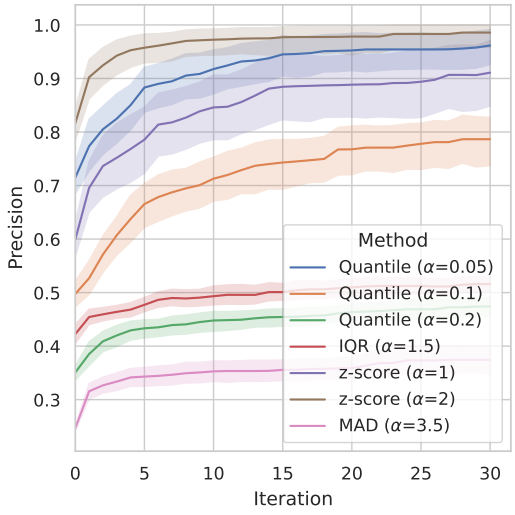


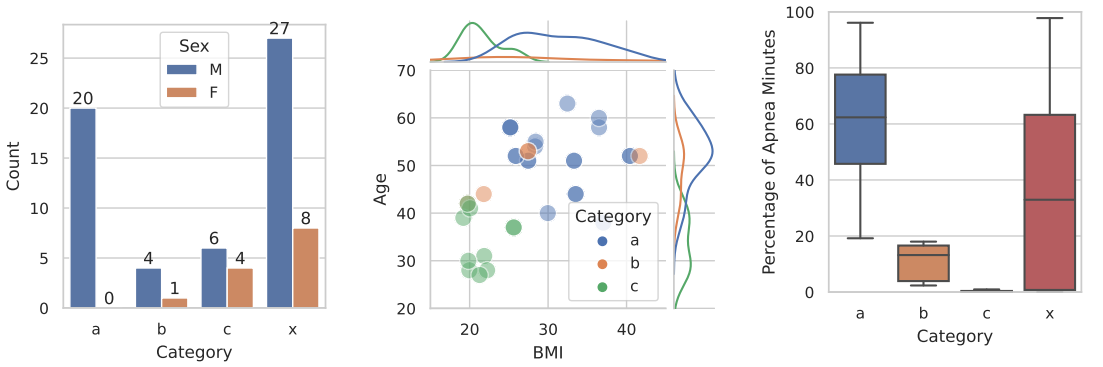
Figure 6.5.8: Precision

Figure 6.5.9: On currupted synthetic data, evolution of recall (left) and precision (right) of outliers detection. Note that by construction, when no detection method is applied, precision and recall are null

These ECG signals were used by Li et al. 2018, Pathinarupothi et al. 2017b,a, Novak et al. 2008, De Falco et al. 2018 and Dey et al. 2018.

The 70 single-lead ECG records are sampled at 100 Hz with lengths ranging between 400 min to 509 min each, and are meticulously segmented into two sets: a learning set encompassing 35 records (labeled a01 to a20, b01 to b05, and c01 to c10), and a test set consisting of 35 records (denoted x01 to x35). The gender distribution – presented in fig. 6.5.10a – shows 57 males and 13 females.

The recordings are segmented into 1 minute intervals, each tagged as either apneic or normal – *i.e.,* indicating the presence or absence of apnea during that minute – by a human scorer. The distribution of the "apnea minutes" percentage, calculated for each subject as the ratio of minutes annotated as 'A' to the total recorded minutes, is depicted in fig. 6.5.10c. This distribution reveals distinct patterns across categories: patients with apnea (category A) consistently exhibit no less than 19 % apnea minutes, those in the borderline apnea group (category B) range from 2 % to 18 %, and control subjects (category C) show less than 1 % of their minutes as apneic. In contrast to the distributions of Age and Body mass index (BMI), which do not delineate clear thresholds between these categories as illustrated in fig. 6.5.10b, the percentage of apnea minutes emerges as a potentially effective proxy for categorizing subjects within the test set (category X).



(a) Gender distribution (b) Age-BMI correlation (c) Distribution of apnea percentages

Figure 6.5.10: Descriptive statistics on Physionet Apnea-ECG dataset

No preprocessing was applied to the signals, thereby minimizing manual interventions and ensuring the integrity of the raw data. Our experimental procedure entailed learning a dictionary composed of three atoms, each spanning 1 s, from a 10 min segment of ECG data interspersed with blocks of outliers, as shown in fig. 6.3.1. The objective was to evaluate the model's ability to learn an effective dictionary from corrupted data and subsequently apply it for encoding signals free of outliers.

Upon training the model on $X_{\text{train}} \in \mathbb{R}^{10\times 6000}$, we utilized the learned dictionary $\widehat{D} \in \mathbb{R}^{3\times 100}$ to compute sparse codes for $X_{\text{test}} \in \mathbb{R}^{10\times 6000}$. The selection of minutes for X_{train} and X_{test} was manually curated to ensure the presence and absence of outliers, respectively.

Given the inherent uncertainty in the actual proportion of outliers in real data, we applied various detection methods, juxtaposed against a baseline scenario of no detection. The findings, presented in fig. 6.5.11, indicate that all detection methodologies yielded comparable loss values, surpassing the outcomes of learning without any detection mechanism. Illustrations of the dictionaries learned, particularly highlighting the cases of no detection and z-score-based detection, are presented in fig. 6.5.12. We can notice that in the absence of any detection method, the model fails to identify significant atoms, resorting to noise patterns, whereas the incorporation of a z-score-based detection method facilitates the recovery of ECG patterns by the model. This phenomenon is attributable to the fact that outlier blocks, characterized by significantly higher variance than the rest of the signal, are preferentially addressed during the sparse coding phase to minimize reconstruction error, consequently leading to the neglect of non-outlier signal segments that contain relevant ECG patterns.

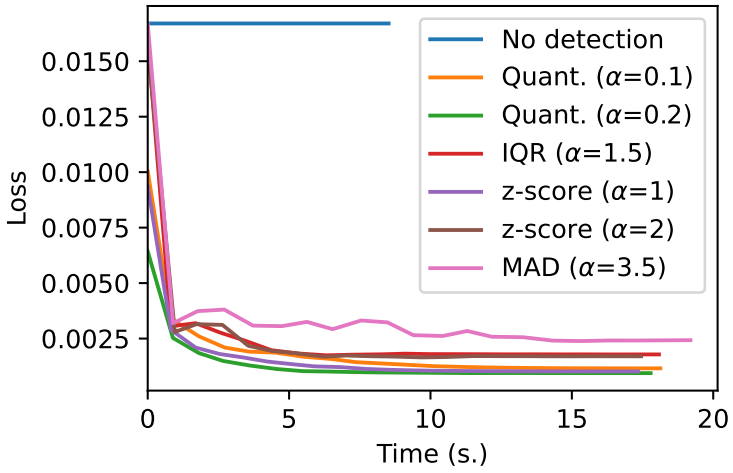


Figure 6.5.11: Loss evolution for different outliers detection method, on 10 good trials of subject a02 of dataset Physionet Apnea-ECG.

This experiment demonstrates our capability to extract pertinent patterns from corrupted data using detection methods, obviating the need for prior data preprocessing and without prior knowledge of the exact outlier percentage.

On empirical time-series with strong artifacts This study evaluates the performance of WinCDL on empirical time-series data characterized by strong artifacts. The data set, derived from a single Local Field Potential (LFP) channel[4] recording on a rodent's striatum, as described in [Dallérac et al., 2017], presents a challenging scenario for signal processing due to the presence of significant artifacts[5]. The initial 200 s of raw data is illustrated in fig. 6.5.13, with the dataset being divided into two segments of 500 s each: a "clean" segment starting from the 100th second and a "dirty" segment from the very start.

Using the `Python` package `AlphaCSC`, a Convolutional Sparse Coding (CSC) model was first fitted on both segments. However, CSC's performance deteriorated when applied to the segment with stronger artifacts. To address this, we utilized an alternative model, α-CSC [Jas et al., 2017], known for its robustness to artifacts.

[4]A *single LFP channel* refers to the recording from one location or electrode within the brain, capturing Local Field Potentials (LFPs). LFPs are the electric potentials recorded in the extracellular space in brain tissue, typically using micro-electrodes (metal, silicon, or glass micropipettes) placed in or near the area of interest in the brain, and measured in millivolts.

[5]Dataset available at: https://github.com/alphacsc/alphacsc/blob/master/examples/rodent_striatum.npy

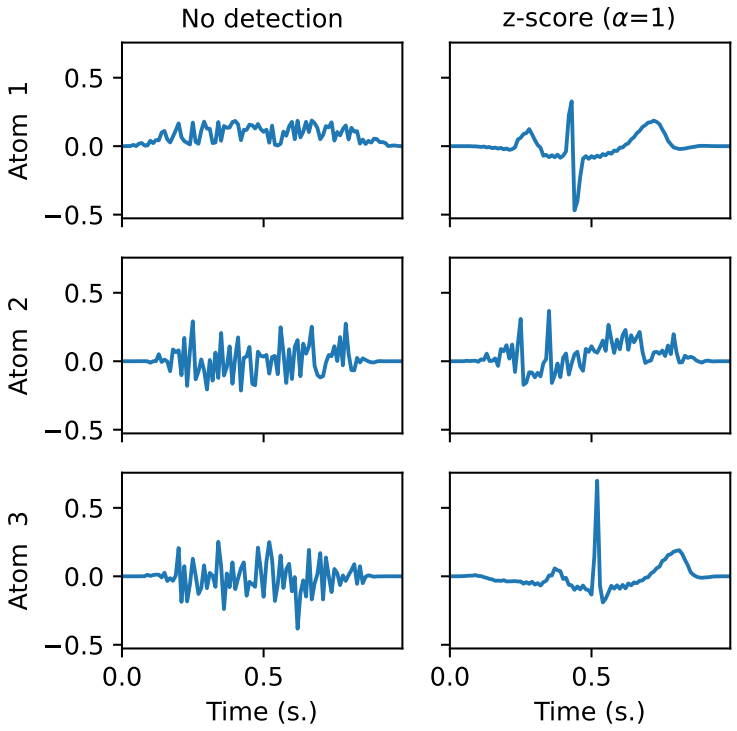


Figure 6.5.12: Learned atoms with and without outliers detection method, on 10 bad trials of subject a02 of dataset Physionet Apnea-ECG.

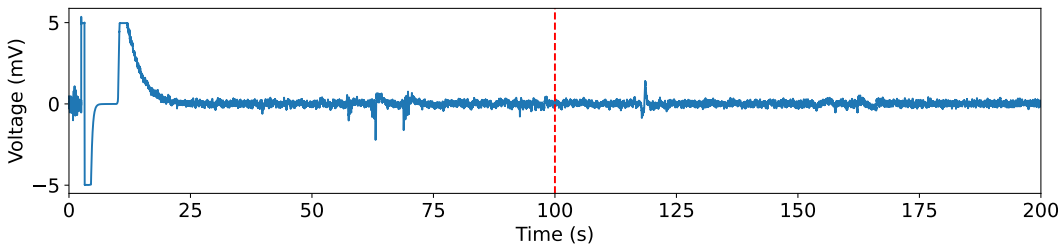


Figure 6.5.13: First 200 s of raw data from *rodent striatum* dataset.

Table 6.1: Global computation time, in seconds, on *rodent striatum* dataset.

	CSC	α-CSC	WinCDL
Clean data	69.90	N/A	**5.60**
Dirty data	23.18	277.34	**8.13**

Further, we applied our WinCDL method to both clean and artifact-laden data. For the clean segment, WinCDL was used without outlier detection, while for the "dirty" data, we implemented it both with and without an outlier detection method, specifically using a quantile at 10 %. The learned atoms, each with a duration of 1 second and totaling 3 in number, are depicted in fig. 6.5.15. In the case of the "dirty" data, we present results only from the most effective methods, namely α-CSC and WinCDL with outlier detection.

The comparative analysis presented in fig. 6.5.14, with global computation time summarized in table 6.1, reveals that WinCDL not only operates faster than the `AlphaCSC` methods but also demonstrates comparable performance in both clean and dirty data when employing outlier detection. Notably, on data with significant artifacts, WinCDL surpasses α-CSC in both computational efficiency and in achieving a lower loss value, underscoring its effectiveness in handling challenging empirical time-series data.

Finally, the experiment was also conducted with WinCDL on "dirty" data, both with and without outlier detection. It clearly demonstrates, as shown in fig. 6.5.14, the efficacy and performance benefits of the outlier detection mechanism. When WinCDL was tested without outlier detection on data containing significant artifacts, its performance was notably inferior compared to when outlier detection was employed. This contrast highlights that the incorporation of outlier detection allows WinCDL to process artifact-laden data with an efficiency akin to that observed in "clean" data, underscoring the practical utility of this feature in enhancing data analysis.

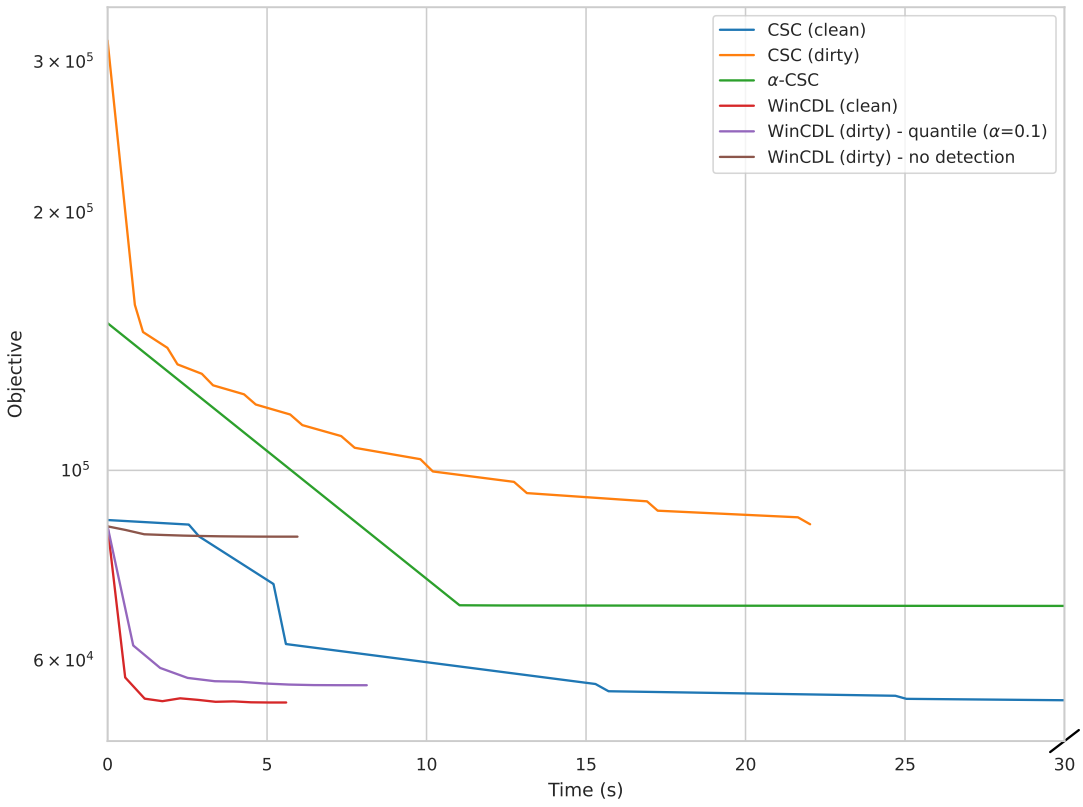
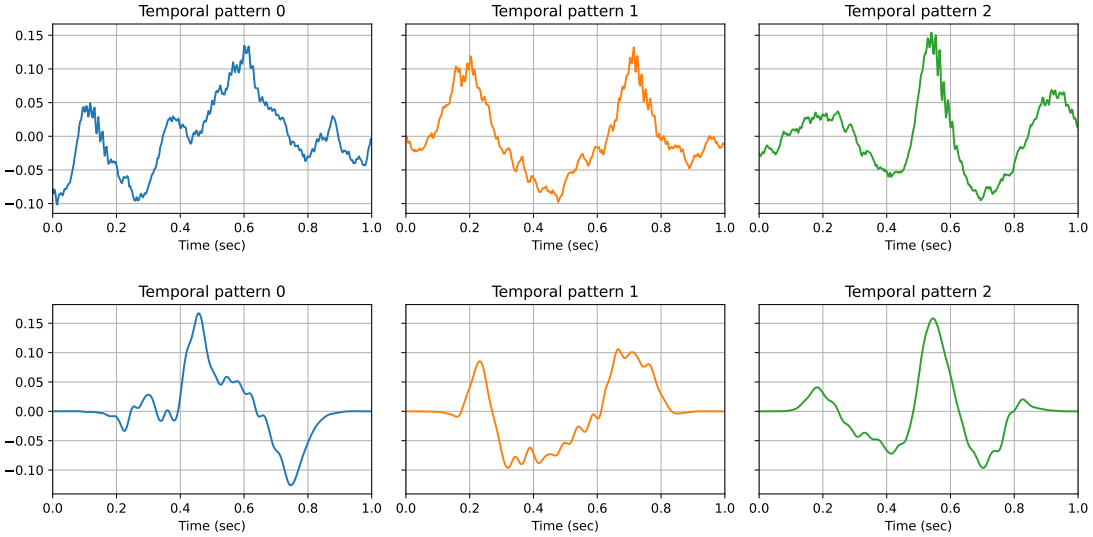
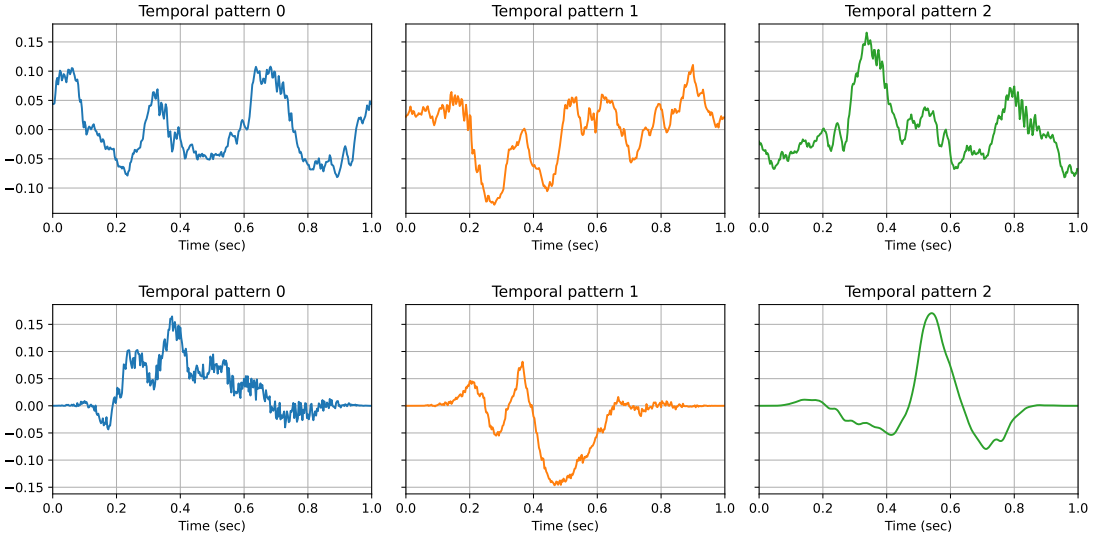


Figure 6.5.14: Loss evolution for multiple methods, on both clean and dirty segments of *rodent striatum* dataset.



(a) Results on clean data.



(b) Results on dirty data.

Figure 6.5.15: Learned atoms on empirical time-series with strong artifacts. For each sub-figure, results obtained with `AlphaCSC` methods (top) and WinCDL (bottom).

On real M/EEG data Finally, to assess qualitative performance on real M/EEG data, we extract 30 atoms of duration 1 s with WinCDL from the MNE *sample* dataset. The results are presented in fig. 6.5.16. We plot the spatial and temporal representations of each pattern, which may correspond either to artefacts or to evoked responses. The results are coherent with what is generally obtained with `AlphaCSC` on the same task with equivalent parameters, as shown in fig. 6.5.17.

6.6 Conclusion

In this study, we have introduced a robust and scalable approach to Convolutional Dictionary Learning (CDL) for unsupervised event detection in physiological signals. This approach effectively addresses the key challenges of scalability and robustness to outliers, which have historically limited the application of CDL in large-scale, population-level biomedical studies. By integrating robust regression, specifically the Least Trimmed Squares (LTS) method, into the CDL framework, we have enhanced the model's ability to learn meaningful dictionaries from data contaminated with outliers.

An advantage of our method, as demonstrated through our experiments with both synthetic and real-world datasets such as the Physionet Apnea-ECG dataset, is its ability to obviate the need for manual data inspection and complex preprocessing to remove outliers. This marks a substantial advancement over traditional approaches, as it simplifies the data preparation process, reducing the time and effort required for preprocessing, especially in scenarios involving large and noisy datasets.

Moreover, the introduction of stochastic gradient approximations and subwindowing strategies has considerably reduced the computational load, making the application of CDL feasible on a much larger scale than was previously possible. While time constraints did not allow for extensive application at the population level, the enhancements in speed and outlier resistance pave the way for such large-scale studies in the future.

However, a specific challenge remains in applying our method to M/EEG datasets, particularly concerning the identification of shared spatial patterns across individuals with varying brain morphologies. This question is pivotal in understanding the generalizability and applicability of our approach in the broader context of neurophysiological studies.

In summary, our work significantly extends the capabilities of CDL in processing physiological signals, reducing the need for manual intervention and enabling

its application on large datasets. Future efforts will focus on addressing the nuances of M/EEG data analysis and exploring the full potential of our method in population-level studies, thereby contributing to the advancement of diagnostic and monitoring tools in healthcare.

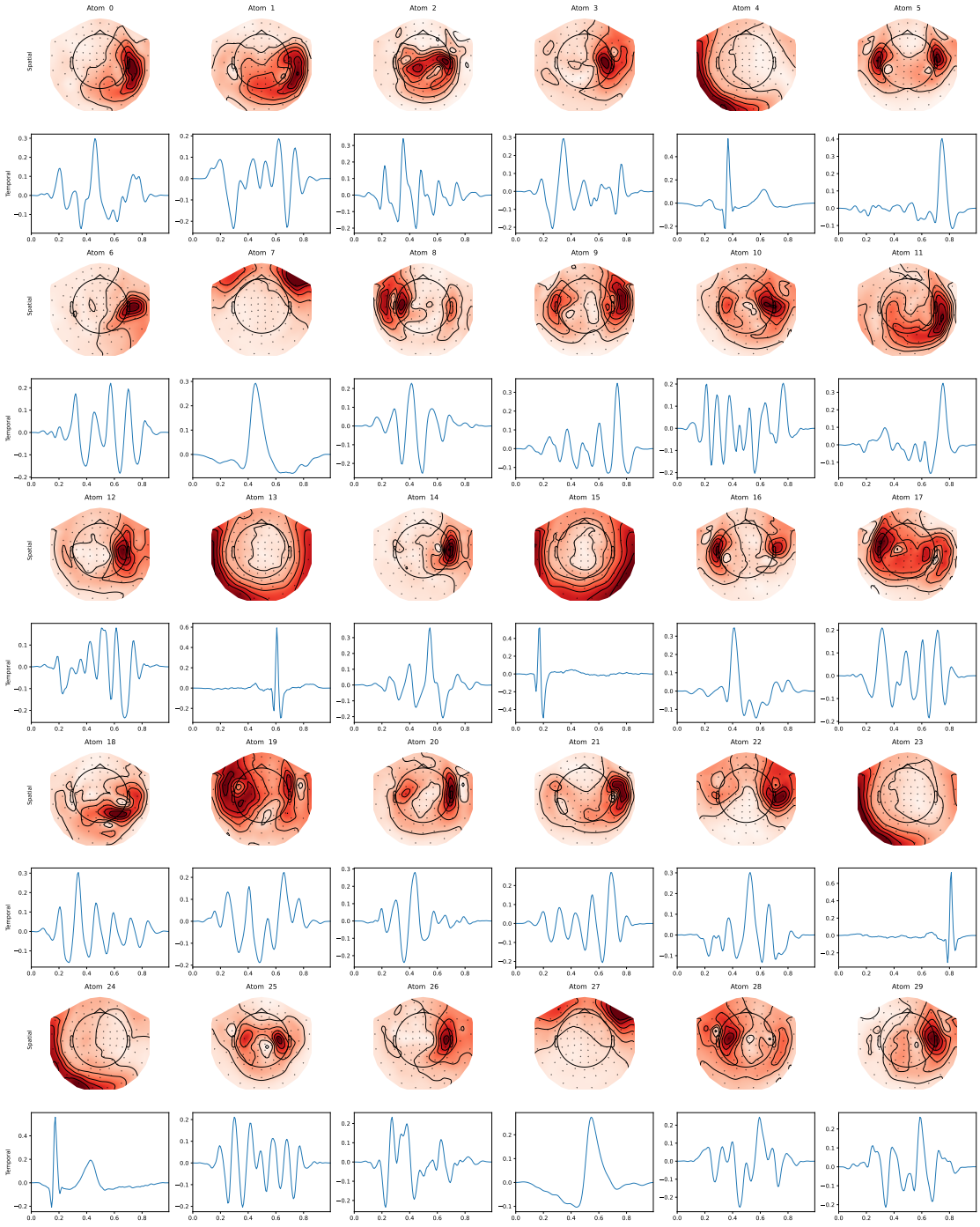


Figure 6.5.16: 30 atoms learned by WinCDL from a MEG recording.

