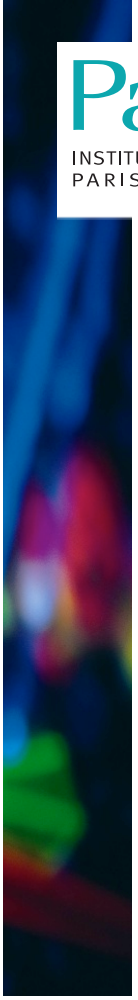
Advances in automating analysis of neural time series data

Mainak Jas



ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2018-ENST-0021

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Signal et Images »

présentée et soutenue publiquement par

Mainak Jas

le 12 Avril 2018

Contributions pour l'analyse automatique de signaux neuronaux

Directeur de thèse : **Alexandre GRAMFORT**

Jury

M. Bradley Voytek, Professeur associé, UC San Diego, Etats-Unis	Rapporteur
M. Arnaud Delorme, Directeur de Recherche, CRNS, France	Rapporteur
Mme. Nathalie George, Directeur de Recherche, CNRS, France	Examinatrice
M. Marco Congedo, Directeur de Recherche, CNRS, France	Examinateur
M. Cédric Gouy-Pailler, Chargé de Recherche, CEA, France	Examinateur
M. Gérard Dreyfus, Professeur émérite, ESPCI, France	Président

TELECOM ParisTech
école de l'Institut Mines-Télécom - membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

THÈSE

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2018-ENST-0021

EDITE - ED 130

Doctorat ParisTech

T H E S I S

submitted in partial fulfilment of the requirements for PhD at

TELECOM ParisTech

Specialized in Image, Data, and Signal Processing

presented and defended publicly by

Mainak JAS

April 2018

Advances in automating analysis of neural time series data

Doctoral advisor: **Alexandre GRAMFORT**

Jury

Mr. Bradley VOYTEK, Prof., UC San Diego	Referee
Mr. Arnaud DELORME, Dr., CNRS	Referee
Ms. Nathalie GEORGE, Dr., CNRS	Examiner
Mr. Marco CONGEDO, Dr., CNRS	Examiner
Mr. Cédric GOUY-PAILLER, CEA	Examiner
Mr. Gerard Dreyfus, Emer. Prof., ESPCI	President

TELECOM ParisTech
école de l'Institut Mines-Télécom - membre de ParisTech

46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

THÈSE

This dissertation is dedicated to the memory
of my friend and colleague Venkat Raghav Rajagopalan (1993 – 2017).

Acknowledgements

First and foremost, my gratitude goes out to Alexandre Gramfort: my advisor, mentor and friend. Through these years, Alex has inspired me with his technical knowledge, his vision for long term impact, philosophy of open science, and wisdom. He introduced me to the welcoming and progressive MNE community which ultimately formed my collaboration network. It is these interactions that often led to new research projects and ideas. A prime example of this is my work on *autoreject* which developed out of my collaboration with Denis Engemann, a core contributor to MNE. It has been a pleasure working with Denis, as he shared his knowledge on the subtleties of MEG signal analysis and motivated me through our mini coding sprints. I would like to express my heartfelt thanks to Riitta Hari, Lauri Parkkonen, and Pavan Ramkumar who introduced me to the world of machine learning in neuroscience, and offered words of support and encouragement over the years.

My gratitude goes out to Matti Hamalainen and Stephanie Jones for inviting me to Boston on our collaborative project. I thank my friend and former colleague, Teon Brooks for working with me on the BIDS project and on the realtime module in MNE, Eric Larson for sharing his extensive experience in MEG and signal processing, and the rest of the MNE team for making me believe in the power of open source. I am also grateful to Chris Gorgolewski and Russ Poldrack for inviting me to the BIDS coding sprint at Stanford. The visit was a very stimulating experience for me as well as a unique networking opportunity. I thank my advisor Alex, and Robert Gower for allowing me to take their optimization course, and for being incredibly patient with my questions. I am sure the skills I learned here will be useful for many years to come.

To my co-authors, I thank them for being so tolerant and supportive in our common projects. None of my papers could have been published without the teamwork and efforts put in by them. I am grateful to Umut Şimşekli and Tom Dupré la Tour for spending their weekends and evenings with me to push forward our sparse coding project. I have benefited a lot from my collaboration with Jaakko Leppakangas who is perhaps one of the most efficient and productive engineers I have met. I am particularly indebted to my colleague, Yousra Bekhti for discussing tricky bugs at work, but more importantly for easing my difficulties with the French administration and with translation. I am grateful to Jean-Baptiste Schiratti for many stimulating technical conversations, and also for helping me with constructive feedback on my writing.

I thank my office mates for sharing the weekly seminars, the Monday cakes, and the latest free food event. I thank my collaborators and friends at the Neurospin laboratory for inviting me to their social events. It is here that I made many like-minded international

friends. To my French teacher Françoise and our Tuesday French lunch group, I am immensely indebted. I thank my international friends: Bianca, David, Fosca, Gabriela, Magdalena, Sokhany, and Sophie. I thank my neighbours, my climbing partners, hiking group, and the *desi* Indian community for making me feel home in France: Aakanksha, Anshuman, Chirag, Nilesh, Niraj, Pratheeban, Praveer, Raghav, Shabbir, Sidharth, and Vamsi.

All of this work would not have been possible without my family. My sister continues to surprise and inspire me with her sense of humour and adventure. Finally, I thank my parents for their continued moral support and motivation.

April 12, 2018
Paris

Abstract

Electrophysiology experiments has for long relied upon small cohorts of subjects to uncover statistically significant effects of interest. However, the low sample size translates into a low power and hence a low rate of reproducibility. To address this issue means solving two related problems: first, how do we facilitate data sharing and reusability to build large datasets; and second, once big datasets are available, what tools can we build to analyze them?

In the first part of the thesis, we introduce a new data standard for sharing data known as the Brain Imaging Data Structure (BIDS), and its extension MEG-BIDS. Next, we introduce the reader to a typical electrophysiological pipeline analyzed with the MNE software package. This will orient them towards the analysis process and the challenges that are often faced when building reproducible pipelines. We consider the different choices that users have to deal with at each stage of the pipeline and provide standard recommendations.

Next, we focus on tools to automate analysis of large datasets such as those offered by the Human Connectome Project (HCP). We propose an automated tool to remove segments of data that are corrupted by artifacts. We develop an outlier detection algorithm based on tuning rejection thresholds with parameter search using Bayesian optimization. More importantly, we use the HCP data, which is manually annotated, to benchmark our algorithm against existing state-of-the-art methods. To our knowledge, this represents the first instance of reanalyzing the dataset using an independent stack of tools as used by the HCP consortium.

Finally, we use convolutional sparse coding to uncover structures in neural time series. The method we propose is inspired by similar algorithms in computer vision, where the goal is to learn the coefficients of a dictionary of atoms (traditionally sinusoidal or wavelets) but also the atoms themselves. We reformulate the existing approach as a maximum a posteriori (MAP) inference to be able to deal with high amplitude artifacts and the heavy tailed noise distributions that is so common in neural time series.

Taken together, this thesis represents an attempt to shift from slow and manual methods of analysis to automated, reproducible analysis.

Keywords: Automation, data sharing, reproducibility, sparse coding, outlier detection, representation learning, electroencephalography, magnetoencephalography

Résumé

Les expériences d'électrophysiologie ont longtemps reposé sur de petites cohortes de sujets pour découvrir des effets statistiquement significatifs d'intérêt. Toutefois, la faible taille de l'échantillon se traduit par une faible puissance statistique, ce qui entraîne un taux élevé de fausses découvertes et, par conséquent, un faible taux de reproductibilité. Pour résoudre ce problème, il faut résoudre deux problèmes : premièrement, comment faciliter le partage et la réutilisation des données pour créer de grands ensembles de données; et deuxièmement, une fois que de grands ensembles de données sont disponibles, quels outils pouvons-nous construire pour les analyser?

Dans la première partie de la thèse, nous introduisons une nouvelle norme de données pour le partage des données connue sous le nom de Brain Imaging Data Structure (BIDS), et son extension MEG-BIDS. Ensuite, nous présentons au lecteur un pipeline typique d'analyse de données d'électrophysiologie analysé avec le progiciel MNE. Cela les orientera vers le processus d'analyse et les défis qui sont souvent rencontrés lors de la construction de pipelines reproductibles. Nous tenons compte des différents choix que les utilisateurs doivent faire à chaque étape du pipeline et nous formulons des recommandations normalisées.

Ensuite, nous concentrons notre attention sur les outils permettant d'automatiser l'analyse de grands ensembles de données tels que ceux offerts par le Human Connectome Project (HCP). Nous proposons un outil automatisé pour supprimer les segments de données corrompus par des artefacts. Nous développons un algorithme de détection d'anomalies basé sur le réglage des seuils de rejet avec recherche de paramètres par optimisation bayésienne. Plus important encore, nous utilisons les données HCP, qui sont annotées manuellement, pour comparer notre algorithme aux méthodes de pointe existantes. À notre connaissance, il s'agit du premier cas de réanalyse de l'ensemble de données à l'aide d'outils indépendants de ceux utilisés par le consortium HCP.

Enfin, nous utilisons le codage convolutionnel parcimonieux pour découvrir les structures des séries chronologiques neuronales. La méthode que nous proposons s'inspire d'algorithmes similaires en vision par ordinateur, où le but est d'apprendre les coefficients d'un dictionnaire d'atomes (traditionnellement sinusoïdaux ou ondelettes) mais aussi les atomes eux-mêmes. Nous reformulons l'approche existante comme une inférence MAP pour être en mesure de faire face aux artefacts d'amplitude élevée et aux distributions de bruits lourds qui sont si courantes dans les séries chronologiques neuronales.

Dans son ensemble, cette thèse représente une tentative de passer de méthodes d'analyse lentes et manuelles à des méthodes d'analyse automatisées et reproductibles.

Contents

List of Figures

List of Tables

List of Algorithms

Abbreviations

ADMM alternating direction method of multipliers

BCI brain-computer interface

BEM boundary element model

BFGS Broyden-Fletcher-Goldfarb-Shanno

BIDS Brain Imaging Data Structure

CFC cross-frequency coupling

CSC convolutional sparse coding

CTPS cross-trial phase statistics

dSPM dynamic statistical parameter mapping

ECG electrocardiography

ECoG electrocorticography

EEG electroencephalography

EM expectation maximization

EMG electromyogram

EOG electrooculogram

ERF event-related field

ERP event-related potential

FDR false discovery rate

FFA fusiform face area

FISTA fast iterative soft thresholding algorithm

FLASH fast low-angle shot

fMRI functional magnetic resonance imaging

fNIRS functional near-infrared spectroscopy

GFP global field power

GP Gaussian process

HCP Human Connectome Project

HPI head position indicator

ICA independent component analysis

ISTA iterative soft thresholding algorithm

JSON JavaScript object notation

LCMV linearly constrained minimum variance

LFP local field potential

MAP maximum a posteriori

MCEM Monte Carlo expectation maximization

MCMC Markov chain Monte Carlo

MEG magnetoencephalography

MH Metropolis-Hastings

MNE minimum norm estimate

MRI magnetic resonance imaging

OFA occipital face area

PCA principal component analysis

PDF probability density function

PSD power spectral density

SNR signal-to-noise ratio

SSP signal space projection

SSS signal space separation

STS superior temporal sulcus

TSV tab-separated value

Chapter 1

Sommaire (en Français)

Comprendre le cerveau humain est l'un des défis les plus importants du 21e siècle. Sans doute l'organe le plus complexe du corps humain, le cerveau est responsable d'un large éventail de fonctions cognitives telles la reconnaissance visuelle, la compréhension du langage, la production de la parole, les interactions sociales et le contrôle exécutif. Les pathologies associées au cerveau demeurent, à ce jour, une problématique extrêmement complexe. Actuellement, les interventions médicales pour les principales maladies infectieuses permettent aux individus de vivre jusqu'à l'âge de 80 et même 90 ans. Malgré les avancées de la médecine, nous n'arrivons pas à cibler efficacement les mécanismes qui engendrent ou qui contribuent à la progression des pathologies mentales, telles le Parkinson, la démence d'Alzheimer, la schizophrénie et l'épilepsie, pour n'en nommer que quelques-unes. Ceci a de graves conséquences sur la société vieillissante, par exemple une personne dans la quarantaine a 50% de risque de développer la maladie d'Alzheimer (Alzheimer's Association et al., 2016).

Néanmoins, notre compréhension actuelle du cerveau est le résultat de décennies d'efforts concertés dans de multiples disciplines allant de la biologie moléculaire, à la génétique et la physiologie, en passant par les neurosciences cognitives et comportementales, la statistique, l'informatique et la science des données. Un sous-domaine en expansion est l'imagerie cérébrale, également connue sous le nom de neuroimagerie. L'imagerie cérébrale fait référence à un ensemble de technologies où l'on mesure l'activité instantanée du cerveau. Ces mesures peuvent être statiques, comme dans le cas des images anatomiques issues de l'imagerie par résonance magnétique (IRM), ou dynamique, comme dans le cas de l'imagerie par résonance magnétique fonctionnelle (IRMf). L'objectif à long-terme serait d'utiliser ces techniques dans les hôpitaux pour aider au niveau du diagnostic et dans les chirurgies, dans les interface neuronale directe (IND) et dans la recherche en neurosciences. Cette thèse sera dédiée à la mesure des courants électriques et/ou du champ magnétique du cerveau par l'entremise de l'électroencéphalographie (EEG), de la magnétoencéphalographie et les potentiels de champ local (LFP). Ces méthodes ont la propriété de posséder une résolution temporelle élevée,ce qui est particulièrement utile pour extraire la dynamique temporelle des signaux du cerveau.

Dans cette thèse, j'élabore sur l'optimisation des analyses de données publiques en neuroimagerie , et ce, au moyen des logiciels libres. À cet effet, j'ai participé à une collaboration internationale pour créer une norme sur les analyses en MEG pour la

structure de données d'imagerie cérébrale (BIDS) (Niso et al., 2018). J'ai écrit le validateur qui a aidé à créer les exemples d'ensembles de données compatibles MEG-BIDS. En tant que collaborateur de MNE (Gramfort et al., 2013a), j'ai dirigé la rédaction d'un tutoriel qui permet d'analyser à nouveau un ensemble de données, nommé Faces (Wakeman and Henson, 2015), pour une reproduire les résultats préalablement retrouvés.Dans le contexte du mouvement de reproductibilité et de partage des données, nous avons commencé à automatiser nos séquences d'étapes d'analyses (pipelines), ce qui nous a amené à développer un algorithme entièrement automatisé pour le rejet et la réparation des artefacts en EEG/MEG (Jas et al., 2016, 2017a). Enfin, nous avons développé des algorithmes pour permettre l'apprentissage automatique de nouveaux motifs (patterns) cérébraux qui n'ont pas pu être démontré par l'utilisation de données extraites de séries temporelles neuronales (Jas et al., 2017c).

La thèse est organisée par chapitres pour mettre en évidence ces quatre principaux domaines de contribution : le partage des données, la reproductibilité, l'automatisation de la détection des artefacts et la découverte automatisée des motifs cérébraux à partir des données EEG/MEG. Un aspect important de cette thèse est que ces contributions ont conduit non seulement à des publications dans des conférences et revues internationales, mais aussi à des implémentations open source reproductibles et à des ensembles de données réutilisables. Une liste complète est donnée ci-dessous.

Publications dans les revues

M. Jas, D. A. Engemann, Y. Bekhti, F. Raimondo, and A. Gramfort. Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, 159:417–429, 2017a

M. Jas, E. Larson, D. A. Engemann, J. Leppakangas, S. Taulu, M. Hamalainen, and A. Gramfort. MEG/EEG group study with MNE: recommendations, quality assessments and best practices. *bioRxiv*, 2017b. doi: 10.1101/240044
(Pending revision at *Frontiers in Neuroscience, Brain Imaging Methods*)

G. Niso, K. J. Gorgolewski, E. Bock, T. L. Brooks, G. Flandin, A. Gramfort, R. N. Henson, M. Jas, V. Litvak, J. T. Moreau, et al. MEG-BIDS, the brain imaging data structure extended to magnetoencephalography. *Scientific data*, 5:180110, 2018

Publications dans les conférences

M. Jas, D. Engemann, F. Raimondo, Y. Bekhti, and A. Gramfort. Automated rejection and repair of bad trials in MEG/EEG. In *6th International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 2016

M. Jas, L. Tour, T. Dupré, U. Şimşekli, and A. Gramfort. Learning the morphology of brain signals using alpha-stable convolutional sparse coding. In *Advances in Neural Information Processing Systems (NIPS)*, 2017c

Documents d'atelier

D. Engemann, F. Raimondo, J. King, M. Jas, A. Gramfort, S. Dehaene, L. Naccache,

and J. Sitt. Automated measurement and prediction of consciousness in vegetative state and minimally conscious patients. In *Workshop on Statistics, Machine Learning and Neuroscience at the International Conference on Machine Learning (ICML)*, Lille, July 2015a

Implémentations Open Source

http://autoreject.github.io/
http://alphacsc.github.io/
http://mne-tools.github.io/mne-biomag-group-demo/
https://jasmainak.github.io/bids-validator/

Datasets

https://openfmri.org/dataset/ds000248/

Dans le résumé en français, je présenterai le contexte de la thèse, suivi d'un bref résumé de chacun des quatre principaux domaines de contribution.

Électrophysiologie

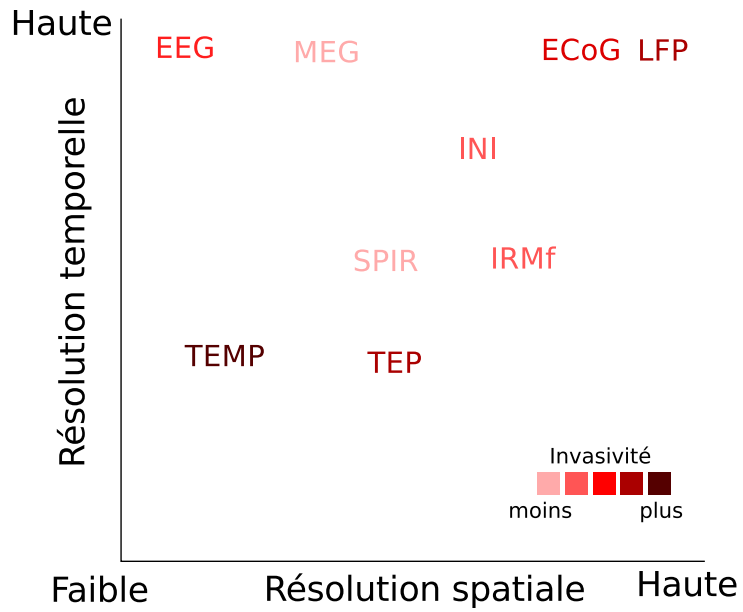


Figure 1.1: Diverses méthodes de neuroimagerie diffèrent en termes d'information qu'elles mesurent. MEG=magnétoencéphalographie, EEG=électroencéphalographie, SPIR=spectroscopie par proche infrarouge, TEP=tomographie par émission de positons, TEMP=tomographie par émission de photons simples, et INI=imagerie inverse, une méthode pour accélérer l'acquisition d'images IRMf, ECoG=Electrocorticographie, LFP=Local Field Potential.

Dans cette thèse, nous discuterons des données enregistrées principalement à l'aide de la MEG et de l'EEG. L'EEG et la MEG sont des méthodes d'électrophysiologie permettant d'étudier les propriétés des cellules et tissus biologiques. Les tissus biologiques ont des propriétés électriques dues à la présence d'ions. Tout comme nous pouvons mesurer les

tensions dans les appareils électriques, il est également possible de mesurer ces tensions dans les tissus vivants.

L'enregistrement électrophysiologique a l'avantage de mesurer directement l'activité cérébrale, par opposition à une mesure indirecte, ce qui est par exemple le cas de l'IRMf. Les techniques d'imagerie cérébrale se caractérisent par leur résolution temporelle et spatiale, c'est-à-dire l'échelle de temps à laquelle elle peut mesurer l'activité cérébrale, ainsi que la précision de la localisation de la source de l'activité, respectivement. La Figure 1.1 résume les différentes méthodes de neuroimagerie en ce qui concerne leur résolution temporelle et spatiale, mais aussi leur caractère invasif. Ceci est utile pour comprendre pourquoi la MEG/EEG est souvent considéré comme la méthode de choix pour comprendre les fondements de l'activité cérébrale humaine. Bien que l'IRMf soit une autre méthode non invasive, elle mesure le flux sanguin qui est une réponse lente en comparaison à l'activité neuronale ce qui lui confère une résolution temporelle peut élevée.

Il existe un certain nombre de méthodes pour mesurer les potentiels électriques dans le corps humain, la plus connue étant peut-être l'électrocardiographie (ECG) qui est utilisée pour mesurer l'activité électrique du cœur. Les trois méthodes dont nous discutons dans la thèse produisent des séries temporelles multivariées. Un bref résumé de chacune de ces méthodes est présenté ci-dessous.

Électroencéphalographie : L'EEG est une technique de mesure portable et non invasive inventée dans les années 1920 et utilisée dans plusieurs contextes tels que les IND, la surveillance médicale et dans l'établissement de diagnostic, ainsi que dans les études cognitives. En électroencéphalographie, un réseau d'électrodes sur un capuchon d'EEG est placé sur le cuir chevelu pour mesurer les tensions par rapport à une électrode de référence. La tension qu'il mesure n'est pas la résultante d'un seul neurone, mais plutôt le résultat de l'activité électrique de populations de neurones lui conférant une résolution temporelle élevée (de l'ordre des ms), quoique la résolution spatiale n'est pas si élevée.

Magnétoencéphalographie : Tout courant électrique est associé à des champs magnétiques en fonction de la théorie de Maxwell. En effet, le cerveau génère de minuscules champs magnétiques qui s'enroulent autour des courants électriques selon la règle de la main droit de Maxwell. Le champ magnétique du cerveau est minuscule ($\sim 10^{-12}T$) comparé au champ magnétique terrestre ($\sim 10^{-4}T$) et au bruit magnétique ambiant ($\sim 10^{-6}T$). . Par conséquent, pour le mesurer, on a besoin d'une mesure électronique très sensible et d'une annulation de bruit important. La mesure elle-même se fait dans une salle blindée magnétiquement, composée de trois couches de métaux. Les capteurs sont des bobines supraconductrices qui capturent le flux magnétique. Ils sont immergés dans de l'hélium liquide à de très basses températures (autour de 4 K), afin de réduire toute perte de signal due à la résistance. Un appareil typique contient deux types de capteurs : les gradiomètres et les magnétomètres. Tandis que le magnétomètre mesure l'amplitude absolue du champ magnétique, le gradiomètre mesure le gradient du champ. La MEG a l'avantage que le crâne ne détériore pas la qualité du signal contrairement à l'EEG.

Potentiel de champ local (LFP) : Le potentiel de champ local est le potentiel électrique qui est enregistré dans l'espace extracellulaire du tissu cérébral. Contrairement à l'EEG, les LFP sont enregistrés en profondeur, à partir du tissu cortical et peuvent donc mesurer

des populations de neurones plus localisées. De petites électrodes intracrâniennes sont généralement utilisées pour mesurer ces potentiels, contrairement aux électrodes de grande surface utilisées dans l'EEG.

Contexte de la thèse

La thèse se focalise sur les récents mouvements de reproductibilité et de partage des données en neuroimagerie. Elle met l'accent sur la simplification de l'analyse des données grâce à de meilleurs outils pédagogiques et à des méthodes automatisées permettant une analyse reproductible à l'ère des grandes données.

La crise de la reproductibilité

Même si des milliers d'articles sont publiés à chaque année sur différents aspects du cerveau, notre compréhension de cet organe complexe n'est pas proportionnelle. Une grande partie de la raison a été attribuée à ce que l'on appelle la crise de la reproductibilité (Ioannidis, 2005a; Simmons et al., 2011; Button et al., 2013). Les progrès de la science reposent sur des expériences reproductibles. La reproductibilité fait référence au fait que les résultats d'une expérience peuvent être régénérés, et ce, de manière indépendante,si le code, les données et les logiciels connexes ont été fournis. Dans de nombreux domaines, cependant, une grande partie des expériences ne peut être reproduite. En psychologie, par exemple, on a estimé que plus de la moitié des articles n'étaient pas reproductibles (Collaboration et al., 2015), et même ceux qui pouvaient être reproduits avaient tendance à avoir un effet plus faible par rapport aux études originales.

Les raisons pour lesquelles les résultats ne sont pas reproductibles peuvent être nombreuses (Baker, 2016), certaines étant : 1) le biais de confirmation, la tendance à ne rapporter sélectivement que les expériences conformes aux croyances préexistantes du chercheur, 2) le "p-hacking" (Simmons et al., 2011), ou la tendance à essayer de multiples hypothèses pour obtenir un résultat positif, 3) le biais de publication ou l'absence d'incitation à publier des résultats négatifs (Rosenthal, 1979), 4) corriger pour les comparaisons multiples, la méthode la plus conservatrice étant la correction de Bonferroni (Dunn, 1961) et 5) la pression à publier. Il existe maintenant un ensemble de recommandations acceptées pour régler bon nombre de ces questions.

L'imagerie cérébrale a ses propres problèmes qui peuvent être liés à la crise de la reproductibilité :

- **Manque de puissance statistique:** C'est sans doute l'une des questions centrales de la crise de la reproductibilité d'aujourd'hui et c'est de loin celle qui a reçu le plus d'attention. La puissance statistique d'une étude fait référence à la probabilité de découvrir un effet intéressant, compte tenu de la taille de l'échantillon. Les petites tailles d'échantillons se traduisent par des études sous-exploitées, ce qui signifie que le risque d'une fausse découverte est élevée. Afin de découvrir l'effet d'intérêt, l'étude doit être menée de façon appropriée.
- **Comparaisons multiples:** Il s'agit essentiellement d'une manifestation de "piratage" qui résulte du grand nombre de voxels ou de points dans le temps en neuroimagerie. Par exemple, dans la célèbre étude sur le saumon mort (Bennett,

2009), un effet significatif a été trouvé même si aucun effet n'était attendu simplement parce que les tests d'hypothèse (comparaisons) ont été effectués sur chaque voxel.

- **Différences dans les versions des logiciels :** Les changements de versions d'un logiciel peut conduire à des résultats différents. Par exemple, dans le cas du logiciel Freesurfer, les différences de volume étaient de l'ordre de $8.8\% \pm 6.6\%$ pour des versions différentes (Gronenschild et al., 2012).
- **Pipelines complexes :** Les pipelines de neuroimagerie impliquent un certain nombre de choix à chaque étape de traitement, et il n'existe actuellement aucun consensus sur la façon de choisir les bonnes étapes d'analyses. Souvent, ces choix méthodologiques ne sont même pas documentés. On estime qu'il y a presque autant de pipelines uniques que d'études (Carp, 2012b).
- **Facteurs confusionnels:** Il existe plusieurs facteurs confusionnels tels que les mouvements de la tête (Yendiki et al., 2014), les différences anatomiques et les changements dans la fréquence et la profondeur de la respiration, qui peuvent conduire à des corrélations fallacieuses.

Dans le Chapitre 4 de la thèse, nous fournirons des lignes directrices concrètes sur la façon de construire des pipelines de traitement des données MEG/EEG. Notre contribution abordera la question des pipelines complexes, des comparaisons multiples et des différences de versions de logiciels dans le cadre de la MEG/EEG. La question de la puissance statistique peut être resolue par le partage des données, ce que nous allons aborder dans la prochaine section.

Partage de données

Le manque de puissance statistique est essentiellement dûe à la taille d'échantillon. Aujourd'hui, dans un environnement scientifique collaboratif et axé sur les données, le partage de données est utile non seulement du point de vue de la reproductibilité, mais aussi pour construire des ensembles de données avec des échantillons de grande taille. Avec de grands ensembles de données, il serait possible de distinguer même les effets subtils (Smith and Nichols, 2017) qui n'étaient pas possibles avec de plus petits ensembles. Le partage de données est bénéfique non seulement du point de vue de la réplication, mais aussi d'un point de vue économique. Plutôt que de recueillir de nouvelles données pour chaque nouvelle hypothèse, les chercheurs peuvent maintenant ré-utiliser les données connues pour vérifier et valider leurs hypothèses.

Les avantages du partage de données remontent à Newton et à sa théorie de la gravitation (Jardine, 2013). Avant que Newton ait développé sa théorie, un autre astronome anglais, John Flamsteed avait été nommé par le roi pour observer les étoiles et produire des cartes précises pour la navigation dans les mers. Sur une période de 40 ans, Flamsteed a créé un catalogue détaillé qui a triplé le nombre d'entrées dans l'atlas du ciel utilisé précédemment. Lorsque la grande comète de 1680 est apparue deux fois de suite dans le ciel, Flamsteed a utilisé ses données pour suggérer que ce n'était pas deux comètes mais bien la même comète qui s'est d'abord dirigée vers le soleil et s'en est ensuite détournée. Newton s'est d'abord opposé à cette théorie, mais il a changé d'avis plus tard en accédant au catalogue inédit de Flamsteed. La comète s'était en effet avérée être une référence importante pour la théorie de la gravitation de Newton.

Il est difficile d'imaginer à notre époque qu'une théorie aussi fondamentale que les lois de la gravitation aurait pu être guidée par les données. Le partage de données est fondamental non seulement pour la reproductibilité scientifique, mais il constitue également la base de l'apprentissage de modèles plus solides et de l'analyse comparative de nouveaux algorithmes. Par conséquent, dans le domaine de l'apprentissage automatique, les percées récentes ont été alimentées par l'augmentation du partage de données et du calcul. Cela inclut la croissance récente de l'apprentissage profond (Deng et al., 2009), l'apprentissage Q (Watkins and Dayan, 1992; Bellemare et al., 2013), le traitement du langage naturel pour la traduction du langage (Halevy et al., 2009), la reconnaissance vocale (Paul and Baker, 1992), et même le modèle du mélange d'experts (Jacobs et al., 1991) pour IBM Watson (Ferrucci et al., 2010). La maxime "plus de données battent un algorithme plus intelligent" (Domingos, 2012) a remarquablement bien résisté à travers les disciplines et les âges.

Bien sûr, les neuroscientifiques commencent à prendre conscience de l'importance du partage de données. Récemment, les données neuronales ont commencé à être partagées par le biais de consortiums internationaux (Van Essen et al., 2013; Ollier et al., 2005), de dépôts de données (Poldrack et al., 2013; Gorgolewski et al., 2015) et d'articles sur les ensembles de données dans des revues ciblées. Pourtant, il y a encore un gros écart entre l'idéal du partage de données et ce qui est pratiqué. Les expériences de neuroimagerie sont souvent très compliquées et il ne suffit pas de partager simplement les données, mais aussi les métadonnées et les informations concernant les protocoles expérimentaux dans un format bien structuré. En l'absence de cette information, les données partagées ne sont pas réutilisables de la même manière que les programmes peu commenté, mal structurés, et compliqués ne sont pas utiles même s'ils sont partagés publiquement. Il n'y a pas de consensus accepté au sein de la communauté sur les pratiques de partage des données et il est nécessaire d'établir une norme. Au Chapitre 4, nous présenterons une nouvelle norme connue sous le nom de BIDS, qui vise à combler cet écart. Il s'agit d'un effort de collaboration entre les développeurs de logiciels et les neuroscientifiques de divers laboratoires afin d'établir un consensus sur les normes et d'élaborer des outils pour faciliter l'adoption de la norme.

Automatisation

En 2014, Nature a publié un article audacieux (Hayden, 2014) qui décrivait une vision de l'avenir de la science : des laboratoires automatisés qui enregistrerait de façon autonome chaque détail d'une expérience, ce qui mènerait à une recherche moins coûteuse, plus efficace et plus fiable. Bien qu'il décrive de nombreux laboratoires de biologie qui automatisent des expériences, les avantages de l'automatisation dans le domaine de la neuroimagerie ne sont pas encore largement reconnus. L'automatisation permet non seulement de gagner du temps, mais aussi de rendre la recherche plus reproductible, comme cela a été noté dans un guide récent pour améliorer la transparence et la reproductibilité de la recherche en neuroimagerie (Gorgolewski and Poldrack, 2016). Les auteurs soulignent que le travail manuel peut sembler facile à première vue, si l'analyse ne doit être effectuée qu'une seule fois. Toutefois, ce n'est pas toujours le cas, car "assez souvent, au cours d'un projet, les paramètres sont modifiés, les sujets sont changés et les étapes de traitement doivent être ré-exécutées. C'est une situation dans laquelle le fait d'avoir un ensemble de scripts capables d'exécuter automatiquement toutes les étapes de traitement au lieu de s'appuyer sur des

interventions manuelles peut être vraiment payant." Comme les grands ensembles de données deviennent de plus en plus courants en neuroimagerie, l'automatisation deviendra en effet une nécessité plutôt qu'un luxe.

En neuroimagerie, il existe en fait plusieurs voies d'automatisation :

- **Réduire l'interactivité :** Si les interfaces graphiques interactives sont d'excellents outils pour la navigation dans les données, elles sont insuffisantes lorsqu'il s'agit d'étendre l'analyse à des dizaines et des centaines de sujets, ce qui est nécessaire pour une étude suffisamment puissante.
- **Réglage des paramètres :** La plupart des algorithmes, bien que scénarisés, nécessitent encore des hyperparamètres à être réglés. Ces hyperparamètres peuvent être le nombre de composants ICA à choisir ou les paramètres de régularisation, et peuvent varier d'un sujet à un autre.
- **Annotation et étiquetage :** Une grande partie des données de neuroimagerie disponibles n'est pas étiquetée ou, au mieux, est faiblement étiquetée. C'est parce que les annotations d'experts sont coûteuses et ne peuvent pas être financées par la foule. Les outils automatisés basés sur l'apprentissage non supervisé peuvent jouer un rôle majeur en ce sujet.
- **Contrôle de qualité :** Actuellement, le contrôle de qualité est effectué manuellement en inspectant les données pour repérer les valeurs aberrantes. Même si l'inspection des données ne peut pas être négligée, elle peut être effectuée plus efficacement grâce à la documentation automatisée des analyses de données et des rapports tels que le carnet de notes Jupyter et le rapport Web MNE (Engemann et al., 2015a). En parallèle, des analyses de tendances statistiques avancées comme dans le projet "Automated Statistician" (Duvenaud et al., 2013) peuvent être utilisées pour créer des résumés.