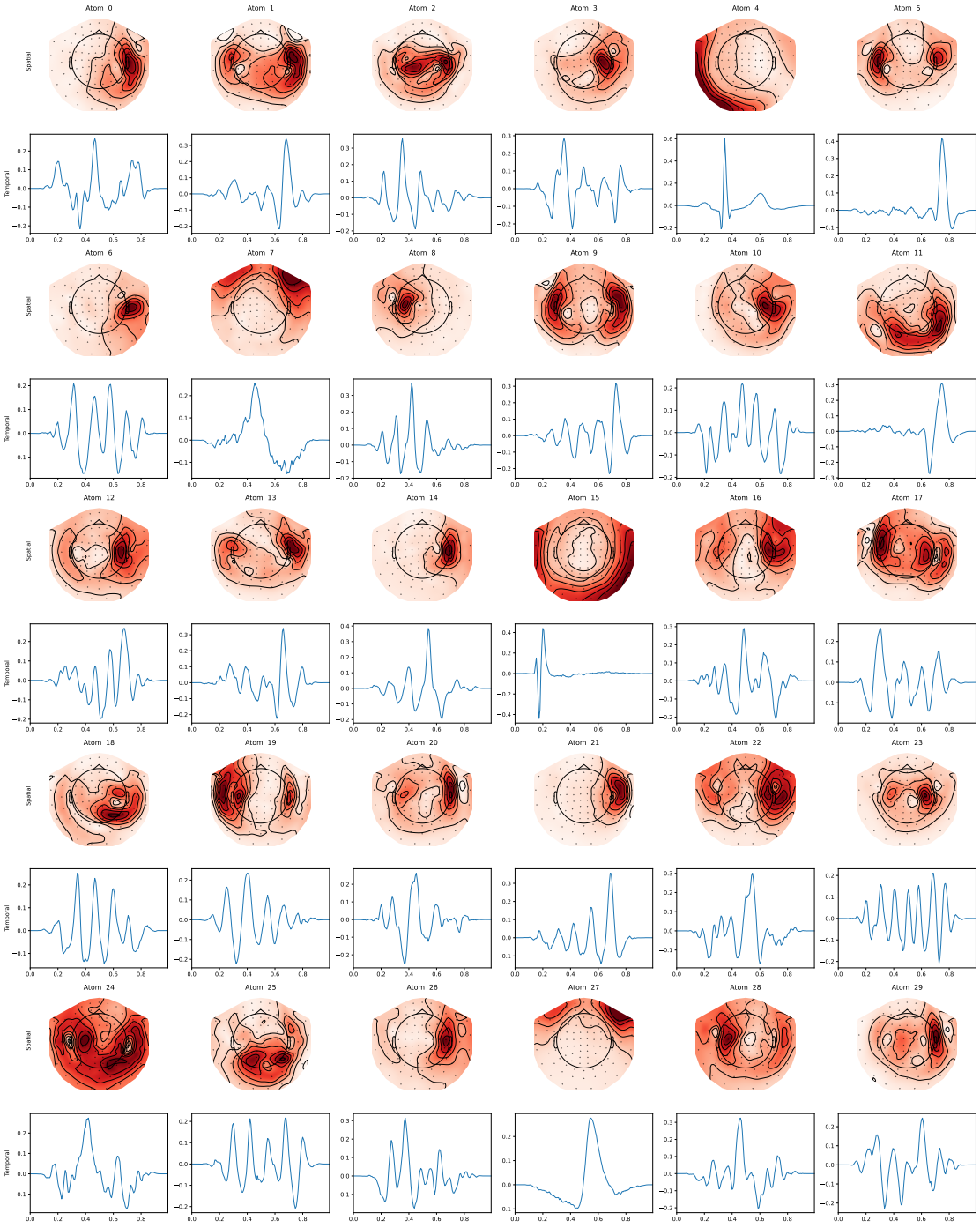


Figure 6.5.17: 30 atoms learned by `AlphaCSC` from a MEG recording.

Chapter 7

Using Population CDL to Detect Task-Related Neuromagnetic Transients and Ageing Trends in a Large Open-Access Dataset

Contents

7.1 Methods **166**
7.1.1 Participants and experimental paradigm 166
7.1.2 Data acquisition and processing 167
7.1.3 Convolutional Dictionary Learning (CDL) 167
7.1.4 Atom clustering 169
7.1.5 Selection of task-related clusters 178
7.1.6 Representative Atom Generation 180
7.1.7 Demographic analysis 181
7.1.8 Supplementary analysis 183
7.2 Results **183**
7.3 Conclusion **195**

The content of this chapter was carried out in collaboration with Lindsey Power and was published in:

Lindsey Power, Cédric Allain, Thomas Moreau, Alexandre Gramfort, and Timothy Bardouille. Using convolutional dictionary learning to detect task-related

neuromagnetic transients and ageing trends in a large open-access dataset. *NeuroImage*, 267:119809, 2023

HUMAN neurophysiological signals recorded by magneto- or electroencephalography (M/EEG) consist of a series of brief bursts of neural activity with variable underlying sources and temporal characteristics [Tal et al., 2020, Jones, 2016]. These data are often analysed by averaging across many repetitions of a task or aggregating over time, resulting in an easily interpretable signal (see section 1.2 for more details on that matter). However, we know that this aggregation is a simplification that results in a loss of important information about the transient burst dynamics (*e.g.,* burst rate, burst power, peak frequency) in the raw signal. The first realisation of "transient bursts" in electrophysiological data dates back to the identification of sleep spindles (spontaneous 12 Hz to 14 Hz transient activity) in human EEG data in the 1930s [Berger, 1929, Loomis et al., 1935]. Since then, transient bursts of various frequencies have been identified in electrophysiological recordings from humans and animal models, and have been linked to physiological and cognitive functions including attention, working memory, arousal and relaxation, and voluntary movement [Herbert and Lehmann, 1977, Lakatos et al., 2004, Feingold et al., 2015, Lundqvist et al., 2016, Shin et al., 2017, Little et al., 2019, Errington et al., 2020, He et al., 2020, Wessel, 2020]. In parallel, analysis methods have been developed to detect, characterise, and identify changes in transient bursts at the single subject and group level.

One set of transient bursts of particular interest are sensorimotor beta and mu bursts, which are modulated by voluntary motor tasks. Feingold et al. [2015] first demonstrated that brief bursts of beta oscillations in the motor and pre-motor cortices could account for virtually all cortical beta-band activity in monkeys. These findings were later replicated in a multi-modal, multi-species study by Shin et al. [2017]. Similarly, Brady et al. [2020] demonstrated that task-related reductions in the inter-trial average beta-band power (*i.e.,* beta suppression) in humans can be explained mainly by a reduction in the rate of occurrence of beta bursts with movement onset. Transient beta and mu bursts have since been shown to play a functional role in movement initiation and cancellation [Wessel, 2020, Errington et al., 2020], and response accuracy and reaction time [Little et al., 2019, He et al., 2020, Wessel, 2020].

Recent work has also shown that sensorimotor beta burst characteristics change with normal healthy ageing. Particularly, it has been demonstrated that transient beta burst characteristics (*e.g.,* burst rate, peak frequency, peak power) show age-related changes [Brady et al., 2020, Brady and Bardouille, 2022] that can explain the previously observed age-related increase in sensorimotor beta suppression in

the average [Bardouille et al., 2019, Rossini et al., 2007]. Furthermore, it has also been shown that the spatial localization of transient beta bursts changes with age, expanding to recruit additional areas, and exhibiting an anterior shift in peak localization with increasing age [Power and Bardouille, 2021].

With the increasing interest in transient burst-based analyses has come a surge in development of analysis methods for detecting and characterising bursts. At present, there is no gold-standard method for detecting and characterising transient bursts in electrophysiological data, and each proposed method comes with associated advantages and limitations. In addition, there is no obvious framework to complete a group-level analysis of the combined spatial and temporal characteristics of identified bursts and most techniques have not been optimized for use with large datasets. The simplest, and most commonly used burst detection method uses amplitude thresholding to detect bursts of high power activity within a predefined frequency range of interest. This method, popularized by Shin et al. [2017] defines bursts as local maxima in the time-frequency representation that exceed a pre-set power threshold (multiple of the median power) and fall within a predefined frequency range [Shin et al., 2017, Brady et al., 2020]. While this method has been widely used to detect mu, beta, and gamma bursts in human and animal models [Herbert and Lehmann, 1977, Lakatos et al., 2004, Feingold et al., 2015, Lundqvist et al., 2016, Shin et al., 2017, Little et al., 2019, Errington et al., 2020, He et al., 2020, Wessel, 2020], it is limited in its applications due to its imposition of assumptions about the frequency, waveform shape, and linearity of the signal of interest. In addition, the method does not effectively account for the aperiodic background activity when applying thresholding. Further, this method operates on a single signal (*e.g.,* channel, source reconstructed time course), and does not take into consideration multi-channel interactions or signal spread, making it difficult to compare spatiotemporal characteristics between subjects. To address these limitations, a number of alternative burst detection methods have been proposed.

The Better Oscillation Detection (BOSC) and Periodic/Aperiodic Parameterization of Transient Oscillations (PAPTO) methods are alternative amplitude thresholding methods that have been designed to account for aperiodic background activity in the signal, and have been shown to increase sensitivity to bursts [Caplan et al., 2001, Whitten et al., 2011, Caplan et al., 2015, Kosciessa et al., 2020, Rayson et al., 2022, Brady and Bardouille, 2022]. These methods, however, still rely on several fundamental assumptions about the approximate frequency, waveform shape, and spatial location of the signal of interest. To reduce these assumptions, many have moved towards data-driven methods of burst detection. Examples of this include Empirical Mode Decomposition (EMD) and cycle-by-cycle analyses which automatically detect approximately sinusoidal waveforms in nonlinear or nonstationary data [Huang et al., 1998, Cole and Voytek, 2019, Fabus et al., 2021], and Brief Amplitude Undulation (BAU) detection which automatically detects stereo-

typical waveforms based on their shape in the temporal domain [Abeles, 2014, Tal and Abeles, 2016, 2017]. Several types of dictionary learning algorithms including MoTIF [Jost et al., 2006, Brockmeier and Príncipe, 2016], Sliding Window Matching [Gips et al., 2017], and Adaptive Waveform Learning [Hitziger et al., 2017] have also been applied to the burst detection problem. These algorithms, which were largely developed for other applications such as image processing, and audio signal segmentation, have shown promise as burst detection methods due to their ability to learn repeating temporal motifs in the signal. While the data-driven nature of all of these methods provides an improvement over traditional amplitude thresholding methods, these methods are still limited in scope as they operate on a single time course and fail to consider the multi-channel dynamics that are critical to understanding electrophysiological signals, and how they change across a population.

Analysis methods that account for multi-channel interactions, such as Hidden Markov Modelling (HMM) and EEG Microstates have been used to detect transient states of brain activity during task-performance and in disease [Baker et al., 2014, Vidaurre et al., 2016, Michel and Koenig, 2018, Quinn et al., 2018, Becker et al., 2020, Seedat et al., 2020, Coquelet et al., 2022]. HMM identifies full graphical networks that exist for 100 ms to 200 ms at a time and exhibit rapid shifts between states, while the EEG Microstates method identifies short periods of stable scalp potentials that reflect sharp events of neural synchronization [Coquelet et al., 2022]. A strength of these data-driven approaches is that they can identify repeating transient states across the whole head. However, the assumption that multiple states cannot coexist in time is a limitation as it is not uncommon to observe independent electrophysiological processes (*e.g.,* occipital alpha and sensorimotor beta) co-occurring in time. Another method that has been proposed for use in the multi-channel detection of transient bursts is the well-known Independent Component Analysis (ICA) algorithm [Vigário et al., 1998, Hyvärinen and Oja, 2000, Himberg et al., 2004, Oja and Zhijian, 2006, Briley et al., 2021]. ICA has been a widely successful workhorse for extracting spatiotemporal components in electrophysiological data (*cf.* section 1.2.2). However, assuming the independence of sources may not be realistic when working with highly correlated task-related brain oscillations such as sensorimotor mu and beta. In addition, ICA considers long time course states of brain activity and does not break the signal into short repeating temporal motifs that are characteristic of transient bursts. Therefore, it is necessary to identify a method that employs a multi-channel, data-driven approach while allowing for the co-occurrence of short, repeating spatiotemporal motifs (*i.e.,* transient bursts), that may or may not have correlated sources.

One such method that meets these criteria is multivariate convolutional sparse coding (CSC) which is a specification of the broader class of convolutional dictionary learning (CDL) algorithms. CDL represents the multivariate neural signals

as a set of spatiotemporal patterns, called *atoms*, with their respective onset times and magnitudes, called *activations*. CDL has emerged as a convenient and efficient tool to extract patterns, in particular due to its ability to easily include physical priors for the patterns to recover. For example, for M/EEG data, Dupré la Tour et al. [2018] have proposed a CDL method which extracts atoms that relate to the current dipoles used to model brain activity by imposing a rank-1 structure to better account for the linear spread of the signal across channels. Each atom is thus associated with an activation vector that provides a record of time points throughout the signal at which the atom is present, and the associated magnitude of the atom at those time points [Jas et al., 2017, Dupré la Tour et al., 2018, Moreau and Gramfort, 2020]. CDL operates similarly to classical Independent Component Analysis (ICA; Winkler et al. 2015), decomposing the signals as a sum of topographies and sources [Dupré la Tour et al., 2018] (*cf.* chapter 2). However, CDL does so not by assuming that the sources are independent, but by assuming that the source time courses are formed by repeated waveforms. CDL has been previously validated on single subject datasets to recover biological artifacts, non-sinusoidal mu patterns with sensorimotor topography, occipital alpha bursts, and evoked-type responses (Dupré la Tour et al. 2018 and chapter 4 of this present manuscript).

Despite the success of CDL and other data-driven methods on single subject studies, the validity of multi-channel, data-driven methods for use in between-subject comparisons and group-level analyses has been largely unexplored. Some work has employed HMM extended to a multi-subject setting by concatenating data across participants to identify common repeating states [Baker et al., 2014, Vidaurre et al., 2016, Quinn et al., 2018, Becker et al., 2020, Seedat et al., 2020], and few studies have demonstrated that ICA and EEG Microstates yield consistent patterns across participants [Himberg et al., 2004, Michel and Koenig, 2018]. However, none of these studies explored variability between subjects, and group-level differences and trends, highlighting the need for methods to explore group-level trends in transient bursts detected by a robust data-driven method. The favourable characteristics and promising preliminary results of the CDL method make it a logical candidate for group-level investigations of transient bursts.

The objective of the current work is thus to use the CDL method to detect and characterise ageing trends in task-related transient bursts at the group level in a large, open-access dataset. Here, we detect (in single subjects) repeating spatiotemporal atoms in sensorimotor MEG data from the Cam-CAN dataset [Shafto et al., 2014, Taylor et al., 2017], and cluster similar atoms across participants to allow for group-level analysis. We then assess clusters for age-related trends in atom characteristics. It is hypothesised that CDL will successfully extract task-related atoms that are biologically plausible, including those that resemble sensorimotor beta and mu transient bursts. This hypothesis is based on the findings of previous

literature that demonstrate a functional role of beta and mu transient bursts in sensorimotor tasks [Herbert and Lehmann, 1977, Lakatos et al., 2004, Feingold et al., 2015, Lundqvist et al., 2016, Shin et al., 2017, Little et al., 2019, Errington et al., 2020, He et al., 2020, Wessel, 2020]. It is further hypothesised that within task-related atom clusters, atoms will show age-related changes in their spatiotemporal characteristics, in line with previous findings. Specifically, it is predicted that for sensorimotor beta-type bursts, burst frequency will decrease with age, spatial position will shift anteriorly with age, and pre-movement activation will increase with age as a result of increasing burst rate with age [Bardouille et al., 2019, Brady et al., 2020]. This work presents, for the first time, the detection of group-level trends in transient bursts using a flexible, multi-channel, data-driven CDL method. By combining this powerful detection algorithm with the big data available in the Cam-CAN dataset, we can increase our understanding of the role of neuromagnetic transients in normal healthy ageing and provide an improved framework for analysing transient bursts at the group level in future work.

7.1 Methods

Text in sections 7.1.1 and 7.1.2 was adapted from Brady et al. [2020] and Power and Bardouille [2021]. Work described in these sections was completed previously, except where specified. See fig. 7.1.1 for a workflow diagram describing the analysis process for this work.

7.1.1 Participants and experimental paradigm

MEG data were collected from 650 participants in Phase 2 of the Cam-CAN examination of healthy cognitive ageing. Participant ages ranged from 18 to 88 years of age, with an equal distribution in age per decile and equal proportions of males and females. All participants provided written, informed consent prior to participating in each phase of the study. The study was conducted in compliance with the Declaration of Helsinki and data collection was approved by local ethics boards [Shafto et al., 2014]. In the current work, we report findings from 563 participants (86.6 % of the original 650 datasets) who had sufficient MEG and anatomical MRI data required for localization. Participants who did not have anatomical MRI data were excluded from analysis to ensure consistent localization procedures were applied across all included participants. Each participant performed a sensorimotor task during the MEG scan [Shafto et al., 2014]. In the task, participants responded with a right index finger button press to unimodal or bimodal audio/visual stimuli. The order of bimodal and unimodal trials was randomized, and the inter-trial

interval varied from 2 s to 26 s. The button press task did not include specific imperatives related to performance, *e.g.,* fast responses. Thus, brain-behaviour interactions focused on response time were not investigated in this report.

7.1.2 Data acquisition and processing

Data were obtained from the Cam-CAN dataset [Shafto et al., 2014, Taylor et al., 2017][1]. MEG data were acquired at 1000 Hz with inline band-pass filtering between 0.03 Hz and 330 Hz using a 306-channel Vectorview system with continuous head position monitoring (Elekta Neuromag, Helsinki, Finland).

All MEG processing was completed in the Python programming environment (v.3.7.7), using the MNE-Python library (v.0.23.0) [Gramfort et al., 2013, 2014]. Data were pre-processed using temporal signal space separation (tSSS) to perform environmental noise reduction, and reconstruction of missing or corrupted MEG channels [Taulu and Simola, 2006]. The task data was then parsed into trials synchronized to each button press, with a duration of 3.4 s, including a 1.7 s pre-movement interval. The 3.4 s window length was selected to ensure a sufficient post-movement interval to capture the entire beta rebound response. Trials were excluded if the button press occurred more than 1 s after the cue (indicating poor task performance) or if the button press occurred within 3 s of the previous button press (which provided insufficient baseline for subsequent analysis). In the current work, data were bandpass filtered between 2 Hz to 45 Hz and resampled with a sample rate of 150 Hz.

7.1.3 Convolutional Dictionary Learning (CDL)

Recall from chapter 2 that the objective of CDL is to decompose a signal into the convolution between a few translationally invariant recurring patterns, called atoms, and their sparse activation vectors. In the application to M/EEG signals, a rank-1 constraint is added to the dictionary to take into account the physics of the signals (*i.e.,* the instantaneous linear spread of the signals across channels; see eq. (2.4.3)). This extra constraint decomposes the atoms with a spatial and a temporal component, which can easily be interpreted by neuroscientists. The result of the optimization is a set of instantaneous spatiotemporal signals and associated sparse activation vectors (see fig. 4.1.2 for a schematic representation of how CDL decomposes raw MEG signals).

[1] Available at https://www.cam-can.org/index.php?content=dataset.

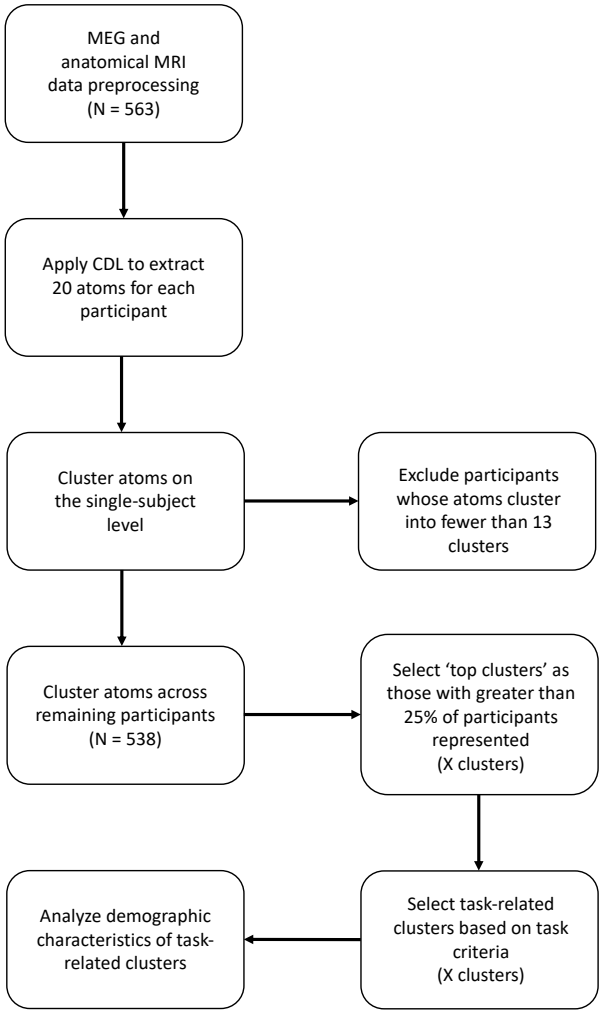


Figure 7.1.1: Workflow diagram.

In the current work, we relied on the `alphacsc`[2] Python package [Dupré la Tour et al., 2018] for CDL with rank-1 constraint. The hyperparameters used in this work are based on those established by Dupré la Tour et al. [2018] and used in the `alphacsc` tutorials[3]. These hyperparameters were used because they have given satisfying results for detecting similar types of induced responses (*e.g.,* somatosensory mu waves) in previous work. For each considered subject, CDL outputs 20 spatiotemporal atoms alongside their respective sparse activation vector z that corresponds to the onsets of the waveforms. For each atom, CDL with rank-1 constraint provides the topography, later referred to as u of size number of MEG sensors, and the temporal waveform v of duration $L = 500\,\text{ms}$.

7.1.4 Atom clustering

After applying CDL to each individual participant, atoms from all participants were clustered into groups, based on their spatiotemporal similarity. The approach of detecting atoms in individuals and then clustering across participants was employed to ensure that individual variability in atoms was preserved to allow for between-subject comparisons of atom characteristics. A correlation-based clustering approach taking into consideration both the u (spatial) and v (temporal) vectors was applied to the atoms. The clustering relies on a simple iterative approach that groups atoms together on the basis of high correlation without prior specification of the number of clusters. This method is similar to that described by Bansal et al. [2004], and has been widely used for biomedical clustering applications in the past [Bhattacharya and De, 2008, 2010, Miljkovic et al., 2010]. This method was selected over other traditional clustering methods because it allowed for simultaneous consideration of multiple clustering metrics (*i.e.,* both spatial and temporal vectors), which is not possible with other "out of the box" clustering methods.

To compare two atoms (u_1, v_1) and (u_2, v_2), we used the Pearson correlation coefficient of the u vectors

$$R_u = R_u\left(u_1, u_2\right) \coloneqq \frac{\langle u_1, u_2 \rangle}{\|u_1\|_2 \cdot \|u_2\|_2} \ , \tag{7.1.1}$$

and the maximum cross-correlation coefficients of the v vectors

$$R_v = R_v\left(v_1, v_2\right) \coloneqq \max_{k \in [\![0, L-1]\!]} \left(\frac{C(k)}{\|v_1\|_2 \cdot \|v_2\|_2} \right) \tag{7.1.2}$$

$$\text{with} \quad C(k) \coloneqq \sum_{i=1}^{L-k} v_1[i] \cdot v_2[i+k] \ . \tag{7.1.3}$$

[2] Available at `alphacsc.github.io/`.

[3] Available at `https://alphacsc.github.io/auto_examples/index.html`.

Cross-correlation was used for the v vector to account for phase differences in otherwise similar waveforms. The algorithm then stepped through each atom of interest sequentially (in no particular order), and clustered atoms with one another based on the magnitude of their correlation coefficients (R values). Each atom was compared to pre-existing clusters by calculating R_u and R_v values between the current atom and each atom in the cluster, and then averaging across all R values (for u and v separately) obtained from the cluster. If, for a given cluster, both (R_u and R_v) average values exceeded a pre-determined threshold ρ, the atom was considered highly correlated to the cluster. For atoms that were highly correlated to more than one pre-existing cluster, the atom was added to the cluster to which it had the highest cumulative correlation (average R_u + average R_v). If the atom was not highly correlated (*i.e.,* the average R_u and R_v values did not exceed the threshold) to any of the pre-existing clusters, then a new cluster was created. The algorithm proceeded through all atoms of interest, yielding a number of clusters not a priori defined. The pseudo-code of this clustering algorithm is provided in algorithm 5.

Algorithm 5: Atoms clustering

input : The threshold ρ and the set of atoms $\{(u_i, v_i)\}_{i=1}^{N}$
output: The dictionary of clusters $\mathcal{K}$

Initialize $K = 1, \mathcal{K}[1] = \{(u_1, v_1)\}$ `// A first cluster with the first atom`
for $i = 2, \ldots, N$ **do**
  for $k = 1, \ldots, K$ **do**
    Compute $\overline{R_u^{(k)}} = \frac{1}{\#\mathcal{K}[k]} \sum_{(u,v)\in\mathcal{K}[k]} R_u(u_i, u)$
    and $\overline{R_v^{(k)}} = \frac{1}{\#\mathcal{K}[k]} \sum_{(u,v)\in\mathcal{K}[k]} R_v(v_i, v)$
  end
  Define $\mathcal{C} = \left\{ k \in [\![1, K]\!],\ \overline{R_u^{(k)}} \geq \rho, \overline{R_v^{(k)}} \geq \rho \right\}$ `// Set of candidates`
  if $\mathcal{C} = \emptyset$ **then**
    $K = K + 1$
    $\mathcal{K}[K] = \{(u_i, v_i)\}$ `// Create new cluster`
  else
    Compute $k' = \arg\max_{k\in\mathcal{C}} \overline{R_u^{(k)}} + \overline{R_v^{(k)}}$
    Append $\mathcal{K}[k']$ with (u_i, v_i) `// Append existing cluster`
  end
end
return $\mathcal{K}$

Single subject exclusion

Preliminary correlation-based analysis of single participants' atoms revealed that a few participants had numerous highly correlated atoms (*i.e.,* little variability in the spatiotemporal features of their atoms) as presented in fig. 7.1.4. Visual inspection of the atom data revealed that these participants tended to have atom profiles dominated by artifacts (*e.g.,* eyeblink and other global artifacts, see fig. 7.1.2 compared to fig. 7.1.3) or by a persistent slow (alpha frequency) rhythm with variable topographic representation. This observation suggested that these participants have abnormal and/or artifactual data that should be excluded from further analysis to avoid skewing effects in the whole-group clustering process. This prompted the development of an exclusion process based on the correlation-based clustering methods described above, by which participants with low atom variability were identified and excluded from further analysis.

Here, the 20 atoms computed for a given participant were compared to one another and clusters of highly similar atoms were created within participants. In order to select the optimal R value threshold ρ for clustering, the threshold was varied from $\rho = 0.2$ to $\rho = 0.9$ and clustering was performed for each value of ρ. Histograms illustrating the number of clusters yielded per participant were then created and examined for each value of ρ (see fig. 7.1.5). The goal of the analysis was to select a threshold that yielded maximum separation between participants with few groups (high degree of similarity between atoms) and those with many groups (dissimilar atoms). Therefore, histograms were examined for a bimodal distribution with maximal separation between peaks. On this basis, a R value threshold of $\rho = 0.8$ was selected. All participants with less than 13 distinct groups of atoms were excluded from subsequent analyses. The value of 13 clusters was selected as the point that best separated the first and second peaks of the distribution. Based on these criteria, 25 participants were excluded at this step, resulting in a total of 538 participants who were used for the remaining analyses, see table 7.1 for the details.

Global clustering

The correlation-based clustering methods described above were then applied on the whole-group level to create clusters of atoms of the same type across participants, which would facilitate atom comparisons between participants. The selection of an R value threshold ρ for the global clustering was conducted separately from the single subject clustering due to the differing objectives of the two analyses. While the single subject clustering aimed to exclude participants with an abnormally high degree of similarity between atoms, the global clustering aimed to create

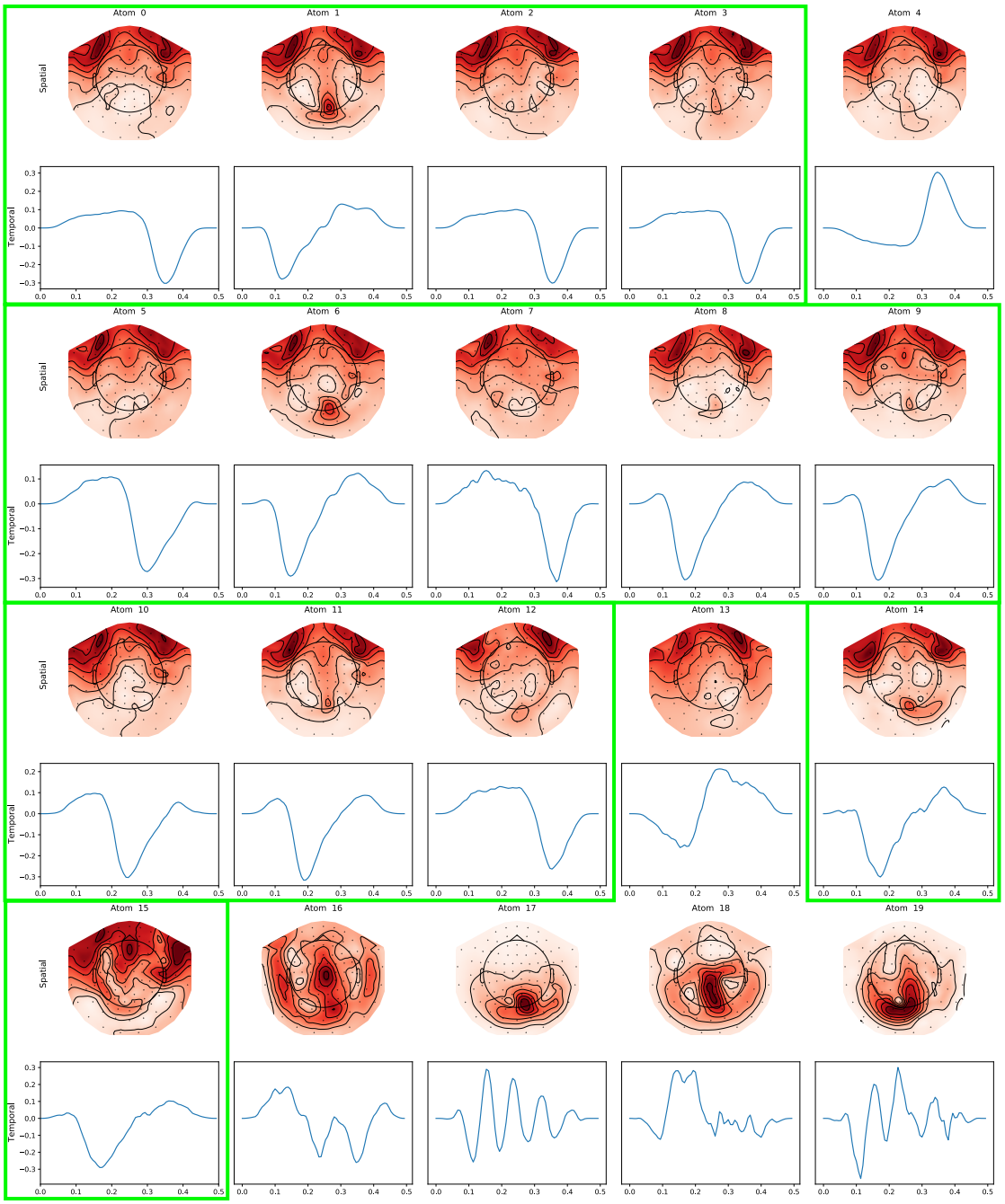


Figure 7.1.2: Spatial and temporal representation of the 20 atoms extracted from the subject CC121428, that obtained 7 clusters. Framed atoms are part of a single intra-subject cluster. One can observe the low variability in the atoms obtained from the CDL step, showing the prevalence of artifacts in the recording.

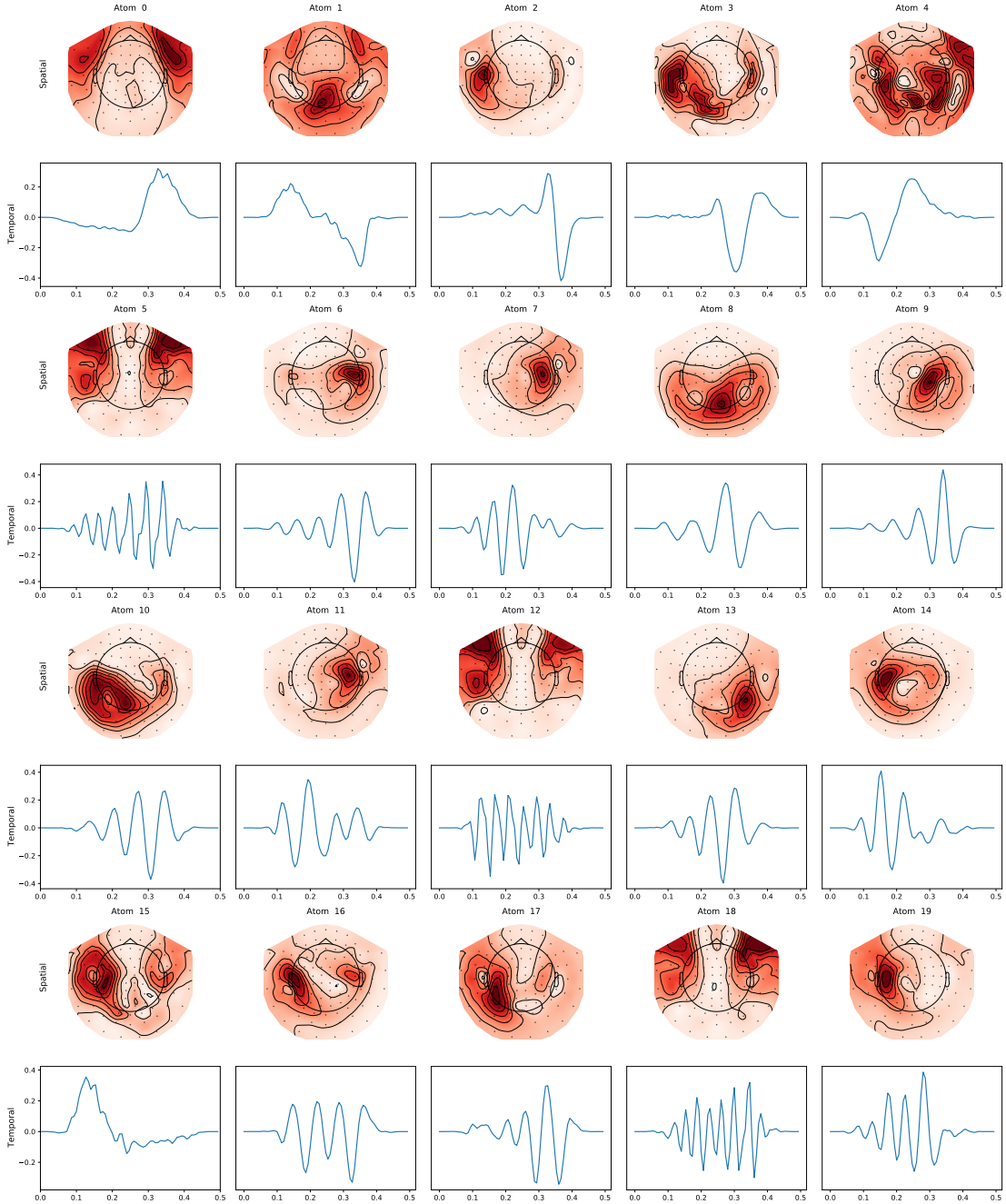


Figure 7.1.3: Spatial and temporal representation of the 20 atoms extracted from the subject CC723395, that obtained 20 clusters. One can observe the high variability in the atoms.

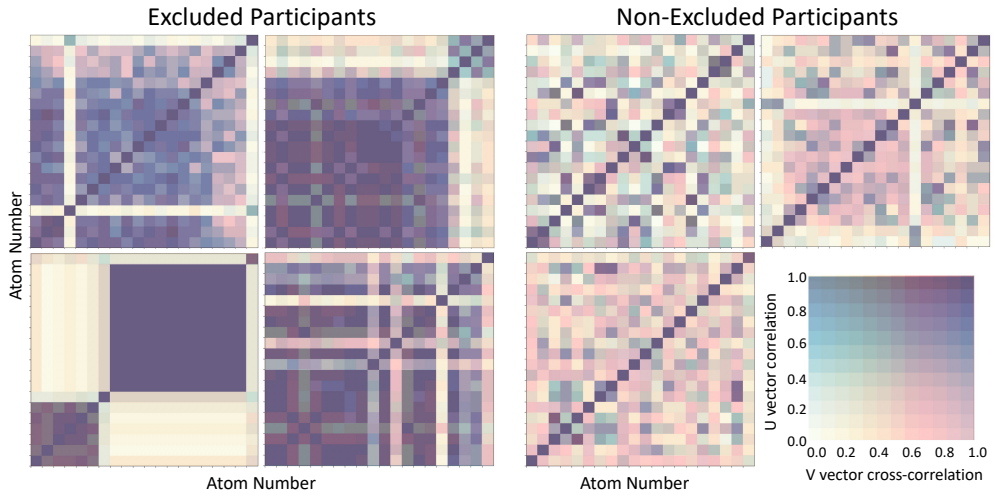


Figure 7.1.4: Two-variable correlation matrices from a representative sample of participants who were excluded (left) or not excluded (right) from the dataset. Each 20×20 cell correlation matrix shows data from a single participant, comparing each of the participant's 20 atoms to each other. The colour of the cells represents the magnitude of the spatiotemporal correlation between each pair of atoms. As indicated in the legend on the bottom right, u vector correlation is represented by a white to blue colour bar (low to high correlation), and the v vector cross-correlation is represented by a yellow to red colour bar. Atom pairs with a high correlation in both the u and v vector are thus indicated by dark purple colouration in the correlation matrices. It can be observed from the examples given here that excluded participants have numerous highly correlated atoms presenting as many dark purple cells in their matrices. This is highly dissimilar from other non-excluded participants, for whom the correlation matrices have few highly correlated atoms.

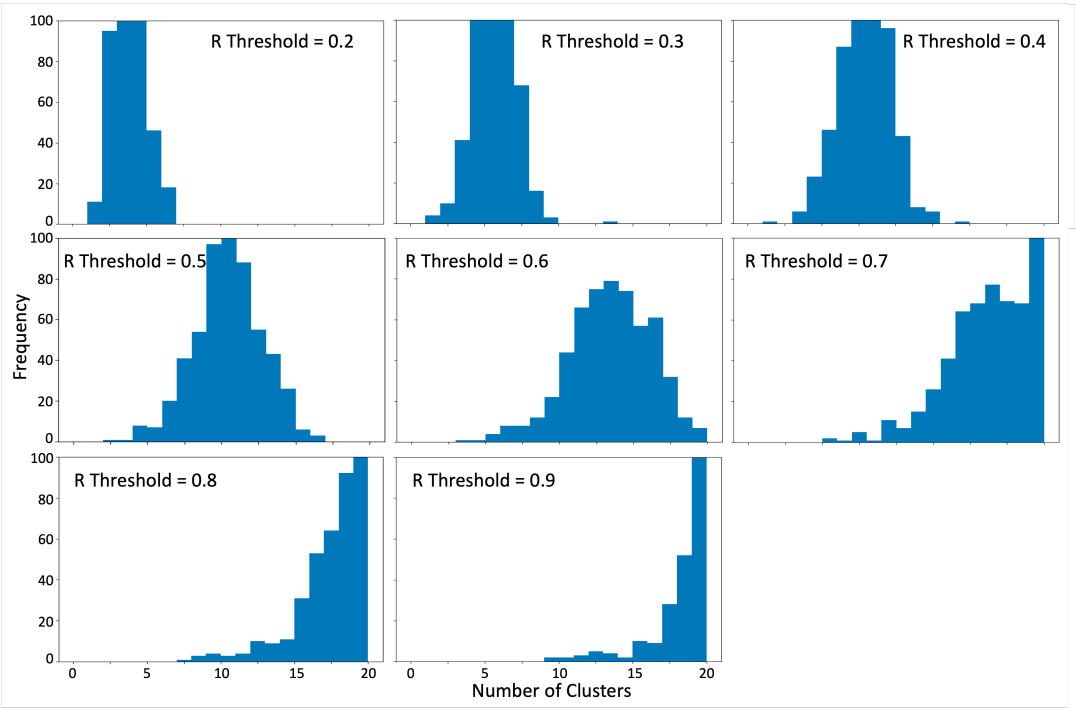


Figure 7.1.5: Histograms showing the distribution of the number of groups per participant as the correlation coefficient is increased for the single subject clustering methods. Thresholds of 0.2 to 0.5 show approximately normal distributions with a single mode that shifts to a higher number of clusters as the threshold increases. Thresholds of 0.6 and above begin to show a left-skewed distribution. Thresholds of 0.8 and 0.9 show abnormal behaviour in the tail of the distribution such that a second small peak emerges that likely represents those participants with abnormal data that should be excluded.

Table 7.1: Summary table of subjects excluded as they do not show enough variety in their extracted atoms as describe in section 16.

Subject ID	Age	Sex	Nb Clusters
CC110037	18	MALE	12
CC110182	18	FEMALE	12
CC121397	27	MALE	10
CC121428	26	FEMALE	8
CC220506	35	FEMALE	8
CC220610	32	FEMALE	10
CC221209	29	FEMALE	12
CC320850	47	FEMALE	11
CC322186	47	MALE	12
CC410325	54	FEMALE	11
CC420061	57	MALE	10
CC420167	51	FEMALE	9
CC420261	54	FEMALE	9
CC420348	57	FEMALE	7
CC420396	53	MALE	9
CC510043	58	MALE	12
CC520517	65	MALE	12
CC521040	63	FEMALE	11
CC610052	77	MALE	9
CC610292	72	FEMALE	12
CC610469	73	FEMALE	12
CC620129	75	MALE	12
CC620490	74	FEMALE	8
CC621642	73	MALE	11
CC720497	80	FEMALE	12

clusters of atoms between participants that had a lower level of similarity but could be presumed to be representative of similar neural processes. Therefore, it was predicted that the ρ used for the global clustering would be lower than that of the subject level clustering.

To select the optimal ρ for the global clustering, the threshold was once again varied from $\rho = 0.2$ to $\rho = 0.9$ and clustering was performed for each threshold. Because of the high computational time associated with clustering on a 538-person dataset, threshold selection was performed using 10 randomly selected 50-person datasets. The R value threshold ρ was selected based on the analysis of both qualitative and quantitative metrics. Firstly, the u and v vectors of a selection of atoms in each cluster were manually inspected to qualitatively assess the success of the clustering. Functional labels (*e.g.,* "occipital alpha", "left central beta", "eyeblink artifact", etc.) were assigned to atoms in each cluster to assess whether similar types of atoms were being appropriately clustered together for various R thresholds. This qualitative analysis suggested that $\rho = 0.4$ yielded the most appropriately grouped atom clusters. This selection was supported by quantitative metrics comparing the number of clusters detected at each R threshold to the number of "top clusters" (*i.e.,* common clusters, defined as clusters for which a minimum of 25 % of participants had atoms present in the cluster; see fig. 7.1.6). fig. 7.1.6 shows that $\rho = 0.4$ yielded the highest number of top clusters relative to the overall number of clusters, suggesting that the top clusters were most representative of the group. These findings were consistent across 10 random selections of data, suggesting that the choice of threshold was stable, and that the selection and ordering of subjects did not have a large effect on the overall results. $\rho = 0.4$ was thus selected as the optimal clustering threshold for the global analysis and was used in subsequent analysis of the entire dataset.

Global clustering on the entire dataset yielded 226 clusters of atoms, 11 of which were considered "top clusters" by the criteria that a minimum of 25 % of participants had atoms present in the cluster. The 25 % value was selected as a reasonable trade-off between maximizing the number of participants in the top clusters and ensuring that movement-related atoms of interest were being captured. fig. 7.1.7 shows representative atoms for each of the top clusters identified when the minimum percentage was varied to values of 50, 35, 25, 20, and 15 %. Minimum percentages above 25 % primarily captured eyeblink artifacts as these were the most stereotypical atoms in the population. The 25 % cutoff was the greatest cutoff that provided insight into movement-related atoms of interest (*e.g.,* contralateral sensorimotor beta) and was therefore selected as an appropriate threshold to define top clusters. To ensure adequate sample sizes for assessing cross-sectional ageing trends, only these top clusters are analysed in subsequent sections.

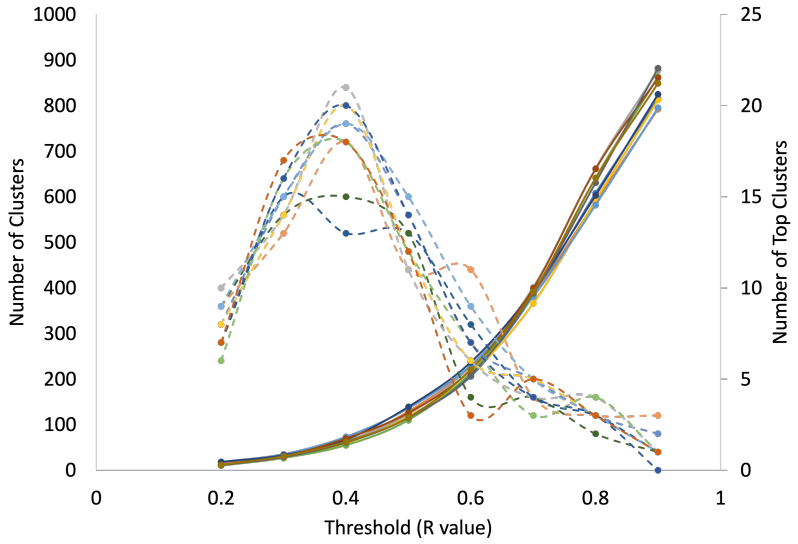


Figure 7.1.6: Plot showing the total number of clusters (solid lines) and the number of top clusters (dashed lines) identified with varying R value thresholds. Each colour of line is a different random sample of 50 participants from the Cam-CAN dataset. For all samples, the total number of clusters increases as the threshold increases, while the number of top clusters shows a clear peak at approximately $\rho = 0.4$.

7.1.5 Selection of task-related clusters

Each of the top clusters were then analyzed to reveal which were "task-related". Clusters were classified as task-related based on criteria related to the average characteristics of their component atoms, *i.e.,* the individual atoms that make up the cluster. In particular, atoms in a cluster had to have, on average, a focal source and a task-related reduction in activation. These criteria are based on previous findings that task-related transients – particularly the movement-related beta transients hypothesized to be present in this work – have a focal localization pattern [Power and Bardouille, 2021] and a marked reduction and rebound in their rate of occurrence with the onset and offset of task performance [Brady et al., 2020].

The focality of source was determined for each atom by calculating an equivalent current dipole (ECD) from the spatial representation of u. The dipole was then projected onto the participant's MRI to determine the anatomical position and orientation of the source. If the average goodness of fit for atoms in a given cluster exceeded 90 %, then the cluster source was considered to be focal.

The task-related reduction in activity criteria was assessed by segmenting the activation z vector into pre-task, task, and post-task intervals (where the task was an unimanual button press) and calculating the percent change in activation

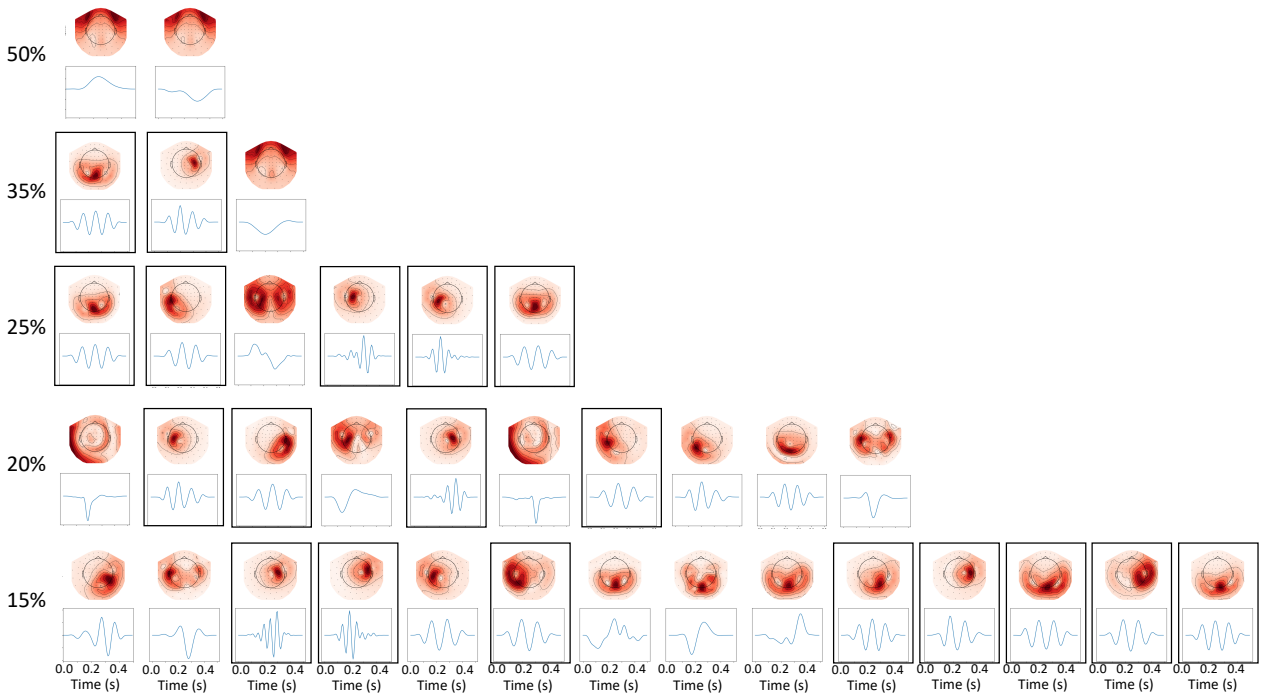


Figure 7.1.7: Representative atoms (spatial topographies and temporal waveforms) for the top clusters that are returned when the top cluster cutoff is varied. The top clusters identified at values of 50, 35, 25, 20, and 15 % are shown. Note that each percentage captures all the clusters in its row, as well as all of the above clusters. Black boxes indicate the clusters that were determined to be "task-related" (as defined in section 7.1.5). It can be observed that task-related clusters resemble common induced responses (*e.g.,* occipital alpha, and sensorimotor beta and mu), while non-task-related clusters resemble common artifacts (*e.g.,* eyeblinks and heartbeats), and evoked responses (*e.g.,* auditory and visual).

between intervals. The pre-movement, movement, and post-movement intervals were set to $-1.25\,\mathrm{s}$ to $-0.25\,\mathrm{s}$, $-0.25\,\mathrm{s}$ to $0.25\,\mathrm{s}$, and $0.25\,\mathrm{s}$ to $1.25\,\mathrm{s}$ relative to movement onset, respectively. Percent changes were then calculated between the movement and pre-movement, movement and post-movement, and post-movement and pre-movement intervals. More specifically, the percent change between period A and period B corresponds to the sum of the activations during period A divided by the sum of activations during period B, subtracted by 1. A positive value for this metric corresponds to a relative increase in the average number of activations in A compared to B while a negative value corresponds to a decrease.