Des mesures ont été prises dans ce sens, notamment la plate-forme Neurosynth (Yarkoni et al., 2011) qui facilite les méta-analyses à grande échelle. Les méta-analyses combinent généralement les résultats de plusieurs études et, dans ce cas, ce sont les cartes d'activation cérébrale de différentes études qui sont combinées à l'aide de méthodes d'apprentissage automatique. Du côté logiciel, le logiciel Freesurfer (Dale et al., 1999; Fischl et al., 1999) fournit une commande `recon-all` qui effectue automatiquement la segmentation corticale sans intervention humaine. Dans les entreprises multinationales, cette philosophie est en train d'être adoptée, en commençant par l'estimation automatisée de la covariance (Engemann and Gramfort, 2015).

Dans cette thèse, nous allons considérer un algorithme qui annote automatiquement les artefacts dans les données (Jas et al., 2016, 2017a). Il s'agit d'une première étape que tout pipeline de traitement MEG/EEG doit franchir, mais elle se fait souvent manuellement. Cela s'explique par le fait que les algorithmes existants ne sont pas conçus pour être *transparents*. Puisque pour la plupart des scientifiques, la clé des nouvelles connaissances est un ensemble de données sans artefact, ils préféreraient consacrer des efforts supplémentaires à le faire manuellement plutôt que de dépendre d'un algorithme générique qui est difficile à interpréter. Basé uniquement sur des rapports anecdotiques, ce processus peut prendre jusqu'à une semaine, même pour une étude de taille moyenne de 10 à 20 sujets.

C'est ce qui nous a conduit à proposer *autoreject*, que nous décrivons au Chapitre 5. C'est

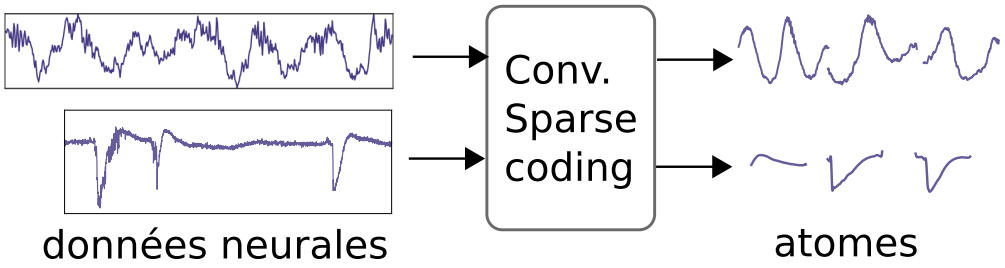


Figure 1.2: Une illustration de la façon dont le codage à faible densité convolutionnelle peut être utilisé pour extraire automatiquement les formes d'onde prototypiques.

un algorithme qui peut être utilisé pour marquer les mauvais segments des données. L'idée clé est que, souvent, certains capteurs dans l'appareil sont corrompus par intermittence plutôt que de façon continue. Nous validons notre algorithme en le comparant à trois benchmarks sur l'ensemble de données du Human Connectome Project (HCP) (Larson-Prior et al., 2013) dont les mauvais segments sont déjà annoté manuellement. Ceci dit, notre travail représente l'une des premières tentatives de réanalyse de la composante MEG de l'ensemble de données HCP.

Apprentissage de la représentation axée sur les données

Depuis l'invention de l'EEG dans les années 1920, les scientifiques ont découvert plusieurs modèles différents d'oscillation cérébrale tels que les ondes alpha, les complexes K et les rythmes mu. Les oscillations et les interactions entre eux ont servi de biomarqueurs pour différentes fonctions et pathologies du cerveau. Les ondes alpha ont été impliquées dans l'attention, les complexes K dans le sommeil et le rythme mu dans l'activité motrice.

Compte tenu de la complexité du cerveau humain, il est clair que ces formes d'ondes ne représentent qu'une fraction des fonctions cognitives que le cerveau peut accomplir. En raison de la richesse des données maintenant disponibles grâce au mouvement de partage des données décrit à la Section 2.2.2, le futur neuroscientifique sera en mesure d'extraire de telles formes d'onde à partir de grands ensembles de données. Imaginez si les neuroscientifiques disposaient d'un outil similaire à Google Photos[1]. De la même manière que Google Photos peut trouver automatiquement des visages et des photos de groupe, ces outils pourront trouver des oscillations prototypiques et regrouper les données en les utilisant. Cliquer sur l'une de ces formes d'onde permet de récupérer les données qui y sont associées.

Cependant, les photos sont intrinsèquement différentes des données neuronales. Premièrement, les données neuronales peuvent être enfouies dans le bruit et corrompues par des artefacts de grande amplitude. Deuxièmement, les images sont étiquetées en raison de données provenant de la foule comme dans le cas d'Imagenet (Deng et al., 2009), mais les données neuronales ne le sont pas. Les annotations d'experts dans le cas de données neuronales ne sont pas facilement disponibles. Enfin, ce sont les données spatio-temporelles avec des dynamiques différentes du monde 3D que les photos capturent.

C'est là que le codage convolutif à faible densité (CSC) peut jouer un rôle en extrayant des

[1] https://photos.google.com/

caractéristiques prototypiques des données, comme le montre la Figure 1.2. Il s'agit d'un algorithme non supervisé de la vision par ordinateur, qui peut apprendre des dictionnaires de formes d'onde prototypiques (atomes) à partir des données en utilisant les opérations de convolution. Pour en savoir plus sur le CSC, le lecteur peut lire la Section 2.3.4 de ce chapitre.

Les algorithmes CSC n'approximent pas le signal en utilisant la base de Fourier (ou sinusoïdale). Bien qu'il s'agisse de la technique conventionnelle d'extraction de signaux enfouis dans le bruit, l'approximation peut dégrader la forme du signal, ce qui peut être un biomarqueur dans de nombreuses maladies cliniques (Cole and Voytek, 2017). . Par exemple, même avec un grand nombre de sinusoïdes de la base de Fourier, les bords d'une onde carrée ne peuvent pas être bien approximés. En effet, l'approximation imparfaite autour de ces bords est ce qu'on appelle souvent des artefacts d'oscillations parasites dans les contextes de traitement du signal. Bien sûr, les transitoires peuvent être mieux approximés en utilisant des ondelettes, mais ce n'est clairement pas suffisant pour d'autres formes de données. Plutôt que de fixer la base à Fourier ou à ondelettes, l'approche CSC consiste à apprendre à la fois la base et les coefficients.

Dans notre travail présenté au Chapitre 6, nous étendons les algorithmes CSC conventionnels pour les bruits de queue lourds. Nous reformulons le problème d'optimisation sous forme d'inférence MAP avec une distribution "alpha-stable" pour remplacer la perte de reconstruction. Nos résultats montrent que ce type d'algorithme est robuste à la présence d'artefacts et peut être utilisé pour découvrir des structures temporelles à partir de signaux neuronaux, même ceux qui impliquent des oscillations imbriquées.

Chapitre 3: Brain Imaging Data Structure (BIDS)

Du point de vue de la reproductibilité, le partage de données est d'une importance capitale. Le code de partage en soi ne permet pas la reproductibilité si les données qui l'accompagnent ne sont pas disponibles et sont coûteuses à acquérir. La réanalyse d'un ensemble de données est cependant utile non seulement du point de vue de la reproductibilité, mais aussi pour découvrir de nouveaux effets qui étaient auparavant négligés. En même temps, plus les données sont partagées, plus la taille de nos échantillons sera grande, ce qui nous permettra de mener des études avec une plus grande puissance statistique. En effet, la faible puissance statistique est l'une des principales raisons de la crise de la reproductibilité.

Bien que le partage de données en neurosciences soit à la hausse, la quantité de données ré-utilisées est encore limitée. Par exemple, depuis la publication des données du projet Human Connectome Project (HCP) (Larson-Prior et al., 2013) du MEG en 2013, il y a eu très peu de cas de réutilisation de ces données. Au moment de la rédaction de cette thèse, nous n'avions qu'un ou deux cas documentés (Jas et al., 2017a) de réutilisation des données HCP. Même dans ces cas, l'effort s'est surtout limité à reproduire les résultats plutôt que de tester de nouvelles hypothèses. Cela représente clairement un écart entre l'idéal et la pratique du partage de données.

Les expériences de neuroimagerie sont souvent compliquées et impliquent différentes tâches cognitives (auditives, visuelles, somatosensorielles, *etc.*), différents paramètres d'acquisition

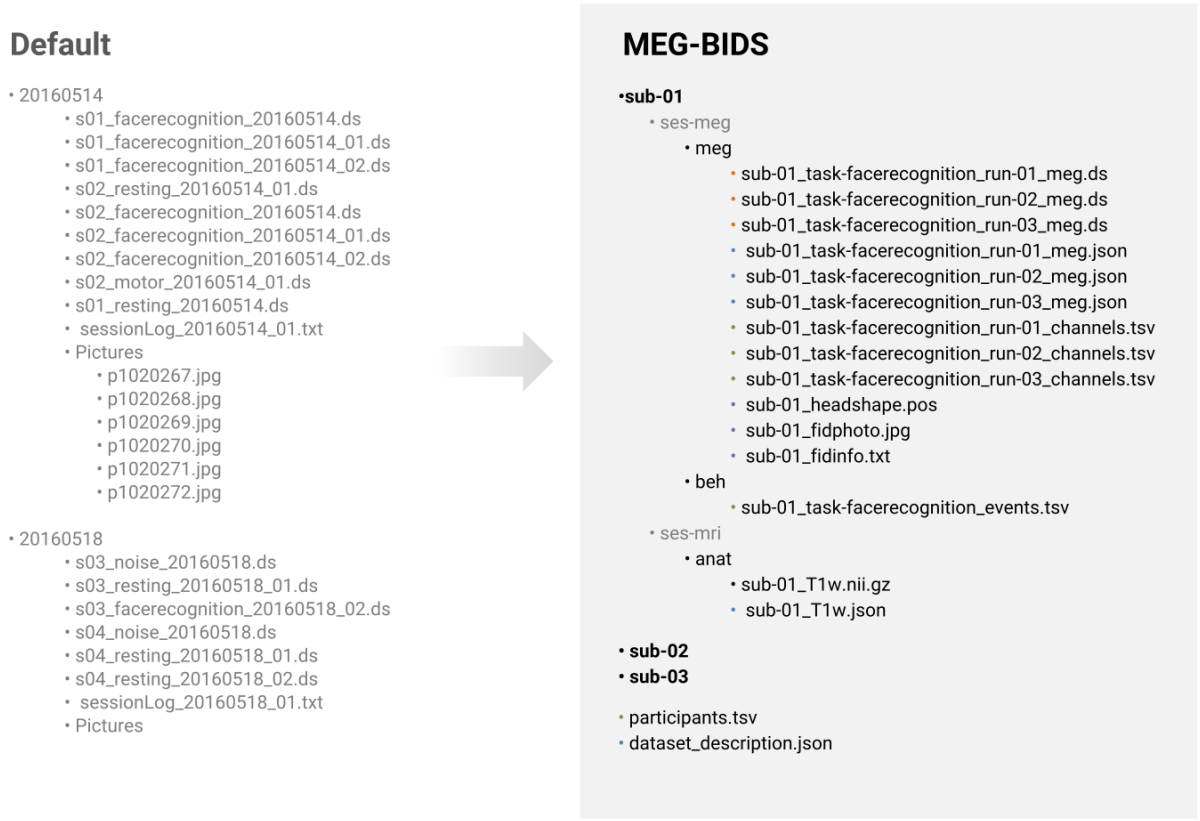


Figure 1.3: Schéma d'organisation des données BIDS-MEG : Gauche : un schéma typique d'organisation de données par défaut où les dossiers sont organisés par date de session et contiennent différentes exécutions pour un participant donné dans une étude. Droite: BIDS-MEG organise les données par étude, puis par participant (sujet), suivi de la modalité, puis des sessions et finalement des runs. Notez les fichiers latéraux qui sont présents à tous les niveaux de la hiérarchie des données et documentez facilement le contenu des métadonnées.

(fréquence d'échantillonnage, nombre de capteurs et leur emplacement, dispositif de mesure, etc.) et des paramètres de population (sexe, âge, *etc.*). Toutes ces métadonnées sont des informations nécessaires pour réanalyser avec succès les données. Malheureusement, dans le passé, il n'y a pas eu de consensus entre les différents laboratoires et fabricants industriels sur ce qui constitue des métadonnées utiles. Cela souligne la nécessité d'établir des normes. Bien qu'à première vue, cela peut sembler être de la paperasserie bureaucratique inutile, en fait, des normes existent dans presque tous les aspects de notre vie.

Outre les méta-informations qui sont stockées avec les données, les données elles-mêmes sont stockées dans l'un des 10 à 20 formats de fichiers différents et à différents stades de traitement. Bien que des efforts aient déjà été déployés pour normaliser les structures de données (Gibson et al., 2009; Grewe et al., 2011; Stoewer et al., 2013; Teeters et al., 2015; Bigdely-Shamlo et al., 2016), il n'a pas été largement accepté. La conception d'une nouvelle norme est délicate car elle nécessite l'obtention d'un consensus communautaire. En même temps, elle doit trouver le juste équilibre entre la rigidité pour l'efficacité et la flexibilité pour s'adapter aux technologies futures.

Le format BIDS (Gorgolewski et al., 2016) est en effet conçu en tenant compte de ces

considérations. La norme comporte une hiérarchie de dossiers pour décrire la technologie d'imagerie utilisée, le nom du sujet et la date de l'expérience. A chaque niveau de hiérarchie, les fichiers sont accompagnés de fichiers "json" latéraux décrivant les métadonnées. Un fichier json est un fichier texte facile à analyser qui contient des paires de clés et de valeurs, de sorte qu'il a l'avantage d'être lisible par l'homme et par la machine en même temps. Ces fichiers suivent un principe d'héritage, c'est-à-dire qu'un champ décrit dans un fichier json à un niveau supérieur de la hiérarchie sera automatiquement propagé en aval. La spécification principale de BIDS est accompagnée de spécifications d'extension qui décrivent des aspects spécifiques pour décrire différentes modalités.

En même temps, la norme n'existe pas isolément. Le consortium BIDS fournit également un écosystème croissant d'outils pour convertir les ensembles de données dans un format compatible BIDS et pour valider les données afin qu'elles soient conformes à la norme.

Dans ce travail, nous avons présenté une extension significative de BIDS pour soutenir les aspects spécifiques des données MEG. Comme nous le savons, la MEG fournit une mesure directe de l'activité cérébrale avec une résolution temporelle en millisecondes et des capacités d'imagerie à source unique. Jusqu'à présent, BIDS a fourni une solution pour structurer l'organisation des données d'IRM. Malgré l'absence d'un format de données standard pour la MEG, BIDS-MEG est une solution de principe pour stocker, organiser et partager les volumes de données typiquement importants produits. Il s'appuie sur BIDS pour l'IRM et permet donc une organisation multimodale des données par construction. Ceci est particulièrement utile pour l'enregistrement anatomique et fonctionnel de l'imagerie source MEG avec l'IRM. Avec BIDS-MEG et une gamme croissante de logiciels adoptant la norme, la communauté MEG dispose d'une solution pour minimiser les frais généraux de conservation, réduire les erreurs de traitement des données et optimiser l'utilisation des ressources informatiques pour l'analyse des données. La norme comprend également des métadonnées bien définies pour faciliter les efforts futurs d'harmonisation et de partage de données, ainsi que des extensions à d'autres modalités de données électrophysiologiques.

Outils logiciels

Dans le cadre de cet effort, j'ai participé aux discussions pour décider de la norme. En même temps, j'ai développé l'extension MEG du validateur BIDS. Développé en javascript à l'aide de "nodejs", il peut être packagé pour fonctionner dans le navigateur (Google Chrome). Une version en ligne de commande est également fournie afin qu'elle puisse être utilisée dans l'analyse par script. Le validateur effectue plusieurs contrôles de santé mentale sur les ensembles de données pour s'assurer qu'ils sont compatibles avec BIDS. Cela comprend l'utilisation d'expressions régulières pour vérifier les noms de fichiers et de schémas de notation d'objets JavaScript (JSON) pour s'assurer que les fichiers de métadonnées sont normalisés par type de données.

Afin de faciliter l'adoption du format BIDS, j'ai également contribué au code Python qui est utilisé pour convertir les jeux de données existants en jeux de données compatibles BIDS. Le code peut être trouvé ici : https://github.com/mne-tools/mne-bids.

Pour une description plus détaillée de la spécification BIDS-MEG, des exemples de données, des ressources et des commentaires, veuillez visiter le site http://bids.neuroimaging.

io.

Chapitre 4: Une étude de groupe M/EEG reproductible

ans la section précédente, nous avons discuté de la façon dont le partage des données peut être facilité à l'aide du BIDS pour la MEG permettant de se rapprocher de l'objectif de reproductibilité. Cependant, la reproductibilité n'est pas obtenue en partageant simplement plus de données dans l'espoir que cela résoudra tous les problèmes. Comme l'indique Baker Baker (2016), l'une des meilleures solutions pour favoriser la reproductibilité scientifique n'est pas technique, mais éducative. C'est bien sûr vrai pour les statistiques, où il y a un besoin urgent de clarifier et d'éduquer les chercheurs ces outils requis en neurosciences. Cela devient maintenant tout aussi important pour les logiciels académiques.

Au cours des dernières années, les boîtes à outils universitaires gratuites ont pris de plus en plus d'importance dans l'analyse du MEG pour diffuser des méthodes de pointe, partager les meilleures pratiques entre différents groupes de recherche et de mettre en commun des ressources pour développer des outils essentiels pour la communauté du MEG. Des événements pédagogiques sont régulièrement organisés dans le monde entier où les bases de chaque boîte à outils sont expliquées par ses développeurs et des utilisateurs expérimentés. Il y a toutefois des lacunes dans les connaissances qui doivent être comblées. Premièrement, la plupart des exemples d'enseignement ne montrent que l'analyse d'un seul 'participant typique', alors que la plupart des études réelles de MEG impliquent l'analyse de données de groupe. Il revient ensuite aux chercheurs sur le terrain de déterminer eux-mêmes comment faire la transition et obtenir des résultats significatifs pour le groupe. Deuxièmement, nous ne connaissons pas d'exemples d'analyse complète d'un même ensemble de données de groupe avec différentes boîtes à outils académiques pour évaluer le degré d'accord dans les conclusions scientifiques et comparer les forces et les faiblesses des diverses méthodes d'analyse et de leur mise en œuvre indépendante.

Les méthodes

Pour répondre à cette question, un atelier a été organisé par les principaux développeurs des six boîtes à outils MEG les plus populaires à Biomag 2017. Ce travail fait suite à l'atelier, qui présente la contribution de l'équipe du logiciel MNE, et sera publié dans *Frontiers in Neuroscience, section Brain Imaging Methods*. Cette étude présente les résultats obtenus par la réanalyse d'un ensemble de données ouvertes de Wakeman and Henson (2015) à l'aide du progiciel MNE. L'analyse couvre les étapes de prétraitement, l'assurance qualité, l'analyse de l'espace capteur des réponses évoquées, la localisation de source et les statistiques dans l'espace capteur et l'espace source. Les résultats avec les stratégies alternatives possibles sont présentés et discutés à différentes étapes telles que l'utilisation du filtrage passe-haut par rapport à la correction de base, le tSSS par rapport à la séparation spatiale du signal (SSS), l'utilisation d'une norme minimale inverse par rapport à la variance minimale linéairement contrainte (LCMV), et l'utilisation de statistiques univariées ou multivariées.

Résultats

L'objectif est de fournir une étude comparative des différentes étapes du pipeline d'analyse MEG/EEG sur le même ensemble de données, avec un accès ouvert (https://github.com/mne-tools/mne-biomag-group-demo) to all of the scripts necessary to reproduce this analysis. An example of such a reanalysis figure is shown in Figure 1.4 qui montre la réponse évoquée pour un capteur EEG. En effet, nous sommes en mesure non seulement de reproduire les résultats de Wakeman and Henson (2015), mais aussi de mettre en évidence les facteurs qui pourraient conduire à des résultats différents. Le travail se termine par un ensemble de recommandations basées sur les leçons que nous avons tirées de cet exercice de réanalyse.

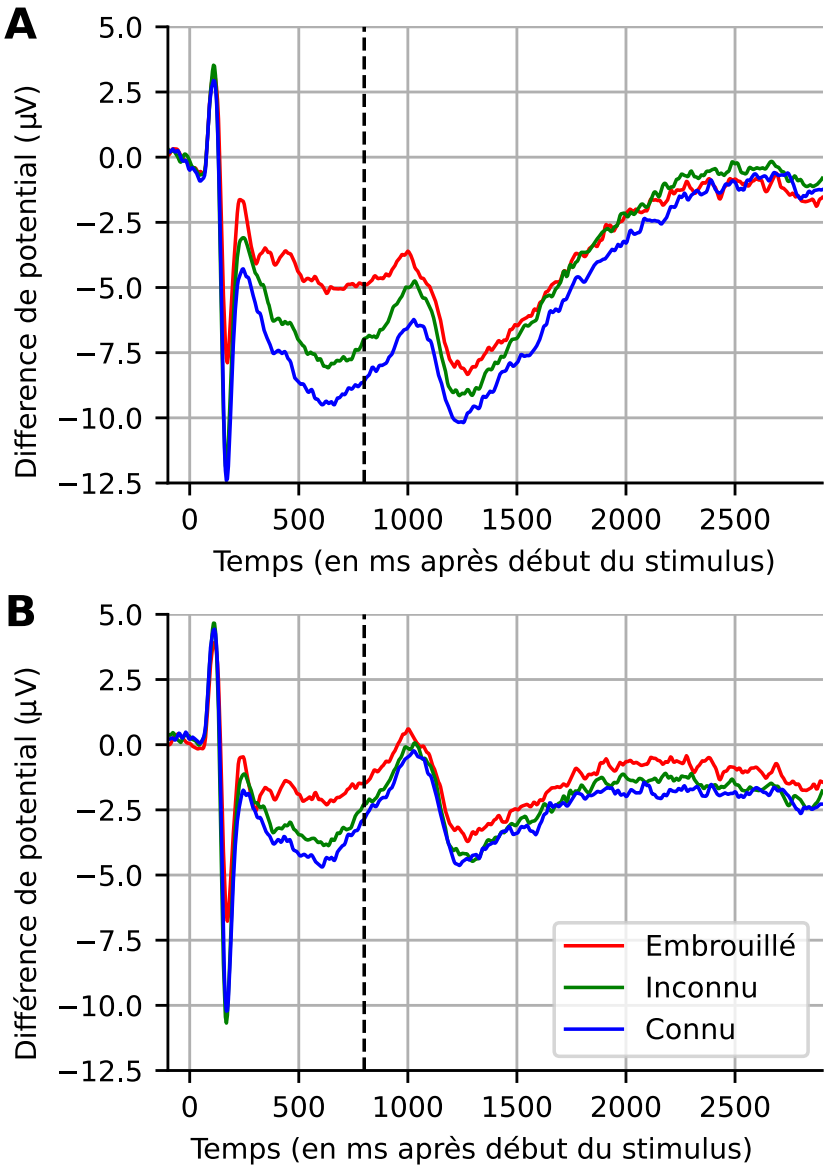


Figure 1.4: Moyenne générale des réponses évoquées sur 16 sujets pour le canal EEG065. (A) Pas de filtre passe-haut. (B) Filtré passe-haut à 1,0 Hz. Il est à noter que, comme (A), les résultats rapportés par Wakeman and Henson (2015) (la ligne en pointillés à 800 ms indique où leur placette s'est arrêtée) montrent de grandes dérives, mais celles-ci reviennent à des niveaux proches de la ligne de base vers la fin d'un intervalle suffisamment long (ici, 2,9 secondes) même sans appliquer un filtre passe-haut.

Chapitre 5: Rejet automatisé des artefacts pour M/EEG

Dans la section précédente, nous avons discuté des problèmes de reproductibilité lors de la réalisation d'études de groupe en MEG et EEG. L'automatisation est un moyen d'améliorer la reproductibilité. et nous avons brièvement évoqué un algorithme permettant d'automatiser la détection de segments de données incorrects, appelé *autoreject*.

Dans cette section, nous allons présenter cet algorithme qui rejette et répare les mauvais essais dans les signaux MEG et EEG. L'annotation de mauvais segments dans les données est peut-être l'un des aspects les plus chronophages du traitement de données électrophysiologiques. Actuellement, il est fait soit manuellement, soit en utilisant des algorithmes automatisés mais qui présentent l'inconvénient d'être des boîtes noires (par exemple, RANSAC (Bigdely-Shamlo et al., 2015), FASTER (Nolan et al., 2010), SNS (De Cheveigné and Simon, 2008)). L'approche manuelle est souvent subjective et sans consensus clair notamment sur ce qui compose un segment de données corrompu. Par conséquent, la réanalyse est non seulement exigeante manuellement, mais peut également entraîner des problèmes de reproductibilité. En revanche, les méthodes automatisées sont contrôlées par des paramètres difficiles à régler. En cas d'échec, il n'est pas toujours évident de savoir ce qui a causé l'échec de la méthode et comment elle a pu être corrigée. Par conséquent, il n'y a pas d'autre choix que d'exclure les données de l'analyse ultérieure.

Les méthodes

Cela nous a conduit à développer une méthode basée sur des choix de conception motivés par la facilité d'interprétation et de diagnostic. La méthode que nous proposons capitalise sur la validation croisée (Figure 1.5) en conjonction avec une métrique d'évaluation robuste pour estimer le seuil optimal pic à pic, une quantité couramment utilisée pour identifier les mauvais essais en MEG / EEG. L'idée clé est que les ensembles de formation et de validation sont d'autant plus similaires lorsque les données ne contiennent pas de valeurs aberrantes. Ainsi, un seuil qui minimise l'erreur de validation croisée est un seuil qui supprime les valeurs aberrantes. En effet, comme le montre la Figure 1.5, pour des seuils très bas, trop d'essais sont supprimés, ce qui se traduit par une moyenne bruitée, alors que pour des seuils très élevés, les valeurs aberrantes ne sont pas éliminées. L'optimum se situe entre les deux, ce qui peut être estimé en minimisant l'erreur de validation croisée.

Cette approche est ensuite étendue à un algorithme plus sophistiqué qui estime ce seuil au niveau des capteurs. Il en résulte une liste des mauvais capteurs par essais. En fonction du nombre de capteurs défectueux, l'essai est ensuite corrigé par interpolation ou exclue de l'analyse. Pour des raisons d'efficacité, nous utilisons l'optimisation bayésienne, technique bien connue pour l'optimisation hyperparamétrique. Toutes les étapes de l'algorithme sont entièrement automatisées, d'où le nom *autoreject*. Il est important de noter que l'algorithme est même capable de traiter des capteurs qui sont localement corrompus, ce qui est souvent le cas pour les données EEG.

Résultats

Afin d'évaluer l'importance pratique de l'algorithme, nous avons effectué une validation approfondie et des comparaisons avec des méthodes de pointe sur quatre ensembles de

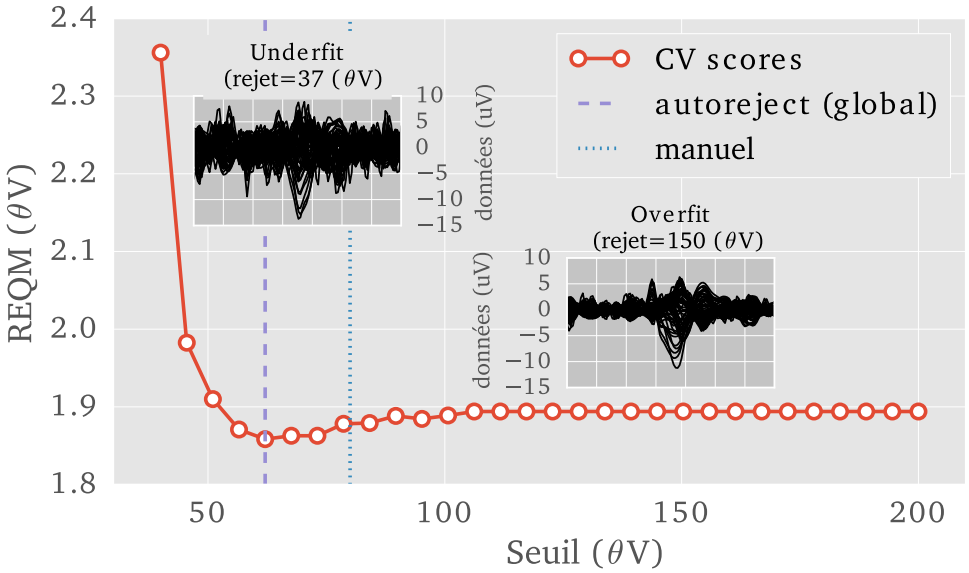


Figure 1.5: Erreur de validation croisée en fonction du seuil de rejet crête à crête sur un ensemble de données EEG. Le racine de l'erreur quadratique moyenne (REQM) entre la moyenne de l'ensemble de formation (après avoir retiré les essais marqués comme étant mauvais) et la médiane de l'ensemble de validation a été utilisée comme mesure de validation croisée (Section 5.2.1). Les deux cartons montrent la moyenne des essais sous forme de "parcelles papillon" (chaque courbe représentant un capteur) respectivement pour des seuils très bas et élevés. Pour les seuils bas, le REQM est élevé parce que la plupart des essais sont rejetés (*underfit*). Pour des seuils élevés, le modèle n'abandonne aucun essai (*overfit*). Le seuil optimal basé sur les données *(autoreject, global)* avec un minimum de REQM se situe quelque part entre les deux. Il correspond approximativement au seuil humain.

données publiques contenant des enregistrements MEG et EEG de plus de 200 sujets. Les comparaisons comprennent des efforts purement qualitatifs (Figure 1.7) ainsi que des analyses comparatives quantitatives par rapport à des pipelines de prétraitement sous surveillance humaine (Figure 1.6) et semi-automatique. Nos comparaisons qualitatives ont montré qu'*autoreject* était capable d'annoter et de réparer les segments défectueux de manière satisfaisante pour ces différents ensembles de données. En effet, l'algorithme permet d'automatiser le prétraitement des données MEG depuis le HCP jusqu'au calcul des réponses évoquées. La nature automatisée de notre méthode minimise le fardeau de l'inspection humaine, ce qui favorise l'extensibilité et la fiabilité exigées par l'analyse des données dans les neurosciences modernes.

Chapitre 6: Temporal representation learning

Jusqu'à présent, nous avons étudié l'automatisation en neuroimagerie dans le but de permettre une analyse évolutive des données, ainsi qu'une reproductibilité. Malgré le fait que la reproductibilité et l'analyse de données à grande échelle nous permettent de consolider les études existantes, elles ne sont pas en soi des outils faits pour découvrir des phénomènes nouveaux et intéressants. Dans cette section, nous explorerons cette dimension de l'automatisation à l'aide de ce qu'on appelle *l'apprentissage de la représentation.*

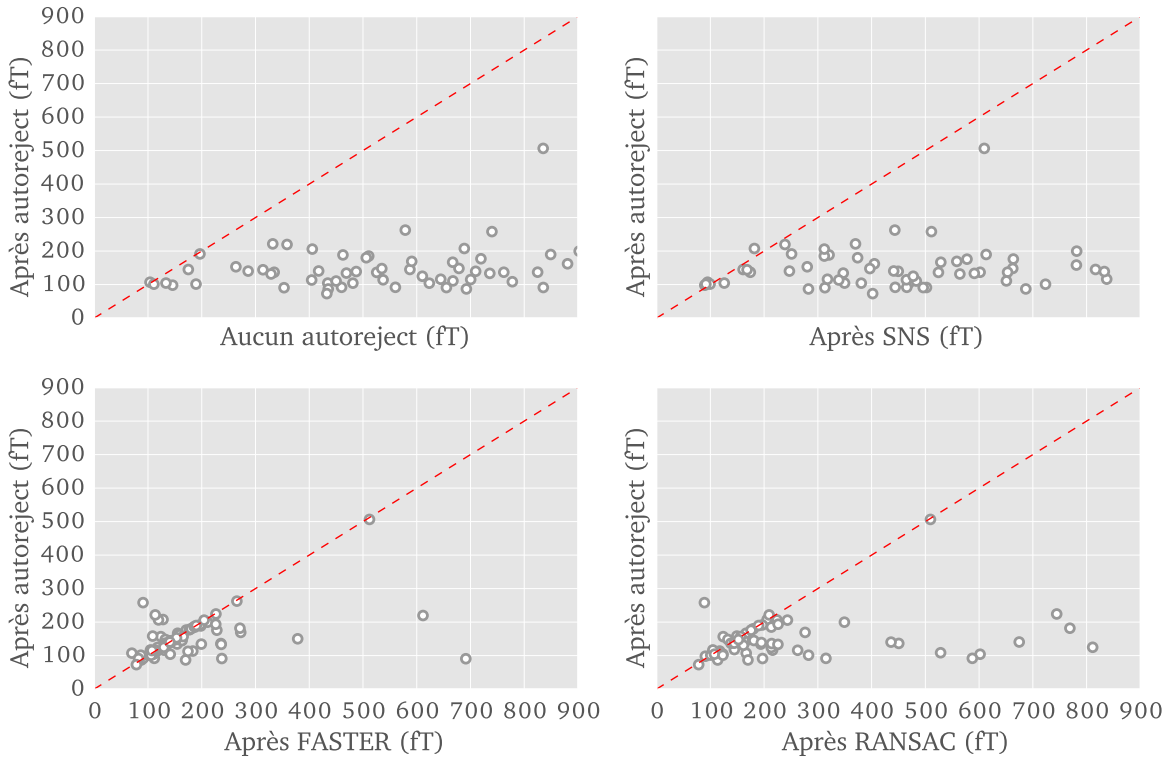


Figure 1.6: Diagrammes de dispersion pour les résultats avec les données HCP. Pour chaque méthode, on utilise la norme $\|\cdot\|_\infty$ de la différence entre la vérité terrain HCP et la méthode. Chaque cercle représente un sujet. (A) *autoreject (local) vs.* aucun rejet, (B) *autoreject (local) vs.* Sensor Noise Suppression (SNS) (SNS), (C) *autoreject vs.* FASTER, (D) *autoreject (local) vs.* RANSAC. Les points de données sous la ligne rouge pointillée indiquent les sujets pour lesquels l'*autoreject (local)* surpasse la méthode alternative.

Les représentations sont les bases du traitement de signal. Il est assez facile de se convaincre de ce fait, si nous utilisons simplement une Transformée de Fourier Rapide (TFR) pour filtrer les données. Lorsque nous utilisons une TFR, nous sommes effectivement en train de décomposer le signal en une somme de sinusoïdes de fréquences variables. Si nous sommes intéressés par une analyse temps-fréquence, un choix commun de représentation pour des signaux en neuroscience consiste en l'utilisation des ondelettes de Morlet.