If, on average, activations for an atom in a given cluster had a decrease from the pre-movement to movement interval that exceeded 10 % and an increase from the movement to post-movement interval that exceeded 10 %, the cluster was considered to meet the criteria for a task-related change in activity. Clusters were also assessed for an additional "rebound" criteria, to determine whether clusters exhibited an increase in activity above pre-movement levels, characteristic of movement-related beta transients [Brady et al., 2020]. If, on average, atoms in the cluster had a difference between post-movement and pre-movement that exceeded 10 %, the cluster was also considered to have a rebound component. The 10 % threshold used to compare task intervals was selected based on the magnitude of task-related changes in beta activity reported in previous analyses of the Cam-CAN dataset [Bardouille et al., 2019, Brady et al., 2020]. Using both average power [Bardouille et al., 2019], and burst-based [Brady et al., 2020] analyses of beta activity, task-related changes on the order of 10-30 % were observed. Therefore, to capture beta events along with other task-related event types for which the magnitude of this change is not defined, the threshold was set to the lower end of this range.

Of the 11 clusters included in this analysis, seven met the criteria for both focality and activity changes and were therefore classified as task-related. Of these, only one did not meet the additional rebound criteria.

7.1.6 Representative Atom Generation

For each of the task-related clusters identified in previous steps, a representative atom was generated using a modified version of the CDL process described in section 7.1.3. The creation of a representative atom allowed the cluster to be characterised and visualised, and provided a basis of comparison for atoms within and between clusters. For each of the N atoms (u_i, v_i) in a given cluster, representative MEG data X_i for the atom was recreated by convolving the activation vector $z_i \in \mathbb{R}^{P\times(T-L+1)}$ with the outer product of the $u_i \in \mathbb{R}^P$ and $v_i \in \mathbb{R}^L$ vectors to

yield MEG time course data (channels × time):

$$\forall i = 1, \ldots, N, X_i = u_i \cdot v_i^\top * z_i \in \mathbb{R}^{P \times T} \quad . \tag{7.1.4}$$

The MEG data from all component atoms in the cluster was then concatenated to create a single representative signal for the cluster: $X = [X_1 \ldots X_N] \in \mathbb{R}^{P \times NT}$. CDL was then applied to the concatenated signal to learn a single, 500 ms atom that would be representative of the most highly repetitive spatiotemporal signal in the cluster. An ECD was also computed for each representative atom using spatial representation of the u vector projected onto an average template brain.

The concatenated signal for each cluster was also used to generate a time-frequency representation (TFR) to show each atom clusters' frequency-specific behaviour relative to the movement task, averaged over tasks and participants. The concatenated signal was epoched based on stimulus onset as described in section 7.1.2. TFRs were then generated using a Morlet wavelet transform with a 500 ms wavelet, and were used as a basis of comparison between the atom types detected in this work and traditional band-limited power analyses.

7.1.7 Demographic analysis

The demographic characteristics of each task-related cluster were first investigated by creating histograms depicting the age and sex distribution of the participants whose atoms were included in the cluster. Participants were only counted once per cluster, regardless of how many atoms the participant had assigned to a given cluster. The demographic distributions for each cluster were then compared to the demographic distribution of the overall dataset (538 participants). The cluster and overall distributions were then quantitatively compared by conducting a Chi-squared test with a Bonferroni-corrected [Dunn, 1961] $\alpha = 0.007$ (as the result of 0.05 divided by 7 clusters) to determine whether the real cluster demographics were significantly different from what would be expected if clusters were created by random sampling. The results of the Chi-squared test provided information on the presence of age- or sex-related biases within clusters.

To further investigate demographic trends within each cluster, a series of regression analyses were conducted relating the component atoms' spatiotemporal characteristics to participant age. Atom characteristics including peak frequency of the power spectrum, activation sum (in the pre-movement and post-movement intervals; described in section 7.1.5), and dipole position and orientation were regressed with age using both linear and quadratic models. To determine the best fit model type, goodness of fit of each model (linear and quadratic) to the data was assessed using a Chi-squared test. An F-test was then employed to decide

the more appropriate model for each regression. A quadratic model was selected if $F > 6.635$, indicating 99 % confidence. Otherwise, a linear model was deemed most appropriate. The appropriate model was then plotted, and significant trends were assessed using Bonferroni-corrected $\alpha = 0.007$.

Additional regression analyses were also implemented to disambiguate the effects of burst rate and burst power on changes in the activation sum for each task interval and cluster. The distinction between these underlying factors is important because they are related to fundamentally different activity of the neural network (*e.g.,* burst rate is related to neural firing rates while burst power is related to neural network size). Therefore, the independent analysis of burst rate and burst power was performed to determine whether burst rate (frequency) or burst power (intensity) underlies age-related changes in atom activation sum, for each cluster.

Burst rate was defined as the number of non-zero activation values in the interval of interest divided by the length of the interval. This was calculated for each component atom during each task interval (pre-movement and post-movement) and regressed against age using the linear and quadratic models as described above.

Burst power was defined as the magnitude of the non-zero activations and was assessed as a distribution of values for each atom in each task interval. The role of burst power in the activation sum trends was investigated by calculating the distribution of activation values for each atom and assessing the shift in the distribution with age. A Gaussian function was fit to the distribution of activation values for each atom, and the μ (mean) and σ (standard deviation) values for each distribution were regressed against age to assess for age-related changes in the distribution.

In addition to the atoms' characteristics, the relationship between participant age and the correlation of their atoms to the cluster's representative atom was also assessed. A significant age-related change in correlation would indicate that similarity to the representative atom changed with age, and would provide particular insight into whether atom characteristics may be converging towards or deviating from the mean with age. The correlation of the u vectors and the maximum cross-correlation of the v vectors were calculated between each atom in a given cluster and the cluster's representative atom to provide a measure of atom similarity to the representative atom for the group. As above, linear and quadratic regression analyses relating the correlation values of each atom to the age of the participant to whom the atom belonged were conducted. All linear and quadratic regression analyses were assessed for significance with a Bonferroni-corrected $\alpha = 0.007$. All results are reported to two decimal places.

7.1.8 Supplementary analysis

While large cluster sizes were necessary for the current work to assess cross-sectional ageing trends, other applications of CDL and the associated clustering methods may not require this. An alternative method for presenting CDL cluster results is thus described below. This approach may be preferred for smaller datasets or in cases where you wish to appreciate more of the variability between atom clusters.

Following global clustering as described in section 7.1.4, all 226 detected clusters underwent task-based filtering as described in section 7.1.5, resulting in a total of 79 task-related clusters. In this case no top cluster criteria was imposed to dictate a minimum cluster size. A representative atom was then generated for each of the 79 task-related clusters using the procedure described in section 7.1.6. The representative atoms for each of the task-related clusters then underwent an additional round of clustering, using the methods described in section 7.1.4 to roughly group clusters into sets based on similarity. This resulted in several sets of clusters each associated with a different class of brain activity (*e.g.,* "left central beta", "occipital alpha", etc.) We could then select sets of interest and analyse all clusters within those sets to appreciate additional inter-subject variability in the atoms. The results from 3 sets identified by this method (right-central beta, left-central beta, and occipito-temporal alpha) are shown in fig. 7.1.8.

Here, the frequency, activation strength, and age distribution of each cluster can be compared within a set. The age distribution in the Cam-CAN dataset is approximately flat, therefore investigating deviations from the flat age distribution provides meaningful information about age dynamics within clusters. This type of analysis allows for between-cluster comparison and can provide insight into the characteristics of participants who tend to have certain variations of atoms.

7.2 Results

The global clustering methods described above resulted in seven task-related clusters of atoms across participants. fig. 7.2.1 shows the spatial topographies and temporal waveforms for each of the representative atoms for the task-related clusters. Of the seven representative atoms, four had waveforms resembling alpha waves (*i.e.,* 8 Hz to 12 Hz sinusoid), one had a waveform resembling a mu wave (*i.e.,* complex waveform with a peak frequency of 8 Hz to 12 Hz), and two had waveforms resembling beta waves (*i.e.,* complex waveform with a peak frequency of 15 Hz to 30 Hz). Three of the clusters characterized by alpha-type waveforms had topographies resembling occipital activation. These clusters were distinguished

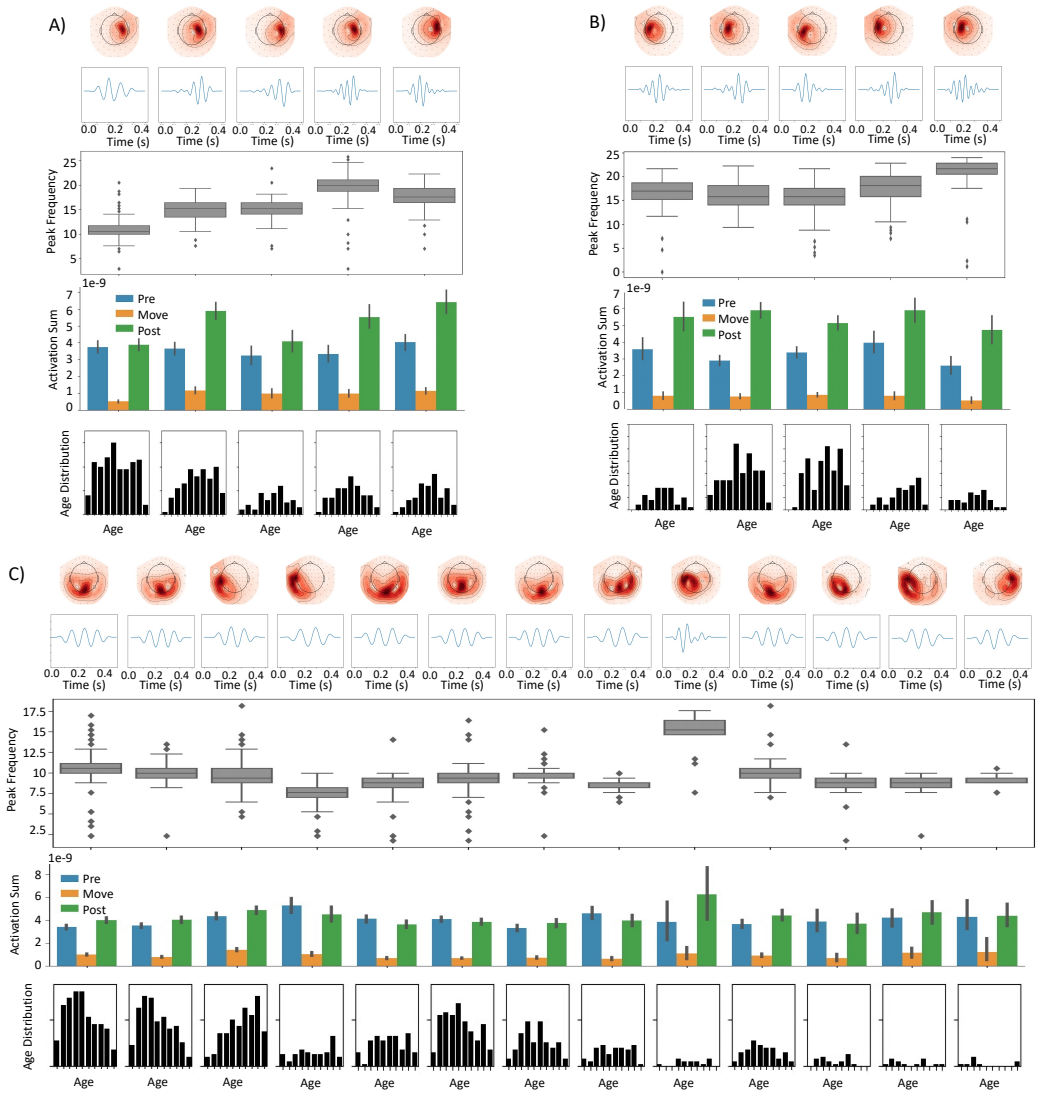


Figure 7.1.8: Cluster sets identified resembling (A) right central beta, (B) left central beta, and (C) occipito-temporal alpha activity. For each set, representative atoms (spatial topographies and temporal waveforms) are shown for each cluster within the set. Box plots show the distribution of frequencies of the atoms composing each cluster. Bar plots show the summed activation in the pre-movement (blue), movement (orange), and post-movement (green) phases for each cluster. Finally, to faciliate demographic comparisons, the age distribution of each cluster is shown.

spatially by their tendency to activate sensors either in the left occipital lobe (LO_alpha), right occipital lobe (RO_alpha) or more anteriorly and medially in the medial occipitoparietal region (MOP_alpha). The fourth alpha-type waveform had an associated spatial topography resembling left temporal lobe activity (LT_alpha). The cluster characterized by a mu waveform had a spatial topography resembling right central (sensorimotor) activation (RC_mu). Finally, both clusters characterized by beta waveforms had topographies suggesting left central (sensorimotor) activity. One such cluster showed peak activity just anterior to the center of the topography, near the primary motor area (LPreC_beta), and the other showed peak activity just posterior to the center of the topography, near the primary somatosensory area (LPostC_beta).

The distribution of the peak frequencies shown in fig. 7.2.1 suggests that there is more variability in the frequency content of the atoms making up the beta-type clusters than the mu- or alpha-type clusters. In terms of activation sums, all task-related clusters had, on average, a decrease in activation from the pre-movement to movement time intervals, and a subsequent increase in activation from movement to post-movement, as this was one of the criteria required to classify the cluster as "task-related". However, it should be noted that for all clusters, there was a large amount of variability in the level of activation of the component clusters during each task interval, as indicated by the error bars in fig. 7.2.1. In addition, 6 of the task-related clusters[4] also had a rebound component, meaning that there was an increase in activation from pre-movement to post-movement intervals. Notably, on average, the beta-type clusters have the largest difference between post-movement and pre-movement activation. This characteristic "rebound" of activation is in line with existing literature that notes a post-movement rebound of beta power, surpassing baseline – *i.e.,* pre-movement – levels, in primary sensorimotor areas contralateral to the movement.

For each component atom, and each representative atom, an equivalent current dipole (ECD) was calculated to infer the approximate source of the atom. The representative atom dipoles (fit to an average template brain) are shown in fig. 7.2.2, and table 7.2 presents the positions, orientations, and goodness of fit values for each of the representative atom ECDs. In general, the ECDs for the representative atom localize to the expected regions based on their spatial topographies.

Average TFRs were created for each atom cluster to allow the atoms' behaviour to be compared to traditional average spectral power analyses (see fig. 7.2.3). In an average power analysis of an audio-visual cued simple movement task, we would expect to see suppression of occipital alpha and sensorimotor mu and beta activity, and a rebound in sensorimotor beta activity. As shown in fig. 7.2.3, each atom type shows a distinct reduction in activity near the onset of the movement (at time = 0 s)

[4] LO_alpha, RO_alpha, RC_mu, LT_alpha, LPreC_beta, and LPostC_beta.

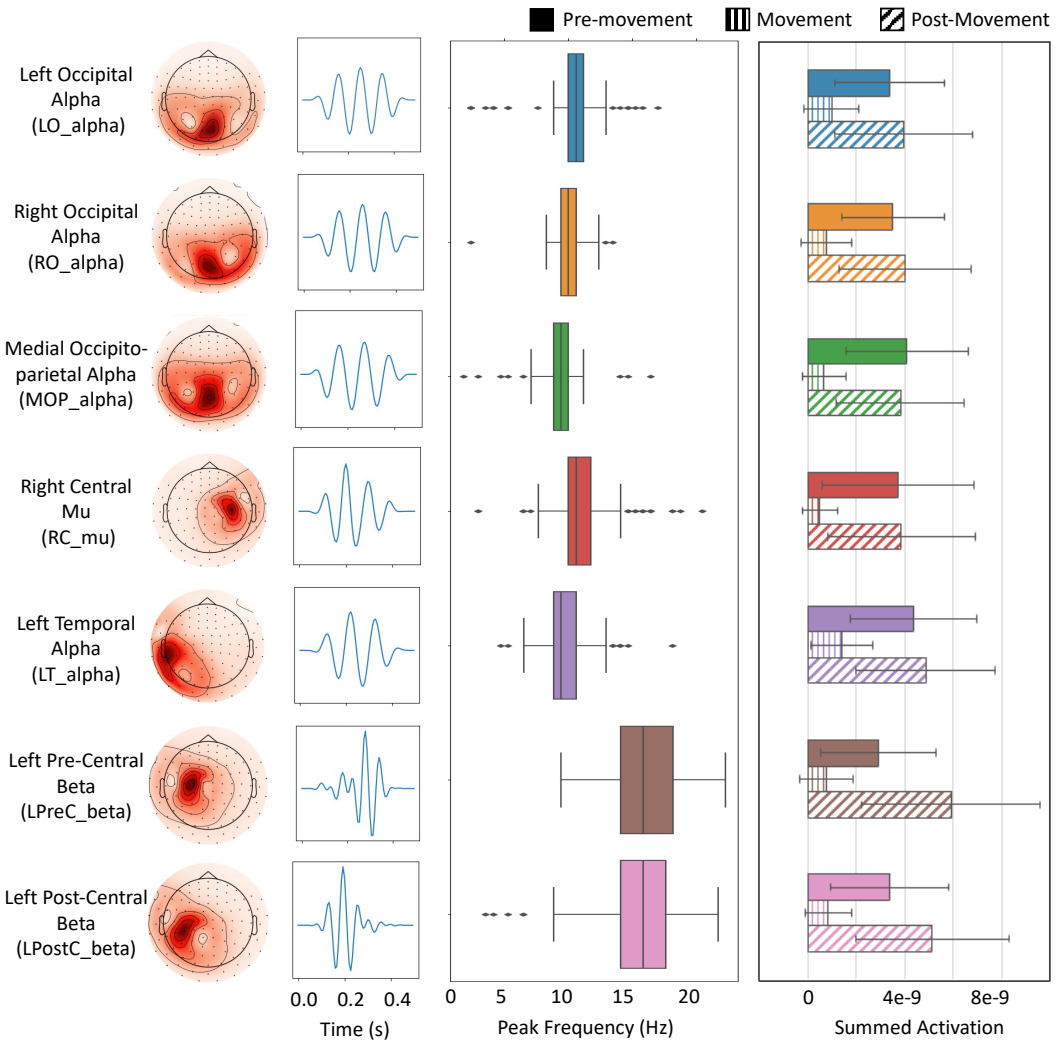


Figure 7.2.1: The seven task-related clusters identified in this work. (Left) The shift-invariant spatial and temporal vectors of the representative atoms created for each of the seven clusters identified. Clusters are given a functional label based on the spatial and temporal representation of the atom. (Center) Box plots depicting the distribution of peak frequencies of the component atoms for each cluster. (Right) The mean and standard deviation (error bars) of the summed activation during the pre-movement (solid bar), movement (vertical striped bar), and post-movement (diagonal striped bar) intervals for the component atoms of each of the seven clusters.

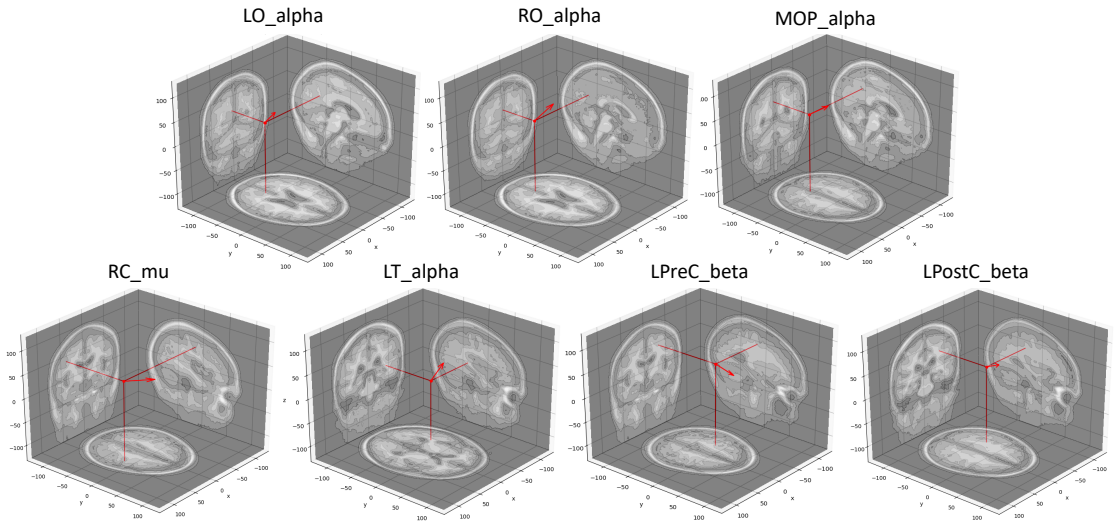


Figure 7.2.2: The dipole fits for each of the seven representative atoms. Axial, sagittal, and coronal slices are shown for each fit, with the red lines indicating the position of the dipole in each plane and red arrows indicating the orientation of the dipole. Representative atom dipoles were fit to an average template brain.

Table 7.2: Summary of the attributes of each cluster.
Nb part.: Number of unique participants; ECD pos.: equivalent current dipole position; GoF: Goodness of Fit

Cluster	**Nb atoms**	**Nb part.**	**Peak freq.** (Hz)	**ECD Pos.** (mm)	**ECD Orientation Unit Vector**	**GoF** (%)
LO_alpha	366	210	10.5	(0.52, -2.7, 6.6)	(0.23, 0.67, 0.71)	90.5
RO_alpha	311	183	10.0	(-.017, -3.1, 6.6)	(-0.26, 0.60, 0.76)	86.7
RC_mu	305	194	10.5	(4.1, 2.1, 8.3)	(-0.40, 0.90, 0.18)	98.3
LT_mu	230	168	9.4	(-4.0, -0.51, 4.2)	(-0.065, 0.50, 0.86)	96.2
LPrecC_beta	243	146	17.6	(-3.0, 0.20, 8.7)	(0.30, 0.95, -0.072)	98.7
LPostC_beta	219	144	16.4	(-3.5, 0.26, 7.7)	(0.34, 0.82, 0.47)	99.1
OP_alpha	324	171	9.4	(0.19, -1.9, 8.0)	(0.00011, 0.81, 0.58)	79.1

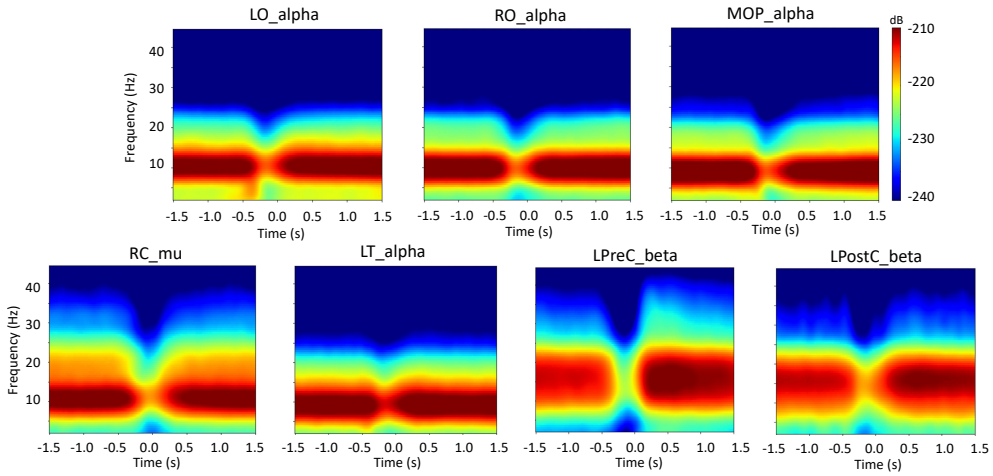


Figure 7.2.3: Time-frequency representations generated from the concatenated signal for each of the 7 task-related atom clusters. The signal is averaged across epochs and atoms, and activity is shown relative to movement onset (time=0 s). Color represents the magnitude of the activity in dB.

and an increase to baseline activity following the task. In addition, clear rebound behaviour is evident in the beta-type atoms, reflecting the post-movement beta rebound seen in traditional analyses, *e.g.,* in Bardouille et al. [2019]. This suggests that CDL reconstructs the expected brain responses. Finally, the right central mu atoms show a frequency profile that is distinctly different from the occipital and temporal alpha atoms, providing further confidence in the relationship between the CDL-detected waveforms and traditional analyses.

The demographic composition of each cluster was assessed by comparing the age and sex distributions of the cluster to the distribution generated from the overall dataset using a Chi-squared test. Clusters LT_alpha and LPostC_beta were found to have an age distribution that was significantly different from that of the overall dataset, with Chi-squared values of 27.47 ($p = 1.2 \times 10^{-3}$) and 36.29 ($p = 3.5 \times 10^{-5}$), respectively. Both clusters had age distributions that were skewed towards older participants. Histograms showing the distributions of the two significant clusters compared to the overall dataset are presented in fig. 7.2.4. No other clusters had age or sex distributions that were significantly different from that of the overall dataset.

All clusters were examined for age-related changes in their component atom characteristics including peak frequency, activation sum (in the pre-movement and post-movement intervals), dipole position and orientation, and correlation of the u and v vectors to the mean atom's vectors were regressed with age for each task-related cluster. For each cluster, for pre-movement and post-movement intervals, age-related trends in activation sum and burst rate are presented in fig. 7.2.5,

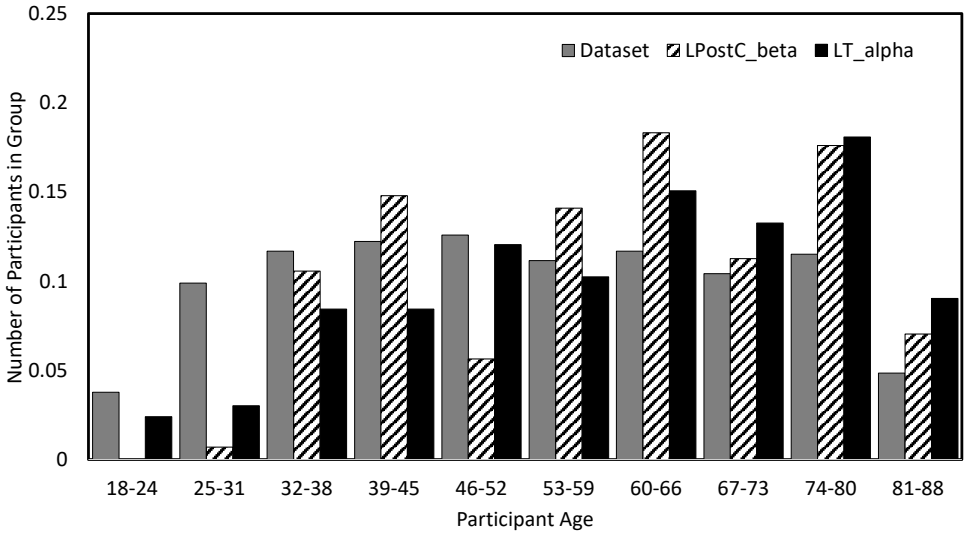


Figure 7.2.4: The age distribution of participants in the LPostC_beta (striped) and LT_alpha (black) clusters compared to the age distribution of the overall dataset (grey).

and fig. 7.2.6 presents the age-related trends for the distribution's parameters of activation values for each atom. The corrected p-values and RMSE values for all trends can be found in table 7.3.

In the pre-movement interval, regression results of the activation sum and the burst rate are similar, with positive linear trends evident in most clusters[5]. All significant effects in the post-movement interval matched those found in the activation sum regression except for clusters LO_alpha and RO_alpha which were better fit with a positive linear model in the burst rate regression rather than the positive quadratic model fit in the activation sum regression. Despite the discrepancy in the best model fit, LO_alpha, and RO_alpha show similar increasing trajectories for both activation sum and burst rate. These results suggest that burst rate can account for most of the trends observed in activation sum with age.

No significant age-related changes in the pre-movement interval were found in the distribution mean. The lack of effects in the mean value suggested that the magnitude of burst power does not change significantly with age. Only one significant effect of the distribution standard deviation was found for cluster LPreC_beta, and its standard deviation increased linearly with age, suggesting that older participants had more variable burst power values than younger participants.

[5]LO_alpha, RO_alpha, MOP_alpha, LPreC_beta, and LPostC_beta.

Table 7.3: p-values and RMSE values for age-related regression of the summed activation vector, activation burst rate, and mu and sigma of the activation distribution. Data is shown for the pre-movement and post-movement activation intervals. Significant p-values are indicated by an asterisk.

Cluster	Pre-Movement							
	Sum		Rate		Mu		Sigma	
	p-value	RMSE	p-value	RMSE	p-value	RMSE	p-value	RMSE
LO_alpha	5.09e-7*	2.2e-9	2.77e-8*	4.6e-1	2.66e-2	2.9e-11	7.63e-2	1.6e-11
RO_alpha	2.76e-7*	2.0e-9	2.14e-6*	4.5e-1	8.35e-2	3.1e-11	1.63e-1	1.6e-11
MOP_alpha	1.10e-8*	2.4e-9	2.54e-7*	4.5e-1	4.28e-1	3.4e-11	6.05e-2	1.8e-11
RC_mu	8.56e-3	3.1e-9	2.19e-2	6.1e-1	9.63e-1	3.8e-11	8.42e-1	1.7e-11
LT_alpha	9.09e-3	2.6e-9	2.41e-3*	5.3e-1	3.91e-1	2.8e-11	4.77e-1	1.4e-11
LPreC_beta	2.53e-5*	2.3e-9	1.36e-3*	5.4e-1	8.87e-3	4.1e-11	4.05e-5*	1.4e-11
LPostC_beta	4.81e-5*	2.3e-9	4.80e-4*	6.2e-1	7.01e-1	3.4e-11	5.68e-1	1.3e-11
	Post-Movement							
	Sum		Rate		Mu		Sigma	
	p-value	RMSE	p-value	RMSE	p-value	RMSE	p-value	RMSE
LO_alpha	1.71e-7*	2.7e-9	8.30e-7*	5.9e-1	2.18e-1	3.0e-11	1.65e-1	1.6e-11
RO_alpha	1.13e-9*	2.6e-9	6.07e-8*	5.4e-1	5.55e-1	3.2e-11	3.33e-1	1.6e-11
MOP_alpha	4.36e-5*	2.6e-9	3.97e-5*	4.8e-1	7.85e-1	3.3e-11	5.06e-1	1.8e-11
RC_mu	6.23e-3*	3.0e-9	3.93e-3*	6.1e-1	5.97e-1	3.7e-11	4.61e-1	1.6e-11
LT_alpha	3.44e-3*	2.8e-9	1.54e-3*	6.0e-1	2.26e-2	2.8e-11	8.08e-1	1.3e-11
LPreC_beta	9.87e-4*	3.6e-9	3.34e-5*	8.7e-1	1.39e-3*	4.4e-11	2.26e-1	1.5e-11
LPostC_beta	3.33e-1	3.2e-9	2.61e-1	8.2e-1	3.19e-2	3.2e-11	1.54e-1	1.3e-11

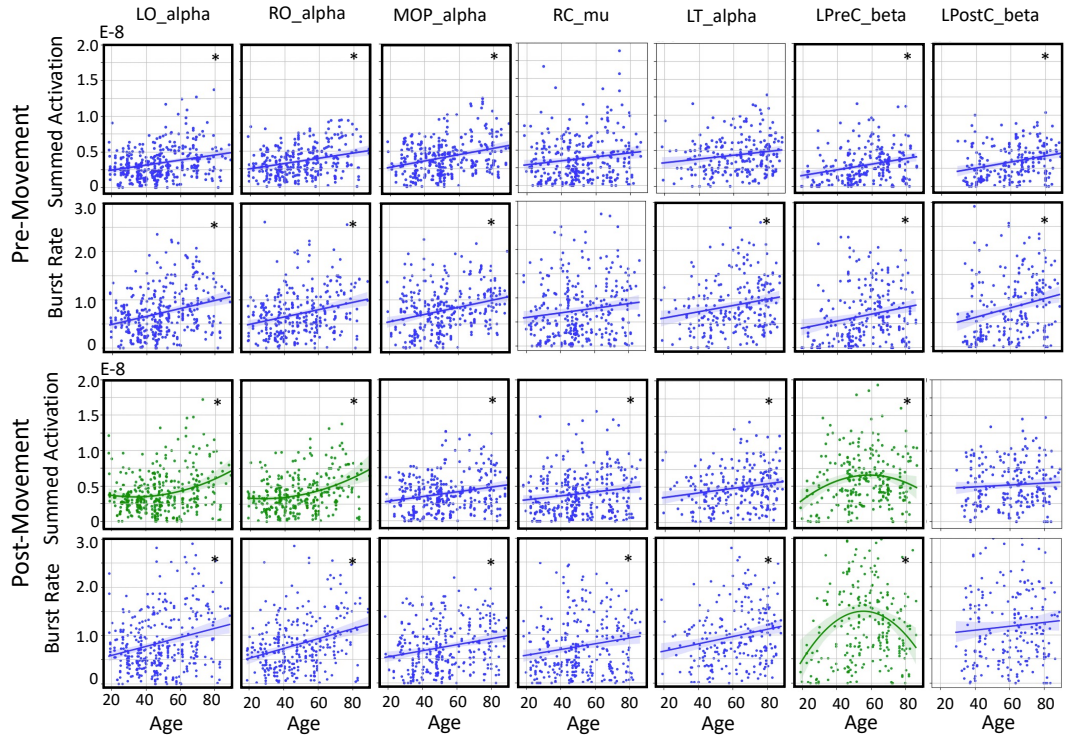


Figure 7.2.5: Results of linear and quadratic regression of summed activation and burst rate with age during pre-movement and post-movement. Summed activation and burst rate during each interval are plotted against participant age for the component atoms of each cluster. Blue plots represent those that were modelled by a linear fit, and green plots were modelled by a quadratic fit. Asterisks indicate clusters and intervals for which the best fit regression was significant (Bonferroni corrected alpha < 0.007).

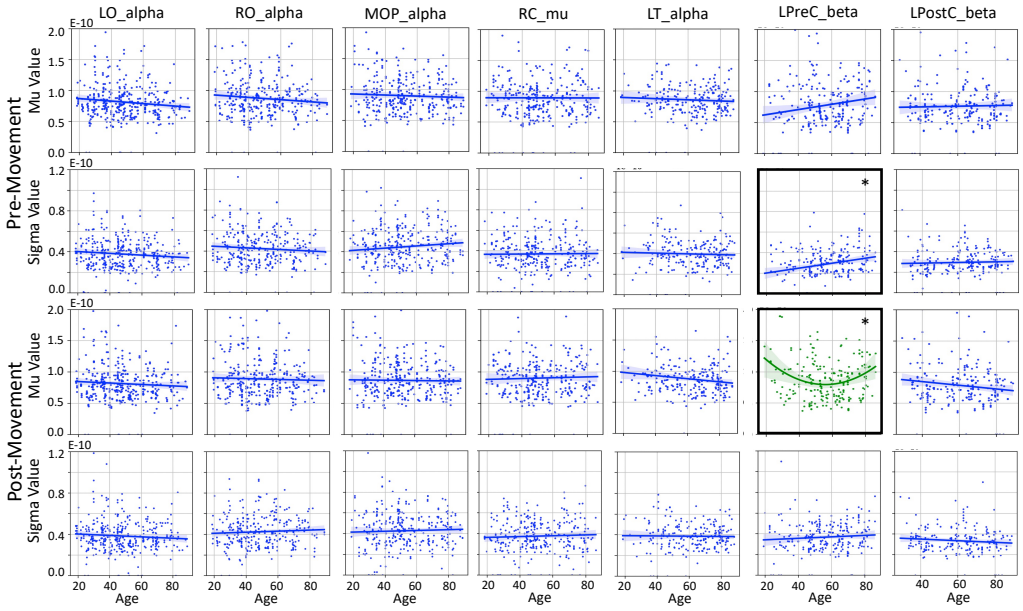


Figure 7.2.6: The mean value (mu) and the standard deviation (sigma) of the distribution of activation values for each atom as a function of age for the pre-movement interval (top) and the post-movement interval (bottom). Mu and Sigma values were regressed against age. Blue plots represent those that were modelled by a linear fit and green plots were modelled by a quadratic fit. Asterisks indicate clusters and intervals for which the best fit regression was significant (Bonferroni corrected alpha < 0.007).

Similarly, in the post-movement interval, there was a single significant age-related quadratic effect in the mean of the distribution for cluster LPreC_beta. The effect was such that the atoms from the youngest and oldest participants had a distribution that was shifted towards larger activation values, while those atoms belonging to middle-aged participants tended to come from a distribution with a lower mean activation value. This suggests that, for cluster LPreC_beta, burst power is highest in young and old participants. This contrasts the results of the activation sum regression for cluster LPreC_beta which showed that young and old participants had a reduced activation sum compared to their middle-aged counterparts. These results suggest that burst power plays a lesser role in the overall activation sum trends, and in some cases may even contradict the overall effects driven by burst rate.

In addition to the dominant age-related effects demonstrated in activation sum, several spurious age-related trends in other atom characteristics were found. fig. 7.2.7 shows plots of the age-related linear and quadratic effects of peak frequency, y position of the dipole, and correlation of the u and v vector to the mean atom, and highlights the clusters for which these effects were significant.

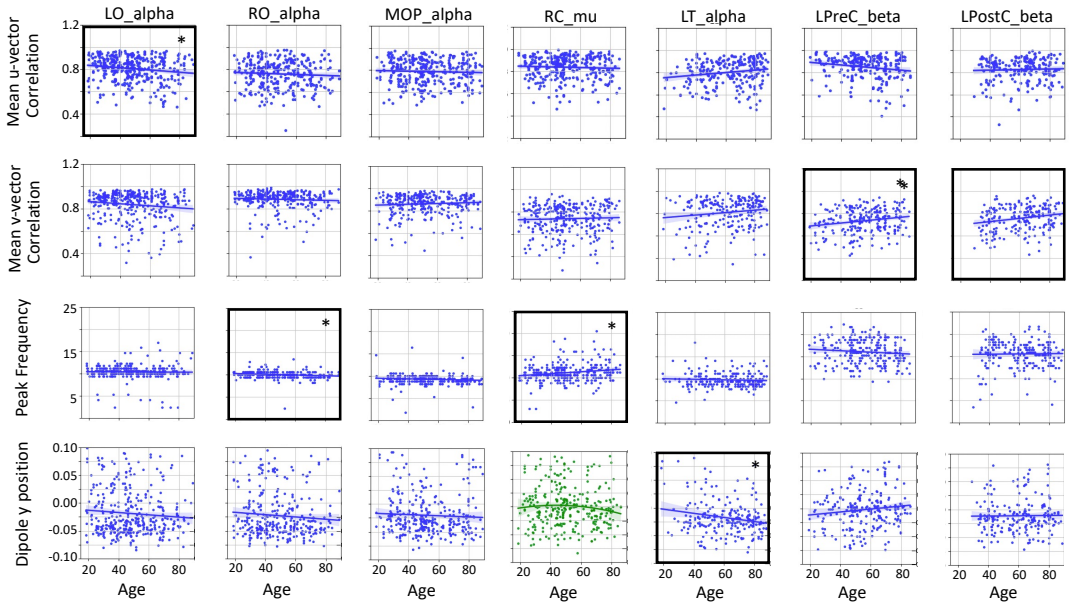


Figure 7.2.7: Results of linear and quadratic regression of several burst characteristics with age. Blue plots represent those that were modelled by a linear fit and green plots were modelled by a quadratic fit. Asterisks indicate clusters and intervals for which the best fit regression was significant (Bonferroni corrected alpha < 0.007).

There were no significant age-related effects found in the x or z position or x, y, or z orientation of the dipole for any cluster. The correlation of the u and v vector of the component atoms in each cluster to the cluster's mean atom were calculated and regressed with age to assess whether atoms were converging to or diverging from the mean atom with age. The correlation of the u vector (spatial topography) to the mean atom showed a significant negative linear trend with age for cluster LO_alpha ($p = 0.0035$). This indicates that the spatial topography for this cluster becomes increasingly variable, or dissimilar to the mean atom with age. The correlation of the v vector (temporal waveform) to the mean atom showed a significant positive linear trend with age for clusters LPreC_beta ($p = 0.0029$) and LPostC_beta ($p = 0.0021$) indicating that these clusters showed a convergence in their temporal waveform with age. In addition, peak frequency of the component atoms showed a significant negative linear trend in cluster RO_alpha with age ($p = 0.0016$) and a significant positive linear trend in cluster RC_mu with age ($p = 0.0014$). Finally, the y position of the dipole showed a significant negative linear trend with age in cluster LT_alpha ($p = 0.0018$), indicating that there may be a posterior shift in the dipole position of cluster LT_alpha with age.

A summary of all age-related effects observed in each cluster is shown in table 7.4. The variability in the effects observed in different clusters suggests that all clusters are distinct and have unique individual relationships to the ageing process.

Table 7.4: Summary of the age-related effects for each cluster

Cluster	Age-related effects
Left Occipital Alpha	Linear increase in pre-movement activation Positive quadratic effect in post-movement activation Linear increase in pre-movement and post-movement burst rate Linear decrease in u vector correlation to mean atom
Right Occipital Alpha	Linear increase in pre-movement activation Positive quadratic effect in movement and post-movement activation Linear increase in pre-movement, movement and post-movement burst rate Linear decrease in peak frequency
Medial Occipitoparietal Alpha	Linear increase in pre-movement and post-movement activation Positive quadratic effect in movement activation Linear increase in pre-movement and post-movement burst rate
Right Central Mu	Linear increase in movement and post-movement activation Linear increase in post-movement burst rate Linear increase in peak frequency
Left Temporal Alpha	Age distribution skewed to older adults Linear increase in post-movement activation Linear increase in pre-movement and post-movement burst rate Linear decrease in y position of dipole (posterior shift)
Left Pre-Central Beta	Linear increase in pre-movement activation Negative quadratic effect in post-movement activation Linear increase in pre-movement burst rate Negative quadratic effect in post-movement burst rate Linear increase in pre-movement σ property of activation distribution Positive quadratic effect in post-movement μ property of activation distribution Linear increase in v vector correlation to mean atom
Left Post-Central Beta	Age distribution skewed to older adults Linear increase in pre-movement activation Linear increase in pre-movement burst rate Linear increase in v vector correlation to mean atom

7.3 Conclusion

In this work, we employed Convolutional Dictionary Learning (CDL) to successfully delineate spatiotemporal atoms, thereby shedding light on age-related variations in neural network firing rates as evidenced through M/EEG data. Notably, occipital alpha-type clusters consistently showed an increase in activation sum with age across different movement intervals. This accords with existing literature indicating age-associated dominant alpha peaks [Chiang et al., 2011] and emphasizes that age-related changes in transient bursts are predominantly driven by alterations in neuronal firing rate. Our results on beta and mu transient activation trends, however, partly contradict the prevailing hypotheses and extant literature. Contrary to expectations, we observed no significant anterior shift in dipole positions or a consistent decrease in burst frequency with age [Power and Bardouille, 2021]. This necessitates a re-evaluation of the established paradigms and may also be attributed to potential misclassifications arising from classic frequency band limits.

The utilization of CDL combined with unsupervised clustering presents several advantages over traditional burst detection methodologies like amplitude thresholding or other data-driven approaches such as Hidden Markov Modelling (HMM) [Quinn et al., 2021]. It permits the extraction of various transient burst types without stringent assumptions about their frequency composition, waveform shape, or spatial distribution. This makes CDL a robust analytical tool for both task-related and resting-state applications, as well as for burst-based neurofeedback interventions [Ossadtchi et al., 2017, Karvat et al., 2020].

Future work may extend this approach towards the formulation of a standard dictionary of transient bursts, capitalizing on large datasets like the Cam-CAN repository [Tal et al., 2020]. The potential for CDL in investigating the emergent field of traveling cortical waves [Hindriks et al., 2014, Muller et al., 2018] also represents an intriguing avenue for future exploration.

Nonetheless, the current work is not devoid of limitations. The bandpass filtering between 2 Hz and 45 Hz restricts the analysis to mu and beta frequency bands, thereby overlooking other transient activities like gamma bursts. The dependency on user-defined hyperparameters in the clustering algorithms also necessitates further refinement. Additionally, the pre-defined atom window size in CDL could potentially result in exclusion of some types of transient bursts, highlighting that even a data-driven approach like CDL comes with its own set of assumptions and constraints.

In conclusion, while CDL offers a robust and nuanced method for the identification of spatiotemporal transient bursts, its applicability should be carefully

weighed against the research question at hand and the specific demands of the dataset being analyzed. It serves as a valuable alternative to existing methods when traditional assumptions about waveform shape or frequency are not met or when a data-driven approach is required.

Conclusion and perspectives

This thesis aimed to address key challenges in the analysis of neural activity, focusing on advanced signal processing and statistical modeling. The work was centered on two primary objectives: enhancing temporal modeling in neuroscientific data and improving computational scalability for large-scale data analysis. While the contributions of this work have been substantial, they also open avenues for future exploration and development in these two key areas.

Temporal modeling The thesis introduced novel point process models for interpreting task-specific neural activities, DriPP and FaDIn. Both contribute to a unified analytical framework for M/EEG data, offering robust methods to identify and interpret temporal patterns influenced by external stimuli. Indeed, by allowing multivariate neural signals to be represented as event-based, Convolutional Dictionary Learning (CDL) opens the doors of point processes to M/EEG data. Therefore, these models present a new approach to modeling prototypical neural waveforms, going beyond conventional analyses based on epoch averaging and time-frequency planes.

Future work could extend these models, particularly FaDIn, by incorporating *marked Hawkes processes* to include amplitude values of atoms' activations. This advancement would enable the analysis to transcend the binary framework of activation by introducing a factor $f(m_k)$ into the intensity function, where m_k represents the mark – *i.e.,* the amplitude value – associated with each activation t_k. Such an enhancement is particularly beneficial in contexts like electrocardiogram (ECG) analysis, where it could facilitate the extraction of critical metrics, such as cardiac frequency and variability, essential during surgical procedures. Furthermore, applying these models to EEG data recorded during surgeries with global anesthesia presents an opportunity to identify patterns associated with undesirable neural responses linked to post-operation complications. Detecting these early warning signs would empower surgeons to adjust anesthetic drug levels in real-time, enhancing patient safety and surgical outcomes.

Computational scalability and robustness in M/EEG data The second axis was aimed at processing large M/EEG data sets, by making CDL faster and extending its results to population analyses. The proposed Stochastic Robust Windowing CDL (WinCDL) algorithm addressed challenges with artifact-laden signals and the needs of population-level studies. This novel implementation of CDL tailored for large-scale M/EEG data improved the efficiency of data processing, enabling the handling of large datasets, eventually corrupted with artifacts, more effectively. In addition, a new aggregation method for multiple single-subject CDL outputs was proposed, allowing the discovery of age-related insights at a population level.

Looking towards future work, there are promising avenues to explore. For instance, enhancing robustness with inline outlier detection could involve developing methods to adjust quantile estimation during the learning process. Techniques such as the Randomized Update-based Multiplicative Incremental Quantile Estimator (RUMIQE), introduced by Yazidi and Hammer [2017], offer a potential pathway. Another key area of future research is the direct application of WinCDL to a group of subjects to derive a common dictionary of recurrent patterns. This approach, however, must address the challenge posed by the variability in brain morphology across individuals, which implies that only the temporal patterns, not the spatial components, can be consistently shared across a population. Such advancements could significantly refine the CDL process, making it even more adaptable and effective for broader neuroscientific investigations.

Bibliography

M. Abeles. Revealing instances of coordination among multiple cortical areas. *Biological Cybernetics*, 108(5):665–675, 2014. doi: https://doi.org/10.1007/s00422-013-0574-2.

Massil Achab. *Learning from Sequences with Point Processes*. PhD thesis, Université Paris-Saclay, 2017.

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, and Praneeth Netrapalli. Learning sparsely used overcomplete dictionaries via alternating minimization. *SIAM Journal on Optimization*, 26(4):2775–2799, 2016.

Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

Cédric Allain, Alexandre Gramfort, and Thomas Moreau. DriPP: Driven point processes to model stimuli induced patterns in M/EEG signals. *International Conference on Learning Representations*, 2022.