Traditionnellement, le choix de la représentation était principalement motivé par la préoccupation analytique et la facilité de la manipulation mathématique. Toutefois, l'essor récent de l'apprentissage profond a suscité un intérêt pour les représentations axées sur des données. C'est parce que de bonnes représentations qui capturent de manière compacte les propriétés des données sont essentielles pour des systèmes d'apprentissage efficaces et précis. Par exemple, en vision par ordinateur, les attributs artisanals telles que les descripteurs SIFT (Lowe, 1999) et GIST (Oliva and Torralba, 2001), Deformable Parts Model (DPM) (Felzenszwalb et al., 2010), Histogram of Oriented Gradient (HOG) (Dalal and Triggs, 2005) *etc.* étaient la norme, avant qu'on ne réalise que l'apprentissage non supervisé et les auto-encodeurs performent bien mieux.

Aujourd'hui, l'apprentissage non supervisé est utilisé comme première étape dans une tâche

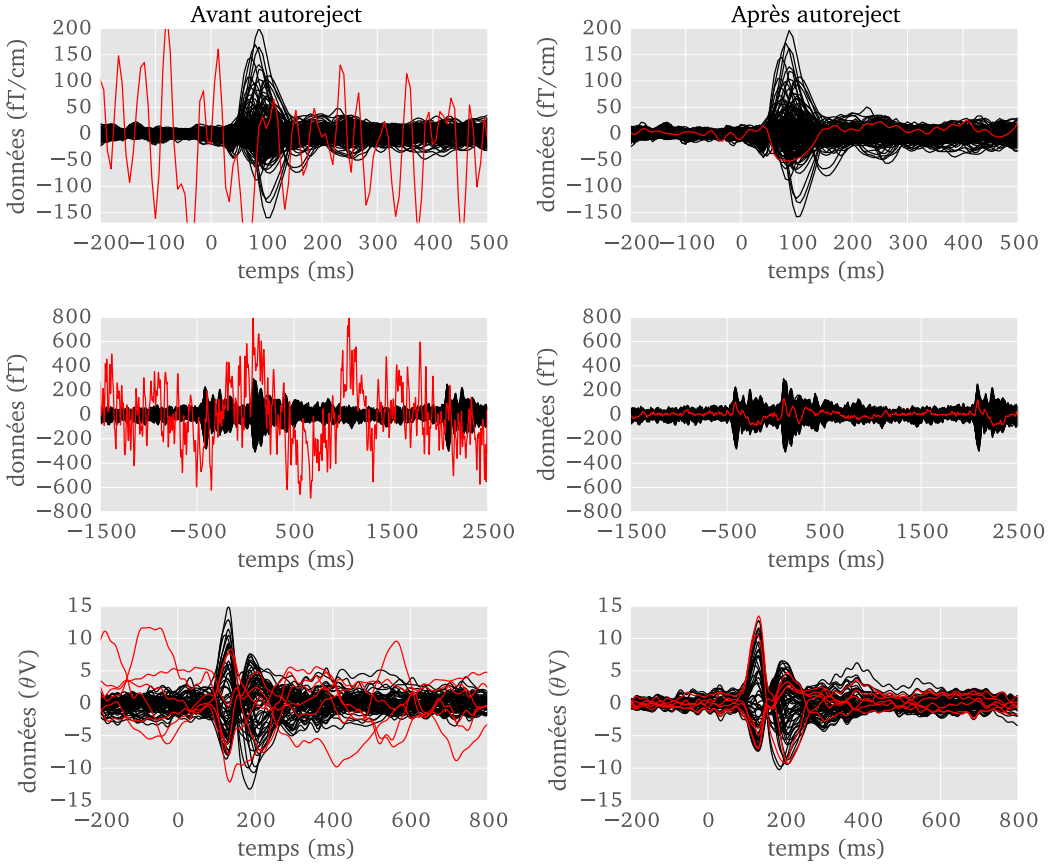


Figure 1.7: La réponse évoquée (moyenne des données d'un essai à l'autre) sur trois ensembles de données différents avant et après l'application de l'*autoreject* — es données de l'échantillon MNE, les données HCP et les données EEG Faces. Chaque capteur est une ligne sur les tracés. A gauche, les mauvais capteurs annotés manuellement sont affichés en rouge. L'algorithme trouve automatiquement les mauvais capteurs et les corrige pour les essais en question. A noter que l'algorithme peut également corriger plusieurs capteurs à la fois et fonctionne pour tout type de modalités d'acquisition.

d'apprentissage supervisé en vision par ordinateur. L'apprentissage de la représentation, en soi, n'est peut-être pas aussi intéressant, hormis pour les visualisations diagnostiques en apprentissage profond (Zeiler and Fergus, 2014). Malgré cela, il y a toujours eu un intérêt à comprendre les représentations dans le cerveau humain (en particulier le système visuel), puisqu'il était pensé que cela nous aiderait à construire de meilleurs systèmes d'apprentissage. L'un des pionniers dans ce domaine de recherche est Bruno Olshausen, dont les travaux sur l'apprentissage de dictionnaire (Olshausen and Field, 1996)ont démontré que les patches de Gabor sont effectivement fondamentaux pour les images naturelles, similaires à ceux que Hubel et Wiesel (Hubel and Wiesel, 1962; Marĉelja, 1980) ont trouvés dans le cortex visuel du chat, et à ce qui est utilisé dans les caractéristiques de GIST. Hormis ces études, la représentation apprise en elle-même n'est pas considérée comme aussi significative que les mesures de performance comme le score de prédiction ou la perte de reconstruction. Cependant, lorsqu'il s'agit de signaux neuronaux, nous avons réalisé que ce n'est pas le cas et que la fidélité de la représentation est en soi intéressante. En effet, la forme du signal est un biomarqueur crucial dans de

nombreuses applications cliniques des neurosciences (Cole and Voytek, 2017).

Un développement parallèle dans le domaine de la neuroimagerie a été la hausse d'intérêt pour l'apprentissage de formes prototypiques qui sont shift-invariantes (Jost et al., 2006; Barthélemy et al., 2013; Brockmeier and Príncipe, 2016; Hitziger et al., 2017). Elle est motivée par le fait que les approximations existantes utilisant la base de Fourier déforment souvent le signal. Il y a, par exemple, un débat au niveau du type de filtre qui devrait être utilisé (voir Section 4.3.3 et Widmann et al. (2015); Parks and Burrus (1987); Ifeachor and Jervis (2002); Götz et al. (2015)). Même s'il y a certains succès qui ont été rapportés en neuroimagerie avec ces algorithmes, leur applicabilité est limitée en raison de leur nature heuristique. Étonnement, jusqu'à présent, il y a eu très peu de pollinisation croisée d'idées entre les communautés de vision par ordinateur et de neuroimagerie sur ces aspects. Notre travail est une tentative de combler cette lacune. Nous proposons un modèle qui s'appuie sur les creuses modèles de codage shift-invariants existants, de manière à pouvoir traiter les bruits de queue lourds et les artefacts. Il suppose la positivité des coefficients, pour tenir compte du fait qu'un atome ne change pas de polarité à travers le temps.

Les méthodes

Un modèle CSC, tel qu'introduit pour la première fois par Grosse et al. (2012), exprime le signal estimé $\hat{x}_n$ comme une somme de convolution des atomes d^k et de leurs activations correspondantes z_n^k:

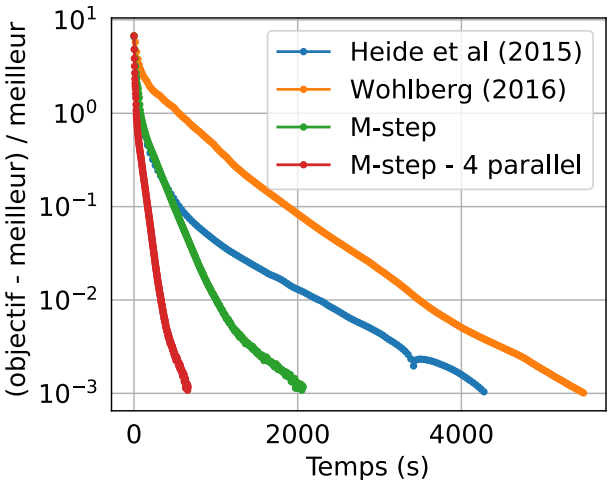
$$z_{n,t}^k \sim \mathcal{E}(\lambda), \quad x_{n,t}|z,d \sim \mathcal{N}(\hat{x}_{n,t},1), \quad \text{where,} \quad \hat{x}_n \triangleq \sum_{k=1}^{K} d^k * z_n^k \ . \tag{1.1}$$

Dans le modèle traditionnel, le bruit est supposé être de distribution gaussienne $\mathcal{N}(\cdot)$ et les activations sont supposées être creuse, tirées d'une distribution exponentielle $\mathcal{E}(\cdot)$. Les activations et les atomes sont estimés à l'aide d'un autre schéma de minimisation. Le modèle que nous présentons est un nouveau modèle probabiliste du CSC pour l'apprentissage d'atomes shift-invariants, à partir de données de séries temporelles neuronales non traitées contenant des artefacts potentiellement importants. Par conséquent, un modèle de bruit gaussien n'est plus suffisant. Au cœur de notre modèle, que nous appelons αCSC, se trouve une famille de distributions à queue lourdes appelées α-stable distributions. Le paramètre α contrôle la lourdeur de la distribution.

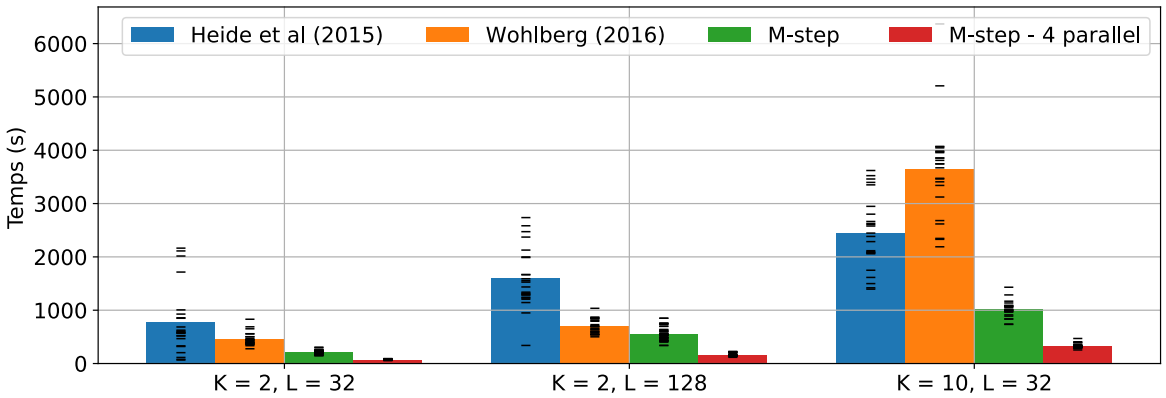
Nous développons un algorithme nouveau et efficace sur le plan informatique, pour la maximisation de l'espérance de Monte Carlo (MCEM), pour l'inférence. L'étape de maximisation se résume à un problème CSC pondéré (Equation 1.2), pour lequel nous développons un algorithme d'optimisation efficace.

$$-\max_{z,d} \sum_{n=1}^{N} \left(\|\sqrt{w_n} \odot (x_n - \sum_{k=1}^{K} d^k * z_n^k)\|_2^2 + \lambda \sum_k \|z_n^k\|_1 \right) \quad \text{s.t. } z_n^k \geq 0, \forall n,k \ . \tag{1.2}$$

Les poids w_n dans l'étape de maximisation sont estimés dans l'étape d'attente par l'usage d'une procédure de Monte Carlo à chaîne de Markov (MCMC). En effet, les poids nous



(a) $K = 10$, $L = 32$.



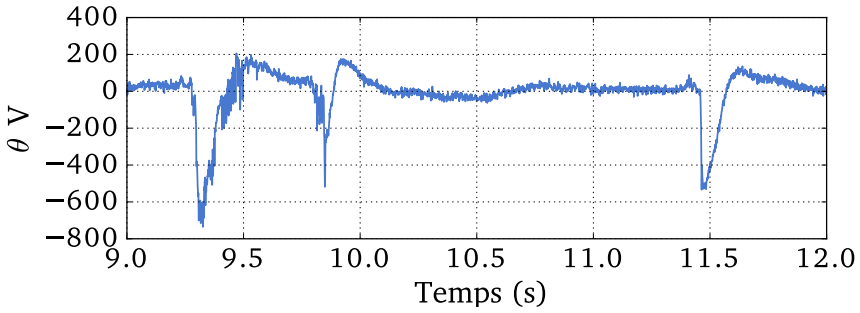
(b) Temps nécessaire pour atteindre une précision relative de 0.01.

Figure 1.8: Comparaison des méthodes de pointe avec notre approche. (a) Diagramme de convergence avec la fonction objectif par rapport au minimum obtenu, en fonction du temps de calcul. (b) Temps nécessaire pour atteindre une précision relative de 10^{-2}, pour différents réglages du nombre d'atomes K et de la longueur d'un atome L.

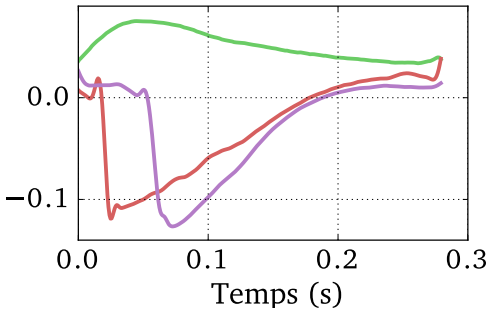
aident à traiter les artefacts en les supprimant dans la fonction d'objectif de l'étape de maximisation.

Résultats

Dans notre travail, nous évaluons rigoureusement l'efficacité informatique de notre algorithme par rapport aux benchmarks concurrents. Parce que le problème CSC est non convexe, la procédure d'optimisation implique des boucles imbriquées et l'analyse théorique est souvent insuffisante pour traiter la complexité des fonctions non convexes. La procédure d'optimisation est imbriquée car le problème est convexe lorsque l'une des variables est fixe : les atomes ou les activations. La boucle extérieure alterne entre ces deux variables, tandis que la boucle intérieure les apprend lorsque l'autre est fixe. Le résultat final dépend de l'initialisation et, par conséquent, les algorithmes ne peuvent être comparés que s'ils sont testés pour de nombreuses initialisations aléatoires différentes et



(a) Données LFP spike de Hitziger et al. (2017)



(b) Atomes appris

Figure 1.9: Atomes appris par αCSC sur les données LFP contenant des pointes épileptiformes avec $\alpha = 2$.

leurs résultats moyennés. Notre analyse qualitative va également au-delà du récit de la vérification de l'existence de formes d'onde connues pour découvrir des structures plus complexes dans les données.

Nos résultats démontrent que l'algorithme proposé atteint des vitesses de convergence de pointe (Figure 1.8). En effet, notre algorithme utilise des solveurs quasi-Newton, ce qui permet une convergence plus rapide par rapport aux méthodes ADMM (Heide et al., 2015; Wohlberg, 2016). De plus, αCSC est significativement plus robuste aux artefacts, lorsqu'on le compare aux trois algorithmes concurrents : il peut extraire des *spike bursts* (Figure 1.9), des oscillations, et même révéler des phénomènes plus subtils tels que le couplage de fréquences croisées lorsque appliqué à des séries temporelles neuronales bruyantes.

Conclusion

La recherche en neuroimagerie est un mariage entre l'informatique et les neurosciences. Il s'agit d'une collaboration entre deux disciplines complémentaires – le but est d'apporter à la table des outils de calcul qui peuvent aider les scientifiques à faire de nouvelles découvertes. Certains aspects de ce sous-domaine interdisciplinaire consistent bien sûr à développer progressivement les outils existants : par exemple, ceux qui peuvent aider à obtenir un meilleur score de prédiction ou une meilleure précision de localisation dans l'estimation des sources neuronales. Cependant, un aspect orthogonal mais tout aussi important de la recherche méthodologique est de développer des outils qui permettent à de nouvelles façons fondamentales d'interagir avec les données. Cette thèse est une

tentative de faire avancer cet objectif en développant des outils pour l'analyse automatisée en électrophysiologie.

Il est devenu évident pour nous que, pour atteindre l'objectif d'une recherche reproductible, les grands ensembles de données publiques sont la clé et les méthodes automatisées pour les analyser sont indispensables. Alors que l'ambition de tout neuroscientifique est de générer de nouvelles connaissances et de pousser les frontières de notre connaissance du cerveau, cela n'est souvent pas possible en raison de la faible taille de l'effet dans de petits ensembles de données. Lorsque l'hypothèse nulle ne peut être rejetée, il est pratique courante de commencer partir à la recherche de résultats significatifs en testant de multiples hypothèses et en rapportant les plus favorables. Cela a résulté en un corpus de littérature où une grande partie des résultats repose sur des bases incertaines.

Dans cette thèse, nous avons développé une nouvelle spécification connue sous le nom de BIDS, qui facilite le partage de données entre neuroscientifiques en promouvant des normes communes pour le stockage des métadonnées relatives aux mesures. Nous avons également donné un aperçu des défis que pose l'analyse de données reproductibles, en ce qui concerne les données MEG/EEG. En tant que contributeurs au progiciel MNE, nous nous sommes sentis particulièrement bien placés pour relever les défis liés aux logiciels : pipelines complexes, versions logicielles, initialisation aléatoire, etc. et recommandations normalisées pour chaque étape de ces pipelines. Pour ce faire, nous avons ré-analysé une étude de groupe sur l'ensemble de données Faces (Wakeman and Henson, 2015).Pour assurer la reproductibilité des résultats, l'ensemble de l'analyse a été scriptée et les tracés ont été générés automatiquement à l'aide du logiciel `sphinx_gallery`[2].

Afin de pousser encore plus loin l'objectif de reproductibilité via l'automatisation, nous avons développé deux nouvelles méthodes d'analyse des données électrophysiologiques. La première méthode, appelée *autoreject*, vise à améliorer la suppression des segments de données contenant des artefacts, qui est une étape de prétraitement de base dans presque toutes les chaînes d'analyse. Nous développons une méthode efficace qui utilise une méthode de recherche de paramètres connue sous le nom d'optimisation bayésienne. Notre approche a été en mesure de faciliter la réanalyse des données du Human Connectome Project (HCP) aux fins de l'analyse comparative. Notre deuxième méthode, connue sous le nom *alphacsc*, permet d'extraire des séries temporelles neuronales pour de nouvelles structures oscillatoires. Non seulement cela, c'est un outil pour estimer des formes d'onde plus précises que ce qui est possible en utilisant l'analyse de Fourier traditionnelle. Dans notre travail, nous avons démontré qu'il était capable de découvrir des oscillations imbriquées à partir des données.

Ces technologies peuvent encore être considérées comme étant à leurs balbutiements à bien des égards. Tout comme les méthodes de localisation des sources dans le MEG/EEG ont évolué des modèles basés sur les dipôles vers des méthodes distribuées vers des modèles plus sophistiqués mettant en œuvre une esparsité structurée, ces nouvelles méthodes sont susceptibles de subir un processus évolutif d'améliorations incrémentielles. Si l'on considère l'exemple du CSC, notre modèle basé sur des distributions alpha-stables a étendu les modèles de vision par ordinateur de manière à pouvoir gérer des distributions à queue lourde, ce qui est caractéristique des données neuronales. Évidemment, ce n'est pas la fin

[2]https://sphinx-gallery.github.io

du chemin. Ajuster les hyperparamètres dans les modèles CSC est encore notoirement difficile, mais ce n'est pas impossible s'il y a une tâche supervisée à la fin du pipeline. Les dictionnaires multi-échelles peuvent être critiques pour les signaux cérébraux, étant donné que les oscillations peuvent avoir un support variable. Comme le problème n'est pas convexe, des stratégies d'initialisation plus intelligentes comme celles basées sur MCMC pourraient conduire à des estimations plus précises (Bachem et al., 2016). Il sera bientôt nécessaire de construire des algorithmes CSC basés sur des approximations stochastiques, pour traiter de plus grands ensembles de données.

Contrairement à la vision par ordinateur ou au traitement du langage naturel, les industries à haut risque comme les soins de santé exigent des algorithmes transparents. Il ne suffit plus de se contenter d'une plus grande précision de prédiction. Dans nos travaux sur l'*autoreject* et l'*alphacsc*c, nous utilisons ces données publiques pour développer des algorithmes faciles à interpréter et à diagnostiquer. *Autoreject* identifie les segments de données à supprimer sur la base d'un paramètre unique, facile à comprendre et réglé automatiquement. De la même manière, *alphacsc* exploite directement les formes d'onde prototypiques, afin de remplacer les mesures indirectes pour mettre au jour des phénomènes d'intérêt.

Dans cette thèse, j'ai décrit une stratégie pour une recherche reproductible dans le futur : des ensembles de données publiques avec des échantillons de grande taille et de l'automatisation. Cependant, certains de mes travaux se sont limités à l'automatisation à l'échelle d'un seul sujet. Même si cela nous permet d'analyser de grands ensembles de données, cela peut parfois être limitatif, puisque ca ne nous permet pas de mettre en commun les données entre les sujets, de manière à pouvoir découvrir des effets plus subtils. Alors que nous entrons dans une ère de science rapide, de tels outils basés sur les données deviendront indispensables. Bien qu'un grand nombre de méthodes de recherche aient été axées sur l'amélioration du rapport signal/bruit dans chaque ensemble de données, cela peut s'avérer moins important lorsqu'il s'agit de traiter d'ensembles de données plus importants. Pour l'avenir, nous préférerons de plus en plus de grands ensembles de données qui ne sont pas parfaitement débruités, plutôt qu'un plus petit ensemble de données parfaitement débruité. De nouveaux outils devront être développés afin de permettre aux cliniciens de sonder rapidement le cerveau, de manière à identifier les signaux et les structures d'intérêt, de quantifier les incertitudes ainsi que les scores de précision, d'effectuer des contrôles de qualité et de visualiser interactivement leurs données.

Chapter 2

Introduction

"Progress in science depends on new techniques, new discoveries and new ideas, probably in that order."

—Sidney Brenner

Contents

Understanding the human brain is one of the most significant challenges of the 21st century. The human brain is arguably the most complex organs of the human body, which performs a wide range of cognitive functions: from visual recognition to language understanding, speech, social interaction, and executive control. Pathologies of the brain are perhaps one of the biggest challenges for medicine today. Medical interventions and drugs for major infectious diseases are available today, and individuals can expect to live up to the mid 80s and even into their 90s. Yet, we still do not have a good grasp over most mental pathologies: Parkinson's, Alzheimer's, dementia, epilepsy to name a few. This is despite the fact that someone today who lives into their mid 80s has a 50% chance of contracting Alzheimer's (Alzheimer's Association et al., 2016).

Our current understanding of the brain is a result of decades of concerted efforts across multiple disciplines ranging from molecular biology, genetics, physiology, cognitive and behavioral neuroscience, to statistics, computer science, and data science. A relatively new subfield here is brain imaging, also known as neuroimaging. Brain imaging refer to a set of technologies where snapshots of the brain are taken. These snapshots can be static, as in the case of anatomical images from magnetic resonance imaging (MRI), or an evolving image, as in the case of functional magnetic resonance imaging (fMRI). The grand vision is to utilize them in hospitals to help diagnosis, in surgeries, in brain-computer interfaces (BCIs), or in research for neuroscientists to better understand the brain. In this thesis, we will focus our attention on measuring the electric currents and/or corresponding magnetic field from the brain, using electroencephalography, magnetoencephalography, and local field potentials. These methods have the property of possessing a high temporal resolution, which is particularly useful for extracting the temporal dynamics of brain signals.

In this manuscript, I will describe several methodological advances that we achieved in brain imaging using our expertise in machine learning and open source software. In the forthcoming sections, I will first describe the context of the thesis. After a brief introduction to electrophysiology, I will delve into how the field has been shaped in recent years by the reproducibility crisis, and how data sharing is going to help alleviate this problem to a large extent. However, the rise of data sharing and large sample sizes makes it difficult to still rely on manual data analysis which does not scale and is not reproducible. To cope with this challenge, we must start relying on automated and data-driven methods for discovering new effects. To this end, I will introduce two new algorithms *autoreject* and *alphacsc*. This is followed by a refresher on background material that might be useful to refer to when reading the thesis. The chapter ends with the contributions of this thesis and the list of papers published during the PhD.

2.1 Electrophysiology

The study of electrical properties of the biological cells and tissues is known as electrophysiology. Biological tissues have electrical properties due to the presence of ions. Just as we can measure voltages in electrical appliances, it is possible also to measure these voltages in living tissues. Electrophysiological recording have the advantage of directly measuring the brain activity, as opposed to an indirect measure, which is for example the case in fMRI. Brain imaging techniques are characterized by their temporal and spatial resolution, *i.e.,* the time scale at which it can measure brain activity, and also the accuracy

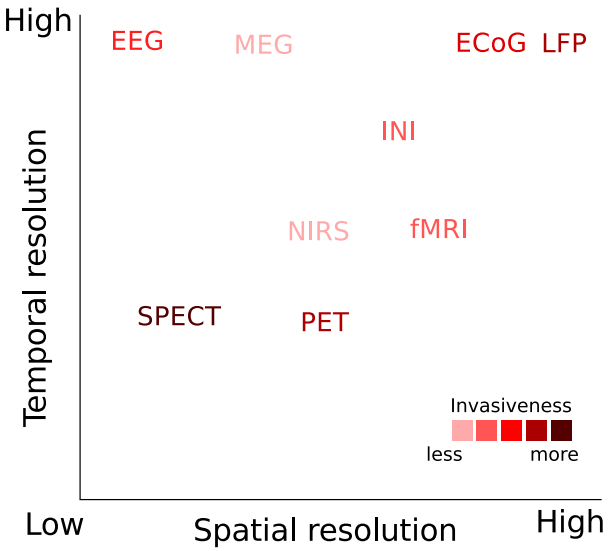


Figure 2.1: Various neuroimaging methods differ in terms of the information they measure. MEG=magnetoencephalography, EEG=electroencephalography, NIRS=near-infrared spectroscopy, PET=positron emission tomography, SPECT=single photon emission tomography, and INI=Inverse imaging, a method to speed up acquisition of fMRI images, ECoG=Electrocorticography, LFP=Local Field Potential.

of localizing the source of the activity. Figure 2.1 summarizes different neuroimaging methods with respect to their temporal and spatial resolution. In the case of fMRI, as it measures the blood flow which is a slow response to neural activity, its temporal resolution cannot be high.

There are a number of methods to measure the electrical potentials in the human body, the most well-known being perhaps electrocardiography (ECG) which is used to measure the electrical activity of the heart. However, in our work, we will focus on only three which are relevant for studying the brain. Each of these methods produces a multivariate time series.

Electroencephalography: Electroencephalography (EEG) is a portable and non-invasive measurement technique invented in the 1920s that is used in several contexts such as BCIs, monitoring and diagnosis, and in cognitive studies. In electroencephalography, an array of electrodes on an EEG cap is placed on the scalp to measure the voltages with respect to a reference electrode. The voltage it measures is not the result of a single neuron but instead a result of the electrical activity of populations of neurons. It has a high temporal resolution (in the order of ms), however the spatial resolution is not so high.

Magnetoencephalography: Any electric current is associated with magnetic fields as a consequence of Maxwell's theory. Therefore, the brain generates tiny magnetic fields which wrap around the currents according to Maxwell's right hand thumb rule. The field is tiny ($\sim 10^{-12}T$) compared to the earth's magnetic field ($\sim 10^{-4}T$) and ambient magnetic noise ($\sim 10^{-6}T$). Therefore, to measure it, one would need very sensitive electronics and heavy noise cancellation. The measurement itself is done in a magnetically shielded room made of three layers of metals. The sensors are superconducting coils which capture the magnetic flux. They are immersed in liquid Helium at very low temperatures (around

4 K), so as to lower any loss in signal due to resistance. A typical device contains two types of sensors: gradiometers and magnetometers. While the magnetometer measures the absolute magnitude of magnetic field, the gradiometer measures gradient of the field. Magnetoencephalography (MEG) has the advantage that the skull does not deteriorate the signal quality as in EEG.

Local field potential (LFP) The Local Field Potential is the electric potential that is recorded in the extracellular space of the brain tissue. In contrast to EEG, LFP are recorded in depth, from within the cortical tissue and can therefore measure more localized populations of neurons. Small intracerebral electrodes are typically used to measure these potentials as opposed to large surface electrodes used in EEG.

2.2 Context of the thesis

This emerges out of the recent movements in reproducibility and data sharing in neuroimaging. It focuses on simplifying data analysis through better educational tools and automated methods to enable reproducible analysis in the age of big data.

2.2.1 The reproducibility crisis

Even though thousands of papers are published every year about different aspects of the brain, our understanding of this complex organ has not scaled in proportion. A large part of the reason has been attributed to what is known as the reproducibility crisis (Ioannidis, 2005a; Simmons et al., 2011; Button et al., 2013). Progress in science rests on reproducible experiments. Reproducibility refers to the fact that the findings of an experiment can be regenerated independently if the code, data, and related software was provided. In many fields, however, a large fraction of experiments cannot be reproduced. In psychology, for instance, it was estimated that over half of the papers were not reproducible (Collaboration et al., 2015), and even those which could be reproduced tended to have a weaker effect size compared to the original studies.

The reasons for unreproducible results can be numerous (Baker, 2016), some being: 1) confirmation bias, the tendency to selectively report only experiments that conform to the researcher's pre-existing beliefs, 2) "p-hacking" (Simmons et al., 2011), or the tendency to try multiple hypothesis to get a positive result, 3) publication bias or the absence of incentives to publish negative results (Rosenthal, 1979), and 4) pressure to publish. There is now an accepted set of recommendations to address many of these issues: 1) pre-registering research plans to avoid confirmation bias and even report negative results, 2) correct for multiple comparisons, the most conservative method being the Bonferroni correction (Dunn, 1961).

Brain imaging has its own set of issues which can be linked to reproducibility crisis:

- **Power failure:** This is arguably one of the central issues in the reproducibility crisis today and has received by far the most attention. The statistical power of a study refers to the likelihood of discovering an effect of interest, given the sample size. Small sample sizes translate into underpowered studies which means that the chance of a false discovery is high. In order to discover the effect of interest, the study must be appropriately powered.

- **Multiple comparison:** This is essentially a manifestation of "p hacking" that is a result of of the large number of voxels or time points in neuroimaging. For instance, in the famous dead salmon study (Bennett, 2009), a significant effect was found even if none was expected simply because the hypothesis testing (comparisons) was done over each voxel.
- **Differences in software versions:** Changing software versions can lead to different results. For instance in the case of Freesurfer software, differences in volume were found in the range of $8.8\% \pm 6.6\%$ (Gronenschild et al., 2012).
- **Complex pipelines:** Neuroimaging pipelines involve a number of choices at each processing stage, and there is currently no consensus on how to choose the right pipeline. Often, these methodological choices are not even documented. It is estimated that there are almost as many unique pipelines as there are studies (Carp, 2012b).
- **Confounds:** There are several methodological confounds such as head movements (Yendiki et al., 2014), anatomy differences, and changes in breathing rate and depth, which can lead to spurious correlations.

In Chapter 4 of the thesis, we will provide concrete guidelines on how to build processing pipelines for MEG/EEG data. Our contribution will touch upon the issue of complex pipelines, multiple comparison, and differences in software versions in the context of MEG/EEG. The issue of power failure can be alleviated through data sharing as I will discuss in the next section.

2.2.2 Data sharing

Power failure is essentially a consequence of small datasets. In today's collaborative and data-driven scientific environment, data sharing is useful not only from the perspective of reproducibility but also to build datasets with large sample sizes. With large datasets, it would be possible to tease apart even subtle effects (Smith and Nichols, 2017) that were not possible with smaller datasets. Data sharing is beneficial not just from the perspective of replication but also from an economic perspective. Rather than collect new data for every new hypothesis, researchers can now reuse known data for testing the validity of their hypotheses.

The benefits of data sharing can be traced back to Newton and his theory of gravitation (Jardine, 2013). Before Newton had developed his theory, another English astronomer, John Flamsteed had been appointed by the king to observe the stars and produce accurate charts for navigation in the seas. Over a period of 40 years, Flamsteed created a detailed catalogue that tripled the number of entries in the previously used sky atlas. When the great comet of 1680 appeared in the sky twice in close succession, Flamsteed used his data to postulate that it was not two comets but in fact the same comet which first went towards the sun and then turned away from it. Newton initially opposed this theory, but later changed his mind as he gained access to Flamsteed's unpublished catalogue. The comet had indeed turned out to be an important benchmark for Newton's theory of gravitation.

It is hard to imagine in this day and age that a theory as fundamental as the laws of gravitation could have been data driven. Data sharing is fundamental not only to reproducible science, but also it forms the cornerstone for learning stronger models and

benchmarking new algorithms. Consequently, in machine learning, recent breakthroughs have been powered by the increase in data sharing and computation. This includes the recent growth of deep learning (Deng et al., 2009), Q learning (Watkins and Dayan, 1992; Bellemare et al., 2013), natural language processing for language translation (Halevy et al., 2009), speech recognition (Paul and Baker, 1992), and even the mixture of experts model (Jacobs et al., 1991) for IBM Watson (Ferrucci et al., 2010). The maxim, "more data beats a cleverer algorithm" (Domingos, 2012) has held up remarkably well across disciplines and over the ages.

Of course, neuroscientists are beginning to realize the importance of sharing data. In recent times, neural data has started being shared through international consortiums (Van Essen et al., 2013; Ollier et al., 2005), data repositories (Poldrack et al., 2013; Gorgolewski et al., 2015) and dataset papers in targeted journals. Yet, there is an unaddressed gap in terms of the ideal of data sharing and the practice of data sharing. Neuroimaging experiments are often very complicated, and it is not enough to share simply the data, but also the metadata and information regarding the experimental protocols in a well-structured format. In the absence of this information, shared data is not *reusable* in the same way that uncommented and poorly structured, complicated programs are not useful even if shared publicly. There is not an accepted consensus in the community on the practices of sharing data and there is a need to establish a standard. In Chapter 4, we will present a new standard known as the Brain Imaging Data Structure (BIDS), which is intended to address this gap. It is a collaborative effort between software developers and neuroscientists across various laboratories to establish a consensus on the standards and build tools to facilitate adoption of the standard.

2.2.3 Automation

Back in 2014, Nature published a bold article (Hayden, 2014) which described a vision for the future of science: automated labs that would autonomously record every detail in an experiment, which in turn would lead to cheaper, more efficient and reliable research. While it goes on to describe many biology labs which are automating experiments, the benefits of automation in the neuroimaging community are yet to be widely recognized. Automation not only saves time but also makes the research more reproducible, as was noted in a recent guide to improve the transparency and reproducibility of neuroimaging research (Gorgolewski and Poldrack, 2016). The authors point out that manual work may seem easy at first, if the analysis has to be performed only once. However, this is not always the case as "quite often in the course of a project, parameters are modified, subjects are changed, and processing steps need to be rerun. This is a situation in which having a set of scripts that can perform all of the processing steps automatically instead of relying on manual interventions can really pay off." As large datasets become more common in neuroimaging, automation will indeed become a necessity rather than a luxury.

In neuroimaging, there are in fact several avenues for automation:

- **Reducing interactivity:** While interactive graphical user interfaces are excellent tools for browsing the data, they fall short when it comes to scaling up the analysis to tens and hundreds of subjects, which is necessary for a sufficiently powered study.
- **Parameter tuning:** Most algorithms, although scripted, still require hyperparameters to be tuned. These hyperparameters could be the number of ICA components

to choose or the regularization parameters, and can vary from one subject to the next.

- **Annotation and labeling:** A large fraction of neuroimaging data that is available is unlabeled or at best weakly labeled. This is because expert annotations are expensive, and cannot be crowdsourced. Automated tools based on unsupervised learning can play a major role in this regard.
- **Quality control:** Currently, quality control is performed manually by inspecting the data to spot outliers. While data inspection cannot be overlooked, it can be performed more efficiently through automated documentation of data analysis and log reports such as the Jupyter notebook and the MNE web report (Engemann et al., 2015a). At the same time, advanced statistical trend analyses as in the Automated Statistician project (Duvenaud et al., 2013) can be used for creating summaries.

There have been some steps taken in this direction, most notably the Neurosynth platform (Yarkoni et al., 2011) which facilitates large-scale meta analysis. Meta analysis typically combine results from multiple studies, and in this case, it is the brain activation maps from different studies which are combined by using machine learning methods. On the software side, the Freesurfer software package (Dale et al., 1999; Fischl et al., 1999) provides a `recon-all` command that performs cortical segmentation automatically without any human intervention. In MNE, this philosophy is now being adopted starting with automated covariance estimation (Engemann and Gramfort, 2015).

In this thesis, we will consider an algorithm that automatically annotates artifacts in the data (Jas et al., 2016, 2017a). This is a first step that any MEG/EEG processing pipeline has to go through but it is often done manually. A reason for this is that existing algorithms are not designed to be *transparent*. Since for most scientists, the key to new insights is an artifact-free dataset, they would rather spend extra effort in doing this manually rather than depend on a generic algorithm which is difficult to interpret. Merely based on anecdotal reports, this process can take up to a week even for a moderately sized study of 10–20 subjects.

This is what led us to propose *autoreject*, which we describe in Chapter 5. It is an algorithm which can be used to mark bad segments of the data. The key insight is that, often certain sensors in the device are intermittently corrupted rather than continuously. We validate our algorithm against 3 benchmarks on the HCP dataset (Larson-Prior et al., 2013) which is manually annotated with bad segments. In the process, our work also represents one of the first attempts at reanalyzing the MEG component of the HCP dataset.

2.2.4 Representation learning for data-driven discovery

Since the invention of EEG in the 1920s, scientists have discovered several different brain oscillation patterns such as alpha waves, K-complexes, and mu rhythms. The oscillations and interactions between them have served as biomarkers for different brain functions and pathologies. Alpha waves have been implicated in attention, K-complexes in sleep, and mu rhythm in motor activity.

Considering the complexity of the human brain, clearly these waveforms represent only a fraction of the cognitive functions that the brain may perform. As a result of the wealth

of data now available through the data sharing movement described in Section 2.2.2, the future neuroscientist will be able to mine such waveforms from large datasets. Imagine if neuroscientists had at their disposal a tool similar to Google Photos[1]. In the same way that Google Photos can automatically find faces and group photos, such tools will be able to find prototypical oscillations and cluster the data using them. Clicking on any of these waveforms would retrieve the data associated with them.

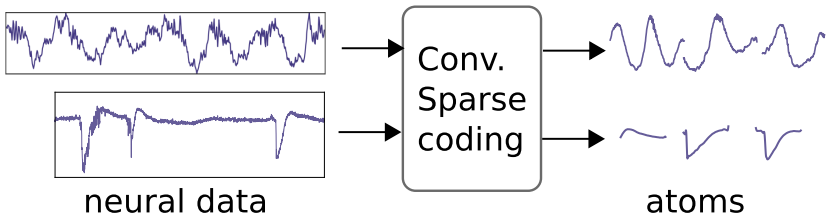


Figure 2.2: An illustration of how Convolutional sparse coding can be used to automatically mine prototypical waveforms

However, photos are inherently different from neural data. First, neural data can be buried in noise and corrupted by high amplitude artifacts. Second, images are labelled owing to crowdsourced data as in the case of Imagenet (Deng et al., 2009), but neural data is not. Expert annotations in the case of neural data are not easily available. Finally, it is spatiotemporal data with different dynamics from the 3D world that photos capture. This is where convolutional sparse coding (CSC) can play a role by extracting prototypical features from the data, as shown in Figure 2.2. It is an unsupervised algorithm from computer vision, which can learn shift-invariant dictionaries of prototypical waveforms (atoms) from the data using the convolution operations. For a more comprehensive background on CSC, the reader may read Section 2.3.4 later in this chapter.

CSC algorithms do not approximate the signal using Fourier (or sinusoidal) basis. While this is the conventional technique for extracting signals buried in noise, the approximation can degrade the shape of the signal, which can be a biomarker in many clinical diseases (Cole and Voytek, 2017). As an example, even with a large number of sinusoids from the Fourier basis, the edges of a square wave cannot be approximated well. Indeed, the imperfect approximation around such edges is what is often termed as ringing artifacts in signal processing contexts. Of course, transients can be better approximated using wavelets but it is clearly not sufficient for other shapes of data. Rather than fix the basis to be Fourier or wavelet, the CSC approach is to learn *both* the basis and the coefficients.

In our work presented in Chapter 6, we extend conventional CSC algorithms for heavy-tailed noise. We reformulate the optimization problem as a MAP inference with an alpha-stable distribution to replace the reconstruction loss. Our results show that this kind of algorithm is robust to the presence of artifacts and can be used to uncover temporal structures from neural signals, even those involving nested oscillations.

[1] https://photos.google.com/

2.3 Mathematical Background

Here, we will introduce some basic linear algebra, optimization, and machine learning concepts that will be useful particularly in Chapters 5 and 6 on *autoreject* and *alphacsc*.

2.3.1 Norms

Informally speaking, a norm is used to measure the length or size of a vector. It must also satisfy some properties, but it will not be of concern for us in this thesis. It is sufficient to know the mathematical expression, how it behaves, and the physical property that it captures.

Definition 2.1. (ℓ_p norm.) For $1 \leq p < \infty$, the ℓ_p norm of a vector x is defined by:

$$\|x\|_p = \Big(\sum_n |x_n|^p \Big)^{1/p}. \tag{2.1}$$

Definition 2.2. (ℓ_0 "norm".) The ℓ_0 "norm" of a vector x is defined by:

$$\|x\|_0 = \sum_n |x_n|^0. \tag{2.2}$$

Definition 2.3. (ℓ_∞ norm.) The ℓ_∞ norm of a vector x is defined by:

$$\|x\|_\infty = \sup_n |x_n|. \tag{2.3}$$

As we are taking the supremum, this norm is sensitive to large values in the vector, which is needed for measuring artifacts.

Definition 2.4. (Frobenius norm.) The Frobenius norm of a real-valued matrix A is defined by:

$$\|A\|_{\text{Fro}} = \|\text{vec}(A)\|_2 = \sqrt{\text{trace}(AA^\top)}. \tag{2.4}$$

The Frobenius norm is a matrix norm that is simply the ℓ_2 vector norm of the vectorized matrix $\text{vec}(A)$.

2.3.2 Cross validation

Cross validation is a statistical technique to estimate how well a predictive model will *generalize* to unseen data. We will be using cross validation in our *autoreject* work presented in Chapter 5. Cross validation is normally performed by partitioning the data into two and learning the model from one part (typically the larger one) while validating it on the other. These two parts are known as the *training* and the *validation* sets respectively. In order to get reliable estimates of the model performance, this procedure is repeated multiple times and the results are averaged.

Depending on the data, different types of partitioning schemes are preferred. In *K-fold* cross validation, the data is divided into K equal parts (with or without shuffling the samples), where $K-1$ parts are used for training and the Kth part is used for validation. In a *stratified* cross-validation scheme, the partitions are done such that each partition has roughly equal number of samples from each class.

Underfitting and overfitting: When a statistical model explains the data too closely, it results in overfitting. An overfitted model is often unable to make accurate predictions of unseen data. The converse is also true. When the model is unable to explain much structure in the data, it is known as underfitting. As we will see in Chapter 5, these concepts are closely related to model selection. The goal is typically to find a model with hyperparameters (*e.g.* regularization constant or the width of a Gaussian) that neither underfit nor overfit. We will next discuss some strategies to tune these hyperparameters.

Grid search: Quite often, a machine learning model contains hyperparameters which need to be tuned to get optimal performance on the data. This tuning to select the best model can be done automatically using cross validation. The idea is to exhaustively search over the parameter space by trying an equally spaced grid of hyperparameters that the model can admit and selecting the one which performs the best.

Random search: Sometimes grid search is too slow, and it is better to try parameters sampled randomly without a considerable loss in performance. This is known as random search.

Nested cross validation: When cross validation has been used for finding the best model, a separate unseen partition, known as the *test* set, is needed to assess the generalization power of the best model. The test set must be different from all the data that was seen so far, including *both* the training and the validation sets. This is done to avoid a circularity bias as noted in Varoquaux et al. (2017); Cawley and Talbot (2010); Friedman et al. (2001). Therefore a nested cross-validation scheme is necessary where in each outer loop, the best model is found, and its performance is computed.

2.3.3 Bayesian optimization

Sometimes grid search is too slow and random search is not sufficiently accurate for tuning hyperparameters. In such cases, one may have to rely on more efficient black box optimization techniques. Bayesian optimization (Snoek et al., 2012) is one such method which is a sequential method, and iteratively improves the objective function that we are optimizing.

Since these methods must work for any kind of function which may not be necessarily convex, we cannot rely on classical gradient based methods. The central idea in these types of methods is to search as much of the hyperparameter space as possible, while keeping a record of the objective values evaluated so far. A very natural strategy is therefore to either search around points which have high objective values, or in those areas where it is unknown. This is indeed a case for the classic "exploration" *vs.* "exploitation" dilemma that is well known in computer science.

In order solve this dilemma, an *acquisition function* is typically used, as shown in Figure 2.3. While the "exploitation" component of this function is easy to compute from the objective values, in order to compute the "exploration" component, we can fall back upon Gaussian processs (GPs) (Rasmussen, 2004). Given the points evaluated until now, a GP can be used to estimate the uncertainty in the rest of the search space.

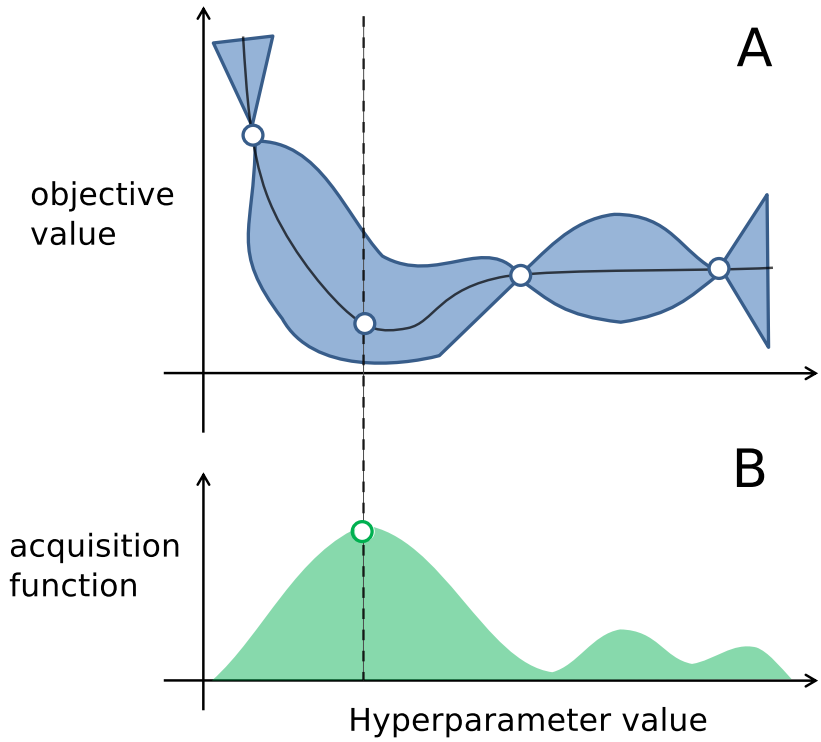


Figure 2.3: An illustration of how Bayesian optimization is used to sequentially select new points (A) Gaussian process to estimate uncertainties on the objective function for the parameter space, and (B) Acquisition function to balance the "exploration" *vs* "exploitation" dilemma.

2.3.4 Dictionary learning

As we described in Section 2.2.4, we will be using the framework of dictionary learning to automatically mine prototypical waveforms from neural time series. Our work will build upon existing work in shift-invariant sparse coding and data decomposition methods. In this section, I introduce some concepts, which are well known in the dictionary learning community, but might be useful for anyone not so familiar with the area.

Dictionaries and sparsity: A *dictionary* is a set of *atoms* (also known as filters sometimes) which can be combined with certain *coefficients* (or activations) to approximate the data. The atoms could be fixed (for example wavelets or Gabor), or they could be learned directly from the data itself. As such, there is no requirement of orthogonality on the atoms. To learn a representation of the data boils down to estimating the coefficients used in the approximation, which are often assumed to be sparse, *i.e.*, they have very few non-zero values. The learning algorithm is therefore called *sparse coding* and the coefficients learned are known as the *sparse code*.

Sparsity can be promoted by using an ℓ_0 or ℓ_1 penalty, which helps maintain a compact representation. Note that the ℓ_0 "norm" simply counts the number of elements in the vector (c.f., Definition 2.2) , therefore adding it as a regularizer will favour solutions that have fewer non-zero elements. In the traditional approach, it is quite common to start off with an *overcomplete* dictionary, *i.e.,* to have more atoms than would be needed, and only estimate the coefficients (keeping the atoms fixed). This can be done, for example,

using matching pursuit (Mallat and Zhang, 1993), which is a greedy algorithm for sparse approximations.

Learning atoms: It is easy to see that the overcomplete approach is memory intensive as it has more atoms that necessary, but also this approach requires making assumptions about the shape of the atoms in the dictionary. Nowadays, it is more common to learn the atoms in addition to the sparse code, which is known as *dictionary learning*. Dictionary learning can be thought of as a data decomposition method (or matrix factorization) technique (like principal component analysis (PCA) or independent component analysis (ICA)), but with a sparse regularization. As for obtaining sparse solutions, the goal is to have as few non-zero coefficients as possible, an ℓ_0 penalty is typically used. However, this leads to a non-convex formulation of the problem. One strategy to deal with this is using a convex relaxation by replacing the non-convex ℓ_0 penalty using the ℓ_1 penalty which is convex. This is often a reasonable approximation which makes the problem biconvex, and can therefore be solved by alternate minimization. Dictionary learning is now being used for denoising (Elad and Aharon, 2006), inpainting (Mairal et al., 2009a), and classification (Mairal et al., 2009b).

Coding for shift invariance: In traditional dictionary learning, the signal or the image is divided into patches and these patches are used as samples for the learning. The main disadvantage of this method is that it results in redundant atoms that are shifted versions of each other. This is the reason a shift-invariant version of dictionary learning is needed. One of the earliest paper in this regard can be credited to Lewicki and Sejnowski (1999). In their work, the dictionary was fixed and the shift-invariance was encoded using convolutions. This is possible because the convolution operator is defined by an integration (correspondingly summation in the discrete domain) as below.

Definition 2.5. (Convolution) A convolution of two functions f and g is written $f * g$ and computed as:

$$(f * g)(t) \triangleq \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau \tag{2.5}$$

In the discrete case where f and g are vectors with finite support $[0, .., M]$ and $[0, .., N]$ and $N > M$, this can be rewritten as

$$(f * g)[n] = \sum_{m=0}^{M-1} f[m]g[n-m], \tag{2.6}$$

so that $f * g$ has a support of $[0, .., (N - M + 1)]$ if the edges are truncated. As we can see, if either f or g is zero padded, the other function can be shifted along these zeros yielding the same result.

As a result of the convolutional approach, the atoms learned are not redundant and the location of the atom is encoded in the activations. This type of approach is refered to as CSC or shift-invariant dictionary learning in the literature. In practice, an overcomplete dictionary in the shape of a Toeplitz matrix is constructed by shifting the atoms across time. This is the same formulation which is still being used to code shift invariance even if the learning algorithms are more sophisticated now. More recent work by Grosse et al. (2012) have made use of the Fourier transform to solve the problem in Frequency domain

and compute an inverse transform of the learned dictionary in the end. Another approach that has proved to be quite efficient is the so-called predictive sparse coding which uses neural networks (Kavukcuoglu et al., 2010). It is hard to summarize all the work in this area, but it suffices to say that shift-invariant dictionary learning has been gaining popularity in audio signals, images, music, and now for neural data (See Section 6.1 for a more comprehensive list of references).

2.3.5 Iterative solvers for convex problems

Much of the dictionary learning techniques in literature leverage convex optimization methods, which are iterative techniques to solve minimization problems involving convex functions. In our *alphacsc* work presented in Chapter 6, we will also be using such methods. CSC being a biconvex problem (*i.e.,* convex if either the atoms or the activations are fixed), is amenable to such methods.

Therefore, let us first recall what is a convex function before diving into the details of these methods.

Definition 2.6. (Convex function) A function f is convex if it satisfies

$$\forall x_1, x_2 \in X, \forall t \in [0,1] : f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2). \tag{2.7}$$

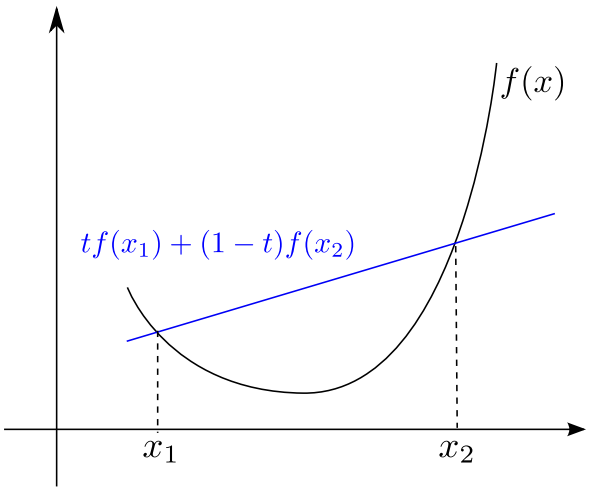


Figure 2.4: A graphical illustration of Definition 2.6 for convex functions.

In other words, for any two points on the function, the line joining those two points lies above the graph of the function (as illustrated in Figure 2.4). A convex function lends itself to the property that the global minimum is also necessarily the local minimum. This concept can also be extended to define strong convexity which guarantees that the local minimum is not only the global minimum, but also it is unique.

Throughout the thesis, we will be solving unconstrained minimization problems of the form

$$x^\star = \underset{x}{\operatorname{argmin}} \, \|Ax - b\|_2^2 + \Omega(x), \tag{2.8}$$

where typically $A \in \mathbb{R}^{n \times p}$ is the design matrix with n rows (samples) and p columns (features), $x \in \mathbb{R}^p$, and $b \in \mathbb{R}^n$. A penalty or regularization term $\Omega(x)$ is added to prevent

solutions that overfit. If there was no regularization term $\Omega(x)$, we could solve this in closed form and get $x^\star = (A^\top A)^{-1} A^\top b$. However, this can be computationally prohibitive as inverting $A^\top A$ is $\mathcal{O}(p^3)$, and iterative methods based on gradient descent are more efficient as they require only a matrix-vector dot product. These are what we shall refer to as *solvers*. More importantly, these gradient based methods can work for arbitrary functions as long as we have access to the gradient (or even its approximation).

In gradient based methods, we make updates of the form:

$$x_{k+1} = x_k + \rho_k d_k, \tag{2.9}$$

where ρ_k is a scalar which defines the step size and $d_k \in \mathbb{R}^p$ is the search direction.

Gradient descent: In the case of gradient descent, the search direction d_k is the negative gradient $-g(x_k)$ and the typical step size $\rho_k = 1/L$ where L is the Lipschitz constant of the gradient, which upper bounds the rate at which it changes. The intuition behind this method is that if we follow the negative gradient direction, we will progressively reach a point where the gradient is 0 and this corresponds to the minimum for convex functions.

However, often the regularization term $\Omega(x)$ is not smooth, and therefore it does not have a unique derivative at all points. In these cases, we must resort to proximal gradient methods.

Proximal algorithms: An example of a non-smooth regularizer which is encountered in practice, and will also be used in our work, is the ℓ_1 norm (Tibshirani, 1996). It induces sparsity using $\Omega(x) = \lambda \|x\|_1$, where λ controls the sparsity level. The higher the λ, sparser the solution. In such situations, we can use what are known as proximal methods. The idea behind proximal methods is to take a gradient step using only the smooth part of the function, and then apply a proximal operator (which depends on the non-smooth part) on the resulting iterate.

Definition 2.7. (Proximal operator) A proximal operator associated with a convex function f is defined as

$$\text{prox}_f(v) = \underset{x}{\text{argmin}} \left(f(x) + \frac{1}{2}\|x - v\|_2^2 \right). \tag{2.10}$$

The proximal operator can be thought of as a generalized projection operator (Parikh et al., 2014). For the ℓ_1 norm with $f(x) = \|x\|_1$, it is the soft thresholding function $\mathcal{S}_\lambda(\cdot)$ which induces sparsity, and is given by

$$\mathcal{S}_\lambda(v) = \begin{cases} v - \lambda & \text{if } v > \lambda, \\ 0 & \text{if } -\lambda \leq v < \lambda, \\ v + \lambda & \text{if } v < -\lambda. \end{cases} \tag{2.11}$$

The resulting algorithm is known as iterative soft thresholding algorithm (ISTA) (Daubechies et al., 2004; Bach et al., 2012), which has a convergence rate of $\mathcal{O}(1/k)$ for k iterations. Proximal algorithms are also helpful when dealing with constraints. For example, in

Equation 2.8, if we had a norm-1 constraint $\|x\|_2^2 \leq 1$, this could be recast using an indicator function $i(\cdot)$:

$$i(x) = \begin{cases} \infty & \text{if } \|x\|_2^2 \leq 1 \\ 0 & \text{if } \|x\|_2^2 > 1 \end{cases} \tag{2.12}$$

In this case, the proximal operator can be shown to be the projection $\pi(x)$ on to the unit ball which is expressed as

$$\pi(x) = \frac{x}{\max(1, \|x\|_2)}. \tag{2.13}$$

This is what is known as *projected gradient descent*. We will encounter such constraints in our dictionary learning problem in Section 6.3.4. It will be used to handle the scale ambiguity between the atoms and activations when performing optimization.

Acceleration: Gradient descent has slow convergence due to oscillations if the condition number of A is high as it leads to pathological curvature. A faster rate of convergence can be achieved using Nesterov accelerations (Nesterov, 1983), which adds a momentum term to the update in Equation 2.9. The momentum term takes into account the update vector from the past iterations so as to dampen the oscillatory behaviour observed in classical gradient descent and ISTA. This results in an algorithm known as fast iterative soft thresholding algorithm (FISTA) (Beck and Teboulle, 2009) which has a faster convergence rate of $\mathcal{O}(1/k^2)$. It has been proved theoretically that this is the fastest rate possible if we had access to only the gradient and function evaluations. The full algorithm is summarized in Algorithm 1.

Algorithm 1 Fast iterative soft thresholding algorithm

Require: Regularization: $\lambda \in \mathbb{R}_+$, Design matrix A, b

$x_1 = 0 = z_1, \beta_1 = 1$
for $k = 1$ to K **do**
  $x_{k+1} = \mathcal{S}_{\lambda/L}(z_k - \frac{1}{L} g(z_k))$ */* Proximal step */*
  $\beta_{k+1} = (1 + \sqrt{1 + 4(\beta_k)^2})/2$
  $z_{k+1} = x_{k+1} + \frac{\beta_k - 1}{\beta_{k+1}}(x_{k+1} - x_k)$ */* Momentum update */*
end for
return x_{K+1}

Quasi-Newton methods: If in the updates of Equation 2.9, instead of using the negative gradient for the search direction, we used $d_k = -H^{-1}(x_k)g(x_k)$, with $H(x_k)$ being the Hessian, it would be known as Newton's method. As opposed to first order methods which use only gradient information, Newton-based methods also make use of the curvature. Newton's method has a locally quadratic convergence, but it may diverge when the current iterate is far from the optimum. Even if this is not the case, a Hessian that is not positive definite can also cause diverging behaviour. Therefore, the step size ρ_k must be chosen carefully, for example using a line search strategy.

Of course, the Hessian is costly to compute and to invert, therefore quasi-Newton methods can be used which approximate it using a matrix B_k. Starting with $B_k = I$, we can

update this matrix at each step using a cheap rank 1 or rank 2 correction. For instance, in the Broyden formula, we would do

$$B_{k+1} = B_k + vv^\top. \tag{2.14}$$

The vector-vector outer product $vv^\top$ is the rank-1 correction in the above formula. The more advanced David-Fletcher-Powell formula uses a rank-2 correction and so does the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. When the entire Hessian matrix cannot be stored in memory, a limited memory version of BFGS (Wright and Nocedal, 1999) can be used where only a few vectors v from the last few iterations are stored for the approximation.

Coordinate descent: In coordinate descent, we minimize one coordinate at a time. The idea behind coordinate descent is rather simple, yet it performs surprisingly well in practice. Thus Equation 2.9 would be coordinatewise, *i.e.,* for each iteration, we would choose a coordinate i_k and perform an update:

$$\begin{cases} x_{k+1}^{(i)} = x_k^{(i)} - \rho_k^{(i)} g^{(i)}(x_k) & \text{if } i = i_{k+1} \\ x_{k+1}^{(i)} = x_k^{(i)} & \text{if } i \neq i_{k+1} \end{cases} \tag{2.15}$$

Of course, we could have also solved each coordinate in closed form instead of using gradients per coordinate. However, in this formulation, if the non-smooth part of the objective (typically the regularizer) is *separable* (*i.e.,* $f(x) = \sum_i f^{(i)}(x_i)$), it allows us to even extend it to proximal coordinate descent. While in coordinate descent, we iterate over the p coordinates, each coordinate update is computationally less demanding than a full gradient update. The coordinates can be chosen either cyclically or at random. In *block* coordinate descent, groups of coordinates are selected for updating rather than one coordinate at a time. It is what we will use for updating the atoms and activations in *alphacsc* (Chapter 6).

2.4 Contributions

In this thesis, I attempt to synthesize the lessons learned from analysing public neuroimaging data with open source software. To this effect, I participated in an international collaboration to create an MEG standard for BIDS (Niso et al., 2018). I wrote the validator which helped create the MEG-BIDS compatible example datasets. As a contributor to MNE (Gramfort et al., 2013a), I led an effort to write a tutorial paper which reanalyzes the Faces dataset (Wakeman and Henson, 2015) for a reproducible group study. In the backdrop of the reproducibility and data sharing movement described in Sections 2.2.1 and 2.2.2, we started automating our pipelines which led us to develop a fully automated algorithm for artifact rejection and repair (Jas et al., 2016, 2017a). Finally, we develop algorithms to learn new undiscovered motifs automatically from neural time series data (Jas et al., 2017c).

The thesis is organized by chapters to highlight these four main contribution areas: data sharing, reproducibility, automation for artifact detection, and automated data-driven motif discovery. Each chapter contains the text from the original paper (which has

been minimally edited in some parts) prefaced by a 1-2 page description of the context surrounding the work.

Journal publications

M. Jas, D. A. Engemann, Y. Bekhti, F. Raimondo, and A. Gramfort. Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, 159:417–429, 2017a

M. Jas, E. Larson, D. A. Engemann, J. Leppakangas, S. Taulu, M. Hamalainen, and A. Gramfort. MEG/EEG group study with MNE: recommendations, quality assessments and best practices. *bioRxiv*, 2017b. doi: 10.1101/240044
(Pending revision at *Frontiers in Neuroscience, Brain Imaging Methods*)

G. Niso, K. J. Gorgolewski, E. Bock, T. L. Brooks, G. Flandin, A. Gramfort, R. N. Henson, M. Jas, V. Litvak, J. T. Moreau, et al. MEG-BIDS, the brain imaging data structure extended to magnetoencephalography. *Scientific data*, 5:180110, 2018

Conference publications

M. Jas, D. Engemann, F. Raimondo, Y. Bekhti, and A. Gramfort. Automated rejection and repair of bad trials in MEG/EEG. In *6th International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 2016

M. Jas, L. Tour, T. Dupré, U. Şimşekli, and A. Gramfort. Learning the morphology of brain signals using alpha-stable convolutional sparse coding. In *Advances in Neural Information Processing Systems (NIPS)*, 2017c

Workshop papers

D. Engemann, F. Raimondo, J. King, M. Jas, A. Gramfort, S. Dehaene, L. Naccache, and J. Sitt. Automated measurement and prediction of consciousness in vegetative state and minimally conscious patients. In *Workshop on Statistics, Machine Learning and Neuroscience at the International Conference on Machine Learning (ICML)*, Lille, July 2015a

Open source implementations

http://autoreject.github.io/
http://alphacsc.github.io/
http://mne-tools.github.io/mne-biomag-group-demo/
https://jasmainak.github.io/bids-validator/

Datasets

https://openfmri.org/dataset/ds000248/

Chapter 3

Brain Imaging Data Structure (BIDS)

"Data! data! data! I can't make bricks without clay."

—Sherlock Holmes

Contents

From the perspective of reproducibility, data sharing is of paramount importance. Sharing code by itself does not enable reproducibility if the accompanying data is not available, and expensive to acquire. Reanalysis of a dataset is however useful not just from the perspective of reproducibility but also for discovering new effects that were previously overlooked. At the same time, the more data is shared, the larger our sample sizes will be and this will enable us to conduct studies with higher statistical power. Low statistical power, as we discussed in Chapter 2 is one of the main reasons for the reproducibility crisis.

While data sharing in neuroscience is on the rise, the amount of data reuse is still limited. For example, since the release of the Human Connectome Project (HCP) (Larson-Prior et al., 2013) MEG data in 2013, there have been very few instances of reusing this data. At the time of writing this thesis, we had only one or two documented cases (Jas et al., 2017a) of reusing the HCP data. Even in these cases, the effort has mostly been limited to reproducing results rather than testing new hypotheses. This clearly represents a gap between the ideal and the practice of data sharing.

Neuroimaging experiments are often complicated involving different cognitive tasks (auditory, visual, somatosensory *etc.*), different acquisition parameters (sampling frequency, number of sensors and their location, measurement device *etc.*), and population parameters (subject's gender, age *etc.*). All of this metadata is necessary information to successfully reanalyze the data. Unfortunately, historically there has been a lack of consensus amongst different labs and industrial manufacturers as to what constitutes useful metadata. This points to the need for establishing standards. While on a first glance, this may appear to be unnecessary bureaucratic red tape, in fact standards exist in almost all facets of our life.

Apart from the meta information that is stored with the data, the data itself is stored amongst one of 10–20 different file formats and at different stages of processing. While there have been some efforts previously to standardize data structures (Gibson et al., 2009; Grewe et al., 2011; Stoewer et al., 2013; Teeters et al., 2015; Bigdely-Shamlo et al., 2016), it has not gained wide acceptability. Designing a new standard is tricky as it requires gaining a community consensus. At the same time it must strike the right balance between rigidity for efficiency and flexibility for adapting to future technologies.

The Brain Imaging Data Structure (BIDS) format (Gorgolewski et al., 2016) is indeed designed with these considerations in mind. The standard involves a hierarchy of folders to describe the imaging technology used, the name of the subject, and the date of the experiment. At each level of hierarchy, files are accompanied by sidecar `json` files describing the metadata. A `json` file is an easy to parse text file that contains key and value pairs, so that it has the advantage of being machine and human readable at the same time. These files follow an *inheritance principle*, that is, a field described in a `json` file in a higher level of the hierarchy will be automatically propagated downstream. The main BIDS specification is accompanied by extension specifications which describe specific aspects to describe different modalities. At the same time, the standard does not exist in isolation. The BIDS consortium is also providing a growing ecosystem of tools to convert datasets into BIDS compatible format as well as to validate data to conform to the standard.

In the present work, we present a significant extension of BIDS to support the specific aspects of magnetoencephalography (MEG) data. MEG, as we know from Chapter 2, provides direct measurement of brain activity with millisecond temporal resolution and unique source imaging capabilities. So far, BIDS has provided a solution to structure the organization of magnetic resonance imaging (MRI) data. Despite the lack of standard data format for MEG, BIDS-MEG is a principled solution to store, organize and share the typically-large data volumes produced. It builds on BIDS for MRI, and therefore readily yields a multimodal data organization by construction. This is particularly valuable for the anatomical and functional registration of MEG source imaging with MRI. With BIDS-MEG and a growing range of software adopting the standard, the MEG community has a solution to minimize curation overheads, reduce data handling errors and optimize usage of computational resources for data analysis. The standard also includes well-defined metadata to facilitate future data harmonization and sharing efforts, and extensions to other electrophysiological data modalities.

My contributions to the effort:

- Development of a new tool MNE-BIDS to automatically convert existing datasets into BIDS compatible format.
- Development of the BIDS javascript validator to be compatible with MEG data.

Section 3.1 to Section 3.5 was published in:

- G. Niso, K. J. Gorgolewski, E. Bock, T. L. Brooks, G. Flandin, A. Gramfort, R. N. Henson, M. Jas, V. Litvak, J. T. Moreau, et al. MEG-BIDS, the brain imaging data structure extended to magnetoencephalography. *Scientific data*, 5:180110, 2018

3.1 Introduction

The Brain Imaging Data Structure (BIDS) is an emerging standard for the organization of neuroimaging data (Gorgolewski et al., 2016). The significance of BIDS is timely: there is increasing availability of open neuroimaging data resources, and strong interest in aggregating large, heterogeneous datasets to harness machine learning techniques and address a new range of scientific questions, with greater statistical power. A single neuroimaging study by itself can represent a large and intricate volume of data, with multiple protocols and modalities, and several categories of participants possibly enrolled in repeated sessions. These aspects pose challenges to data organization, harmonization and sharing. The situation is aggravated by the lack of a unique neuroscience standard for digital data across, and sometimes within, modalities such as electrophysiology. Consequently, present data management practices are often based on solutions that do not generalize between labs, or even between persons within the same group. This leads to suboptimal usage of human (time lost retrieving data), infrastructure (data storage space) and financial (limited longevity and value of disorganized data after first publication) resources. Poor or lacking data management strategies also negatively affects the reproducibility of results, even within the lab where the data were collected. BIDS is a standard to describe the organization of MRI data. It is based on a simple, hierarchical folder structure, with key study parameters documented in text-based metadata files. One benefit is the handling of multiple MRI data sequences, with minimal curation overheads, which reduces the possibility of data-handling errors. An important secondary outcome is the facilitation of interoperability between tools for data analytics, provided that software and pipelines adopt BIDS for data inputs.