Cédric Allain, Benoît Malezieux, and Thomas Moreau. Fast and robust convolutional dictionary learning for large m/eeg data. *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. Under review.

Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005, 2015.

Emmanuel Bacry, Martin Bompaire, Philip Deegan, Stéphane Gaïffas, and Søren V. Poulsen. Tick: A python library for statistical learning, with an emphasis on hawkes processes and time-dependent models. *Journal of Machine Learning Research*, 18(1):7937–7941, 2017a.

Emmanuel Bacry, Martin Bompaire, Stéphane Gaïffas, and Soren Poulsen. Tick: a python library for statistical learning, with a particular emphasis on time-dependent modelling. *arXiv preprint arXiv:1707.03003*, 2017b.

Emmanuel Bacry, Martin Bompaire, Stéphane Gaïffas, and Jean-Francois Muzy. Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research*, 21(50):1–32, 2020.

S. Baillet. Magnetoencephalography for brain electrophysiology and imaging. *Nature Neuroscience*, 20:327 EP –, 02 2017.

Adam P. Baker, Matthew J. Brookes, Iead A. Rezek, Stephen M. Smith, Timothy Behrens, Penny J. Probert Smith, and Mark Woolrich. Fast transient networks in spontaneous human brain activity. *eLife*, 3:e01867, 2014. doi: http://dx.doi.org/10.7554/eLife.01867.001.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine Learning*, 56:89–113, 2004. doi: https:/doi.org/10.1023/B:MACH.0000033116.57574.95.

Timothy Bardouille, Lyam Bailey, and CamCAN Group. Evidence for age-related changes in sensorimotor neuromagnetic responses during cued button pressing in a large open-access dataset. *NeuroImage*, 193:25–34, 2019. doi: https://doi.org/10.1016/j.neuroimage.2019.02.065.

Q. Barthélemy, C. Gouy-Pailler, Y. Isaac, A. Souloumiac, A. Larue, and J. I. Mars. Multivariate temporal dictionary learning for EEG. *J. Neurosci. Methods*, 215 (1):19–28, 2013.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2:183–202, 2009.

Niels Becker. Estimation for discrete time branching processes with application to epidemics. *Biometrics*, pages 515–522, 1977.

R. Becker, D. Vidaurre, A.J. Quinn, R.G. Abeysuriya, O. Parker Jones, S. Jbabdi, and M.W. Woolrich. Transient spectral events in resting state meg predict individual task responses. *NeuroImage*, 215:116818, 2020. doi: https://doi.org/10.1101/419374.

H. Berger. Uber das elektroenkephalogramm des menschen. *Archiv fur Psychiatrie und Nervenkrankheiten*, 87:527–570, 1929.

Anindya Bhattacharya and Rajat K. De. Divisive correlation clustering algorithm (dcca) for grouping of genes: detecting varying patterns in expression profiles. *Bioinformatics*, 24(11):1359–1366, 2008. doi: https://doi.org/10.1093/bioinformatics/btn133.

Anindya Bhattacharya and Rajat K. De. Average correlation clustering algorithm (acca) for grouping of co-regulated genes with similar pattern of variation in their

expression values. *Journal of Biomedical Informatics*, 43(4):560–568, 2010. doi: https://doi.org/10.1016/j.jbi.2010.02.001.

Philippe Bolon, Jean-Marc Chassery, Jean-Pierre Cocquerez, Didier Demigny, Christine Graffigne, Annick Montanvert, Sylvie Philipp, Rachid Zéboudj, Josiane Zerubia, and Henri Maître. Analyse d'images: filtrage et segmentation, 1995.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146, 2013. doi: 10.1007/s10107-013-0701-9.

M. Bompaire. *Machine learning based on Hawkes processes and stochastic optimization*. Theses, Université Paris Saclay (COmUE), July 2019.

Kimiko O Bowman and LR Shenton. Estimation: Method of moments. *Encyclopedia of statistical sciences*, 3, 2004.

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1): 1–122, 2011.

Brendan Brady and Timothy Bardouille. Periodic/aperiodic parameterization of transient oscillations: Implications for healthy ageing. *NeuroImage*, 251:118974, 2022. doi: https://doi.org/10.1016/j.neuroimage.2022.118974.

Brendan Brady, Lindsey Power, and Timothy Bardouille. Age-related trends in neuromagnetic transient beta burst characteristics during a sensorimotor task and rest in the cam-can open-access dataset. *NeuroImage*, 222:117245, 2020. doi: https://doi.org/10.1016/j.neuroimage.2020.117245.

Pierre Brémaud and Laurent Massoulié. Stability of nonlinear hawkes processes. *The Annals of Probability*, pages 1563–1588, 1996.

Paul M. Briley, Elizabeth B. Liddle, Molly Simmonite, Marjie Jansen, Thomas P. White, Vajender Balain, Lena Palaniyappan, Richard Bowtell, Karen J. Mullinger, and Peter F. Liddle. Regional brain correlates of beta bursts in health and psychosis: A concurrent electroencephalography and functional magnetic resonance imaging study. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 6(12):1145–1156, 2021. doi: https://doi.org/10.1016/j.bpsc.2020.10.018.

H. Bristow, A. Eriksson, and S. Lucey. Fast convolutional sparse coding. In *Computer Vision and Pattern Recognition (CVPR)*, pages 391–398, 2013.

A. J. Brockmeier and J. C. Príncipe. Learning recurrent waveforms within EEGs. *IEEE Transactions on Biomedical Engineering*, 63(1):43–54, 2016.

Emery N Brown, Riccardo Barbieri, Uri T Eden, and Loren M Frank. Likelihood methods for neural spike train data analysis. *Computational neuroscience: A comprehensive approach*, pages 253–286, 2003.

Raiha Browning, Judith Rousseau, and Kerrie Mengersen. A flexible, random histogram kernel for discrete-time Hawkes processes. *arXiv preprint arXiv:2208.02921*, 2022.

Jeremy B. Caplan, Joseph R. Madsen, Shridhar Raghavachari, and Michael J. Kahana. Distinct patterns of brain oscillations underlie two basic parameters of human maze learning. *Journal of Neurophysiology*, 86(1):368–380, 2001. doi: http://doi.org/10.1152/jn.2001.86.1.368.

Jeremy B. Caplan, Monica Bottomley, Pardeep Kang, and Roger A. Dixon. Distinguishing rhythmic from non-rhythmic brain activity during rest in healthy neurocognitive aging. *Neuroimage*, 112:341–352, 2015. doi: http://doi.org/10.1016/j.neuroimage.2015.03.001.

Marcos Costa Santos Carreira. Exponential Kernels with Latency in Hawkes Processes: Applications in Finance. *preprint ArXiv*, 2101.06348, 2021.

Shizhe Chen, Ali Shojaie, Eric Shea-Brown, and Daniela Witten. The multivariate hawkes process in high dimensions: Beyond mutual excitation. *arXiv preprint arXiv:1707.04928*, 2017.

A.K.I. Chiang, C.J. Rennie, P.A. Robinson, S.J. Van Albada, and C.C. Kerr. Age trends and sex differences of alpha rhythms including split alpha peaks. *Clinical Neurophysiology*, 122(8):1505–1517, 2011. doi: http://doi.org/10.1016/j.clinph.2011.01.040.

Alexandr Chvátal and Alexei Verkhratsky. An early history of neuroglial research: personalities. *Neuroglia*, 1(1):245–281, 2018.

Mike X Cohen. *Analyzing Neural Time Series Data: Theory and Practice*. The MIT Press, 01 2014. ISBN 9780262319553. doi: 10.7551/mitpress/9609.001.0001.

S. R. Cole and B. Voytek. Brain Oscillations and the Importance of Waveform Shape. *Trends in Cognitive Sciences*, 21(2):137–149, 2017.

Scott Cole and Bradley Voytek. Cycle-by-cycle analysis of neural oscillations. *Journal of Neurophysiology*, 122(2):849–861, 2019. doi: https://doi.org/10.1152/jn.00273.2019.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

N. Coquelet, X. De Tiège, L. Roshchupkina, P. Peigneux, S. Goldman, M. Woolrich, and V. Wens. Microstates and power envelope hidden markov modeling probe bursting brain activity at different timescales. *NeuroImage*, 247:118850, 2022. doi: http://doi.org/10.1016/j.neuroimage.2021.118850.

David R Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):129–157, 1955.

Alexander Craik, Yongtian He, and Jose L. Contreras-Vidal. Deep learning for electroencephalogram (EEG) classification tasks: A review. *Journal of Neural Engineering*, 16(3):031001, April 2019.

R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.

David F Crouse. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016.

D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Volume I: Elementary theory and methods*. Probability and Its Applications. Springer-Verlag New York, 2003.

D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Volume II: general theory and structure*. Probability and Its Applications. Springer-Verlag New York, 2007.

Glenn Dallérac, Michael Graupner, Jeroen Knippenberg, Raquel Chacon Ruiz Martinez, Tatiane Ferreira Tavares, Lucille Tallot, Nicole El Massioui, Anna Verschueren, Sophie Höhn, Julie Boulanger Bertolus, et al. Updating temporal expectancy of an aversive event engages striatal plasticity under amygdala control. *Nature communications*, 8(1):13920, 2017.

John M. Danskin. *Theory of Max-Min and Its Application to Weapons Allocation Problems*. Springer Berlin Heidelberg, Berlin/Heidelberg, 1967.

Ingrid Daubechies, Michel Defrise, and Christine Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constrains. *Communications on Pure and Applied Mathematics*, 57, 2004.

Justin Dauwels, Francois Vialatte, and Andrzej Cichocki. Diagnosis of alzheimer's disease from eeg signals: where are we standing? *Current Alzheimer Research*, 7(6):487–505, 2010.

Abir De, Utkarsh Upadhyay, and Manuel Gomez-Rodriguez. Temporal point processes. Technical report, Notes for Human-Centered ML, Saarland University, 2019.

Ivanoe De Falco, Giuseppe De Pietro, Giovanna Sannino, Umberto Scafuri, Ernesto Tarantino, Antonio Della Cioppa, and Giuseppe A Trunfio. Deep neural network hyper-parameter setting for classification of obstructive sleep apnea episodes. In *2018 IEEE symposium on computers and communications (ISCC)*, pages 01187–01192. IEEE, 2018.

Stanislas Dehaene. *Face à face avec son cerveau.* Odile Jacob, 2021.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Debangshu Dey, Sayanti Chaudhuri, and Sugata Munshi. Obstructive sleep apnoea detection using convolutional neural network based deep learning framework. *Biomedical engineering letters*, 8:95–100, 2018.

Sophie Donnet, Vincent Rivoirard, and Judith Rousseau. Nonparametric Bayesian estimation for multivariate Hawkes processes. *The Annals of Statistics*, 48(5): 2698 – 2727, 2020.

T. Donoghue, M. Haller, E. J. Peterson, P. Varma, P. Sebastian, R. Gao, T. Noto, A. H. Lara, J. D. Wallis, R. T. Knight, A. Shestyuk, and B. Voytek. Parameterizing neural power spectra into periodic and aperiodic components. *Nature Neuroscience*, 23(12):1655–1665, 2020.

D. P. Drennan and E. C. Lalor. Cortical tracking of complex sound envelopes: modeling the changes in response with intensity. *eneuro*, 6(3), 2019.

Olive Jean Dunn. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, 1961.

T. Dupré la Tour, T. Moreau, M. Jas, and A. Gramfort. Multivariate convolutional sparse coding for electromagnetic brain signals. *Advances in Neural Information Processing Systems*, 31:3292–3302, 2018.

Michael Eichler, Rainer Dahlhaus, and Johannes Dueck. Graphical modeling for multivariate Hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242, 2017.

Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer, 2010.

Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006.

Kjersti Engan, Sven Ole Aase, and J Hakon Husoy. Method of optimal directions for frame design. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 5, pages 2443–2446. IEEE, 1999.

Steven P. Errington, Geoffrey F. Woodman, and Jeffrey D. Schall. Dissociation of medial frontal beta-bursts and executive control. *The Journal of Neuroscience*, 40(48):9272–9282, 2020. doi: https://doi.org/10.1523/JNEUROSCI.2072-20.2020.

Marco S. Fabus, Andrew J. Quinn, Catherine E. Warnaby, and Mark W. Woolrich. Automatic decomposition of electrophysiological data into distinct nonsinusoidal oscillatory modes. volume 126, pages 1670–1684, 2021. doi: https://doi.org/10.1152/jn.00315.2021.

Joseph Feingold, Daniel J. Gibson, Brian DePasquale, and Ann M. Graybiel. Bursts of beta oscillation differentiate postperformance activity in the striatum and motor cortex of monkeys performing movement tasks. *Proceedings of the National Academy of Sciences*, 112(44):13687–13692, 2015. doi: https://doi.org/10.1073/pnas.1517629112.

R Douglas Fields, Alfonso Araque, Heidi Johansen-Berg, Soo-Siang Lim, Gary Lynch, Klaus-Armin Nave, Maiken Nedergaard, Ray Perez, Terrence Sejnowski, and Hiroaki Wake. Glial biology in learning and cognition. *The neuroscientist*, 20(5):426–431, 2014.

Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368, 1922.

Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. 2013. doi: 10.1007/978-0-8176-4948-7.

Andreas Galka, Okito Yamashita, Tohru Ozaki, Rolando Biscay, and Pedro Valdés-Sosa. A solution to the dynamical inverse problem of eeg generation using spatiotemporal kalman filtering. *NeuroImage*, 23(2):435–453, 2004.

A. Galves and E. Löcherbach. Modeling networks of spiking neurons as interacting processes with memory of variable length. *arXiv preprint arXiv:1502.06446*, 2015.

R Gil, L Zai, J-Ph Neau, Th Jonveaux, C Agbo, T Rosolacci, P Burbaud, and P Ingrand. Event-related auditory evoked potentials and multiple sclerosis. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 88(3):182–187, 1993.

B. Gips, A. Bahramisharif, E. Lowet, M. Roberts, P. de Weerd, O. Jensen, and J. van der Eerden. Discovering recurring patterns in electrophysiological recordings. *J. Neurosci. Methods*, 275:66–79, 2017.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23): e215–e220, 2000.

Camillo Golgi. *Opera omnia*, volume 3. Ulrico Hoepli, 1903.

A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, et al. MEG and EEG data analysis with MNE-Python. *Frontiers in neuroscience*, 7:267, 2013.

A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, and M. S. Hämäläinen. MNE software for processing MEG and EEG data. *Neuroimage*, 86:446–460, 2014.

Alexandre Gramfort. Lecture notes in meeg: Functional brain imaging with meg, eeg and seeg, http://bit.ly/meeg_course.

Roberta Grech, Tracey Cassar, Joseph Muscat, Kenneth P Camilleri, Simon G Fabri, Michalis Zervakis, Petros Xanthopoulos, Vangelis Sakkalis, and Bart Vanrumste. Review on solving the inverse problem in eeg source analysis. *Journal of neuroengineering and rehabilitation*, 5(1):25, 2008.

R. Grosse, R. Raina, H. Kwong, and A. Y. Ng. Shift-invariant sparse coding for audio classification. In *23rd Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 149–158. AUAI Press, 2007. ISBN 0-9749039-3-0.

Eric C. Hall and Rebecca M. Willett. Tracking dynamic point processes on networks. *IEEE Transactions on Information Theory*, 62(7):4327–4346, 2016.

Matti Hämäläinen, Riitta Hari, Risto J Ilmoniemi, Jukka Knuutila, and Olli V Lounasmaa. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of modern Physics*, 65(2):413, 1993.

N. R. Hansen, P. Reynaud-Bouret, V. Rivoirard, et al. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015.

C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585 (7825):357–362, September 2020.

Hastie, Tibshirani, and Wainwright. *Statistical learning with sparsity: The lasso and generalizations*. 2015. doi: 10.1201/b18401.

A. G. Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443, 1971.

A. G. Hawkes and D. Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, pages 493–503, 1974.

Shenghong He, Claudia Everest-Phillips, Andrew Clouter, Peter Brown, and Huiling Tan. Neurofeedback-linked suppression of cortical beta bursts speeds up movement initiation in healthy motor control: A double-blind sham-controlled study. *The Journal of Neuroscience*, 40(20):4021–4032, 2020. doi: https://doi.org/10.1523/JNEUROSCI.0208-20.2020.

F. Heide, W. Heidrich, and G. Wetzstein. Fast and flexible convolutional sparse coding. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5135–5143. IEEE, 2015.

R. Herbert and D. Lehmann. Theta bursts: An eeg pattern in normal subjects practising the transcendental meditation technique. *Electroencephalography and Clinical Neurophysiology*, 42(3):397–405, 1977. doi: http://doi.org/10.1016/0013-4694(77)90176-6.

Johan Himberg, Aapo Hyvärinen, and Fabrizo Esposito. Validating the independent components of neuroimaging time series via clustering and visualization. *Neuroimage*, 22(3):1214–1222, 2004. doi: http://doi.org/10.1016/j.neuroimage.2004.03.027.

Rikkert Hindriks, Michel J.A.M. van Putten, and Gustavo Deco. Intra-cortical propagation of eeg alpha oscillations. *Neuroimage*, 103:444–453, 2014. doi: https://doi.org/10.1016/j.neuroimage.2014.08.027.

S. Hitziger, M. Clerc, S. Saillet, C. Bénar, and Papadopoulo T. Adaptive waveform learning: A framework for modeling variability in neurophysiological signals. *IEEE Transactions on Signal Processing*, 65(16):4324–4338, 2017.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.

Norden E. Huang, Zheng Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H. Lui. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1971):903–995, 1998. doi: https://doi.org/10.1098/rspa.1998.0193.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science Engineering*, 9(3):90–95, 2007.

A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13:411–430, 2000. doi: https://doi.org/10.1016/S0893-6080(00)00026-5.

Boris Iglewicz and David C Hoaglin. *Volume 16: how to detect and handle outliers*. Quality Press, 1993.

Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

Kazufumi Ito and Bangti Jin. *Inverse problems: Tikhonov theory and algorithms*, volume 22. World Scientific, 2014.

M. Jas, T. Dupré La Tour, U. Simsekli, and A. Gramfort. Learning the morphology of brain signals using alpha-stable convolutional sparse coding. In *Advances in Neural Information Processing Systems*, pages 1099–1108, 2017.

H. H. Jasper. Charting the sea of brain waves. *Science*, 108(2805):343–347, 1948.

S. R. Jones. When brain rhythms aren't 'rhythmic': implication for their mechanisms and meaning. *Curr. Opin. Neurobiol.*, 40:72–80, 2016.

P. Jost, P. Vandergheynst, S. Lesage, and R. Gribonval. MoTIF: an efficient algorithm for learning translation invariant dictionaries. In *Acoustics, Speech and Signal Processing (ICASSP)*, volume 5. IEEE, 2006.

Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven Siegelbaum, A James Hudspeth, Sarah Mack, et al. *Principles of neural science*, volume 4. McGraw-hill New York, 2000. URL https://archive.org/details/isbn_9780838577011/page/n7/mode/2up.

N Kannathal, U Rajendra Acharya, Choo Min Lim, and PK Sadasivan. Characterization of EEG—a comparative study. *Computer methods and Programs in Biomedicine*, 80(1):17–23, 2005.

Golan Karvat, Artur Schneider, Mansour Alyahyay, Florian Steenbergen, Michael Tangermann, and Ilka Diester. Real-time detection of neural oscillation bursts allows behaviourally relevant neurofeedback. *Communications Biology*, 3(72), 2020. doi: https://doi.org/10.1038/s42003-020-0801-z.

S. Kim, D. Putrino, S. Ghosh, and E. N. Brown. A granger causality measure for point process models of ensemble neural spiking activity. *PLoS Comput Biol*, 7 (3):e1001110, 2011.

Matthias Kirchner. Hawkes and INAR(∞) processes. *Stochastic Processes and their Applications*, 126(8):2494–2525, August 2016.

Matthias Kirchner and A Bercher. A nonparametric estimation procedure for the hawkes process: comparison with maximum likelihood estimation. *Journal of Statistical Computation and Simulation*, 88(6):1106–1116, 2018.

Glenn T Konopaske, Karl-Anton Dorph-Petersen, Robert A Sweet, Joseph N Pierri, Wei Zhang, Allan R Sampson, and David A Lewis. Effect of chronic antipsychotic exposure on astrocyte and oligodendrocyte numbers in macaque monkeys. *Biological psychiatry*, 63(8):759–765, 2008.

Julian Q. Kosciessa, Thomas H. Grandy, Douglas D. Garrett, and Markus Werkle-Bergner. Single-trial characterization of neural rhythms: Potential and challenges. *Neuroimage*, 206:116331, 2020. doi: https://doi.org/10.1016/j.neuroimage.2019.116331.

Michael Krumin, Inna Reutsky, and Shy Shoham. Correlation-based analysis and generation of multiple spike trains using hawkes models with an exogenous input. *Frontiers in computational neuroscience*, 4:147, 2010.

Daisuke Kurisu. Discretization of self-exciting peaks over threshold models. 2016. doi: 10.48550/ARXIV.1612.06109.

Peter Lakatos, Nóra Szilàgyi, Zsuzsanna Pincze, Csaba Rajkai, István Ulbert, and György Karmos. Attention and arousal related modulation of spontaneous gamma-activity in the auditory cortex of the cat. *Cognitive Brain Research*, 19 (1):1–9, 2004. doi: https://doi.org/10.1016/j.cogbrainres.2003.10.023.

T. A. Lasko. Efficient inference of gaussian-process-modulated renewal processes with application to medical event data. In *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, volume 2014, page 469. NIH Public Access, 2014.

L. Le Magoarou and R. Gribonval. Chasing butterflies: In search of efficient dictionaries. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015.

L. Le Magoarou, R. Gribonval, and A. Gramfort. FAμST: speeding up linear transforms for tractable inverse problems. In *European Signal Processing Conference (EUSIPCO)*, Nice, France, August 2015.

Re Lemonnier and Nicolas Vayatis. Nonparametric markovian learning of triggering kernels for mutually exciting and mutually inhibiting multivariate hawkes processes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 161–176. Springer, 2014.

Sabine Leske and Sarang S Dalal. Reducing power line noise in eeg and meg data via spectrum interpolation. *Neuroimage*, 189:763–776, 2019.

M. S. Lewicki and T. J. Sejnowski. Coding time-varying signals using sparse, shift-invariant representations. In M. J. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 730–736. MIT Press, 1999.

PA W Lewis and Gerald S Shedler. Simulation of nonhomogeneous Poisson processes by thinning. *Naval research logistics quarterly*, 26(3):403–413, 1979.

R. Lewis and G. Mohler. A nonparametric EM algorithm for multiscale Hawkes processes. *Journal of Nonparametric Statistics*, 1(1):1–20, 2011.

Kunyang Li, Weifeng Pan, Yifan Li, Qing Jiang, and Guanzheng Liu. A method to detect sleep apnea based on deep neural network and hidden markov model using single-lead ecg signal. *Neurocomputing*, 294:94–101, 2018.

W. Lian, R. Henao, V. Rao, J. Lucas, and L. Carin. A multitask point process predictive model. In *International Conference on Machine Learning*, pages 2030–2038. PMLR, 2015.

Scott Linderman and Ryan Adams. Discovering latent network structure in point process data. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1413–1421, 2014.

Scott W Linderman and Ryan P Adams. Scalable bayesian inference for excitatory point process networks. *arXiv preprint arXiv:1507.03228*, 2015.

Simon Little, James Bonauito, Gareth Barnes, Sven Bestmann, and Karunesh Ganguly. Human motor cortical beta bursts relate to movement planning and response errors. *PLOS Biology*, 17(10):e3000479, 2019. doi: https://doi.org/10.1371/journal.pbio.3000479.

T. Long, F. Mehrdad, S. Le, and Z. Hongyuan. *NetCodec: Community Detection from Individual Activities*, pages 91–99. SIAM International Conference on Data Mining (SDM), 2015.

Alfred L. Loomis, E. Newton Harvey, and Garret Hobart. Potential rhythms of the cerebral cortex during sleep. *Science*, 81:597–598, 1935. doi: http://doi.org/10.1126/science.81.2111.597.

Mikael Lundqvist, Jonas Rose, Pawel Herman, Scott Brincat, Timothy Buschman, and Earl Miller. Gamma and beta bursts underlie working memory. *Neuron*, 90 (1):152–164, 2016. doi: https://doi.org/10.1016/j.neuron.2016.02.028.

Luz Luz, Eduardo José da S, William Robson Schwartz, Guillermo Cámara-Chávez, and David Menotti. ECG-based heartbeat classification for arrhythmia detection: A survey. *Computer Methods and Programs in Biomedicine*, 127: 144–164, 2016.

Julien Mairal, Francis Bach, J. Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11, 2009.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(1), 2010.

Julien Mairal, Francis Bach, Jean Ponce, et al. Sparse modeling for image and vision processing. *Foundations and Trends® in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

Benoît Malézieux, Thomas Moreau, and Matthieu Kowalski. Understanding approximate and unrolled dictionary learning for pattern recovery. *International Conference on Learning Representations*, 2022.

Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

Stéphane Mallat and Zhifeng Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41:3397 – 3415, 1994.