We describe here a key extension of BIDS to electrophysiology data. The technical sophistication of MEG makes it the most challenging electrophysiology data type for standardization (Baillet, 2017b). For this reason, BIDS-MEG can readily be generalized to electroencephalography, multiunit recordings, and local field potentials. Further to strengthening and rationalizing data management in MEG labs, BIDS-MEG provides a common structure to present and future large MEG open-data repositories (Larson-Prior et al., 2013; Taylor et al., 2015; Niso et al., 2016b). The absence of unique data file format in MEG is compensated by BIDS-MEG's standard data organization: the sharing and processing of large and complex data hierarchies are simplified, and made compatible and reproducible across tools for data analytics.

To derive the BIDS-MEG specifications, we have combined perspectives from investigators, technical support staff and data managers. We also involved the expertise of leading academic software developers for MEG science (Baillet et al., 2011), including Brainstorm (Tadel et al., 2011), FieldTrip (Oostenveld et al., 2011), MNE (Gramfort et al., 2014), and SPM (Litvak et al., 2011). The proposed BIDS-MEG specifications are presently compatible with these software applications and toolboxes.

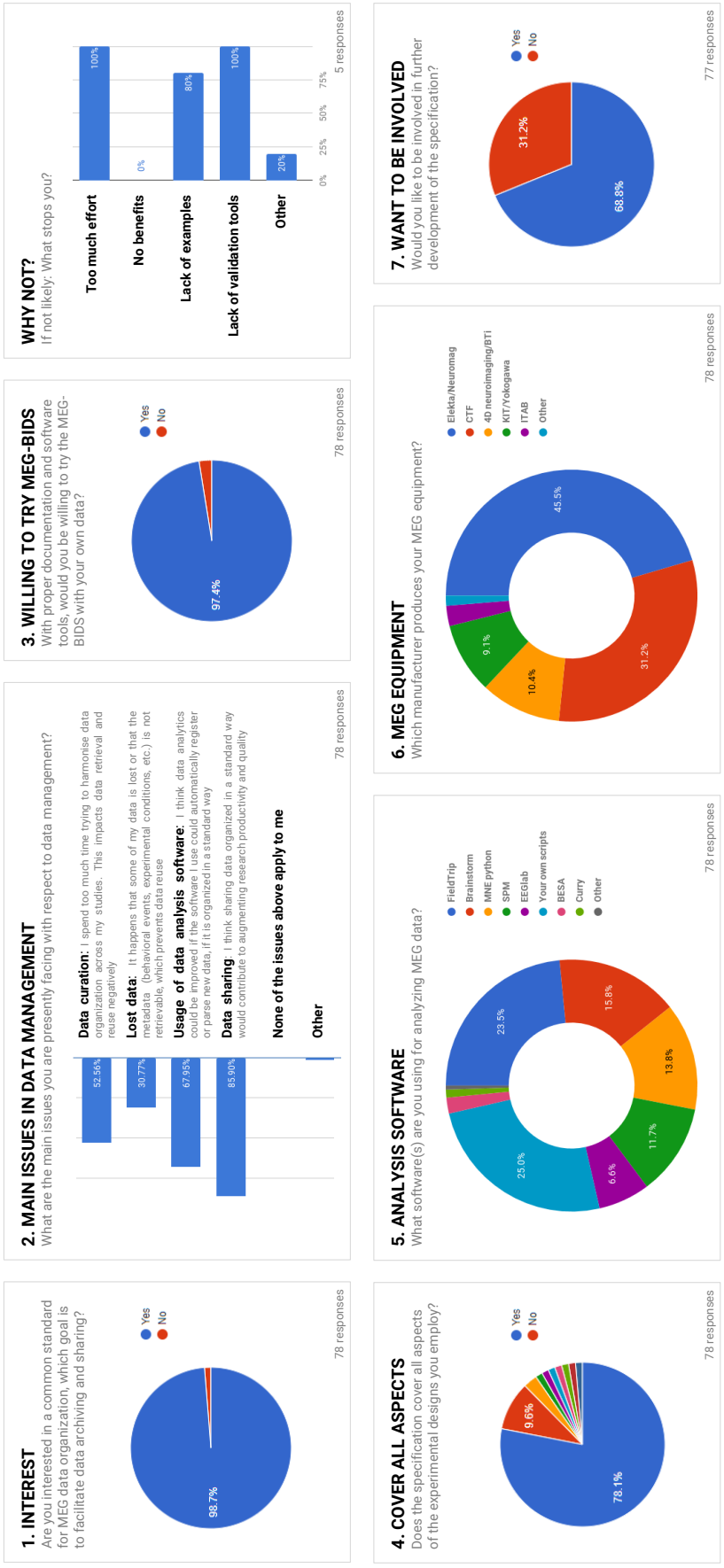


Figure 3.1: Results from the BIDS-MEG poll

3.2 Technical specification

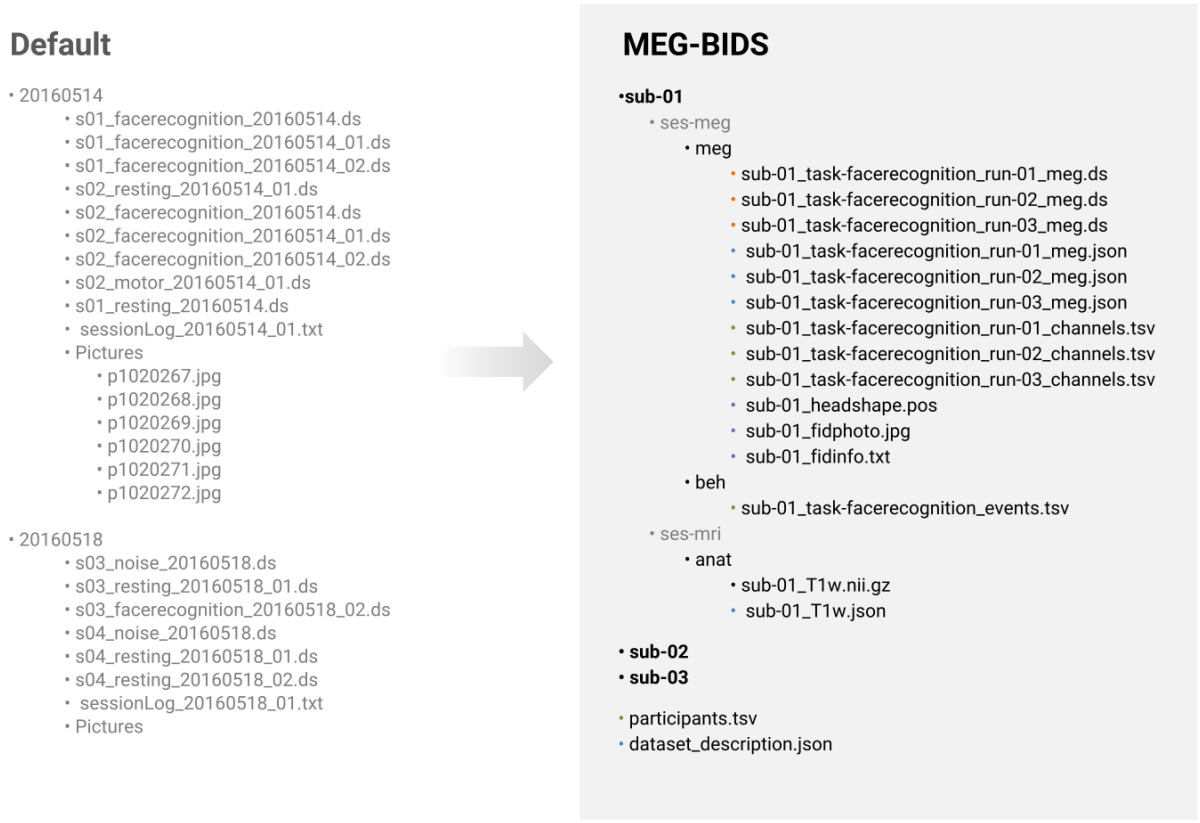


Figure 3.2: BIDS-MEG data organization scheme: Left: a typical default data organization scheme where folders are organized by date of session and contain different runs for a given participant in a study. Right: BIDS-MEG organizes data per study, then participant (subject), followed by modality, then sessions and eventually, runs. Note the sidecar files that are present at all levels of the data hierarchy, and document conveniently the metadata contents.

The BIDS-MEG fields and data organization were defined, bearing in mind best-practice guidelines for conducting MEG research (Gross et al., 2013b). The initiative fostered contributions from multiple MEG experts, with a face-to-face discussion at the 2016 International Conference on Biomagnetism, where the first incarnation of BIDS-MEG was introduced (Niso et al., 2016a). The first set of feedback comments and the minutes from further group discussions are publicly available (`https://groups.google.com/d/msg/bids-discussion/xTHBsGhu0hk/MN25xbxRBwAJ`).

A first version of the present manuscript was shared via the preprint server bioRxiv, from where more comments stemming from the community were collected and considered for improvement of BIDS-MEG. A poll survey to probe the interest of the concerned community was also conducted (see Figure 3.1).

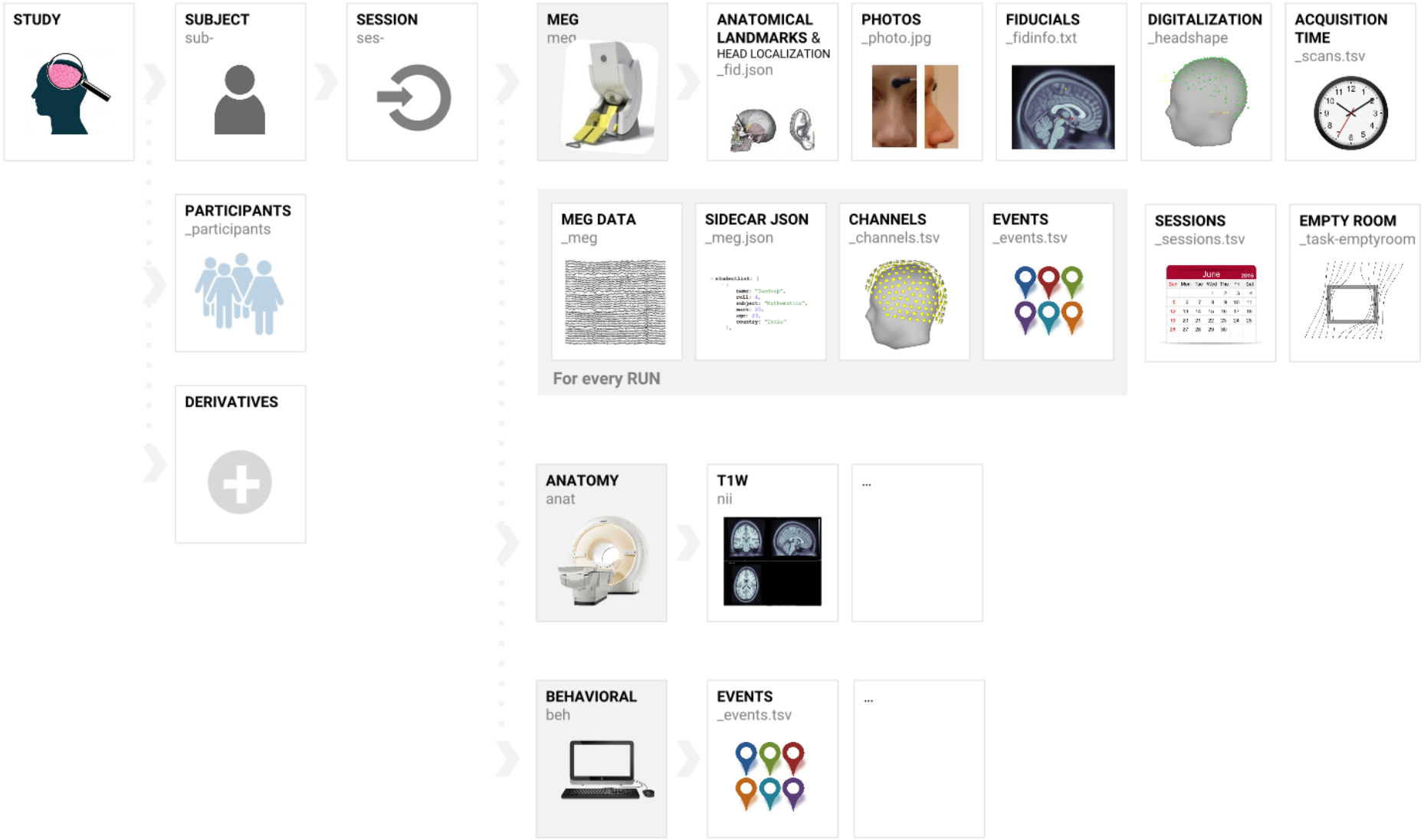


Figure 3.3: BIDS-MEG general overview

BIDS-MEG builds on the BIDS hierarchical data structure (see Figure 3.2 and 3.3). For instance data descriptors such as subject, session, technique, run are BIDS notions that were re-used in BIDS-MEG. Similarly, the simple although extensively used human and machine readable file formats (JavaScript object notation (JSON), and tab-separated value (TSV) text files) that contributed to the versatility and practicality of documenting metadata elements in BIDS, were expanded with BIDS-MEG. BIDS-MEG employs a straightforward terminology, cautiously defined in line with the general BIDS specifications, although adapted to the unique requirements of MEG. For further reference, the BIDS-MEG specifications are detailed in an open-access online document: https://docs.google.com/document/d/1FWex_kSPWVh_f4rKgd5rxJmxlboAPtQlmBc1gyZlRZM)

The terms used refer to notions that were defined by reaching a consensus amongst the BIDS-MEG contributors and the MEG community. For example, 'Subject' refers to the scanned participant. Note that from a technical standpoint, MEG is not a scanning technique. Yet, we used this terminology for convenience, affinity with other neuroimaging modalities, and to reflect the language used in most MEG labs. A 'Session' defines a non-intermittent period of time during which the subject is in the scanner. A 'Run' is a period of time during which empty-room (for noise characterization) or brain activity is recorded continuously, with no interruptions. It is typical with MEG that a session consists of multiple runs: task instructions can change and/or participants can take a break between runs. The notion of 'Task' refers to the instructions (and corresponding stimulus material) that are performed by the participant. 'Responses' is a feature to indicate the recorded behaviour of the subject in relation to the task.

As with the general BIDS specifications (Gorgolewski et al., 2016), BIDS-MEG file names are constituted by a series of key-value pairs, with multiple possible file types. Some typological aspects are mandatory, while others remain optional, although required to abide to the BIDS guidelines. BIDS-MEG can therefore register data of any kind, including but not limited to task-based, resting-state, and empty-room MEG recordings (e.g., for noise estimation purposes). We emphasize that all of the above notions apply also to EEG, and all modalities of electrophysiology, for which BIDS-MEG serves as template for standardization.

There is no common, open or standard file format in MEG, equivalent to DICOM or NIfTI in MRI. MEG systems manufacturers (CTF, Elekta/Neuromag, BTi/4D Neuroimaging, KIT/Yokogawa/Ricoh, Tristan Technologies, ITAB, KRISS/Compumedics Neuroscan, York Instruments) all cater a vendor-specific format. With BIDS-MEG, unprocessed (raw) data is stored in the native file format (see Discussion for further consideration of common MEG format initiatives) and users can still rely on their preferred data analysis application, possibly provided by the MEG vendor, to browse and read in the MEG file contents. Software can also extract meta information elements from the raw data files. They concern e.g., data collection parameters and other study descriptors, which are eventually transcribed into sidecar JSON files by said BIDS-MEG compatible applications. One major benefit of metadata extraction is the facilitation of subsequent data searches and indexation, without the handling and repeated parsing of large raw data files. Additional relevant files can be included alongside the MEG raw data: some propositions are detailed in the online specifications.

For a given investigator or research group, BIDS-MEG describes a hierarchical structure

that descends from a 'Study' folder. Multiple 'Subject' subfolders contain the data from the participants enrolled. They are arranged by 'Session', each session subfolder containing 'Run' folders and eventually, data and metadata files.

The 'Run' folder includes a variety of files: MEG recording files in native format, a sidecar JSON document (`*_meg.json`), a channel description table (`*_channel.tsv`), and other general BIDS files, such as task events tables (`*_events.tsv`) that are likely to be specific of each run. 'Session' specific files include the coordinates of anatomical landmarks and head-localization coils stored in a JSON document (`*_fid.json`), optional photographs of the anatomical landmarks and/or head localization coils (`*_photo.jpg`), fiducials information (`*_fidinfo.txt`), 3-D scalp digitalization files (`*_headshape.<manufacturer_specific_format>`) and acquisition times (`scans.tsv`). The 'Subject' and 'Study' specific files are inherited directly from the general BIDS specifications (e.g., participants.tsv). Note that in case of conflict between fields of different runs/sessions, the inheritance principle should be applied: the description file closer to the data prevails (see Section '3.5 The Inheritance Principle' of the BIDS specifications (Gorgolewski et al., 2016)).

One issue that required special attention was the multiplicity of coordinate systems and units between MEG systems. To impose a unique coordinate system for BIDS based on the subjects' brain anatomy (e.g., MNI coordinates or equivalent) was an appealing solution, which however would lack generalizability in the MEG practice. MEG data can be collected without anatomical information, such as empty-room noise recordings, which are important to optimal source modeling (Gross et al., 2013a). BIDS-MEG therefore associates all recordings with a coordinate file defined according to the MEG system used. Again, BIDS-MEG compatible software can read and interpret this information properly.

Akin to MRI, we anticipate that the systematic data organization enabled by BIDS-MEG will be supported by an increasing number of neuroimaging tools, and that more shared data repositories will be organized accordingly. The straightforward design of BIDS-MEG makes it an interoperable common exchange format for transferring data between investigators and community repositories e.g., OMEGA (Niso et al., 2016b) and OpenfMRI (Poldrack and Gorgolewski, 2017). It also facilitates multimodal integration (between MRI, fMRI, MEG, etc), as the data from multiple modalities follow the same organization scheme.

3.3 Open BIDS-MEG datasets

We provide four different publically-available datasets in BIDS-MEG format (∼200GB). They are freely available for download in the public domain. Note that for educational and demonstration purposes, the International Neuroinformatics Coordinating Facility's GitHub also hosts a lighter version (data structure only, no actual MEG data provided) of the MEG-BIDS dataset examples (`https://github.com/INCF/BIDS-examples`)

The BIDS-MEG sample data release includes:

OMEGA Resting-State samples: Five minutes of eyes-open, resting-state MEG data is available for 5 subjects from The Open MEG Archive (OMEGA) (Niso et al., 2016b). The data are available from the Brainstorm Tutorial: MEG resting state

& OMEGA database. The first release of data in BIDS-MEG format (∼10.5GB) is available here: https://box.bic.mni.mcgill.ca/s/omega?path=%2FContributions%20(in%20BIDS%20format)%2Fsample_BIDS_omega (access to these datasets require registration to OMEGA, https://www.mcgill.ca/bic/omega-registration).

Brainstorm Auditory Example dataset: Brainstorm Auditory tutorial dataset8 (∼2.3GB): https://box.bic.mni.mcgill.ca/s/omega?path=%2FContributions%20(in%20BIDS%20format)%2Fsample_BIDS_auditory (released in Public Domain; includes defaced anatomical T1 of participant, access to these datasets require registration to Brainstorm, http://neuroimage.usc.edu/bst/register.php).

MNE Sample data: Sample data with visual and auditory stimuli described in11: https://drive.google.com/drive/folders/0B_sb8NJ9KsLUQ3BMS0dxZW5nSHM?usp=sharing (released in Public Domain; includes anatomical T1 of participant as well as flash MRI sequences).

OpenfMRI study ds000117: A multi-subject, multi-modal human neuroimaging dataset of 19 subjects participating in a visual task16 (∼178GB): https://openfmri.org/dataset/ds000117/. This dataset is used in one of the SPM tutorials, for training purposes: http://www.fil.ion.ucl.ac.uk/spm/doc/manual.pdf#Chap:data:multimodal

3.4 Software

Widely used MEG software packages have already added functionality to support BIDS-MEG:

Brainstorm (Tadel et al., 2011): Brainstorm is an application with rich graphical-user interactions and analytic pipeline designs for MEG, EEG, NIRS, and electrophysiology recordings. BIDS-formatted MEG/EEG datasets can be imported automatically into the Brainstorm database, as described in the OMEGA tutorial: http://neuroimage.usc.edu/brainstorm/Tutorials/RestingOmega

FieldTrip (Oostenveld et al., 2011): FieldTrip is an open-source MATLAB toolbox for the analysis of MEG, EEG, and other electrophysiological data. Like most other tools listed herewith, FieldTrip can implement full analysis pipelines, starting from coregistration, preprocessing, time- and spectral analysis, source reconstruction, connectivity and statistics. Among others, FieldTrip has been used for the MEG part of the Human Connectome Project. Since a FieldTrip analysis pipeline is represented as a MATLAB script, its application on BIDS structured data implies that the BIDS details are represented in the analysis scripts that users write.

MNE (Gramfort et al., 2013a, 2014): MNE (http://martinos.org/mne) is a software package, whose name stems from its capability to compute cortically-constrained minimum-norm current estimates from M/EEG data. It provides comprehensive analysis tools and workflows including preprocessing (Maxwell filtering, ICA, signal space projectors), source estimation (eg. using MNE, beamformers or mixed-norm sparse solvers), time-frequency analysis, statistical analysis including multivariate decoding, and several methods to estimate functional connectivity between distributed brain regions. The core of MNE is written in Python and is distributed under the permissive BSD Licence. MNE will use the

BIDS data structure to distribute all its tutorial datasets and the documented analysis scripts. MNE provides Python code to read and write files in BIDS compatible format, as well as summary reports automatically generated via the MNE report command. A preliminary version is available at https://github.com/mne-tools/mne-bids.

SPM (Litvak et al., 2011): SPM (http://www.fil.ion.ucl.ac.uk/spm) is a free and open source software written in MATLAB where many widely used methods for the analysis of PET, fMRI and for computational neuroanatomy have been originally developed and implemented. More recently SPM has been extended to perform M/EEG analyses, including topological inference for neurophysiological data, empirical Bayesian framework for source reconstruction and Dynamic Causal Modeling (DCM), an approach combining neural modeling with data analysis. SPM (Litvak et al., 2011) includes a library, spm_BIDS.m, to parse and query BIDS-formatted datasets, as well as low-level functions to read/write JSON and TSV metadata files. A complete pipeline for the analysis of a group MEG dataset in BIDS format is presented in preparation.

MEG-BIDS JSON file generator: Python scripts to produce the JSON sidecar files and reduce the overhead of adopting MEG-BIDS: publically available (https://github.com/INCF/pybids).

MEG-BIDS validator: Developed in javascript using nodejs, it can be packaged to work in the browser (Google Chrome). A command line version is also provided so that it can be used in scripted analysis. The validator performs several sanity checks on datasets to ensure they are compatible to BIDS. This includes the use of regular expressions to check filenames, and JSON schemas to ensure that the metadata files are standardized by data types.

For a more detailed description of the BIDS-MEG specification, example datasets, resources and feedback, please visit http://bids.neuroimaging.io.

3.5 Discussion

Although BIDS-MEG is a proposal to establish a standard framework for the organization of electrophysiology data, it does not impose the standardization of the data file format per se. Some initiatives are aiming towards the definition of a new common binary file format for electrophysiology (MNE python group). Akin to DICOM in MRI, one single file format would be beneficial, when considering the diversity of native raw file formats in electrophysiology. Yet, MEG data volumes are typically very large (several GBs), hence their duplication into a standard format may be impractical. Our position is rather to promote BIDS-MEG and ascertain that tools for data analysis continue to be equipped with the necessary readers for all existing vendor formats. We believe the capacity of BIDS-MEG to organise data without requiring a common data format is actually a strength: the standard is flexible in the sense that any data parameters can be extracted and stored as metadata in sidecar json files at the time of creating a new data entry, regardless of the file format for the raw data. Therefore, any new data format for electrophysiology, including emerging standards, is by construction compatible with BIDS-MEG.

Along the same lines, special attention was given to the handling of the various coordinate

systems used by the different MEG vendors and toolboxes. BIDS-MEG is also flexible in that respect, as long as the conventions used are characterized and documented in the `*_fid.json` file. The coordinate systems presently handled by BIDS-MEG are detailed in the Specification document. The coordinate systems used for MEG and EEG sensors, MRI volumes, locations of fiducials, anatomical landmarks and digitized head points, need to be described following this principle, as some are likely to be different.

We aim at extending BIDS-MEG further towards the handling of processed data, which for now and akin to MRI-BIDS, are simply stored in a data Derivatives folder.

BIDS-MEG represents a significant effort towards a common standard for MEG. We anticipate that the BIDS-MEG software ecosystem and the variety of publicly available BIDS-MEG datasets will grow and incentivize the research community towards adoption.

Chapter 4

A reproducible M/EEG group study

"Extraordinary claims require extraordinary evidence."

—Carl Sagan

Contents

In the previous chapter, we discussed how data sharing can be facilitated using the Brain Imaging Data Structure (BIDS) for magnetoencephalography (MEG). This is taking us one step closer to the goal of reproducibility. However, reproducibility is not achieved by merely sharing more data with the hope that this will solve all problems. As noted in Baker (2016), one of the best solutions to foster reproducible science is not a technical one, but an educational one. This is of course true for statistics, where there is an urgent need to clarify and educate researchers about the statistical tools required in neuroscience. But it is now increasingly important also for academic software.

In recent years, free academic toolboxes have gained increasing prominence in MEG analysis as a means to disseminate cutting edge methods, share best practices between different research groups and pool resources for developing essential tools for the MEG community. Teaching events are regularly held around the world where the basics of each toolbox are explained by its developers and experienced power users. There are however, knowledge gaps that need to be addressed. First, most teaching examples only show analysis of a single 'typical best' subject whereas most real MEG studies involve analysis of group data. It is then left to the researchers in the field to figure out for themselves how to make the transition and obtain significant group results. Secondly, we are not familiar with any examples of fully analyzing the same group dataset with different academic toolboxes to assess the degree of agreement in scientific conclusions and compare strengths and weaknesses of various analysis methods and their independent implementations.

To address this very issue, a workshop was organised by the lead developers of six most popular free academic MEG toolboxes at Biomag 2017. This work is a follow up to the workshop, which presents the contribution of the MNE software team, and will be published in *Frontiers in Neuroscience, section Brain Imaging Methods*. This study presents the results obtained by the reanalysis of an open dataset from Wakeman and Henson (2015) using the MNE software package. The analysis covers preprocessing steps, quality assurance, sensor-space analysis of evoked responses, source localization, and statistics in both sensor and source space. Results with possible alternative strategies are presented and discussed at different stages such as the use of high-pass filtering versus baseline correction, tSSS versus signal space separation (SSS), the use of a minimum norm inverse versus linearly constrained minimum variance (LCMV) beamformer, and the use of univariate or multivariate statistics. This aims to provide a comparative study of different stages of MEG/electroencephalography (EEG) analysis pipeline on the same dataset, with open access to all of the scripts necessary to reproduce this analysis.

Section 4.1 to Section 4.6 was published in:

- M. Jas, E. Larson, D. A. Engemann, J. Leppakangas, S. Taulu, M. Hamalainen, and A. Gramfort. MEG/EEG group study with MNE: recommendations, quality assessments and best practices. *bioRxiv*, 2017b. doi: 10.1101/240044

4.1 Introduction

MEG and EEG are neuroimaging technologies with a high temporal resolution, which provide non-invasive access to population-level neuronal dynamics on virtually any temporal scale currently considered relevant to cognition. While MEG can recover spatial patterns at a higher signal-to-noise ratio (SNR) and enjoys a more selective cortical resolution than EEG (Baillet, 2017a), EEG is more portable and less expensive, and thus supports the study of cognition in a wider range of situations. Processing M/EEG recordings, however, is inherently challenging due to the multi-dimensional nature of the data, the low SNR of brain-related M/EEG signals, and the differences in sensitivity of these measurement techniques. This can give rise to complex sequences of data processing steps which demand a high degree of organization from the investigator.

In an effort to address reproducibility issues recently shown to affect neuroimaging studies (Ioannidis, 2005a; Button et al., 2013; Carp, 2012a,b), a number of community-led efforts have begun developing data sharing (Poldrack and Gorgolewski, 2017) and data organization (Gorgolewski et al., 2016; Niso et al., 2018) projects. These efforts are necessary first steps, but are not sufficient to solve the problem—they must be complemented by educational tools and guidelines that establish good practices for M/EEG analysis (Gross et al., 2013a). However, putting guidelines into practice is not always straightforward, as researchers in the M/EEG community rely on several software packages (Tadel et al., 2011; Delorme and Makeig, 2004; Delorme et al., 2011; Oostenveld et al., 2011; Dalal et al., 2011; Litvak et al., 2011), each of which is different. Even though these packages provide tutorials for single subject data analysis, it is typically left up to the investigator to coordinate and implement multi-subject analyses. Here, we try to address this gap by demonstrating a principled approach to the assembly of group analysis pipelines with publicly available code[1] and extensive documentation.

As members and maintainers within the MNE community, we will present analyses that make use of the MNE software suite (Gramfort et al., 2014). Historically, MNE was designed to calculate minimum-norm estimates from M/EEG data, and consisted in a collection of C-routines interfaced through bash shell scripts. Today, the MNE software has been reimplemented in (Gramfort et al., 2013a) and transformed into a general purpose toolbox for processing electrophysiology data. Built on top of a rich scientific ecosystem that is open source and free, MNE now offers state-of-the-art inverse solvers and tools for preprocessing, time-frequency analysis, machine learning (decoding and encoding), connectivity analysis, statistics, and advanced data visualization. MNE, moreover, has become a hub for researchers who use it as a platform to collaboratively develop novel methods or implement and disseminate the latest algorithms from the M/EEG community (Engemann et al., 2015b; Smith and Kutas, 2015a,b; Haufe et al., 2014; King and Dehaene, 2014; Gramfort et al., 2013b; Schurger et al., 2013; Khan and Cohen, 2013; Larson and Lee, 2013; Hauk et al., 2011; Gramfort et al., 2010; Rivet et al., 2009; Kriegeskorte et al., 2008; Maris and Oostenveld, 2007). With this work, we not only share best practices to facilitate reproducibility, but also present these latest advances in the MNE community which enable automation and quality assessment.

Here, we demonstrate how to use MNE to reanalyze the OpenfMRI dataset ds000117

[1] https://github.com/mne-tools/mne-biomag-group-demo

by Wakeman and Henson (2015). This requires setting the objectives for the data analysis, breaking them down into separate steps and taking a series of decisions on how to handle the data at each of those steps. While there may be several interesting scientific questions that have not yet been addressed on this dataset, here we confine ourselves to the analysis of well-studied time-locked event-related M/EEG components, i.e, event-related fields (ERFs) and event-related potentials (ERPs). This is motivated by educational purposes to help facilitate comparisons between software packages and address reproducibility concerns. To this end, we will lay out all essential steps from single subject raw M/EEG recordings to group level statistics. Importantly, we will highlight the essential options, motivate our choices and point out important quality control objectives to evaluate the success of the analysis at every step.

We will first analyze the data in sensor space. We will discuss the best practices for selecting filter parameters, marking bad data segments, suppressing artifacts, epoching data into time windows of interest, averaging, and doing baseline correction. Next, we turn our attention to source localization: the various steps involved in the process starting from defining a head conductivity model, source space, coregistration of coordinate frames, data whitening, lead field computation, inverse solvers, and transformation of source-space data to a common space. Along the way, we will present various diagnostic visualization techniques that assist quality control at each processing step, such as channel-wise power spectral density (PSD), butterfly plots with spatial colors to facilitate readability, topographic maps, and whitening plots. Finally, we will attempt to distill from our analysis, guiding principles that should facilitate successfully designing *other* reproducible analyses rather than blindly copying the recipes presented here.

4.2 Preliminaries

In this work, we describe a full pipeline using MNE to analyze the OpenfMRI dataset ds000117 by Wakeman and Henson (2015). The data consist of simultaneous M/EEG recordings from 19 healthy participants performing a visual recognition task. Subjects were presented images of famous, unfamiliar and scrambled faces. The dataset provides a rich context to study different neuroscientific and cognitive questions, such as: Which brain dynamics are characteristic of recognizing familiar as compared to unfamiliar faces? How do commonly studied face-responsive brain regions such as the superior temporal sulcus (STS), the fusiform face area (FFA) and the occipital face area (OFA) interact when processing the familiarity of the face? At the same time, it presents a well-studied paradigm which can be particularly beneficial for the development of methods related to connectivity and source localization.

4.2.1 Data description

The subjects participated in 6 runs, each 7.5 minutes in duration. In the original study, three subjects were discarded due to excessive artifacts in the data. To produce comparable results, the same subjects are also discarded from the group results in this study. The data were acquired with an Elekta Neuromag Vectorview 306 system consisting of 102 magnetometers and 204 planar gradiometers. In addition, a 70 channel Easycap EEG system was used for recording EEG data simultaneously.

4.2.2 Reading data

MNE supports multiple file formats written by M/EEG hardware vendors. Apart from Neuromag *FIF* files, which are the default storage format, MNE can natively read multiple other formats ranging for MEG data including 4D Neuroimaging BTI, KIT, and CTF, and for EEG data B/EDF, EGI, and EEGLAB *set*[2]. Despite this heterogeneity of systems, MNE offers a coherent interface to the metadata of the recordings using the so-called *measurement info*[3]. Regardless of the input format, all processed files can be saved as *FIF* files or in the HDF5 format[4].

MNE can handle multimodal data containing different channel types, the most common being magnetometer, gradiometer, EEG, electrooculogram (EOG), electrocardiography (ECG), and stimulus trigger channels that encode the stimulation paradigm. MNE also supports electromyogram (EMG), stereotactic EEG (sEEG) and electrocorticography (ECoG), functional near-infrared spectroscopy (fNIRS) or miscellaneous (misc) channel types. Declaring and renaming channel types is a common step in the preparation of M/EEG datasets before analysis. In our case, once the files were read in, some of the channels needed to be renamed and their channel types corrected in the measurement info (see (Wakeman and Henson, 2015)): the EEG061 and EEG062 electrodes were set as EOG, EEG063 was set as ECG, and EEG064 was set as a miscellaneous channel type as it was a free-floating electrode. If this step is omitted, some preprocessing functions may fall back to potentially less optimal defaults, for example, using the average of the magnetometers instead of the ECG channel when searching for cardiac events.

4.3 MEG and EEG data preprocessing

4.3.1 Maxwell filtering (SSS)

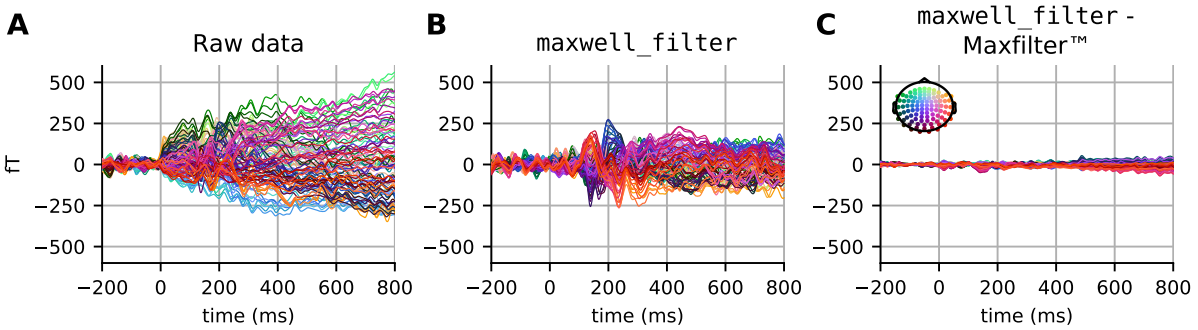


Figure 4.1: Evoked responses (filtered between 1 and 40 Hz) in the magnetometer channels from (A) unprocessed data, (B) data processed with `maxwell_filter` in MNE, and (C) the difference between data processed using `maxwell_filter` and Elekta MaxFilter (TM). The colors show the sensor position, with (x, y, z) sensor coordinates converted to (R, G, B) values, respectively.

Neuromag MEG recordings are often preprocessed first using the SSS method (Taulu and

[2] http://martinos.org/mne/stable/manual/io.html
[3] http://martinos.org/mne/stable/auto_tutorials/plot_info.html
[4] https://support.hdfgroup.org/HDF5/

Simola, 2006), otherwise known as Maxwell filtering. SSS decomposes the data using multipole moments based on spherical harmonics and removes the component of magnetic field originating from outside the MEG helmet. SSS is therefore useful for removing environmental artifacts, and can also be used to compensate for head movements during the recording. In this study, movement compensation is not strictly necessary as the participants managed to stay predominantly still.

The data provided by OpenfMRI (Poldrack and Gorgolewski, 2017) already contain files processed using the proprietary Elekta software MaxFilter, which is what we use in our analysis for the sake of reproducibility. However, MNE offers an open source reimplementation and extension of SSS as well. Before running SSS, it is crucial that bad channels are marked, as otherwise SSS may spread the artifacts from the bad channels to all other MEG channels in the data. This step is preferably done manually with visual inspection. When using the MNE implementation of Maxwell filtering, we reused the list of bad channels available from the Elekta MaxFilter logs in the dataset.

Results comparing raw data, data processed by Elekta MaxFilter, and data processed by the MNE `maxwell_filter` function are provided in Figure 4.1. While the unprocessed data do not show a clear evoked response, the Maxwell filtered data do exhibit clear event-related fields with a clear peak around 100 ms post-stimulus. Note that the results obtained with Elekta implementation and the MNE implementation have minimal differences due to slight differences in computation of component regularization parameters.

Alternatives In principle, SSS can be applied to data acquired with any MEG system providing it has comprehensive sampling (more than about 150 channels). However, so far it has not been tested extensively with other than the 306-channel Neuromag systems. SSS requires relatively high calibration accuracy, and the Neuromag systems are thus carefully calibrated for this purpose. If SSS is not an option, for example due to the lack of fine-calibration information, reasonable noise reduction can be readily obtained from signal space projections (SSPs) (Uusitalo and Ilmoniemi, 1997). This intuitively amounts to projecting out spatial patterns of the empty room data covariance matrix using principal component analysis (PCA). In practice, depending on the shielding of the room, up to a dozen SSP vectors can be discarded to obtain satisfactory denoising.

Caveats. It is important to highlight that after SSS, the magnetometer and gradiometer data are projected from a common lower dimensional SSS coordinate system that typically spans between 64 and 80 dimensions. As a result, both sensor types contain highly similar information, which also modifies the inter-channel correlation structure. This is the reason why MNE will treat them as a single sensor type in many of the analyses that follow.

4.3.2 Power spectral density (PSD)

The PSD estimates for all available data channels provide a convenient way to check for spectral artifacts and, in some cases, bad channels. MNE computes the PSD of raw data using the standard Welch's method (Welch, 1967; Percival and Walden, 1993), whereby the signal for each channel is analyzed over consecutive time segments, with eventually some overlap. Each segment is windowed and then the power of the discrete Fourier transform (DFT) coefficients is computed and averaged over all segments. By making the assumption that each of these segments provides a realization of a stationary process, the

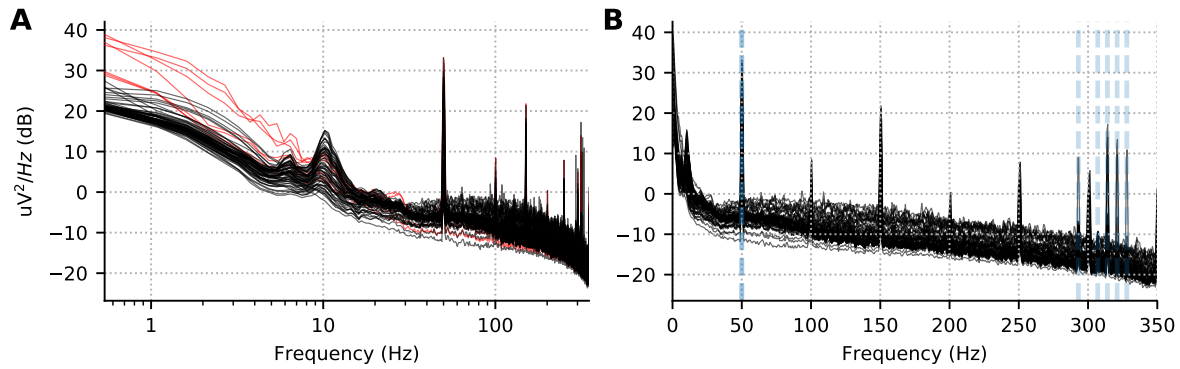


Figure 4.2: Power spectral density per channel for subject 10, run 02. (A) Log scale for the x axis accentuates low frequency drifts in the data. The red lines show the PSD for the bad channels marked manually and provided to us by Wakeman and Henson (2015). (B) The same data with a linear x-axis scale. Five peaks corresponding to HPI coils around 300 Hz are visible and marked in gray dotted lines alongside the power line frequency (50 Hz).

averaging procedure produces an unbiased estimate of the PSD with reduced noise.

Starting from MNE version 0.14, we show channel-wise PSD plots rather than an average across channels, as this facilitates spotting outlier channels. In Figure 4.2, we show the PSD for the EEG channels in one run for one subject. We use windows of length 8192 samples (about 7.4 s given the 1.1 kHz sampling rate) with no overlap. Using a power of 2 for the length and no overlap accelerates computations. Using a logarithmic frequency-axis scaling for the PSD enables quality control by facilitating screening for bad channels. In fact, we found that some potentially bad channels (e.g., EEG024 in subject 14 for run 01) were omitted by the authors of (Wakeman and Henson, 2015), although they are clearly visible in such plots. Concretely we see a few channels with strongly increased low-frequency power below 1 Hz. On the other hand, using a linear frequency-axis scaling, we can convince ourselves easily that the data is unfiltered, as it contains clear peaks from power line at harmonics of 50 Hz, as well as the five head position indicator (HPI) coils used to monitor the head position of the subject, at frequencies of 293, 307, 314, 321, and 328 Hz.

Alternatives The same could have been achieved with the multitaper method (Percival and Walden, 1993; Slepian, 1978), where the data is multiplied element-wise by orthogonal data tapers. However, this method can be an order of magnitude slower than the Welch method for long continuous recordings. The multitaper method is indeed recommended for short data segments. Here we are interested in the PSD for diagnostic purposes on the raw continuous data, and we therefore use the Welch method, a.k.a. averaged periodogram method.

4.3.3 Temporal filtering

In this study, we focused on event-related brain signals below 40 Hz. We low-pass filtered our data at a 40 Hz cutoff frequency with 10 Hz transition band. Such a filter does not affect ERP signals of interest, attenuates the line frequency of 50 Hz and all HPI coil

frequencies. It also limits the effects of temporal ringing thanks to a wide transition band. Because the low-pass was sufficiently low, we did not employ a notch filter separately. Note that such a choice of filters is not necessarily a good default for all studies of event-related brain responses, as ERFs or ERPs can contain rather high frequencies (see for example (Götz et al., 2015)).

When filtering, it is important to take into account the frequency response and impulse response of the filter. In MNE 0.16, the default filter will adapt the filter length and transition band size based on the cutoff frequencies, as done in the EEGLAB software (Widmann et al., 2015; Parks and Burrus, 1987; Ifeachor and Jervis, 2002)[5]. Although no default parameters will fit all analysis requirements, MNE chooses parameters that aim to achieve reasonable stop-band attenuation without excessive filter ringing. To illustrate this point, we compare filters across MNE versions using frequency response and impulse response plots in Figure 4.3. The stop-band attenuation and transition bandwidth in Figure 4.3A and Figure 4.3B are less restricted in the newer versions, which results in less steep attenuation but also less temporal ringing in the impulse response (see Figures 4.3C and D). It can be seen that the previous default parameters gave rise to stronger filtering artifacts as indicated by higher impulse response amplitudes across the time window.

Alternatives and Caveats: If the signal quality is satisfactory, filtering may not be necessary. In the context of this study, we decided to baseline correct our signals rather than high-pass filter them, keeping in mind the ongoing discussion in the community on this topic (Tanner et al., 2015; Rousselet, 2012; Widmann and Schröger, 2012; Acunzo et al., 2012; Maess et al., 2016). Our choice will be motivated in Section 4.3.7 on baseline correction.

4.3.4 Marking bad segments and channels

The next step in the pipeline is to remove bad data segments and bad channels. As data have been processed with Maxwell filter, there are no more bad MEG channel at this stage. For the bad EEG channels, we use the ones provided by the original authors.

To remove bad data segments and bad epochs due to transient artifacts, it is possible in MNE to use the epochs plotter interactively, or to do it via scripting. Either way, the indices of all epochs that are removed from further analysis are logged in the *drop log* attribute of the epochs objects (see online documentation of the Epochs class[6]).

As we are building a reproducible pipeline, here we prefer the scripting route. In MNE, this can be achieved by removing trials whose peak-to-peak amplitude exceeds a certain rejection threshold. Even though this works reasonably well for single subject analysis, it would likely need to be tuned for individual subjects in group studies. Therefore, instead of specifying the thresholds manually, we learn it from the data using the *autoreject* (global) (Jas et al., 2017a) algorithm. *Autoreject* is an unsupervised algorithm which minimizes the cross-validation error, measured by the Frobenius norm between the average signal of the training set and the median signal of the validation set. *Autoreject* not only removes trials containing transient jumps in isolated MEG or EEG channels, but also eyeblink artifacts affecting groups of channels in the frontal area. Since we are dealing

[5] https://martinos.org/mne/stable/auto_tutorials/plot_artifacts_correction_filtering.html

[6] http://martinos.org/mne/stable/auto_tutorials/plot_epoching_and_averaging.html

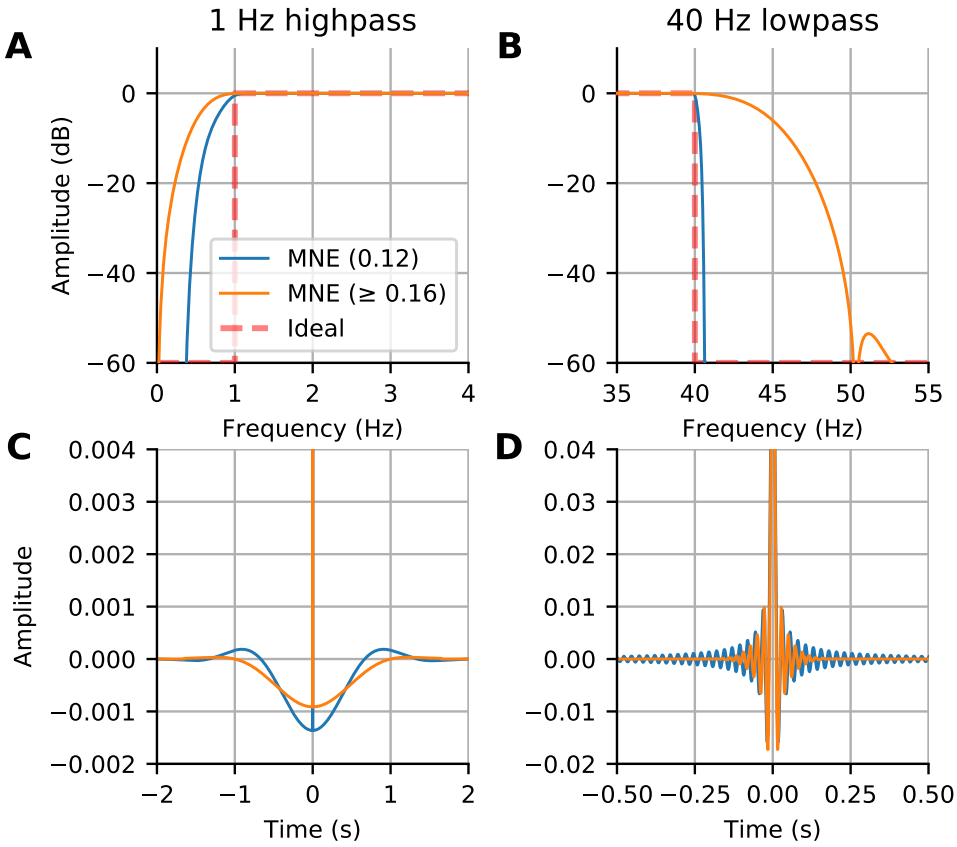


Figure 4.3: Comparison of filters between new (0.16) and old (0.12) MNE versions: (A) The frequency response of the highpass filter; (B) The frequency response of the lowpass filter; (C) The impulse response of the highpass filter; (D) The impulse response of the lowpass filter. The filters in MNE are now adaptive with trade-offs between frequency attenuation and time domain artifacts that by default adapt based on the chosen low-pass and high-pass frequencies.

with visual stimuli, it is preferable to remove the eyeblink trials altogether using the EOG rejection threshold over the stimulus presentation interval rather than suppressing the artifact using a spatial filter such as independent component analysis (ICA) or SSP. Given the large number of trials at our disposal, we can afford to remove some without affecting the results very much.

For the purpose of group averaging, the bad EEG channels were repaired by spherical spline interpolation (Perrin et al., 1989) so as to have the same set of channels for each subject.

4.3.5 Independent Component Analysis (ICA)

Bad channel or segment removal can correct for spatially and temporally isolated artifacts. However, it does not work well for systematic physiological artifacts that affect multiple sensors. For this purpose, ICA is commonly used (Jung et al., 1998). ICA is a blind source separation technique that maximizes the statistical independence between the components. While PCA only requires orthogonal components, ICA looks for independence for example by looking at higher statistical moments beyond (co)variance. In the context of MEG

and EEG analysis, common physiological artifacts have skewed and peaky distributions, hence are easily captured by ICA methods that look for non-Gaussian sources. ICA is therefore popular for removing eye blinks and heart beats, which manifest themselves with prototypical spatial patterns on the sensor array.

In the present study, we use FastICA (Hyvarinen, 1999) to decompose the signal into maximally independent components. We estimate the ICA decomposition on band-pass filtered (1 Hz highpass with 1 Hz transition band, 40 Hz lowpass with 10 Hz transition band) data that has been decimated. In practice, to improve the quality of ICA solution, high-pass filtering is often helpful as it can help to minimize violations of the stationarity assumption made by ICA. Likewise, it is recommended to exclude data segments containing environmental artifacts with amplitudes higher than the artifacts of interest. Finally, generous decimation can save computation time and memory without affecting the quality of the ICA solution, at least, when it comes to separating physiological artifacts from brain signals. Both measures can be implemented using the `reject` and `decim` parameters provided by the ICA fitting routine in MNE. Here we decimated the data by a factor of 11, and excluded time segments exceeding amplitude ranges of $4000 \times 10^{-13}\,\mathrm{fT\,cm^{-1}}$ and $4 \times 10^{-12}\,\mathrm{fT}$ on the magnetometers and gradiometers, respectively.

The ICA component corresponding to ECG activity is then identified using cross-trial phase statistics (CTPS) (Dammers et al., 2008) using the default threshold of 0.8 on the Kuiper statistic. Pearson correlations are used to find EOG related components. As ICA is a linear model, the solution can be estimated on continuous *raw* data and subsequently used to remove the bad components from the *epochs* or *evoked* data.

Alternatives MNE also implements CORRMAP (Viola et al., 2009) which is particularly useful when no ECG or EOG channels are available. This approach uses pattern matching of ICA spatial components. Once templates have been manually defined for one subject, similar patterns can be found for the remaining subjects. If ICA is not an option, SSP projections provide a simple and fast alternative. Here, they can be computed from time segments contaminated by the EOG and ECG artifacts and commonly the first 1 to 2 components are projected out. In our experience, SSP is less precise in separating artifacts from brain components than ICA for the reasons mentioned above, yet, often good enough for a wide class of data analysis scenarios. For analysis of single EEG sensors, multivariate methods cannot be applied. Computing the residuals of a linear regression from the ECG sensor on the EEG is an option in this case.

Caveats. Before blindly applying ICA, it is recommended to estimate the amount of contamination of the MEG and EEG signals. This can be easily achieved by detecting artifact events and epoching and averaging the data accordingly. If, for example, the amplitude range of the average ECG artifact is close to the amplitude range of the brain signals and only few events occur, chances are low to estimate clear cut ECG components using ICA. However, in this case the contamination by ECG is low and therefore no advanced artifact suppression is needed. Second, there is a trade-off between processing time and accuracy. For many analyses, mitigating the artifact contamination by a significant proportion is sufficient and methods like SSP are a reasonable choice. In certain decoding analyses, such preprocessing considerations may have little relevance if any for the final classification results. Indeed, the combination of supervised and multivariate decoding algorithms allows to extract the signals of interest directly in one

step.

4.3.6 Epoching

In event-related M/EEG studies, a trigger channel (in this data STI101) contains binary-coded trigger pulses to mark the onset/offset of events. These pulses can be automatically extracted from the data during analysis and the values on the trigger channel are mapped to the *event IDs*. MNE offers the possibility to extract events when the signal in the trigger channel increases, decreases, or both. It also allows the construction of binary masks to facilitate selecting only the desired events. We masked out the higher order bits in the trigger channel when extracting the events as these corresponded to key presses. After extraction, events can be freely manipulated or created as necessary by the user, as they only require i) the sample number, and ii) some integer code relevant for the experiment or analysis.

As a next step, we extracted segments of data from the continuous recording around these events and stored them as single trials, which are also called epochs, in MNE. The `Epochs` object can store data for multiple events and the user can select a subset of these as `epochs[event_id]`[7]. Moreover, MNE offers the possibility for the user to define a hierarchy of events by using tags (similar in flavor to hierarchical event descriptors by Bigdely-Shamlo et al. (2013)). This is done using `event_id` which is a dictionary of key-value pairs with keys being the tags separated by a forward slash (`/`) and values being the trigger codes[8]. For the paradigm used in this study we used:

```
events_id = {
'face/famous/first': 5,
'face/famous/immediate': 6,
'face/famous/long': 7,
'face/unfamiliar/first': 13,
'face/unfamiliar/immediate': 14,
'face/unfamiliar/long': 15,
'scrambled/first': 17,
'scrambled/immediate': 18,
'scrambled/long': 19,
}
```

At the highest level of hierarchy are 'face' and 'scrambled'. A 'face' can be 'famous' or 'unfamiliar'. And a famous face can be 'first', 'immediate' or 'long' (This distinction between the three categories of famous faces was not used in our analysis). Later on, accessing all the epochs related to the 'face' condition is straightforward, as one only needs to use `epochs['face']` and MNE internally pools all the sub-conditions together. Finally, the epochs were constructed starting 200 ms before stimulus onset and ending 2900 ms after (the earliest possible time of the next stimulus onset).

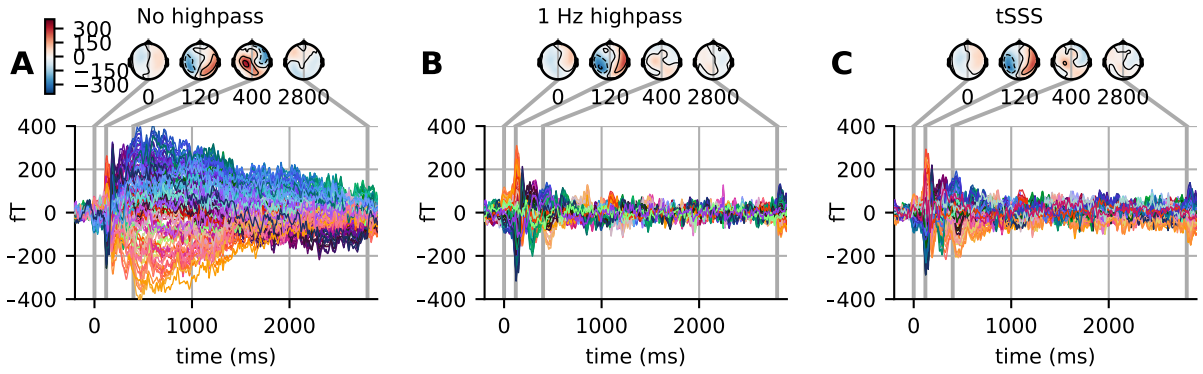


Figure 4.4: (A) Evoked response in magnetometers for subject 3 with baseline correction. Note how signals tend toward the baseline late in the epochs (where the rightmost time point, 2.9 sec, is the earliest possible start time for the next stimulus). (B) The highpass filtered version of the signal and (C) the signal processed with temporal SSS (tSSS). Both reduce the magnitude of the slow and late sustained responses shown in (A).

4.3.7 Baseline correction

It is common practice to use baseline correction so that any constant offsets in the baseline are removed. High-pass filtering achieves similar results by eliminating the low-frequency components in the data. However, when using baseline correction, the low frequency drifts present in the data are not attenuated. Thus it is useful to examine long time-courses of the data, if possible, to determine if low-frequency drifts are present. The difference between the two approaches can be seen in Figure 4.4. The evoked responses in the figure are across-trial averages for the famous face condition. If a maximum time of approximately one second were used, a simple baseline correction would appear to produce an undesired *"fanning"* in the later responses. Indeed one can observe in Figure 4.4A that at one second post-stimulus, the channels still significantly deviate from zero. However, by extending the time window much longer (here to 2.9 seconds) we can see that the signals do mostly return to the baseline level.

Caveats and Alternatives With highpass filter at 1 Hz (and 1 Hz transition band), the signal returns to the baseline level much sooner. Note also the similarities between Figures 4.4B and 4.4C, illustrating how using temporal version of the SSS algorithm (tSSS) acts implicitly as a high-pass filter. For tSSS, we use a buffer size of length 1 s and a correlation limit of 0.95 to reject overlapping inner/outer signals. However, these high-passing effects come at the expense of distorting the sustained responses. We will thus focus on analyses that utilize the baseline-corrected data here.

4.4 Sensor space analysis

An important step in analyzing data at single-subject and group levels is sensor-space analysis. Here we show how several different techniques can be employed to understand the data.

[7] http://martinos.org/mne/stable/auto_tutorials/plot_epoching_and_averaging.html
[8] http://martinos.org/mne/stable/auto_tutorials/plot_object_epochs.html

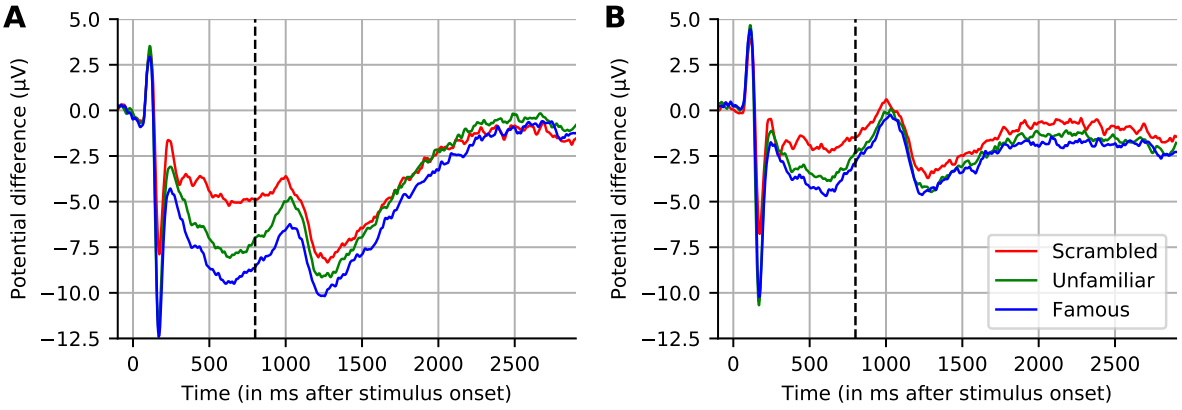


Figure 4.5: Grand averaged evoked response across 16 subjects for channel EEG065. (A) No highpass filter. (B) Highpass filtered at 1.0 Hz. Note that, similar to (A), the results reported by Wakeman and Henson (2015) (dashed line at 800 ms indicates where their plot stopped) show large drifts, but these return to near-baseline levels toward the end of a sufficiently long interval (here, 2.9 seconds) even without applying a highpass filter.

4.4.1 Group average

A classical step in group studies is known as "grand averaging" (Delorme et al., 2015). It is particularly common for EEG studies and it consists in averaging ERPs across all subjects in the study. As not all subjects have generally the same good channels, this step is commonly preceded by an interpolation step to make sure data are available for all channels and for all subjects. Note that grand averaging is more common for EEG than for MEG, as MEG tends to produce more spatially resolved topographies that may not survive averaging due to signal cancellations.

The grand average of the 16 subjects for one EEG sensor (EEG065) is presented in Figure 4.5. We selected this channel to compare with the figure proposed by Wakeman and Henson (2015). We present the grand average for the 'scrambled', 'famous', and 'unfamiliar' conditions using a high-pass filter (cf. Section 4.3.7), and baseline corrected using prestimulus data. This figure replicates the results in (Wakeman and Henson, 2015). We can see the early difference between faces, familiar or unfamiliar, and scrambled faces around 170 ms. We can also notice a difference in the late responses between the two conditions 'unfamiliar' and 'famous'. However, the effect is smaller when using high-pass filtering, as it corrects for the slow drifts.

Caveats For MEG, the grand average may wash out effects or induce spurious patterns due to misalignment between head positions. SSS can be used to align subjects in one common coordinate systems.

4.4.2 Contrasting conditions

Two conditions of interest are often compared using a statistical contrast. A paired contrast between two conditions can be computed by computing the difference in their evoked responses. The difference does not take into account the number of trials used to compute the evoked response – in other words, each condition is weighted equally. Recall that the event IDs were organized hierarchically during epoching (as described in

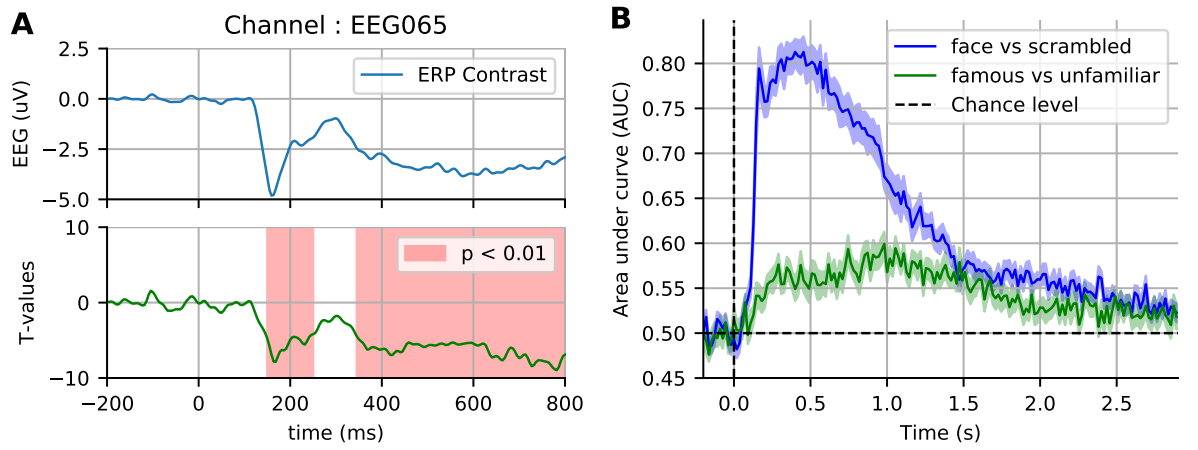


Figure 4.6: Sensor space statistics. (A) A single sensor (EEG065) with temporal clustering statistics. The clustering is based on selecting consecutive time samples that have exceeded the initial paired t-test threshold (0.001), and finding clusters that exceed the size expected by chance according to exchangeability under the null hypothesis ($p < 0.01$, shaded areas). (B) AUC score of time-by-time decoding, averaged across cross validation folds. As opposed to a cluster statistic, time decoding is a multivariate method which pools together the signal from different sensors to find discriminative time points between two conditions.

Section 4.3.6). Such a hierarchical organization is natural for contrasting conditions in the experiment, as we compare not only 'faces' against 'scrambled faces', but also 'famous faces' against 'unfamiliar faces'.

Caveats. Although this is standard in EEG pipelines, historically, for computing the source estimates, weighted averages have sometimes been used. However, MNE provides a mathematically correct estimate for the effective number of trials averaged, so equal-weighted combinations (additions or subtractions) of evoked data are properly accounted for even in the context of unequal trial counts. This logic, however, does not apply when working with experimental protocols (for example, oddball tasks) which, by design, produce many more examples of one than the other conditions.

4.4.3 Cluster statistics

To contrast our conditions of interest, here we use a non-parametric clustering statistical procedure as described by Maris and Oostenveld (2007). This combines neighboring values that are likely to be correlated (here, neighboring time instants) to reduce the problem of multiple comparisons. The contrast score (here the t-values) for each cluster are summed up to compute the mass of each cluster, which serves as our actual statistic. Next, we need to know if the distribution data in our two conditions (here measured using cluster sizes) is significantly different from what would be obtained by chance. For this purpose, we generate a null distribution from the data by randomly swapping our conditions for each subject according to exchangeability under our null hypothesis. In this case, it is equivalent to changing the sign of the contrast data (as we are using a one-sample t-test on the difference between conditions), and then recomputing the maximal cluster size for each permutation. From an estimate of the distribution of the maximum cluster size under

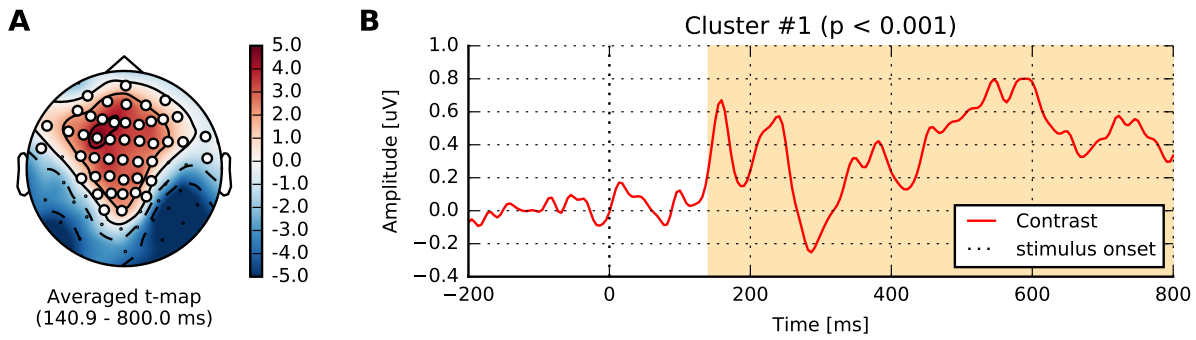


Figure 4.7: Spatiotemporal cluster statistics on the EEG sensors. (A) Topographic map of the t-statistic. (B) Average over the sensors that were part of the significant cluster.

the null-hypothesis, we can compute the probability of observing each cluster relative to this distribution. This gives us a control of the type I error, when reporting a significant difference between the distribution of data in our two conditions.

Running this nonparametric permutation test on the single sensor EEG065 (also used by Wakeman and Henson (2015)) revealed two across-time clusters that allowed us to reject the null hypothesis at the level $p < 0.01$. To perform the clustering, we used an initial thresholding of $p < 0.001$ with a two-sided paired t-test (Figure 4.6). The statistic used was a one-sample t-test on the contrast ERPs using as contrast weights (0.5 for 'familiar', 0.5 for 'unfamiliar' and -1 for 'scrambled'), testing for the condition faces versus scrambled faces. A first cluster appears around the same time as the evoked response, and the other captures the late effects. Running another statistical test, this time incorporating the spatial distribution of the sensors into the clustering procedure, yields one spatiotemporal cluster with $p < 0.05$ for the contrast condition as shown in Figure 4.7.

Alternatives and caveats. It is important to note that this clustering permutation test does not provide feature-wise (vertex, sensor, time point, etc.) but *cluster-level* inference. This is because the test statistic is the cluster size and not any specific t-values used to obtain the cluster in the first place. When inspecting a significant cluster, no conclusion can be drawn on which time point or location was more important. A computationally more expensive alternative is the so-called TFCE method which provides feature-level inference and, moreover, mitigates the problem of having to set the initial threshold on the t-values to define clusters (Smith and Nichols, 2009). When strong *a priori* hypotheses exist considering few regions of interest in either time, frequency or space can be a viable alternative. In that case, the multiple comparisons problem may be readily alleviated by more conventional measures, such as false discovery rates (FDRs) (Genovese et al., 2002).