Henry Markram, Eilif Muller, Srikanth Ramaswamy, Michael W Reimann, Marwan Abdellah, Carlos Aguado Sanchez, Anastasia Ailamaki, Lidia Alonso-Nanclares, Nicolas Antille, Selim Arsever, et al. Reconstruction and simulation of neocortical microcircuitry. *Cell*, 163(2):456–492, 2015.

J. P. Martinez, R. Almeida, S. Olmos, A. P. Rocha, and P. Laguna. A wavelet-based ECG delineator: evaluation on standard databases. *IEEE Transactions on Biomedical Engineering*, 51(4):570–581, 2004.

B. Mehrdad and L. Zhu. On the Hawkes process with different exciting functions. *arXiv preprint arXiv:1403.0994*, 2014.

H. Mei, T. Wan, and J. Eisner. Noise-contrastive estimation for multivariate point processes. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5204–5214. Curran Associates, Inc., 2020.

Hongyuan Mei and Jason M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. *Advances in neural information processing systems*, 30, 2017.

Arthur Mensch, Julien Mairal, Bertrand Thirion, and Gaël Varoquaux. Dictionary learning for massive matrix factorization. In *International Conference on Machine Learning*, pages 1737–1746. PMLR, 2016.

Christoph M. Michel and Thomas Koenig. Eeg microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: A review. *Neuroimage*, 180:577–593, 2018. doi: https://doi.org/10.1016/j.neuroimage.2017.11.062.

Milos Miljkovic, Tatyana Chernenko, Melissa J. Romeo, Benjamin Bird, Christian Matthaus, and Max Diem. Label-free imaging of human cells: algorithms for image reconstruction of raman hyperspectral datasets. *Analyst*, 135:2002–2013, 2010. doi: https://doi.org/10.1039/C0AN00042F.

George Mohler et al. Modeling and estimation of multi-source clustering in crime and security data. *The Annals of Applied Statistics*, 7(3):1525–1539, 2013.

PAP Moran. The random division of an interval—part iii. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 15(1):77–80, 1953.

Thomas Moreau and Alexandre Gramfort. Distributed Convolutional Dictionary Learning (DiCoDiLe): Pattern Discovery in Large Images and Signals. *arXiv preprint arXiv:1901.09235*, 2019.

Thomas Moreau and Alexandre Gramfort. DiCoDiLe: Distributed Convolutional Dictionary Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Thomas Moreau, Laurent Oudre, and Nicolas Vayatis. Dicod: Distributed convolutional coordinate descent for convolutional sparse coding. In *International Conference on Machine Learning*, pages 3626–3634. PMLR, 2018.

Sheikh Shanawaz Mostafa, Fábio Mendonça, Antonio G. Ravelo-García, and Fernando Morgado-Dias. A systematic review of detecting sleep apnea using deep learning. *Sensors*, 19(22):4934, 2019.

Lyle Muller, Frédéric Chavane, John Reynolds, and Terrence J. Sejnowski. Cortical travelling waves: mechanisms and computational principles. *Nature Reviews Neuroscience*, 19:255–268, 2018. doi: http://doi.org/10.1038/nrn.2018.20.

Aaron J. Newman, Danny Godfrey, and Reann Post. Data science for psychology and neuroscience - in python. Online textbook, 2023. URL https://neuraldatascience.io/intro.html. Accessed: 2023-07-22.

Daniel Novak, Karel Mucha, and Tarik Al-Ani. Long short-term memory for apnea detection based on heart rate variability. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5234–5237. IEEE, 2008.

R. Näätänen and T. Picton. The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. *Psychophysiology*, 24(4):375–425, 1987.

Yosihiko Ogata. Estimators for stationary point processes. *Ann. Inst. Statist. Math*, 30(Part A):243–261, 1978.

Yosihiko Ogata. On lewis' simulation method for point processes. *IEEE transactions on information theory*, 27(1):23–31, 1981.

Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401):9–27, 1988.

Yosihiko Ogata. Seismicity analysis through point-process modeling: A review. In *Seismicity patterns, their statistical significance and physical meaning*, pages 471–507. Springer, 1999.

Seiji Ogawa, Tso-Ming Lee, Asha S Nayak, and Paul Glynn. Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic resonance in medicine*, 14(1):68–78, 1990.

E. Oja and Yuan Zhijian. The fastica algorithm revisited: Convergence analysis. *IEEE Transactions on Neural Networks*, 17(6):1370–1381, 2006. doi: http://doi.org/10.1109/TNN.2006.880980.

M. Okatan, M. A. Wilson, and E. N. Brown. Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. *Neural computation*, 17(9):1927–1961, 2005.

Bruno A. Olshausen and David J Field. Sparse coding with an incomplete basis set: A strategy employed by \protect{V1}. *Vision Research*, 37(23):3311–3325, 1997.

Alexei Ossadtchi, Tatiana Shamaeva, Elizaveta Okorokova, Victoria Moiseeva, and Mikhail A. Lebedev. Neurofeedback learning modifies the incidence rate of alpha spindles, but not their duration and amplitude. *Scientific Reports*, 7(3772), 2017. doi: https://doi.org/10.1038/s41598-017-04012-0.

J Pan and W J Tompkins. A real-time {QRS} detection algorithm. *IEEE Transactions on Biomedical Engineering*, 32(3):230–236, 1985.

Zhimeng Pan, Zheng Wang, Jeff M Phillips, and Shandian Zhe. Self-adaptable point processes with nonparametric time decays. *Advances in Neural Information Processing Systems*, 34:4594–4606, 2021.

Liam Paninski. Lecture notes in Statistical analysis of neural data course: Chapter 2 – Introduction to Point Processes, Fall 2019.

Vardan Papyan, Yaniv Romano, and Michael Elad. Convolutional neural networks analyzed via convolutional sparse coding. *The Journal of Machine Learning Research*, 18(1):2887–2938, 2017.

Roberto Domingo Pascual-Marqui. Review of methods for solving the eeg inverse problem. *International journal of bioelectromagnetism*, 1(1):75–86, 1999.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

Rahul K Pathinarupothi, R Vinaykumar, Ekanath Rangan, E Gopalakrishnan, and KP Soman. Instantaneous heart rate as a robust feature for sleep apnea severity detection using deep learning. In *2017 IEEE EMBS international conference on biomedical & health informatics (BHI)*, pages 293–296. IEEE, 2017a.

Rahul Krishnan Pathinarupothi, Ekanath Srihari Rangan, EA Gopalakrishnan, R Vinaykumar, KP Soman, et al. Single sensor techniques for sleep apnea diagnosis using deep learning. In *2017 IEEE international conference on healthcare informatics (ICHI)*, pages 524–529. IEEE, 2017b.

Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302): 157–175, 1900.

Bill Pelz. Introductory psychology – 3. biopsychology, parts of the nervous system. MOOC. URL https://web.archive.org/web/20230521074211/https://courses.lumenlearning.com/sunyhccc-ss-151-1/chapter/parts-of-the-nervous-system/. Accessed: 2023-07-12.

Thomas Penzel, George B Moody, Roger G Mark, Ary L Goldberger, and J Hermann Peter. The apnea-ecg database. In *Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163)*, pages 255–258. IEEE, 2000.

Gert Pfurtscheller and FH Lopes Da Silva. Event-related eeg/meg synchronization and desynchronization: basic principles. *Clinical neurophysiology*, 110(11):1842–1857, 1999.

J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. J. Chichilnisky, and E. P. Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.

John Pollock. Nervous system – informational reading about the nervous system. MOOC, 2023. URL https://web.archive.org/web/20230324152028/https://www.thepartnershipineducation.com/resources/nervous-system. Accessed: 2023-07-08.

Lindsey Power and Timothy Bardouille. Age-related trends in the cortical sources of transient beta bursts during a sensorimotor task and rest. *NeuroImage*, 245: 118670, 2021. doi: https://doi.org/10.1016/j.neuroimage.2021.118670.

Lindsey Power, Cédric Allain, Thomas Moreau, Alexandre Gramfort, and Timothy Bardouille. Using convolutional dictionary learning to detect task-related neuromagnetic transients and ageing trends in a large open-access dataset. *NeuroImage*, 267:119809, 2023.

Darren Price, Lorraine K Tyler, R Neto Henriques, Karen L Campbell, Nitin Williams, Matthias S Treder, Jason R Taylor, and RNA Henson. Age-related delay in visual and auditory evoked responses is mediated by white-and grey-matter differences. *Nature communications*, 8(1):15671, 2017.

Python Software Foundation. Python Language Reference, version 3.8. *http://python.org/*, 2019.

Andrew J. Quinn, Diego Vidaurre, Romesh Abeysuriya, Robert Becker, Anna C. Nobre, and Mark W. Woolrich. Task-evoked dynamic network analysis through hidden markov modeling. *Frontiers in Neuroscience*, 12(603), 2018. doi: https://doi.org/10.3389/fnins.2018.00603.

Andrew J. Quinn, Vítor Lopes-dos Santos, Norden Huang, Wei-Kuang Liang, Chi-Hung Juan, Jia-Rong Yeh, Anna C. Nobre, David Dupret, and Mark W. Woolrich. Within-cycle instantaneous frequency profiles report oscillatory waveform dynamics. *Journal of Neurophysiology*, 126:1190–1208, 2021. doi: https://doi.org/10.1152/jn.00201.2021.

K. Rad and L. Paninski. Information rates and optimal decoding in large neural populations. *Advances in neural information processing systems*, 24:846–854, 2011.

Jakob Gulddahl Rasmussen. Bayesian inference for hawkes processes. *Methodology and Computing in Applied Probability*, 15(3):623–642, 2013.

Holly Rayson, Ranjan Debnath, Sanaz Alavizadeh, Nathan Fox, Pier F. Ferrari, and James J. Bonauito. Detection and analysis of cortical beta bursts in developmental eeg data. *Developmental Cognitive Neuroscience*, 54:101069, 2022. doi: https://doi.org/10.1016/j.dcn.2022.101069.

Patricia Reynaud-Bouret and Vincent Rivoirard. Near optimal thresholding estimation of a poisson intensity on the real line. *Electronic journal of statistics*, 4: 172–238, 2010.

Patricia Reynaud-Bouret and Sophie Schbath. Adaptive estimation for Hawkes processes; application to genome analysis. *The Annals of Statistics*, 38(5): 2781 – 2822, 2010. doi: 10.1214/10-AOS806. URL https://doi.org/10.1214/10-AOS806.

Patricia Reynaud-Bouret, Vincent Rivoirard, Franck Grammont, and Christine Tuleau-Malot. Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. *The Journal of Mathematical Neuroscience*, 4(1):1–41, 2014.

Paolo M. Rossini, Simone Rossi, Claudio Babilioni, and John Polich. Clinical neurophysiology of aging brain: From normal aging to neurodegeneration. *Progress in Neurobiology*, 83(6):375–400, 2007. doi: https://doi.org/10.1016/j.pneurobio.2007.07.010.

Peter J Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984.

Meyer Scetbon, Michael Elad, and Peyman Milanfar. Deep k-svd denoising. *IEEE Transactions on Image Processing*, 30:5944–5955, 2021.

Zelekha A. Seedat, Andrew J. Quinn, Diego Vidaurre, Lucrezia Luizzi, Lauren E. Gascoyne, Benjamin A.E. Hunt, George C. O'Neill, Daisie O. Pakenham, Karen J. Mullinger, Peter G. Morris, Mark W. Woolrich, and Matthew J. Brookes. The role of transient spectral 'bursts' in functional connectivity: A magnetoencephalography study. *Neuroimage*, 209:116537, 2020. doi: https://doi.org/10.1016/j.neuroimage.2020.116537.

Dennis J Selkoe. Alzheimer's disease is a synaptic failure. *Science*, 298(5594): 789–791, 2002.

M. A. Shafto, L. K. Tyler, M. Dixon, J. R. Taylor, J. B. Rowe, R. Cusack, A. J. Calder, W. D. Marslen-Wilson, J. Duncan, T. Dalgleish, et al. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC neurology*, 14(1):1–25, 2014.

Claude E Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

O. Shchur, N. Gao, M. Biloš, and S. Günnemann. Fast and flexible temporal point processes with triangular maps. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 73–84. Curran Associates, Inc., 2020.

Oleksandr Shchur, Marin Biloš, and Stephan Günnemann. Intensity-free learning of temporal point processes. *arXiv preprint arXiv:1909.12127*, 2019.

M. A. Sherman, S. Lee, R. Law, S. Haegens, C. A. Thorn, M. S. Hämäläinen, C. I. Moore, and S. R. Jones. Neural mechanisms of transient neocortical beta rhythms: Converging evidence from humans, computational modeling, monkeys, and mice. *Proceedings of the National Academy of Sciences*, 113(33):E4885–E4894, 2016.

Hyeyoung Shin, Robert Law, Shawn Tsutsui, Christopher I Moore, and Stephanie R Jones. The rate of transient beta frequency events predicts behavior across tasks and species. *eLife*, 6:e29086, nov 2017.

Mahsa Soufineyestani, Dale Dowling, and Arshia Khan. Electroencephalography (eeg) technology applications and available devices. *Applied Sciences*, 10(21): 7453, 2020.

Nelson Spruston. Pyramidal neurons: dendritic structure and synaptic integration. *Nature Reviews Neuroscience*, 9(3):206–221, 2008.

Guillaume Staerman, Cédric Allain, Alexandre Gramfort, and Thomas Moreau. FaDIn: Fast Discretized Inference for Hawkes Processes with General Parametric Kernels. *International Conference on Machine Learning*, 2023.

Jeremias Sulam, Vardan Papyant, Yaniv Romano, and Michael Elad. Projecting on to the multi-layer convolutional sparse coding model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6757–6761. IEEE, 2018.

Deborah Sulem, Vincent Rivoirard, and Judith Rousseau. Bayesian estimation of nonlinear Hawkes process. *arXiv preprint arXiv:2103.17164*, 2021.

Idan Tal and M. Abeles. Imaging the spatiotemporal dynamics of cognitive processes at high temporal resolution. *Neural Computation*, 30(3):610–630, 2017. doi: https://doi.org/10.1162/neco_a_01054.

Idan Tal and Moshe Abeles. Temporal accuracy of human cortico-cortical interactions. *Journal of Neurophysiology*, 115(4):1810–1820, 2016. doi: https://doi.org/10.1152/jn.00956.2015.

Idan Tal, Samuel Neymotin, Stephan Bickel, Peter Lakatos, and Charles E. Schroeder. Oscillatory bursting as a mechanism for temporal coupling and information coding. *Frontiers in Computational Neuroscience*, 14(82), 2020. doi: https://doi.org/10.3389/fncom.2020.00082.

C. Tallon-Baudry, O. Bertrand, C. Delpuech, and J. Pernier. Stimulus specificity of phase-locked and non-phase-locked 40 hz visual responses in human. *Journal of Neuroscience*, 16(13):4240–4249, 1996.

Catherine Tallon-Baudry and Olivier Bertrand. Oscillatory gamma activity in humans and its role in object representation. *Trends in cognitive sciences*, 3(4): 151–162, 1999.

Hideaki Tanaka, Thomas Koenig, Roberto D Pascual-Marqui, Koichi Hirata, Kieko Kochi, and Dietrich Lehmann. Event-related potential and EEG measures in parkinson's disease without and with dementia. *Dementia and geriatric cognitive disorders*, 11(1):39–45, 2000.

Hiroyuki Tasaki. Convergence rates of approximate sums of Riemann integrals. *Journal of Approximation Theory*, 161(2):477–490, 2009. ISSN 0021-9045.

S Taulu and J Simola. Spatiotemporal signal space separation method for rejecting nearby interference in meg measurements. *Physics in Medicine and Biology*, 51 (7):1759–1768, 2006. doi: http://doi.org/10.1088/0031-9155/51/7/008.

Jason R Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A Shafto, Marie Dixon, Lorraine K Tyler, Richard N Henson, et al. The cambridge centre for ageing and neuroscience (cam-can) data repository: Structural and functional mri, meg, and cognitive data from a cross-sectional adult lifespan sample. *Neuroimage*, 144:262–269, 2017. doi: http://doi.or/10.1016/j.neuroimage.2015.09.018.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996.

Bahareh Tolooshams and Demba Ba. Pudle: Implicit acceleration of dictionary learning by backpropagation. *arXiv preprint*, 2021.

W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93 (2):1074–1089, 2005.

Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. *Advances in neural information processing systems*, 32:3732–3745, 2019.

Alejandro Veen and Frederic P Schoenberg. Estimation of space–time branching process models in seismology using an em–type algorithm. *Journal of the American Statistical Association*, 103(482):614–624, 2008.

Jayabal Velmurugan, Sanjib Sinha, and Parthasarathy Satishchandra. Magnetoencephalography recording and analysis. *Annals of Indian Academy of Neurology*, 17(Suppl 1):S113, 2014.

Diego Vidaurre, Andrew J. Quinn, Adam P. Baker, David Dupret, Alvaro Tejero-Cantero, and Mark W. Woolrich. Spectrally resolved fast transient brain states in electrophysiological data. *Neuroimage*, 126:81–95, 2016. doi: https://doi.org/10.1016/j.neuroimage.2015.11.047.

Ricardo Vigário, Veikko Jousmäki, Matti Hämäläinen, Riitta Hari, and Erkki Oja. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. *NeurIPS Proceedings*, 1998.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C J Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

Yichen Wang, Bo Xie, Nan Du, and Le Song. Isotonic hawkes processes. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 2226–2234. PMLR, 2016.

Jan R. Wessel. Beta-bursts reveal the trial-to-trial dynamics of movement initiation and cancellation. *The Journal of Neuroscience*, 40(2):411–423, 2020. doi: https://doi.org/10.1523/JNEUROSCI.1887-19.2019.

Tara A. Whitten, Adam M. Hughes, Clayton T. Dickson, and Jeremy B. Caplan. A better oscillation detection method robustly extracts eeg rhythms across brain state changes: The human alpha rhythm as a test case. *Neuroimage*, 54(2):860–874, 2011. doi: https://doi.org/10.1016/j.neuroimage.2010.08.064.

Andreas Widmann, Erich Schröger, and Burkhard Maess. Digital filter design for electrophysiological data–a practical approach. *Journal of neuroscience methods*, 250:34–46, 2015.

Wikipedia. Action potential — wikipedia, the free encyclopedia, 2023a. URL https://en.wikipedia.org/wiki/Action_potential. Accessed on 2023-07-12.

Wikipedia. Electroencephalography — wikipedia, the free encyclopedia, 2023b. URL https://en.wikipedia.org/wiki/Electroencephalography. Accessed on 2023-07-12.

Wikipedia. Magnetoencephalography — wikipedia, the free encyclopedia, 2023c. URL https://en.wikipedia.org/wiki/Magnetoencephalography. Accessed on 2023-07-12.

I. Winkler, S. Debener, K.-R. Müller, and M. Tangermann. On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4101–4105. IEEE, 2015.

B. Wohlberg. Convolutional sparse representation of color images. In *IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, pages 57–60, 2016a.

B. Wohlberg. Efficient algorithms for convolutional sparse representations. *Image Processing, IEEE Transactions on*, 25(1):301–315, 2016b.

Tong Tong Wu, Kenneth Lange, et al. Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics*, 2(1):224–244, 2008.

Yande Xiang, Zhitao Lin, and Jianyi Meng. Automatic qrs complex detection using two-level convolutional neural network. *Biomedical engineering online*, 17 (1):1–17, 2018.

H. Xu, M. Farajtabar, and H. Zha. Learning granger causality for Hawkes processes. In *International conference on machine learning*, pages 1717–1726, 2016.

Hongteng Xu, Dixin Luo, and Hongyuan Zha. Learning hawkes processes from short doubly-censored event sequences. In *International Conference on Machine Learning*, pages 3831–3840. PMLR, 2017.

Hongteng Xu, Dixin Luo, Xu Chen, and Lawrence Carin. Benefits from superposed Hawkes processes. In *International Conference on Artificial Intelligence and Statistics*, pages 623–631. PMLR, 2018.

Yingxiang Yang, Jalal Etesami, Niao He, and Negar Kiyavash. Online learning for multivariate hawkes processes. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Anis Yazidi and Hugo Hammer. Multiplicative update methods for incremental quantile estimation. *IEEE transactions on cybernetics*, 49(3):746–756, 2017.

Rafael Yuste. From the neuron doctrine to neural networks. *Nature reviews neuroscience*, 16(8):487–497, 2015.

Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. In *2010 IEEE Computer Society Conference on computer vision and pattern recognition*, pages 2528–2535. IEEE, 2010.

Rui Zhang, Christian Walder, Marian-Andrei Rizoiu, and Lexing Xie. Efficient non-parametric bayesian hawkes processes. *arXiv preprint arXiv:1810.03730*, 2018.

Ke Zhou, Hongyuan Zha, and Le Song. Learning triggering kernels for multi-dimensional Hawkes processes. In *International conference on machine learning*, pages 1301–1309. PMLR, 2013a.

Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *Artificial Intelligence and Statistics*, pages 641–649. PMLR, 2013b.

Appendix

A Pioneers of modern neuroscience

Even though neuroscience emerged in the Hellenistic world following the first experimental discoveries of the nerves connecting the brain with the body [Chvátal and Verkhratsky, 2018], the roots of modern neuroscience can be traced back to the late 19th and early 20th century, during which scientists sought to understand the intricate workings of the human brain. The pioneering work of Camillo Golgi and Santiago Ramón y Cajal has dramatically shaped the field.

Camillo Golgi (1843-1926) In 1873, the Italian physician Camillo Golgi introduced the silver-chromate staining technique, revolutionizing microscopic visualization of neural cells [Golgi, 1903]. Golgi's black reaction allowed him to capture the intricate structures of intact neurons using ink and paper. By staining nerve tissue with silver nitrate, Golgi achieved the groundbreaking feat of rendering neurons black, making them distinguishable from the surrounding transparent cells. In 1875, Camillo Golgi published his first scientific drawing, made possible by his chemical reaction. This illustration depicted the neural fibers, gray matter, and other components of a dog's olfactory bulb, showcasing the remarkable detail afforded by Golgi's staining technique (*cf.* fig. A.1).

Santiago Ramón y Cajal (1852-1934) The Spanish neuroscientist, pathologist, and artist Santiago Ramón y Cajal made significant contributions to the field of neuroscience through his intricate and accurate illustrations of the inner workings of the brain[6]. In 1913, Cajal refined the silver staining technique of Golgi by using a gold stain to map the central nervous system. His meticulous drawings depicted the complex structures within the brain, including neurons, with remarkable precision. Cajal's work was based on the assumption, later scientifically proven in the 1950s, that neurons were cellular entities separated by fine

[6]The UNESCO Courier, January-March 2022, p. 12, unesdoc.unesco.org

spaces – a discontinuity that Charles Scott Sherrington coined "synapse" in 1897 – and not the fibers of an uninterrupted network, which challenged Golgi's reticular theory. This groundbreaking insight, known as the neuron doctrine, revolutionized our understanding of the nervous system. Cajal and Golgi were jointly awarded the Nobel Prize in Physiology or Medicine in 1906 "in recognition of their work on the structure of the nervous system"[7]. Cajal is considered one of the founders of neuroscience, and his drawings continue to be used in the field to illustrate the neural architecture underlying memory and human cognition.

Figure A.1: Golgi's illustration of a dog's olfactory bulb. [Golgi, 1903]

[7]The Nobel Prize in Physiology or Medicine 1906. nobelprize.org.

B Adaptation of FISTA for CDL's inner problem

We here adapt the FISTA procedure described in algorithm 1 for the current optimization problem expressed in eq. (6.4.1), to obtain algorithm 6.

Algorithm 6: Adapted FISTA for multivariate CDL's inner problem

Input $X \in \mathbb{R}^{P \times T}$, $\mathbf{D} \in \mathbb{R}^{K \times P \times L}$, $\lambda > 0$ and $\mathcal{L}$;
Set $Z^{(1)} = 0_{\mathbb{R}^{K \times \tilde{T}}}$, $W^{(1)} = 0_{\mathbb{R}^{K \times \tilde{T}}}$, $\beta^{(1)} = 1$;
for $m = 1, \dots, M$ **do**
    $Z^{(m+1)} = \mathrm{ST}_{\frac{\lambda}{\mathcal{L}}} \left(W^{(m)} - \frac{1}{\mathcal{L}} \mathbf{D}^\top \left(\mathbf{D} * W^{(m)} - X \right) \right)$;
    $\beta^{(t+1)} = \frac{1 + \sqrt{1 + 4\left(\beta^{(m)}\right)^2}}{2}$;
    $W^{(m+1)} = Z^{(m+1)} + \frac{\beta^{(m)} - 1}{\beta^{(m+1)}} \left(Z^{(m+1)} - Z^{(m)} \right)$;
end
return $Z^{(M+1)}$

To do so, one needs to compute the gradient of the following function:

$$L(Z) = \frac{1}{2} \left\| X - \sum_{k=1}^{K} \mathbf{z}_k * D_k \right\|_F^2 \quad . \tag{B.1}$$

First, let us focus on the univariate case with one single atom. We have the following result.