4.4.4 Time Decoding

As an alternative to mass-univariate analysis, a event-related brain dynamics can studied using a multivariate decoding approach (Ramkumar et al., 2013; King and Dehaene, 2014). Here, a pattern classifier, often a linear model (e.g. logistic regression) is trained to discriminate between two conditions: 'face' versus 'scrambled', and also 'famous faces' versus 'unfamiliar faces'. The classifier can be trained on single trials, time-point by

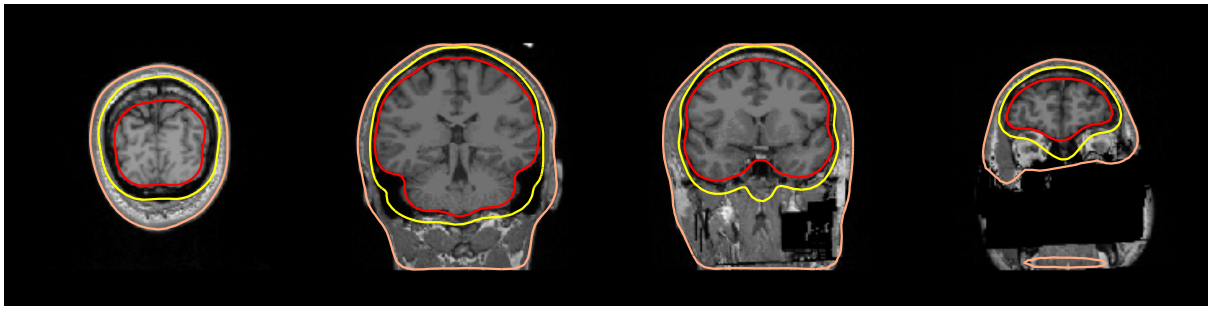


Figure 4.8: BEM surfaces on flash MRI images. The inner skull, outer skull and outer skin are outlined in color.

time-point. The prediction success can then be assessed with cross-validation at every instant, yielding an intuitive display of the temporal evolution of discrimination success. In Figure 4.6B, we display such cross-validation time-series averaged across the 16 subjects. As anticipated, discriminating between faces and scrambled faces is much easier than discriminating between 'famous' and 'unfamiliar' faces, based on information in early components in the first second after stimulus-onset.

For performance evaluation, we use is area under the receiver operating characteristic curve (ROC-AUC), as it is a metric that is insensitive to class imbalance (i.e., differing numbers of trials) therefore allowing us to average across subjects, and also to compare the two classification problems (faces vs. scrambled and familiar vs. unfamiliar). Results on the faces vs. scrambled conditions show that time-resolved decoding reveals decoding accuracy greater than chance around the same time intervals as the non-parametric cluster statistic. The effect although appears here quite sustained over time. Results on familiar *vs.* unfamiliar conditions are also above chance from 200 to 300 ms, however the best decoding performance emerges later for this contrast. This more subtle effect even peaks after 800 ms, which exceeds the time window investigated in the original study.

4.5 Source reconstruction

The MNE software relies on the FreeSurfer package (Dale et al., 1999; Fischl et al., 1999) for the processing of anatomical MRI images. This automatic procedure is run using the command `recon-all` on the T1 MRI of each subject. This provides many useful pieces of information, but the most critical here are the cortical reconstruction (a high resolution triangulation of the interface between the white and the gray matter) and the inner skull surface.

For inverse source reconstruction and beamforming, we must first compute the forward solution, often called a gain or lead field matrix. It describes the sensitivity of the sensors to a given set of dipoles (Mosher et al., 1999). Computing the gain matrix, which is a linear operator, requires having a so-called source space of dipole locations, a conductor model for the head, and the sensor locations relative to those dipoles. This latter requirement in practice means putting in the same coordinate system the MRI (where the source

space and conductor model are defined), the head (where the EEG electrode locations are recorded or digitized), and the MEG device (where the MEG sensors are defined). This step is commonly referred to as *coregistration*. We will cover each of these steps below.

4.5.1 Source space

As we expect most of our activations of interest to be due to cortical currents (Dale et al., 2000a), we position the candidate dipoles on the cortical mantle. We chose a source space obtained by recursively subdividing the faces of an octahedron six times (oct6) for both the left and right hemispheres. This leads, for each subject, to a total of 8196 dipoles evenly spaced on the cortical surface (See Figure 6 in (Gramfort et al., 2014)).

4.5.2 Head conductivity model

MNE can use simple spherical conductor models but when the MRI of subjects are available, the recommended approach is to use a piecewise-constant conductivity model of the head. Tissue conductivities are defined for each region inside and between the segmented interfaces forming the inner skull, outer skull and the outer skin. It corresponds to a so-called three layer model, however a single layer is possible when using only MEG data. The default electrical conductivities used by MNE are 0.3 S/m for the brain and the scalp, and 0.006 S/m for the skull, i.e., the conductivity of the skull is assumed to be 1/50 of that of the brain and the scalp. With such a head model, Maxwell equations are solved with a boundary element model (BEM).

In addition to the T1 MRI image, fast low-angle shot (FLASH) images are provided in the present dataset. Such MRI images allow to automatically extract precise surfaces for the inner skull and outer skull. Note that in the absence of FLASH images, MNE offers a somewhat less accurate solution based on the watershed algorithm. One output of the MNE automatic BEM surface extraction is presented in Figure 4.8. It contains the three surfaces needed for the computation of the EEG gain matrix. In our results shown here, we used only the MEG data for source reconstruction, and consequently only made use of the inner skull surface in a one-layer model. As MRIs shared here are defaced, outer skull and scalp surfaces are anyway quite wrong, so we considered it satisfactory to only use the inner skull surface.

Quality insurance at this stage consists in checking that the three surfaces do not intersect with each other and that they follow the interfaces between the brain, the skull and the skin. A slice-by-slice visual inspection of approximate alignment is best and is conveniently proposed by MNE BEM plotting function that outputs a figure as presented in Figure 4.8.

Here, as the MRIs shared in this dataset were anonymized, the outer skin surface obtained automatically using Freesurfer intersected with the outer skull surface for most subjects. However this is rarely observed with non defaced T1 MRI images.

4.5.3 Coregistration

In order to compute the gain matrix, the sensor locations (and normals), head model, and source space must be defined in the same coordinate system. In practice, this means

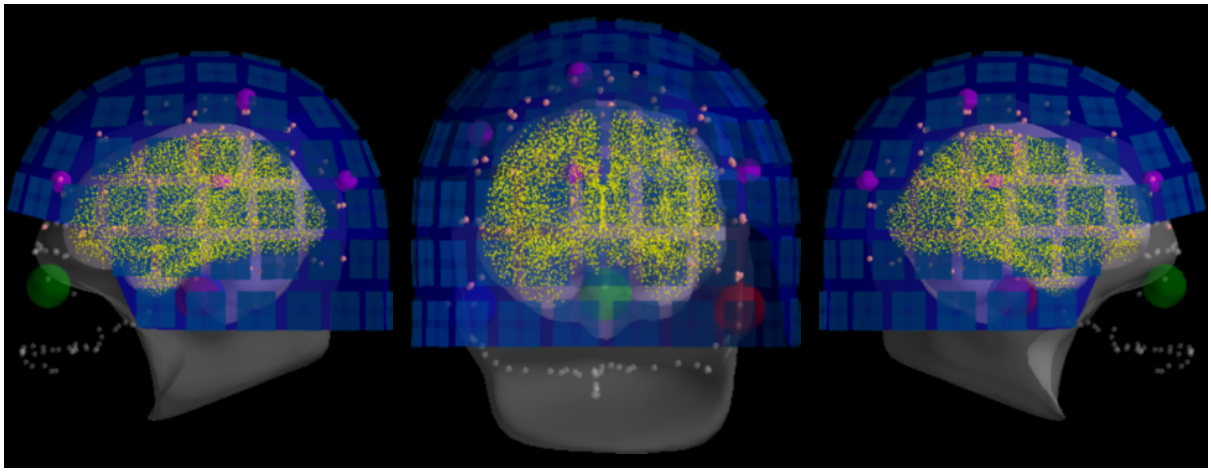


Figure 4.9: The result of head-to-MRI (and MEG-to-head) transformations with inner skull and outer skin surfaces for one subject. Note that the MEG helmet is well aligned with the digitization points. The digitized fiducial points are shown with large dots, EEG electrodes with small pink dots, and extra head digitization points with small gray dots. Note that the anonymization of the MRI produces a mismatch between digitized points and outer skin surface at the front of the head.

that the BEM surfaces and source space (which are defined in MRI coordinates) must be coregistered with the EEG sensors, which are digitized in the Neuromag head coordinate frame (defined by the digitized nasion, LPA, and RPA). The MEG sensor locations and normals are defined in the MEG device coordinate frame. Typically, the MEG-to-head transformation is determined during acquisition using head position indicator (HPI) coils (or redefined using head position transformation using Maxwell filtering), so MEG sensor locations can be easily transformed to head coordinates. The transformation between the MRI and head coordinate frames is typically estimated by identifying corresponding points in the head and MRI coordinate systems, and then aligning them.

The most common points used to provide an initial alignment are the fiducial landmarks that define the Neuromag head coordinate frame. They consist of the nasion and two pre-auricular points which are digitized during acquisition, and are then also identified by offline visual inspection on the MRI images. Additional digitization points on the head surface can also be used to better adjust the coregistration. In this study, on average, 135 digitization points were available per subject. The transformation, which consists of a rotation matrix and a translation vector, is then typically saved to a small file, also called *trans* file, and later used to compute the forward solution.

For quality insurance, MNE offers a simple function to visualize the result of the coregistration. Figure 4.9 shows one example obtained with this function with the defaced, low-resolution MRI head surface. As here the MRI were defaced, many important digitization points close to the nose where useless. To reduce the risk of bad coregistration due to defaced MRI images, we used the trans files kindly provided by the original authors.

4.5.4 Covariance estimation and Whitening

As inverse solvers typically assume Gaussian noise distribution on the sensors with an identity covariance matrix, a whitening step is first necessary (Engemann et al., 2015b). M/EEG signals are indeed highly spatially correlated. Whitening also allows integration of data from different channel types that can have different units and signal amplitudes which differ by orders of magnitudes (cf. planar gradiometers, axial magnetometers, and EEG electrodes). To whiten the data, one must provide an estimate of the spatial noise covariance matrix. This can be computed from empty-room recordings for MEG or pre-stimulus periods (Gramfort et al., 2014). Here, we followed the approach proposed by Engemann et al. (2015b), which consists in picking the best model and estimating the best regularization parameters by computing the Gaussian log-likelihood of left-out data (i.e., a cross-validation procedure). Such an approach has been shown to be particularly robust for scenarios where a limited number of samples is available for covariance estimation.

In this analysis, the noise covariance is estimated from the 200 ms of data before stimulus presentation. During this period, only a fixation color is visible at the center of the screen. Given this covariance matrix and the gain matrix, one can assemble the inverse operator to compute the MNE or dynamic statistical parameter mapping (dSPM) solutions (Dale et al., 2000a).

The quality of the covariance estimation and whitening can have a significant impact on the source localization results. The rank-adjusted global field power (GFP) has been proposed by Engemann et al. (2015b) as a measure that can be used to check the quality of the whitening. It is defined as GFP$= \sum_i x_i^2/P$ where P is the rank of the data and x_i is the signal in the ith sensor at a time instant. The GFP being a normalized sum of Gaussian random variables with an identity covariance matrix, it follows a χ^2 distribution with an expected value of 1. What is not captured by our noise model, e.g. actual brain signals, thereof will pop out in the whitened domain. To understand this better, we show some whitened data and the GFP in Figure 4.10. If the Gaussian assumption has not been violated, we expect the whitened data to contain 95% of the signal within the range of -1.96 and 1.96, which we mark in dotted red lines. The baseline period, where we estimated our noise covariance from, appears to satisfy this assumption. Consequently, the GFP is also 1 during this period. One can observe a strong increase in the GFP just after the stimulus onset, and that it returns slowly to 1 at the end of the time interval. Such a diagnostic plot can in fact be considered essential for quality assurance before computing source estimates. This has as consequence that what appears in the source estimates depends on our noise model. For instance, using a noise covariance obtained from empty room recordings would suggest the presence of "interesting" signals, simply because it contains brain signals that are fundamentally different from the empty room noise.

For the LCMV beamformer, we also need to estimate a signal covariance. For this we use the 30 ms to 300 ms window after the stimulus onset. The data covariance is again regularized automatically following (Engemann et al., 2015b) and is motivated by the results from (Woolrich et al., 2011a; Engemann et al., 2015c).

Caveats. If empty-room data are used to whiten processed signals, one must make sure

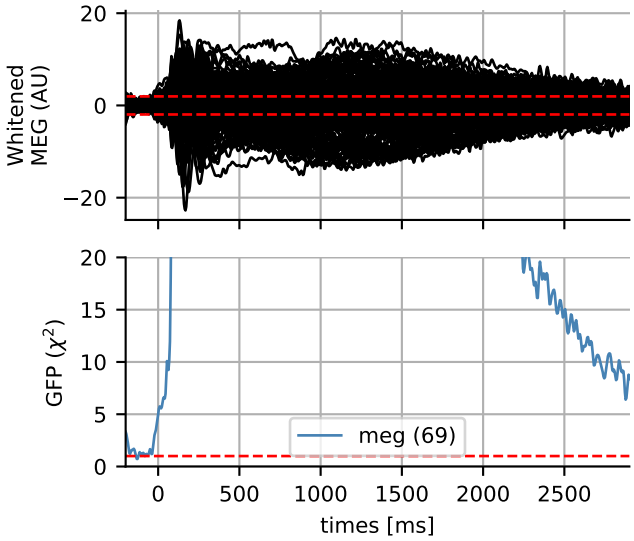


Figure 4.10: Whitened MEG data for subject 4 and the global field power (GFP) which follows a χ^2 distribution if the data is assumed Gaussian. The dotted horizontal red lines represent the expected GFP during the baseline for Gaussian data. Here the data slowly return to baseline at the end of the epoch.

that the obtained noise covariance matrix corresponds to the processed data rather than to the original empty-room data. This is done by processing the empty-room data with exactly the same algorithm and the same parameters as the actual data to be analyzed. For example if SSS, SSP or ICA are applied on processed data, it should be applied to empty room data before estimating the noise covariance. Concretely, SSP vectors and ICA components projected out from the data of interest should also be projected out from the empty room data. SSS should be performed with identical parameters. Also note that magnetometers and gradiometers are whitened jointly. Moreover, if SSS was applied, the display of whitening treats magnetometers and gradiometers as one channel-type. For proper assessment of whitening, a correct assessment of the spatial degrees of freedom is necessary. The number of SSS dimensions is commonly a good estimate for the degrees of freedom. When movement compensation was applied, the estimated data rank may be unreliable and suggest too many independent dimensions in the data. Even the actual number of SSS components can be misleading in such circumstances. It is then advisable to inspect the eigenvalue spectrum of the covariance matrix manually and specify the degrees of freedom manually using the rank parameter.

4.5.5 Inverse solvers and beamforming

The goal of an inverse solver is to estimate the locations and the time courses of the sources that have produced the data. While the data $\mathbf{M}$ can be expressed linearly from the sources $\mathbf{X}$ given the gain matrix $\mathbf{G}$, $\mathbf{M} \approx \mathbf{GX}$, the problem is ill-posed. Indeed $\mathbf{G}$ has many more columns than rows. This means that there are more unknown variables (brain sources) than the number of measured values (M/EEG sensors) at each time point. This also implies that the solution of the inverse problem is not unique.

For this reason, many inverse solvers have been proposed in the past ranging from dipole

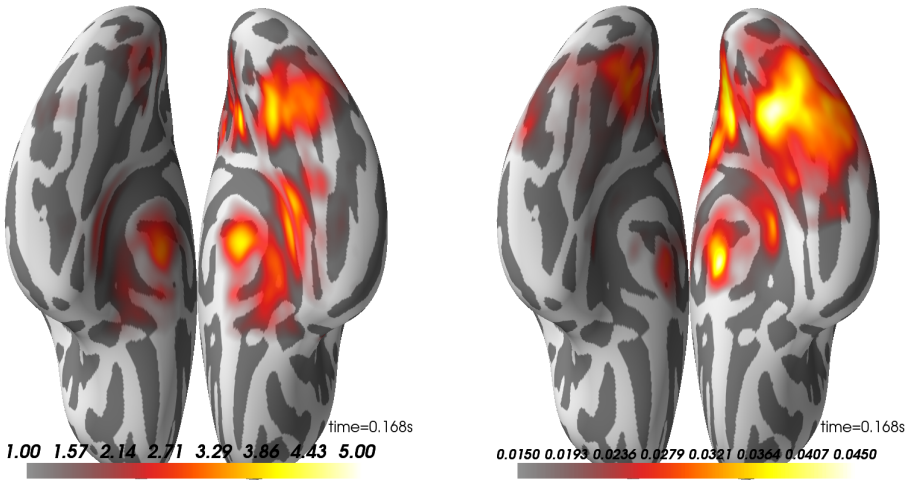


Figure 4.11: Group average on source reconstruction with dSPM (left) and LCMV (right). Here, we have the ventral view of an inflated surface with the anterior-posterior line going from the bottom to top of the image. Right hemisphere is on the right side.

fits (Scherg and Von Cramon, 1985; Mosher et al., 1992), minimum norm estimates (MNEs) (Hämäläinen and Ilmoniemi, 1984), and scanning methods such as RAP-MUSIC or beamformers such as LCMV and DICS (Van Veen et al., 1997; Gross et al., 2001; Sekihara et al., 2005). There is therefore no absolute perfect inverse solver, although some are more adapted than others depending on the data. Some are adapted to evoked data for which one can assume a few set of focal sources. Some also give you source amplitudes in a proper unit, which is nAm for electrical current dipoles, such as MNE, MxNE Gramfort et al. (2013b) or dipole fits. Some give spatially normalized statistical maps such as dSPM (Dale et al., 2000c) or LCMV combined with neural activation index (NAI) filter normalization (Van Veen et al., 1997).

Given the important usage of dSPM and the LCMV beamformer in the cognitive neuroscience literature, we wanted to investigate how much using one of these two most commonly used methods was affecting the source localization results. The dSPM solution was computed with MNE default values: loose orientation of 0.2, depth weighting (Lin et al., 2006) of 0.8, and SNR value of 3. The LCMV used was a vector beamformer with unit-noise-gain normalization (Sekihara et al., 2005) as implemented in MNE 0.15. No specific regularization was used in the beamformer filter estimation.

4.5.6 Group source reconstruction

To analyze data at the group level, some form of data normalization is necessary, whereby data from all subjects is transformed to a common space in a manner that helps compensate for inter-subject differences. This procedure, called *morphing* by the MNE software, exploits the FreeSurfer spherical coordinate system defined for each hemisphere (Dale et al., 1999; Fischl et al., 1999). In our analysis, the data are morphed to the standard FreeSurfer average subject named `fsaverage`. The morphing procedure is performed in three steps. First, the subsampled data defined on the high resolution surface are spread to neighboring vertices using an isotropic diffusion process. Next, registration is used to interpolate the data on the average surface. And finally, the data defined on the average

Figure 4.12: Spatio-temporal source space clusters obtained by nonparametric permutation test that allowed rejection of the null hypothesis that the distribution of data for the "faces" condition was the same as that of "scrambled". The clusters here are collapsed across time such that vertex colors indicate the duration that each vertex was included in its cluster (each cluster here occurring with type I error corrected $p < 0.05$). Hot colors indicate durations for vertices in clusters where response for faces > scrambled (cool colors would be used for scrambled > faces, but no such clusters were found).

surface is subsampled to yield the same number of source locations in all subjects (here, 10242 locations per hemisphere). Once the morphing is complete, the data is simply averaged.

What is presented in Figure 4.11 is the group average of the dSPM and LCMV beamformer solutions on contrast between faces and scrambled at 170 ms post-stimulus.

Looking at these results, one can observe that both methods highlight a peak of activation on the right ventral visual cortex known to be involved in face processing (Grill-Spector et al., 2017, 2004; Wakeman and Henson, 2015). The dSPM peak seems however to be slightly more anterior.

4.5.7 Source-space statistics

Just as we did for the sensor time courses, we can subject the source time courses (here for dSPM only) to a cluster-based permutation test. The null hypothesis is again that there is no significant difference between the data distributions (here measured using cluster size) for faces versus scrambled (paired). Under each permutation, we do a paired t-test across subjects for the difference between the (absolute value of the) faces and scrambled values for each source space vertex and time point. These are clustered, and maximal cluster size for each permutation is selected to form the null distribution. Cluster sizes from the actual data are compared to this null; in this case we find three clusters that lead us to reject the null with $p < 0.05$ (see Figure 4.12).

Alternatives When strong hypotheses exist with regard to spatial, temporal and spectral regions of interest, it may be preferable to test the experimental hypotheses on fewer well-chosen signals. In the context of a group analysis, a linear multilevel modeling approach may provide an interesting option for obtaining joint inference at the single subject and group level Gelman (2006); Baayen et al. (2008).

4.6 Discussion and conclusion

Analyzing M/EEG requires successive operations and transformations on the data. At each analysis stage, the different processing choices can affect the final result in different ways. While this situation encourages tailoring data analysis strategies to the specific demands of the scientific problem, this flexibility comes at a cost and can lead to spurious findings when not handled appropriately (Ioannidis, 2005b; Simmons et al., 2011; Carp, 2012a). In the absence of fully automated data analysis pipelines that can optimize the choice of processing steps and parameters, it is crucial to develop principled approaches to planning, conducting and evaluating M/EEG data analysis.

The present study makes the effort to elucidate common elements and pitfalls of M/EEG analysis. It presents a fully reproducible group analysis of the publicly available dataset from Wakeman and Henson (2015). All code and results are publicly accessible `http://mne-tools.github.io/mne-biomag-group-demo/`. The study provides contextualized in-depth discussion of all major steps of the analysis with regard to alternative options, caveats and quality control measures. As a rare contribution to the M/EEG literature, this study illustrates in comparative figures, the experimental results obtained when changing essential options at different steps in the analysis. In the following, we want to share some insights that we obtained from working together on this study.

Collaborative data analysis. In our experience, high-level planning and hands-on data analysis are commonly divided between, e.g., masters or doctoral students, post-docs, and senior researchers. As a consequence, the results are typically appreciated from figures produced without connection to the research code that generated them. In this study, several authors contributed repeatedly to the code, analyses were repeated on different computers, and results were inspected in an ongoing fashion by many authors. This experience has had as consequence that incoherences, model violations, and other quality concerns were perhaps detected more often than usual, which has greatly contributed to the overall quality of the data analysis. While it is perhaps too extreme or onerous to recommend adopting social interaction habits from open source software development—such as peer review, pair or extreme programming—in scientific data analysis, we believe that data analysis should not be done in isolation. In order to enable full-blown collaborative data analysis in research, analysis must be repeatable, hence, scripted, and a minimum of code organization and readability must be enforced. On the other hand, the best coding efforts will have limited impact if there are not multiple authors with fresh and active data analysis habits. We hope that the example stated by this paper, together with the open source tools and the community it is built upon, can stimulate more collaborative approaches in M/EEG research.

The costs of reproducibility. It is a commonly neglected reality that reproducibility comes at a price. Making an analysis strictly reproducible not only requires intensified

social interactions, hence more time, but also demands more computational resources. It is a combinatorially hard problem if one were to consider all the potential sources of variability. For example, analyses have to be repeated on different computers with different architectures and performance resources. This sometimes reveals differences in results depending on the hardware, operating system, and software packages used. As observed in the past by Glatard et al. (2015), we noticed that some steps in our pipeline such as ICA are more sensitive to these changes, eventually leading to small differences at the end of the pipeline, which is in our case are cluster-level statistics in the source space. Of course, differences due to these changes are harder to control and enforce in the context of today's fast technological progress. Indeed, what we manage to achieve is reproducibility, as opposed to the pure replicability which would be the case if the same results could be achieved even when the computer hardware and software packages were changed.

Also, when code is developed on large desktop computers which are common in many laboratory settings, replication efforts with lower-performance workstations may incur high costs in terms of human processing time. The analysis not only runs slower but may crash, for example due to differences in computer memory resources. We therefore emphasize the responsibility of software developers in providing scalable performance control and the responsibility of hands-on data analysts to design the analysis bearing performance and social constraints in mind. In other words, consider that code needs to run on someone else's computer.

When to stop? Obviously, in the light of the current replication crisis, clear rules need to be established on when to stop improving the data analysis (Simmons et al., 2011; Szucs and Ioannidis, 2017). A particular risk is emanating from the possibility of modifying the analysis code to eventually confirm the preferred hypothesis. This would invalidate inference by not acknowledging all the analysis options explored. Apart from commonly recommended preregistration practices and clean hold out data systems, we want to emphasize the importance of quality criteria for developing the analysis. The bulk of M/EEG preprocessing tasks are either implicitly or explicitly model-based, as shown by the rich battery of quality control visualizations presented in this chapter. Such plots allow to assess if M/EEG analysis outputs can be considered good signals. Consequently, analysis should be stopped when no further improvement on quality control metrics is to be expected, within a reasonable time investment. In other words, not research hypotheses (and statistical significance of results) but rather signal quality metrics are the criterion for constructing M/EEG analyses. Ideally, only when quality control is done, should the contrast(s) of interest be investigated.

With these broader insights in mind, we will make an attempt to extract from our analysis practical recommendations that should facilitate *future* M/EEG analyses. We encourage the reader not to take the analysis presented here as a direct justifications for parameter choices used in their analyses, but instead learn the principles underlying the choices made in our examples. The general rule is: assess your options and chose the optimal measure at each processing step, then visualize and automate as much as you can.

Practical recommendations:

1. **Know your I/O.** Make sure to have a clear idea about the meta-data available

in your recordings and that the software package knows about relevant auxiliary channels, e.g, stim, EOG, ECG. Use custom MNE functions and other libraries to add quick reading support if I/O for a file-type is not readily supported.

2. **Think noise.** Inspect your raw data and power spectra to see if and how much denoising is necessary. When using methods such as SSS, SSP, ICA, or reference-channel correction, be aware of their implications for later processing. Remember also to process your empty room data the same way. The interpretation of sensor types may change. Denoising may implicitly act as a high-pass filter (cf. tSSS). High-pass filtering or baselining may not be a good thing, depending on the paradigm. For calibrating your inverse solution, think of what is an appropriate noise model, it may be intrinsically linked to your hypothesis.

3. **Mind signals of non-interest.** Detect and visualize your physiological artifacts, e.g. ECG, EOG, prior to attempting to mitigate them. Choose an option that is precise enough for your data. There is no absolute removal, only changes in signal-to-noise ratio. Not explicitly suppressing any artifacts may also be a viable option in some situations, whereas a downstream method (e.g., temporal decoding) will not benefit from them. When employing an artifact removal technique, visualize how much of your signal of interest is discarded.

4. **Visually inspect at multiple stages.** Use diagnostic visualizations often to get a sense of signal characteristics, from noise sources, to potential signals of interest. Utilize knowledge of paradigms (e.g., existence of an N100 response) to validate steps. Visual inspection of data quality and SNR is recommended even if the processing is automated. When using the an anatomical pipeline, look at your coregistration and head models to make sure they are satisfactory. Small errors can propagate and induce spurious results. Check for model violations when working with inverse solvers and understand them. Inappropriate noise models will distort your estimated sources in simple or complex ways and may give rise to spurious effects.

5. **Apply statistics in a planned way.** Averaging data is a type of statistical transformation. Make sure that what you average is actually comparable. To handle the multiple-comparisons problem, different options exist. Non-parametric hypothesis-tests with clustering and multivariate decoding are two such options, and they are not mutually exclusive. Keep in mind that MEG/EEG is primarily about time, not space. A whole-brain approach may or may not be the best thing to pursue in your situation. Anatomical labels may provide an effective way of reducing the statistical search space.

6. **Be mindful of non-deterministic steps.** To maximize reproducibility, make sure to fix the random initialization of non-deterministic algorithms such as ICA. Not only does it ensure reproducibility, debugging is also easier when the code is deterministic. Prefer automated scripts as opposed to interactive or manual pipelines wherever possible.

7. **Keep software versions fixed.** In an ideal world, software (and hardware) versions would not matter, as each operation necessary for data analysis should be tested against known results to ensure consistency across platforms and versions. However, this ideal cannot always be met in practice. To limit difficulties, do not

change software versions, hardware or operating system versions in the middle of your analysis. Keep in mind that MNE is based on several other pieces of software. Updating them can have an impact on the outcome of MNE routines. Once data analysis is complete, cross-checking on different platforms or with different software versions can be useful for community feedback and identifying fragile or problematic steps.

In order to facilitate the reproduction of all the results presented in this chapter, all the code used to make the figures in this paper, but also much more, is available at http://mne-tools.github.io/mne-biomag-group-demo/.

Chapter 5

Automated artifact rejection for M/EEG

"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk."

—John von Neumann

Contents

In the last chapter, we discussed the reproducibility challenges when performing group studies in magnetoencephalography (MEG) and electroencephalography (EEG). One way to improve reproducibility is automation, and we briefly touched upon an algorithm for automating detection of bad data segments, known as *autoreject*.

In this chapter, we will present this algorithm which rejects and repairs bad trials in MEG and EEG signals. Annotating bad segments in the data is perhaps one of the most time consuming aspects of data processing in electrophysiology. Currently, it is either done manually, or using automated black-box algorithms. The manual approach is often subjective with often no clear consensus on what constitutes a corrupted data segment. Therefore, reanalysis is not only manually demanding but can also lead to problems in reproducibility. On the other hand, the automated methods are controlled by parameters that are not straightforward to tune. In the case of failure, it is not always obvious what caused the method to fail and how it can be corrected. As a result, one is left with no choice but to exclude the data from further analysis.

This led us to develop a method based on design choices motivated by ease of interpretation and diagnosis. The method we propose capitalizes on cross-validation in conjunction with a robust evaluation metric to estimate the optimal peak-to-peak threshold–a quantity commonly used for identifying bad trials in MEG/EEG. This approach is then extended to a more sophisticated algorithm which estimates this threshold for each sensor yielding trial-wise bad sensors. Depending on the number of bad sensors, the trial is then repaired by interpolation or by excluding it from subsequent analysis. For efficiency reasons, we use Bayesian optimization which is a well-known technique for hyperparameter optimization. All steps of the algorithm are fully automated thus lending itself to the name *autoreject*. Crucially, the algorithm is even able to deal with sensors that are locally corrupted, which is quite often the case for EEG data.

In order to assess the practical significance of the algorithm, we conducted extensive validation and comparisons with state-of-the-art methods on four public datasets containing MEG and EEG recordings from more than 200 subjects. The comparisons include purely qualitative efforts as well as quantitatively benchmarking against human supervised and semi-automated preprocessing pipelines. The algorithm allowed us to automate the preprocessing of MEG data from the Human Connectome Project (HCP) going up to the computation of the evoked responses. The automated nature of our method minimizes the burden of human inspection, hence supporting scalability and reliability demanded by data analysis in modern neuroscience.

Section 5.1 to Section 5.6 was published in:

- M. Jas, D. Engemann, F. Raimondo, Y. Bekhti, and A. Gramfort. Automated rejection and repair of bad trials in MEG/EEG. In *6th International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 2016
- M. Jas, D. A. Engemann, Y. Bekhti, F. Raimondo, and A. Gramfort. Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, 159:417–429, 2017a

5.1 Introduction

Magneto-/electroencephalography (M/EEG) offer the unique ability to explore and study, non-invasively, the temporal dynamics of the brain and its cognitive processes. The M/EEG community has only recently begun to appreciate the importance of large-scale studies, in an effort to improve replicability and statistical power of experiments. This has given rise to the practice of sharing and publishing data in open archives (Gorgolewski and Poldrack, 2016). Examples of such large electrophysiological datasets include the Human Connectome Project (HCP) (Van Essen et al., 2012; Larson-Prior et al., 2013), the Physiobank (Goldberger et al., 2000), the OMEGA archive (Niso et al., 2016b) and Cam-CAN (Taylor et al., 2015). A tendency towards ever-growing massive datasets as well as a shift towards common standards for accessing these databases (Gorgolewski et al., 2016; Bigdely-Shamlo et al., 2013) is clearly visible. The UK Biobank project (Ollier et al., 2005) which currently hosts data from more than 50,000 subjects is yet another example of this trend.

This has however, given rise to new challenges including automating the analysis pipeline (Gorgolewski and Poldrack, 2016). Automation will not only save time, but also allow scalable analysis and reduce the barriers to reanalysis of data, thus facilitating reproducibility. Engemann and Gramfort (2015) have recently worked towards more automation in M/EEG analysis pipelines by considering the problem of covariance estimation, a step commonly done prior to source localization. Yet, one of the most critical bottlenecks that limits the reanalysis of M/EEG data remains at the preprocessing stage with the annotation and rejection of artifacts. Despite being so fundamental to M/EEG analysis given how easily such data can be corrupted by noise and artifacts, there is currently no consensus in the community on how to address this particular issue.

In the presence of what we will refer to as *bad* data, various data cleaning strategies have been employed. A first intuitive strategy is to exclude bad data from analysis, to *reject* it. While this approach is very often employed, for example, because data cleaning is time consuming, or out of reach for practitioners, it leads to a loss of data that are costly to acquire. This is particularly the case for clinical studies, where patients have difficulties staying still or focusing on the task (Cruse et al., 2012; Goldfine et al., 2013), or even when babies are involved as subjects (Basirat et al., 2014).

When working with M/EEG, the data can be bad due to the presence of bad sensors (also known as channels[1]) and bad trials. A trial refers here to a data segment whose location in time is typically related to an experimental protocol. But here we will also call trial any data segment even if it is acquired during a task-free protocol. Accordingly, a bad trial or bad sensor is one which contains bad data. Ignoring the presence of bad data can adversely affect analysis downstream in the pipeline. For example, when multiple trials time-locked to the stimulation are averaged to estimate an evoked response, ignoring the presence of a single bad trial can corrupt the average. The mean of a random vector is not robust to the presence of strong outliers. Another example quite common in practice, both in the case of EEG and MEG, is the presence of a bad sensor. When kept in the analysis, an artifact present on a single bad sensor can spread to other sensors, for example

[1]They are not necessarily equivalent in the case of a bipolar montage in EEG. However, for the sake of simplicity, we shall use these terms interchangeably in this work.

due to spatial projection. This is why identifying bad sensors is crucial for data cleaning techniques such as the very popular Signal Space Separation (SSS) method (Taulu et al., 2004). Frequency filtering (Widmann et al., 2015) can often suppress many low frequency artifacts, but turns out to be insufficient for broadband artifacts. A common practice to mitigate this issue is to visually inspect the data using an interactive viewer and mark manually, the bad sensors and bad segments in the data. Although trained experts are very likely to agree on the annotation of bad data, their judgement is subject to fluctuations and cannot be repeated. Their judgement can also be biased due to prior training with different experimental setups or equipments, not to mention the difficulty for such experts to allocate some time to review the raw data collected everyday.