PROPOSITION B.1. *For $\mathbf{x} \in \mathbb{R}^T$, $\mathbf{z} \in \mathbb{R}^{T-L+1}$ and $\mathbf{d} \in \mathbb{R}^L$, the gradient of the loss function $L(\mathbf{z}) = \frac{1}{2} \|\mathbf{x} - \mathbf{z} * \mathbf{d}\|_2^2$ is:*

$$\nabla L(\mathbf{z}) = (\mathbf{z} * \mathbf{d} - \mathbf{x}) * \mathbf{d}^{\curvearrowleft} \in \mathbb{R}^{T-L+1} \quad , \tag{B.2}$$

where $\mathbf{d}^{\curvearrowleft}$ is obtained by reversal of the temporal dimension, i.e., $\mathbf{d}^{\curvearrowleft}[t] = \mathbf{d}[L-1-t]$.

Proof B.1

The loss function can be rewritten as:

$$L(\mathbf{z}) = \frac{1}{2} \sum_{t=0}^{T-1} \left(\mathbf{x}[t] - \sum_{\tau=0}^{L-1} \mathbf{z}[t-\tau] \mathbf{d}[\tau] \right)^2 \quad . \tag{B.3}$$

Taking the derivative with respect to $\mathbf{z}[t], t \in [\![0\,, T-L]\!]$*, we have:*

$$\frac{\partial L(\mathbf{z})}{\partial \mathbf{z}[t]} = -\sum_{l=0}^{L-1} \frac{\partial (\mathbf{z} * \mathbf{d})[t+l]}{\partial \mathbf{z}[t]} \left(\mathbf{x}[t+l] - \sum_{\tau=0}^{L-1} \mathbf{z}[t+l-\tau]\mathbf{d}[\tau] \right) \tag{B.4}$$

$$= \sum_{l=0}^{L-1} \left(\mathbf{z} * \mathbf{d} - \mathbf{x}\right)[t+l]\mathbf{d}[l] \tag{B.5}$$

$$= \sum_{l=0}^{L-1} \left(\mathbf{z} * \mathbf{d} - \mathbf{x}\right)[t+l]\mathbf{d}^{\text{↰}}[L-1-l] \tag{B.6}$$

$$= \sum_{\tau=0}^{L-1} \left(\mathbf{z} * \mathbf{d} - \mathbf{x}\right)[t+L-1-\tau]\mathbf{d}^{\text{↰}}[\tau], \quad \tau \coloneqq L-1-l \tag{B.7}$$

$$= \left((\mathbf{z} * \mathbf{d} - \mathbf{x}) * \mathbf{d}^{\text{↰}} \right)[t+L-1] \tag{B.8}$$

$$\tag{B.9}$$

as $\frac{\partial (\mathbf{z}*\mathbf{d})[t+l]}{\partial \mathbf{z}[t]} = \mathbf{d}[l]$ *using the definition of the convolution expressed in eq.* (2.3.1)*. Thus, we retrieve the result in eq.* (B.2)*.*

We can then extend this result to the multi-atoms case.

Proposition B.2. *For* $\mathbf{x} \in \mathbb{R}^T$*,* $Z \in \mathbb{R}^{K \times (T-L+1)}$ *and* $D \in \mathbb{R}^{K \times L}$*, the gradient of the loss function* $L(Z) = \frac{1}{2} \left\| \mathbf{x} - \sum_{k=1}^{K} \mathbf{z}_k * \mathbf{d}_k \right\|_2^2$ *is:*

$$\nabla L(Z) = \left(\sum_{k'=1}^{K} \mathbf{z}_{k'} * \mathbf{d}_{k'} - \mathbf{x} \right) * D^{\text{↰}} \in \mathbb{R}^{K \times (T-L+1)} \quad , \tag{B.10}$$

where $\forall t \in [\![0\,, L-1]\!]\,, \mathbf{d}^{\text{↰}}[t] = \mathbf{d}[L-1-t] \in \mathbb{R}$*.*

Proof B.2

We have, by definition of the gradient, that

$$\nabla L(Z) = \left(\frac{\partial L(Z)}{\partial \mathbf{z}_k} \right)_{k=1,\dots,K} \in \mathbb{R}^{K \times (T-L+1)} \quad . \tag{B.11}$$

The loss function can be rewritten as:

$$L(Z) = \frac{1}{2} \sum_{t=0}^{T-1} \left(\mathbf{x}[t] - \sum_{k'=1}^{K} \sum_{\tau=0}^{L-1} \mathbf{z}_{k'}[t-\tau]\mathbf{d}_{k'}[\tau] \right)^2 \quad . \tag{B.12}$$

Taking the derivative with respect to $\mathbf{z}_k[t], t \in [\![0\,, T-L]\!]$*, we have, similarly as the proof of Proposition B.1:*

$$\frac{\partial L(Z)}{\partial \mathbf{z}_k[t]} = \left(\left(\sum_{k'=1}^{K} \mathbf{z}_{k'} * \mathbf{d}_{k'} - \mathbf{x} \right) * \mathbf{d}_k^{\curvearrowleft} \right) [t + L - 1] \tag{B.13}$$

Hence,

$$\nabla L(Z) = \begin{bmatrix} \left(\left(\sum_{k'=1}^{K} \mathbf{z}_{k'} * \mathbf{d}_{k'} - \mathbf{x} \right) * \mathbf{d}_1^{\curvearrowleft} \right)^\top \\ \vdots \\ \left(\left(\sum_{k'=1}^{K} \mathbf{z}_{k'} * \mathbf{d}_{k'} - \mathbf{x} \right) * \mathbf{d}_K^{\curvearrowleft} \right)^\top \end{bmatrix} \in \mathbb{R}^{K \times (T-L+1)} \,. \tag{B.14}$$

Finally, using the definition of convolution between a vector and a matrix in eq. (2.4.1), we obtained the desired result.

Finally, we have the following result for the multivariate general case.

PROPOSITION B.3. *For* $X \in \mathbb{R}^{P \times T}$*,* $Z \in \mathbb{R}^{K \times (T-L+1)}$ *and* $\mathbf{D} \in \mathbb{R}^{K \times P \times L}$*, the gradient of the loss function* $L(Z) = \frac{1}{2} \left\| X - \sum_{k=1}^{K} \mathbf{z}_k * D_k \right\|_F^2$ *is:*

$$\nabla L(Z) = (Z * \mathbf{D} - X) * \mathbf{D}^{\curvearrowleft} \in \mathbb{R}^{K \times (T-L+1)} \,, \tag{B.15}$$

where $Z * \mathbf{D} \coloneqq \sum_{k=1}^{K} \mathbf{z}_k * D_k$

Proof B.3

We have, by definition of the gradient, that

$$\nabla L(Z) = \left(\frac{\partial L(Z)}{\partial \mathbf{z}_k} \right)_{k=1,\dots,K} \in \mathbb{R}^{K \times (T-L+1)} \,. \tag{B.16}$$

The loss function can be rewritten as:

$$L(Z) = \frac{1}{2} \sum_{t=0}^{T-1} \sum_{p=0}^{P-1} \left(X_p[t] - \sum_{k'=1}^{K} \sum_{\tau=0}^{L-1} \mathbf{z}_{k'}[t-\tau] D_{k',p}[\tau] \right)^2 \,. \tag{B.17}$$

Taking the derivative with respect to $\mathbf{z}_k[t], t \in [\![0\,, T-L]\!]$*, we have:*

$$\frac{\partial L(Z)}{\partial \mathbf{z}_k[t]} = -\sum_{l=0}^{L-1}\sum_{p=0}^{P} \frac{\partial (\mathbf{z}_k * D_{k,p})[t+l]}{\partial \mathbf{z}_k[t]} \left(X_p[t+l] - \sum_{k'=1}^{K}\sum_{\tau=0}^{L-1} \mathbf{z}_{k'}[t+l-\tau] D_{k',p}[\tau] \right) \tag{B.18}$$

$$= \sum_{l=0}^{L-1}\sum_{p=0}^{P} \left(\sum_{k'=1}^{K} \mathbf{z}_{k'} * D_{k',p} - X_p \right)[t+l] D_{k,p}[l] \tag{B.19}$$

$$= \sum_{l=0}^{L-1} \left\langle \left(\sum_{k'=1}^{K} \mathbf{z}_{k'} * D_{k'} - X \right)[t+l], D_k[l] \right\rangle \tag{B.20}$$

$$= \sum_{\tau=0}^{L-1} \left\langle \left(\sum_{k'=1}^{K} \mathbf{z}_{k'} * D_{k'} - X \right)[t+L-1-\tau], D_k^{\curvearrowleft}[\tau] \right\rangle, \quad \tau \coloneqq L-1-l \tag{B.21}$$

$$= \left(\left(\sum_{k'=1}^{K} \mathbf{z}_{k'} * D_{k'} - X \right) * D_k^{\curvearrowleft} \right)[t+L-1] \tag{B.22}$$

Hence,

$$\nabla L(Z) = \begin{bmatrix} \left(\left(\sum_{k'=1}^{K} \mathbf{z}_{k'} * D_{k'} - X \right) * D_1^{\curvearrowleft} \right)^\top \\ \vdots \\ \left(\left(\sum_{k'=1}^{K} \mathbf{z}_{k'} * D_{k'} - X \right) * D_K^{\curvearrowleft} \right)^\top \end{bmatrix} \in \mathbb{R}^{K \times (T-L+1)} \,. \tag{B.23}$$

Finally, with a slight abuse of notation, we have that:

$$\nabla L(Z) = \left(\sum_{k=1}^{K} \mathbf{z}_k * D_k - X \right) * \mathbf{D}^{\curvearrowleft} \tag{B.24}$$

We can now exhibit the pseudo-code of the adapted FISTA that will output the different $(\mathbf{Z}^n)^{(M)}(\mathbf{D})$ presented in algorithm 6, where $\mathcal{L}$ denotes the Lipschitz constant of $\mathbf{D}^\top * \mathbf{D}$. Note that this algorithm is performed independently for every recording $n = 1, \ldots, N$. In practice, we observed that choosing between 20 and 30 iterations is enough to get accurate results on MEG data.

C FaDIn – Additional experiments

This section presents additional experimental results supporting the claims of chapter 5. We first compare the ℓ_2 loss involved in FaDIn with the popular negative log-likelihood, then additional comparisons with popular non-parametric approaches are presented.

C.1 Comparison of FaDIn with the negative log-likelihood loss

We compare both approaches' statistical and computational efficiency to highlight the benefit of using the ℓ_2 loss in FaDIn over the log-likelihood (LL). Precisely, we compare the accuracy of the obtained parameter estimators from FaDIn and the minimization of the negative log-likelihood in the same setting as our approach (discretization and finite-support kernels). We conduct the experiment as follows. We place ourselves in the univariate setting for computational simplicity. We sample a set of events in continuous time through the `tick` library. Three sets are sampled from the kernel shapes: Raised Cosine, Truncated Gaussian, and Truncated Exponential. The parameters are set as $\mu = 0.3$, $\alpha = 0.8$, $(u, \sigma) = (0.2, 0.3)$ for the Raised Cosine, $(m, \sigma) = (0.5, 0.3)$ for the Truncated Gaussian and $\gamma = 5$ for the Truncated Exponential. We set the kernel length W to 1 for each setting. Further, we estimate the parameters of the intensity of sampled events using both FaDIn and LL approaches. The experiment is repeated ten times. The median and 25-75% quantiles of the statistical accuracy and the computation time are reported in Figure C.1 for the three different kernels. We can observe an equivalent accuracy of the parameter estimation for both methods along the different kernels, stepsize and number of events. In contrast, the computational performance of FaDIn outperforms the LL approach. Indeed, the computational time is divided by ≈ 5 in a low data regime with $T = 10^2$ and by ≈ 1000 when $\Delta = 0.01$ and $T = 10^5$. This experiment clearly shows the advantages of using the ℓ_2 loss in FaDIn rather than the log-likelihood.

C.2 Qualitative Comparison with a non-parametric approach

We compare FaDIn with the use of a non-parametric kernel by assessing the statistical and computational efficiency of both approaches. To learn the non-parametric kernel, we select the EM algorithm, provided in Zhou et al. [2013a] and implemented in the `tick` library [Bacry et al., 2017a]. The kernel is set with one basis

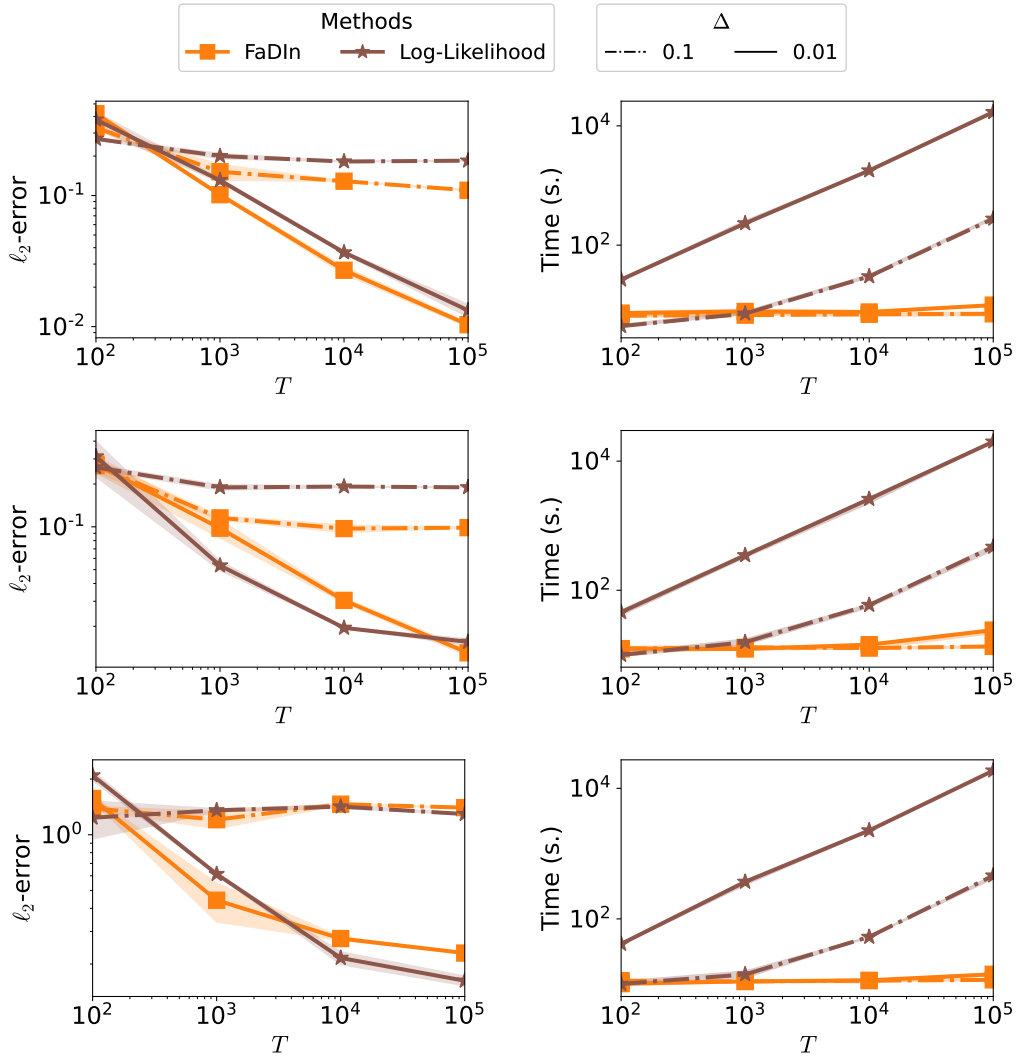


Figure C.1: Comparison of the statistical and computational efficiency of FaDIn with the Log-Likelihood loss. The averaged (over ten runs) statistical error on the intensity function (left) and the computational time (right) are computed regarding the time T for the Raised Cosine (top), the Truncated Gaussian (middle) and the Truncated Exponential (bottom).

function. In addition, we display the running time when computing gradients using PyTorch and automatic differentiation applied to the $\mathcal{L}_G$ discretized loss (5.1.4).

The experiment is conducted as follows. We fix $p = 1$ for simplicity, set $\mu = 1.1$ and choose a Raised Cosine kernel defined by eq. (5.2.1), setting parameters $\alpha = 0.8$, $u = 0.2$ and $\sigma = 0.3$. We simulate events in a continuous time using the `tick` library [Bacry et al., 2017a]. FaDIn and the non-parametric kernel are optimized over 800 iterations (with an early stopping for the EM algorithm). The RMSprop

algorithm is used in FaDIn. The discretization size of the non-parametric kernel is settled as in FaDIn. This experiment is done varying $T \in \{10^3, 10^5, 10^6\}$.

On the one hand, in a relatively small data regime where $T = 10^3$, we evaluate the statistical accuracy of the estimated kernel of both methods with the discretization parameter $\Delta = 0.01$. As we can see in Figure C.2 (top left), the non-parametric approach fails to recover the structure of the kernel. The non-parametric approach results in noisy kernel estimates, with probability mass where the kernel is zero. In contrast, FaDIn can recover the kernel parameters used to simulate data even with a small number of events. On the other hand, we evaluate the computational times varying the discretization steps in a large data regime where $T = 10^5$ and $T = 10^6$ with the same simulation parameters. Figure C.2 (bottom left) reports the average computational times (over 10 runs) regarding the discretization stepsize Δ and the dimension p. Although both approaches can recover the kernel under which we simulate data (see Figure C.2, top right), FaDIn is a great deal more computationally efficient than the non-parametric and the automatic differentiation implementations, improving the computational speed by ≈ 100 when $\Delta \in [0.1, 0.01]$ and by ≈ 10 when $\Delta \approx 0.001$. The computation speed regarding the dimension of the MHP is improved by ≈ 10. It is worth noting that the ℓ_2-Autodiff explodes in memory when $\Delta > 0.01$ or when the dimension grows. Additional shapes of kernels are displayed in Figure C.3 for the Truncated Gaussian and in Figure C.4 for the Truncated Exponential kernels.

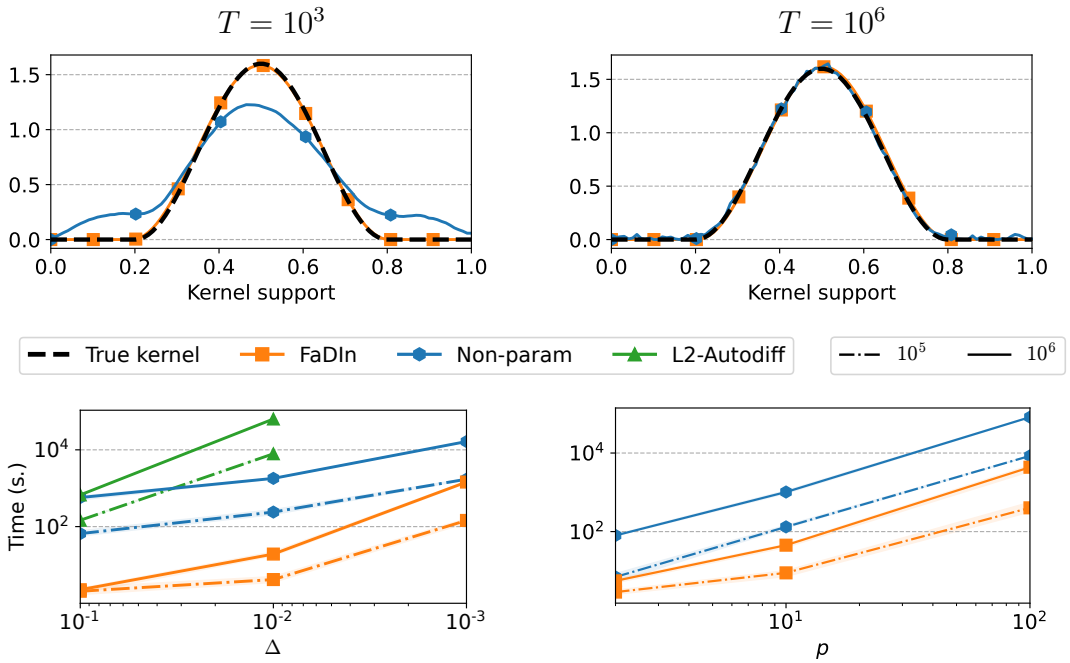


Figure C.2: Comparison between our approach FaDIn and non-parametric approach. Estimated kernels with $\Delta = 0.01$ in a relatively small data setting with $T = 10^3$ (top left), in a large data setting with $T = 10^6$ (top right), and computation time in a large data setting with $T \in \{10^5, 10^6\}$ w.r.t. the stepsize Δ (bottom left) and the dimension p (bottom right). In contrast to non-parametric kernels, FaDIn estimates well the true kernel in a small regime while it is computationally faster than non-parametric kernels in a large regime.

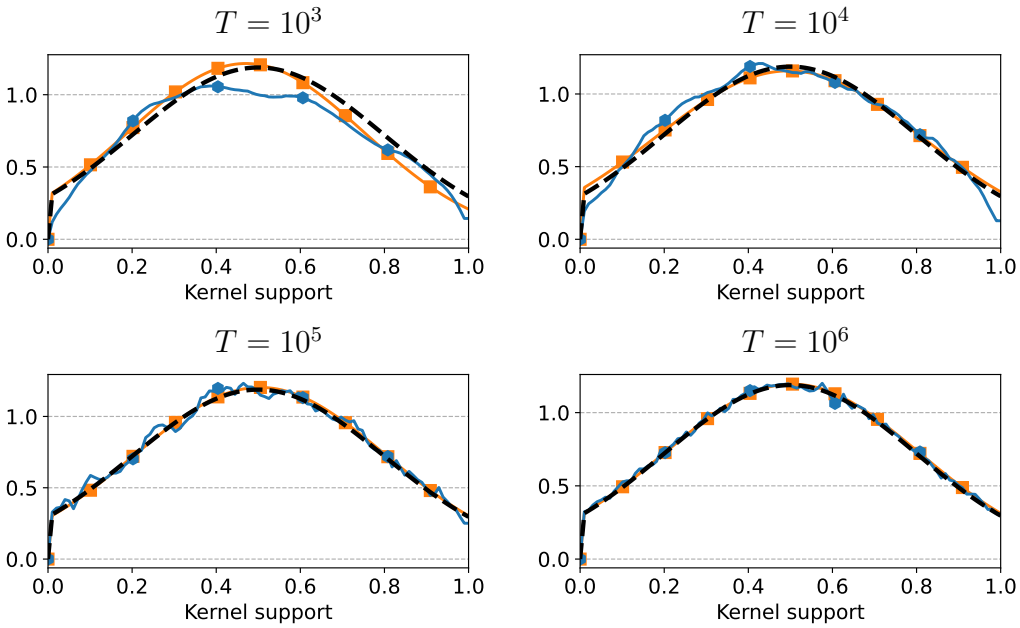


Figure C.3: Comparison between our approach FaDIn and non-parametric approach for a Truncated Gaussian kernel. Estimated kernels with $\Delta = 0.01$ and $T \in \{10^3, 10^4, 10^5, 10^6\}$. The true kernel, FaDIn and the non-parametric approach are depicted in black, orange and blue, respectively.

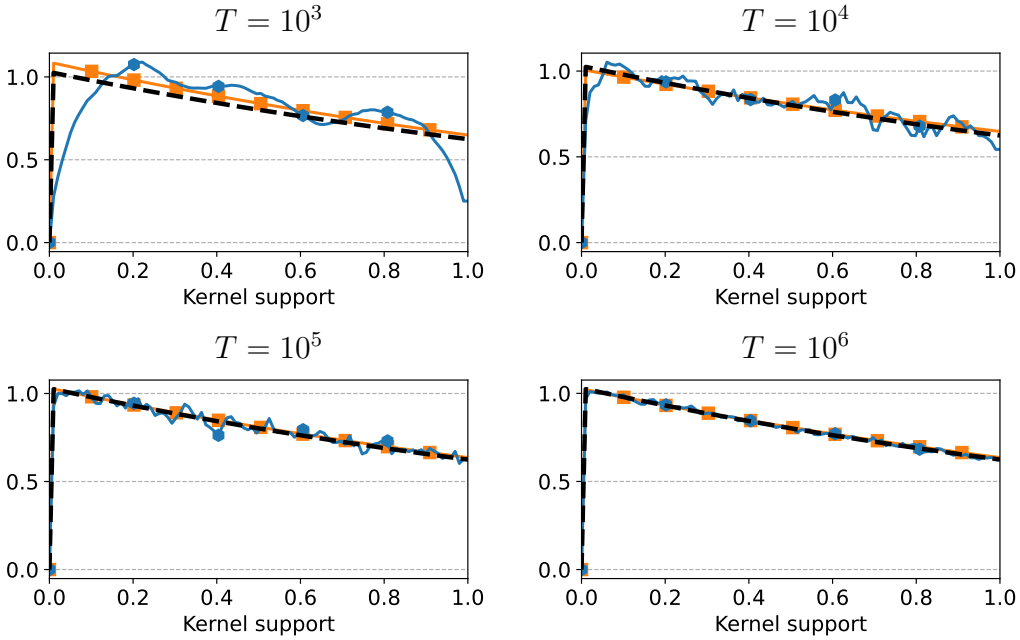


Figure C.4: Comparison between our approach FaDIn and non-parametric approach for a Truncated Exponential kernel. Estimated kernels with $\Delta = 0.01$ and $T \in \{10^3, 10^4, 10^5, 10^6\}$. The true kernel, FaDIn and the non-parametric approach are depicted in black, orange and blue, respectively.