Luckily, popular software tools such as Brainstorm (Tadel et al., 2011), EEGLAB (Delorme and Makeig, 2004), FieldTrip (Oostenveld et al., 2011), MNE (Gramfort et al., 2013a) or SPM (Litvak et al., 2011) already allow for the rejection of bad data segments based on simple metrics such as peak-to-peak signal amplitude differences that are compared to a manually set threshold value. When the peak-to-peak amplitude in the data exceeds a certain threshold, it is considered as bad. However, while this seems quite easy to understand and simple to use from a practitioner's standpoint, this is not always convenient. In fact, a good peak-to-peak signal amplitude threshold turns out to be data specific, which means that setting it requires some amount of trial and error.

The need for better automated methods for data preprocessing is clearly shared by various research teams, as the literature of the last few years can confirm. On the one hand, are pipeline-based approaches, such as Fully Automated Statistical Thresholding for EEG artifact rejection (FASTER by Nolan et al. (2010)) which detect bad sensors as well as bad trials using fixed thresholds motivated from classical Gaussian statistics. Methods such as PREP (Bigdely-Shamlo et al., 2015), on the other hand, aim to detect and clean the bad sensors only. Unfortunately, they do not offer any solution to reject bad trials. Other methods are available to solve this problem. For example, the Riemannian Potato (Barachant et al., 2013) technique can identify the bad trials as those where the covariance matrix lies outside of the "potato" of covariance matrices for good trials. By doing so, it marks trials as bad but does not identify the sensors causing the problem, hence not offering the ability to repair them. It appears that practitioners are left to choose between different methods to reject trials or repair sensors, whereas they are in fact intricately related problems and must be dealt with together.

Robust regression (Diedrichsen and Shadmehr, 2005) also deals with bad trials using a weighted average which mitigates the effect of outlier trials. Trials with artifacts end up with low contributions in the average. A related approach that is sometimes employed to ignore outlier trials in the average is the trimmed mean as opposed to a regular mean. The trimmed mean is a compromise between the mean which offers a high signal-to-noise ratio (SNR) but can be corrupted by outliers, and the median which is immune to outliers of extreme amplitudes but has a low SNR as it involves no averaging. Of course, neither of these strategies are useful when analyses have to be conducted on single trials. Another approach, which is also data-driven, is Sensor Noise Suppression (SNS) (De Cheveigné and Simon, 2008). It removes the sensor-level noise by spatially projecting the data of each sensor onto the subspace spanned by the principal components of all the other sensors. This projection is repeated in leave-one-sensor-out iterations so as to eventually clean all

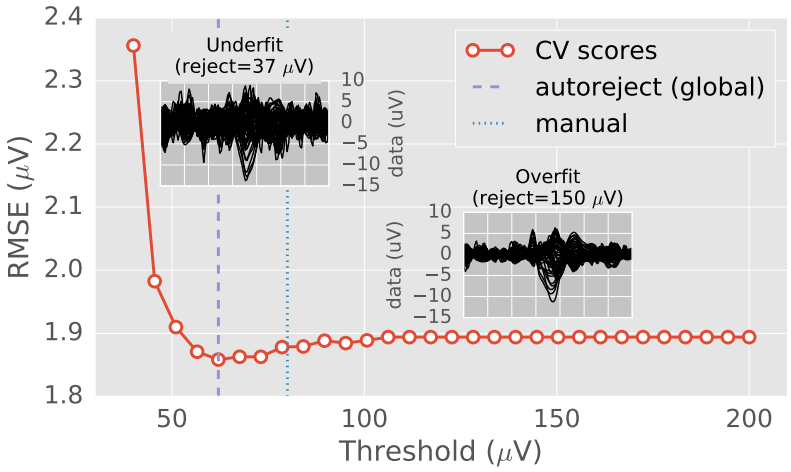


Figure 5.1: Cross-validation error as a function of peak-to-peak rejection threshold on one EEG dataset. The root mean squared error (RMSE) between the mean of the training set (after removing the trials marked as bad) and the median of the validation set was used as the cross-validation metric (Section 5.2.1). The two insets show the average of the trials as "butterfly plots" (each curve representing one sensor) for very low and high thresholds. For low thresholds, the RMSE is high because most of the trials are rejected (underfit). At high thresholds, the model does not drop any trials (overfit). The optimal data-driven threshold (*autoreject, global*) with minimum RMSE is somewhere in between. It closely matches the human threshold.

the sensors. In most of these methods, however, there are parameters which are somewhat dataset dependent and must therefore be manually tuned.

We therefore face the same problem in automated methods as in the case of semi-automated methods such as peak-to-peak rejection thresholds, namely the tuning of model parameters. In fact, setting the model parameters is even more challenging in some of the methods when they do not directly translate into human-interpretable physical units.

This led us to adopt a pragmatic approach in terms of algorithm design, as it focuses on the tuning of the parameters that M/EEG users presently choose manually. The goal is, not only to obtain high quality data but also to develop a method which is transparent and not too disruptive for the majority of M/EEG users. A first question we address below is: can we improve peak-to-peak based rejection methods by automating the process of trial and error? In the following section, we explain how the widely-known statistical method of cross-validation (see Figure 5.1 for a preview) in combination with Bayesian optimization (Snoek et al., 2012; Bergstra et al., 2011) can be employed to tackle the problem at hand. We then explain how this strategy can be extended to set thresholds separately for each sensor and mark trials as bad when a large majority of the sensors have high-amplitude artifacts. This process closely mimics how a human expert would mark a trial as bad during visual inspection.

In the rest of the chapter, we detail the internals of our algorithm, compare it against various state-of-the-art methods, and position it conceptually with respect to these different approaches. For this purpose, we make use of qualitative visualization techniques as well as quantitative reports. In a major validation effort, we take advantage of cleaned up evoked response fields (ERFs) provided by the Human Connectome Project (Larson-Prior

et al., 2013) enabling ground truth comparison between alternative methods. This work represents one of the first efforts in reanalysis of the MEG data from the HCP dataset using a toolkit stack significantly different from the one employed by the HCP consortium. The convergence between our method and the HCP MEG pipelines is encouraging and testifies to the success of the community-wide open science efforts aiming at reproducible research. Naturally, we have therefore made our code available online[2]. In addition to this, we validated our algorithm on the MNE sample data (Gramfort et al., 2013a), the multimodal faces dataset (Wakeman and Henson, 2015), and the EEGBCI motor imagery data (Goldberger et al., 2000; Schalk et al., 2004).

A preliminary version of this work was presented in Jas et al. (2016).

Notations We denote matrices by capital letters $X \in \mathbb{R}^{m \times n}$. The ith row of a matrix is indexed by subscripts, as in X_i, and the entry in the ith row and jth column is indexed as X_{ij}. The matrix X restricted to the rows with indices in the set $\mathcal{G}$ is denoted by $X_{\mathcal{G}}$. All sets $\mathcal{G}$, $\mathcal{T}$ or $\mathcal{V}$ are written in calligraphic fonts.

5.2 Materials and methods

We will first describe how a cross-validation procedure can be used to set peak-to-peak rejection thresholds globally (*i.e.* same threshold for all sensors). This is what we call *autoreject (global)*.

5.2.1 Autoreject (global)

We denote the data matrix by $X \in \mathbb{R}^{N \times P}$ with N trials and P features. These P features are the Q sensor-level time series, each of length T concatenated along the second dimension of the data matrix, such that $P = QT$. We divide the data into K folds (along the first dimension) with training set indices $\mathcal{T}_k$ and validation set indices $\mathcal{V}_k = [1..N] \setminus \mathcal{T}_k$ for each fold k ($1 \leq k \leq K$). For simplicity of notation, we first define the peak-to-peak amplitude for the ith trial and jth sensor as the difference between the maximum and the minimum value in that time series:

$$\mathcal{A}_{ij} = \max_{(j-1)T+1 \leq t \leq jT}(X_{it}) - \min_{(j-1)T+1 \leq t \leq jT}(X_{it}) \ . \tag{5.1}$$

The set of indices of good trials $\mathcal{G}_k$ in which the peak-to-peak amplitude $\mathcal{A}_{ij}$ for any sensor does not exceed the candidate threshold τ are generated as

$$\mathcal{G}_k = \{i \in \mathcal{T}_k \mid \max_{1 \leq j \leq Q} \mathcal{A}_{ij} \leq \tau\}. \tag{5.2}$$

By comparing the peak-to-peak threshold with the maximum of the peak-to-peak amplitudes, we ensure that none of the sensors exceed the given threshold. Once we have applied the threshold on the training set, it is necessary to evaluate how the threshold performs by looking at new data. For this purpose, we consider the validation set. We propose to compare the mean $\overline{X_{\mathcal{G}_k}}(\tau)$ of good trials in the training set against the median $\widetilde{X_{\mathcal{V}_k}}$ of all trials in the validation set. Using root mean squared error (RMSE) the mismatch $e_k(\tau)$ reads as:

$$e_k(\tau) = \|\overline{X_{\mathcal{G}_k}}(\tau) - \widetilde{X_{\mathcal{V}_k}}\|_{\text{Fro}}. \tag{5.3}$$

[2] https://autoreject.github.io

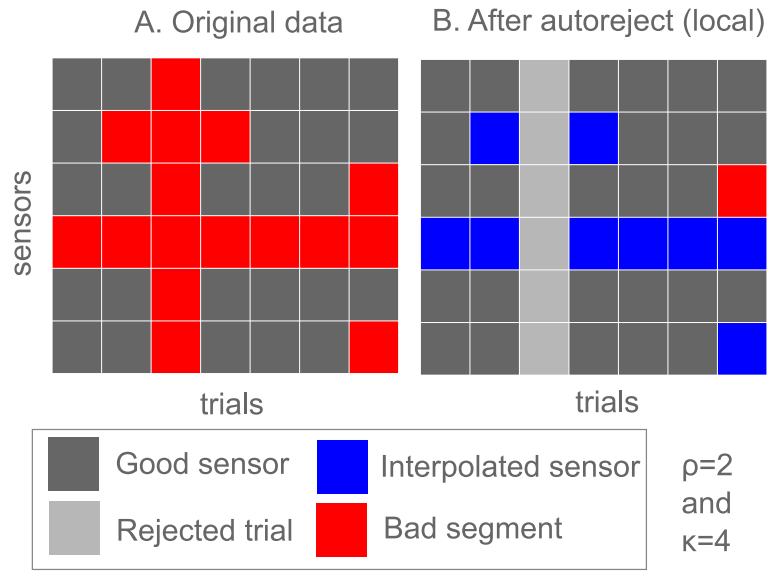


Figure 5.2: A schematic diagram explaining how *autoreject (local)* works. (A) Each cell here is an element of the transposed indicator matrix $C_{ij}^{\top}$ described in Section 5.2.2. Sensor-level thresholds are found and bad segments are marked for each sensor. Bad segments shown in red are where $C_{ij}^{\top} = 1$ (B) Trials are rejected if the number of bad sensors is greater than κ and otherwise, the worst ρ sensors (see Equation 5.7) are interpolated.

Here, $\|\cdot\|_{\text{Fro}}$ is the Frobenius norm. The rationale for using the median in the validation set is that it is robust to outliers. Indeed, it is far less affected by high-amplitude artifacts than the mean. The threshold with the best data quality (lowest mismatch $e_k(\tau)$) on average across the K folds is selected as the optimal threshold. In practice τ is taken in a bounded interval $[\tau_{\min}, \tau_{\max}]$:

$$\tau_\star = \underset{\tau \in [\tau_{\min}, \tau_{\max}]}{\operatorname{argmin}} \frac{1}{K} \sum_{k=1}^{K} e_k(\tau) \tag{5.4}$$

Note, that $\widetilde{X_{\mathcal{V}_k}}$ does not depend on τ. Indeed, it would not be wise to restrict the validation set to good trials according to the value of τ. As τ varies, it would lead to a variable number of validation trials, which would affect the comparison of RMSE across threshold values. The idea of using the median in the context of cross-validation has been previously proposed in the statistics literature in order to deal also with outliers (Zheng and Yang, 1998; Leung, 2005; De Brabanter et al., 2003).

Figure 5.1 (on page 85) shows how the average RMSE changes as the threshold varies for the MNE sample dataset (Gramfort et al., 2013a, 2014). At low thresholds, our model underfits as it drops most of the trials in the data resulting in a noisy average. On the other hand, at high thresholds, the model overfits retaining all the trials in the data including the high-amplitude artifacts. Here the candidate values of τ were taken on a grid. More details on how to solve (5.4) will be given in Section 5.2.3.

5.2.2 Autoreject (local)

A global threshold common to all sensors, however, suffers from limitations. A common case of failure is when a single sensor is affected (locally or globally) by high-amplitude artifacts. In this case, $\max_j \mathcal{A}_{ij}$, which would be the peak-to-peak amplitude that is compared to the threshold, comes from this bad sensor. If the sensor is not repaired or removed, we might end up rejecting a large fraction of otherwise good trials, just because of a single bad sensor. This is certainly not optimal. In fact, a possibly better solution is to replace the corrupted signal in the sensor by the interpolation of the signals in the nearby sensors. A second observation is that sensors can have very different ranges of amplitudes depending on their location on the scalp. A threshold tuned for one sensor may not work as effectively for another sensor. Both of these observations are motivations for estimating rejection thresholds for each sensor separately.

Once we define sensor-wise rejection thresholds $\tau_\star^j$, we can define an indicator matrix $C_{ij} \in \{0, 1\}^{N \times Q}$ which designates the bad trials at the level of individual sensors. In other words, we have:

$$C_{ij} = \begin{cases} 0, & \text{if } \mathcal{A}_{ij} \leq \tau_\star^j \\ 1, & \text{if } \mathcal{A}_{ij} > \tau_\star^j \end{cases} \tag{5.5}$$

The schematic in Figure 5.2A shows a cartoon figure for this indicator matrix C_{ij}. Now that we have identified bad sensors for each trial, one might be tempted to interpolate all the bad sensors in each trial. However, it is not as straightforward since in some trials, a majority of the sensors may be bad. These trials cannot be repaired by interpolation and must be rejected. In some other cases, the number of bad sensors may not be large enough to justify rejecting the trial. However, it might already be too much to interpolate all the sensors reliably. In these cases, a natural idea is to pick the worst few sensors and interpolate them. This suggests an algorithm as described in Figure 5.2B. Reject a trial only if most sensors "agree" that the trial is bad, otherwise interpolate as many sensors as possible. We will denote by κ the maximum number of bad sensors in a non-rejected trial and by ρ the maximum number of sensors that can be interpolated. Note that ρ is necessarily less than κ. The interpolation scheme for EEG uses spherical splines (Perrin et al., 1989) while for MEG it uses a Minimum Norm Estimates formulation with spherical harmonics (Hämäläinen and Ilmoniemi, 1994). The implementation is provided by MNE-Python (Gramfort et al., 2013a).

The set of good trials $\mathcal{G}_k^\kappa$ in the training set $\mathcal{T}_k$ can therefore be written mathematically as:

$$\mathcal{G}_k^\kappa = \{i \in \mathcal{T}_k \mid \sum_{j=1}^{Q} C_{ij} < \kappa\} \ . \tag{5.6}$$

In the remaining trials, if $\rho < \kappa$, one needs to define what are the worse ρ sensors that shall be interpolated. To do this we propose to rank the sensors for "badness" according to a score. A natural strategy to set the score is to use the peak-to-peak amplitude itself:

$$s_{ij} = \begin{cases} \mathcal{A}_{ij} & \text{if } C_{ij} = 1 \\ -\infty & \text{if } C_{ij} = 0 \end{cases} \tag{5.7}$$

The higher the score s_{ij}, the worse the sensor. The $-\infty$ score is for ignoring the good

sensors in the subsequent step. The following strategy is used for interpolation. If the number of bad sensors $\sum_{j'=1}^{Q} C_{ij'}$ is less than ρ we will interpolate all of them. Otherwise, we will interpolate the ρ sensors with the highest scores. In other words, we interpolate at most $\min(\rho, \sum_{j'=1}^{Q} C_{ij'})$ sensors.

Denoting by $X^{\rho}_{\mathcal{G}^{\kappa}_{k}}$ the data in the training set after rejection and cleaning by interpolation, the RMSE averaged over K folds for the parameter pair (ρ, κ) therefore becomes:

$$\bar{e}(\rho, \kappa) = \frac{1}{K} \sum_{k=1}^{K} \| \overline{X^{\rho}_{\mathcal{G}^{\kappa}_{k}}} - \widetilde{X_{\mathcal{V}_k}} \|_{\text{Fro}} \tag{5.8}$$

where $\| \cdot \|_{\text{Fro}}$ is the Frobenius norm. Finally, the best parameters ρ_* and κ_* are estimated using grid search (Hsu et al., 2003).

Data augmentation

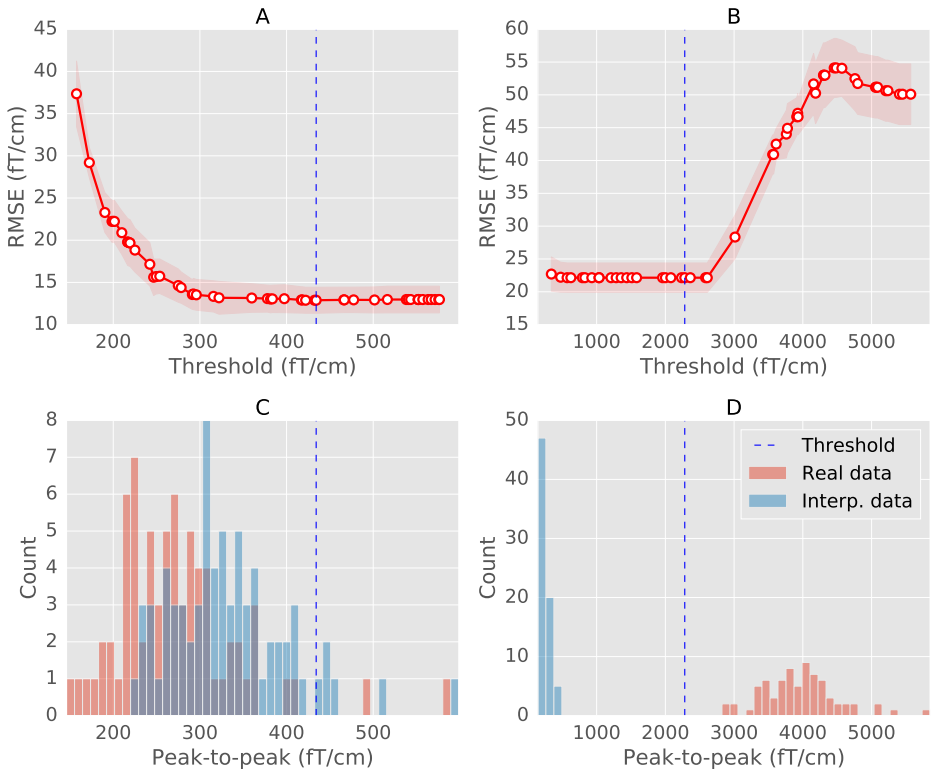


Figure 5.3: (A) and (B) The cross-validation curve obtained with sequential Bayesian optimization (see Section 5.2.3 for an explanation) for a regular (MEG 2523) and a globally bad sensor (MEG 2443) from the MNE sample dataset. The mean RMSE is shown in red circles with the standard deviation in red shades. Vertical dashed line marks the estimated threshold. (C) and (D) Histogram of peak-to-peak amplitudes of trials in the sensor. The histograms are computed separately for the real data (red) and the data interpolated from other sensors (blue). The estimated optimal threshold correctly marks all the trials as bad for the globally bad sensor.

In practice, cross-validation does not work for a globally bad sensor since all the trials are corrupted. In this scenario, the optimal threshold for this bad sensor should be

lower than the lowest peak-to-peak amplitude so that all the trials for that sensor are marked as bad. However, even the median of the validation set has been corrupted. The algorithm therefore attempts to keep as many trials as necessary for the average to be close to the corrupted median. Thus, the estimated threshold ends up being higher than what would have been optimal. Recall from Figure 5.1 that this is the classic case of an overfitting model. A common strategy in machine learning to reduce overfitting is data augmentation (Krizhevsky et al., 2012). It basically boils down to using the properties of the data (in our case, this being the physics of the system) to generate additional plausible data.

To implement data augmentation in our model, we interpolate each sensor from all the other $Q - 1$ sensors and by doing so, we double the number of trials in the data. In the augmented data, half of the trials contain sensor data. The augmented data matrix is $X^{\text{aug}} \in \mathbb{R}^{2N \times P}$. With the augmented data, the median is now closer to the uncorrupted median of the data in that sensor. During cross-validation the folds were stratified so that the number of interpolated trials and original trials in each fold were roughly equal.

5.2.3 Search for optimal thresholds using Bayesian optimization

Now that we have formalized the problem and our approach, we must estimate the threshold $\tau_\star$ which minimizes the error defined in Equation (5.3). A naïve strategy is to define a set of equally spaced points over a range of thresholds $[\tau_{\min}, \tau_{\max}]$. The estimated threshold would be the one which obtains the lowest error among these candidate threshold. This is the approach taken in Figure 5.1. The range of thresholds is easy to set as it can be determined from the minimum and maximum peak-to-peak amplitude for the sensor in the augmented data matrix X^{aug}. However, it is not obvious how to set the spacing between the candidate thresholds, and experiments showed that varying this spacing could impact the results. If the candidate thresholds are far apart, one might end up missing the optimal threshold. On the other hand, if the thresholds are very dense, it is computationally more demanding.

This motivated us to use Bayesian optimization (Snoek et al., 2012; Bergstra et al., 2011) to estimate the optimal thresholds. It is a sequential approach which decides the next candidate threshold to try based on all the observed thresholds so far. It is based on maximizing an acquisition function given an objective function of samples seen so far (data likelihood) and the prior (typically a Gaussian process (GP) (Rasmussen and Williams, 2006)). The objective function in our case is the mean cross-validation error as defined in Equations (5.3). To obtain the next iterate, an acquisition function is maximized over the posterior distribution. Popular choices of the acquisition function include "probability of improvement", "expected improvement" and "confidence bounds of the GP" (Snoek et al., 2012). We pick "expected improvement" as it balances exploration (searching unknown regions) and exploitation (maximizing the improvement) strategies without the need of a tuning parameter. For our analysis, we use the scikit-optimize[3] implementation of Bayesian optimization, which internally uses the Gaussian process module from scikit-learn (Pedregosa et al., 2011).

Figure 5.3A and 5.3B show the cross-validation curve for a regular sensor and a globally

[3] https://scikit-optimize.github.io

Table 5.1: Overview of rejection strategies evaluated

method	statistical scope	parameter defaults
FASTER[a]	univariate	threshold on zscore = 3
SNS[b]	multivariate	number of neighbors = 8
RANSAC[c]	multivariate outlier detection	#resamples = 50, fraction of channels = 0.25, threshold on correlation = 0.75, unbroken time = 0.4
autoreject	univariate with cross-validation	sensor-level thresholds, ρ and κ; learned from data

[a]Nolan et al. (2010), [b]De Cheveigné and Simon (2008), [c]Bigdely-Shamlo et al. (2015)

bad sensor in the MNE sample dataset (Gramfort et al., 2014, 2013a). The RMSE is evaluated on thresholds as determined by the Bayesian optimization rather than a uniform grid. These plots also illustrate the arguments presented in Section 5.2.2 with respect to data augmentation. The histograms in Figure 5.3C for the interpolated data and the real data are overlapping for the regular sensor. Thus, the estimated threshold for that sensor marks a trial as outlier if its peak-to-peak values is much higher than the rest of the trials. However, in the case of a globally bad sensor, the histogram (Figure 5.3D) is bimodal – one mode for the interpolated data and one mode for the real data. Now, the estimated threshold is no longer marking outliers in the traditional sense. Instead, all the trials belonging to that sensor must be marked as bad.

5.3 Experimental Validation Protocol

To experimentally validate *autoreject*, our general strategy is to first visually evaluate the results and thereafter quantify the performance. We describe below the evaluation metric used, the methods we compare against, and finally the datasets analyzed. All general data processing was done using the open source software MNE-Python (Gramfort et al., 2013a).

5.3.1 Evaluation metric

The evoked response from the data cleaned using our algorithm or a competing benchmark is denoted by $\overline{X}(method)$. This is compared to the ground truth evoked response $\overline{X}(clean)$ (See Section 5.3.2 to see how these are obtained for different datasets) using:

$$\|\overline{X}(method) - \overline{X}(clean)\|_{\infty} \tag{5.9}$$

where $\|\cdot\|_{\infty}$ is the infinity norm. The reason for using infinity norm is that it is sensitive to the maximum amplitude in the difference signal as opposed to the Frobenius norm which sums up the squared difference. The $\|\cdot\|_{\infty}$ is a particularly sensitive metric to quantity artifacts which are also visually striking such as those localized on one sensor or at a given time instant.

5.3.2 Competing methods

Here, we list the methods that will be quantitatively compared to *autoreject* using the evaluation metric in Equation 5.9. These methods are also summarized for the reader's convenience in Table 5.1.

- *No rejection*: It is a simple sanity check to make sure that the data quality upon applying the *autoreject (local)* algorithm does indeed improve. This is the data before the algorithm is applied.
- *Sensor Noise Suppression (SNS)*: The SNS (De Cheveigné and Simon, 2008) algorithm, as described in the Introduction (Section 5.1), projects the data of each sensor onto the subspace spanned by the principle components of all the other sensors. What it does is regressing out the sensor noise that cannot be explained by other sensors. It works on the principle that brain sources project onto multiple sensors but the noise is uncorrelated across sensors. In practice, not all the sensors are used for projection, but only a certain number of neighboring sensors (determined by the correlation in the data between the sensors).
- *Fully Automated Statistical Thresholding for EEG artifact Rejection (FASTER)*: It finds the outlier sensor using five different criteria: the variance, correlation, Hurst exponent, kurtosis and line noise. When the z-score of any of these criteria exceeds 3, the sensor is marked as bad according to that criterion. Note that even though FASTER is typically used as an integrated pipeline, here we use the bad sensor detection step, as this is what appears to dominate the bad signals in the case of the HCP data (Section 5.3.2). We take a union of the sensors marked as bad by the different criteria and interpolate the data for those sensors.
- *Random Sample Consensus (RANSAC)*: We use the RANSAC implemented as part of the PREP pipeline (Bigdely-Shamlo et al., 2015). In fact, RANSAC (Fischler and Bolles, 1981) is a well-known approach used to fit statistical models in the presence of outliers in the data. In this approach, adopted for the use case of artifact detection in EEG, a subset of sensors (inliers) are sampled randomly (25% of the total sensors) and the data in all sensors are interpolated from these inliers sensors. This is repeated multiple times (50 in the PREP implementation) so as to yield a set of 50 time series for each sensor. The correlation between the median, computed instant by instant, of these 50 time series and the real data is computed. If this correlation is less than a threshold (0.75 in the PREP implementation), then the sensor is considered an outlier and therefore marked as bad. It is perhaps worth noting that unlike in the classical RANSAC algorithm, the inlier model is not learned from the data but instead determined from the physical interpolation. A sensor which is bad for more than 40% of the trials in succession is marked as globally bad and interpolated. Even though the method was first proposed on EEG data only, we extended it for MEG data by replacing spline interpolation with field interpolation using spherical harmonics as implemented in MNE (Gramfort et al., 2013a; Hämäläinen and Ilmoniemi, 1994). Note that this is the same interpolation method that is used by *autoreject (local)*.

Datasets

We validated our methods on four open datasets with data from over 200 subjects. This allowed us to evaluate experimentally strengths and potential limitations of different rejection methods. The datasets contained either EEG or MEG data. To obtain solid experimental conclusions, diverse experimental paradigms were considered with data from working memory, perceptual and motor tasks.

We detail below how we defined $\overline{X}(clean)$, the cleaned ground-truth data for two of our

Table 5.2: Overview of datasets analyzed

Algorithm	Dataset	Acquisition device	Sensors used	#subjects
autoreject (global)	MNE sample data	Neuromag VectorView	60 EEG electrodes	1
	EEGBCI	BCI2000 cap	64 EEG electrodes	105
autoreject (local)	MNE sample data	Neuromag VectorView	60 EEG electrodes	1
	EEG faces	Neuromag VectorView	60 EEG electrodes	19
	HCP working memory	4D Magnes 3600 WH	248 magnetometers	83

datasets – HCP MEG and EEG faces data. This is perhaps one of the most challenging aspects of this work because the performance is evaluated on real data and not on simulations. An overview of all the datasets used in this study is provided in Table 5.2.

MNE sample data The MNE sample data (Gramfort et al., 2013a) is a multimodal open dataset consisting of MEG and EEG data. It has been integrated as the default testing dataset into the development of the MNE software (Gramfort et al., 2013a). The simultaneous M/EEG data were recorded at the Martinos Center of Massachusetts General Hospital. The MEG data with a Neuromag VectorView system, and an MEG-compatible cap comprising 60 electrodes was used for the EEG recordings. Data were sampled at 150 Hz. In the experiment, auditory stimuli (delivered monoaurally to the left or right ear) and visual stimuli (shown in the left or right visual hemifield) were presented in a random sequence with a stimulus onset asynchrony of 750 ms. The data was low-pass filtered at 40 Hz. The trials were 700 ms long including a 200 ms baseline period which was used for baseline correction.

EEGBCI dataset This is a 109-subject dataset (of which we analyzed 105 subjects which can be easily downloaded and analyzed using MNE-Python (Gramfort et al., 2013a)) containing EEG data recording with a 64-sensor BCI2000 EEG cap (Schalk et al., 2004). Subjects were asked to perform different motor/imagery tasks while their EEG activity was recorded. In the related BCI protocol, each subject performed 14 runs, amounting to a total of 180 trials for hands and feet movements (90 trials each). The data was band-pass filtered between 1 and 40 Hz, and 700 ms long trials were constructed including a 200 ms pre-stimulus baseline period.

EEG faces data (OpenfMRI ds000117) The OpenfMRI ds000117 dataset (Wakeman and Henson, 2015) contains multimodal task-related neuroimaging data over multiple runs for EEG, MEG and fMRI. For our analysis, we restrict ourselves to EEG data. The EEG data was recorded using a 70 channel Easycap EEG with electrode layout conforming to the 10-10% system. Subjects were presented with images of famous faces, unfamiliar faces and scrambled faces as stimuli. For each subject, on average, about 293 trials were available for famous and unfamiliar faces. The authors kindly provided us with run-wise bad sensor annotations which allowed us to conduct benchmarking against human judgement. To generate the ground truth evoked response $\overline{X}(clean)$, we randomly select 80 percent of the total number of trials in which famous and unfamiliar faces were displayed. In these trials, we interpolated the bad sensors run-wise. Then, we removed physiological artifacts (heart beat and eye blinks) using Independent Component Analysis (ICA) (Vigário et al., 2000). Following the ICA pipelines recommended by the MNE-Python software, the bad ICA components were marked automatically using cross-trial phase statistics (Dammers et al., 2008) for ECG (threshold=0.8) and adaptive z-scoring (threshold=3) for EOG

components. The evoked response from the cleaned data $\overline{X}(method)$ is computed from the remaining 20 percent trials cleaned using either *autoreject (local)* or *RANSAC* (see Section 5.4.3 for a description of this method). Computing the ground-truth evoked potential from a large proportion of trials minimized the effect of outliers in the average. However, it is noteworthy that this choice of assigning fewer trials to the estimation with rejection algorithms acts in a conservative sense: each unnoticed bad trial may affect the ensuing evoked potentials more severely.

Human Connectome Project (HCP) MEG data The HCP dataset is a multimodal reference dataset realized by the efforts of multiple international laboratories around the world. It currently provides access to both task-free and task-related data for more than 900 human subjects with functional MRI data, 95 of which have presently also MEG (Larson-Prior et al., 2013). An interesting aspect of the initiative is that the data provided is not only in unprocessed BTi format, but also processed using diverse processing pipelines. These include annotations of bad sensors and corrupted time segments for the MEG data derived from automated pipelines and supplemented by human inspection. The automated pipelines are based on correlation between neighboring sensors, z-score metrics, ratio of variance to neighbors, and independent component analysis (ICA) decomposition. Most significant for our purposes, the clean average response $\overline{X}(clean)$ is directly available. It allows us to objectively evaluate the proposed algorithm against state-of-the-art methods by reprocessing the raw data and comparing the outcome with the official pipeline output.

The HCP MEG dataset provides access to MEG recordings from diverse tasks, *i.e.*, a motor paradigm, passive listening and working memory. Here, we focused on the working memory task for which data is available for 83 subjects out of 95. A considerable proportion of subjects were genetically related, but we can ignore this information as the purpose of our algorithm is artifact removal rather than analyzing brain responses. For each subject two runs are available. Two classes of stimuli were employed, faces and tools. Here, we focused on the MEG data in response to stimulus onsets for the “faces" condition.

The MEG data were recorded with a wholehead MAGNES 3600 (4D Neuroimaging, San Diego, CA) in a magnetically shielded room at Saint Louis University. The system comprises 248 magnetometers and 23 reference sensors to capture environmental signals. Time windows precisely matched values used by the HCP “eravg” pipeline with onsets and offsets at -1.5 s and 2.5 s before and after the stimulus event, respectively. As in the HCP pipeline, signals were down-sampled to 508.63 Hz and band-pass filtered between 0.5–60 Hz. As it is commonly done with BTi systems, reference sensors at the periphery of the head were used to subtract away environmental noise. Given the linearity of Maxwell equations in the quasi-static regime, a linear regression model was employed. More precisely, signals from reference sensors are used as regressors in order to predict the MEG data of interest. The ensuing signal explained by the reference sensors in this model was then removed. The HCP preprocessing pipeline contains two additional steps: ICA was used to remove components not related to brain activity (including eye blinks and heart beats) and then bad trials and bad segments were removed with a combination of automated methods as well as annotations by a human observer. To have a fair comparison and focus on the latter step, the ICA matrices provided by the HCP consortium were applied to the data. We interpolated the missing sensors in $\overline{X}(clean)$ so that it has the same dimensions as

the data from $\overline{X}(method)$. All the algorithms were executed separately on each run and the evoked response of the two runs was averaged to get $\overline{X}(method)$.

To enable easy access of the files along with compatibility in MNE-Python, we make use of the open source MNE-HCP package[4]. For further details on the HCP pipelines, the interested reader can consult the related paper by Larson-Prior et al. (2013) and the HCP S900 reference manual for the MEG3 release.

5.4 Results

We conducted qualitative and quantitative performance evaluations of *autoreject* using four different datasets comparing it to a baseline condition without rejection as well as three different alternative artifact rejection procedures.

5.4.1 Peak-to-peak thresholds

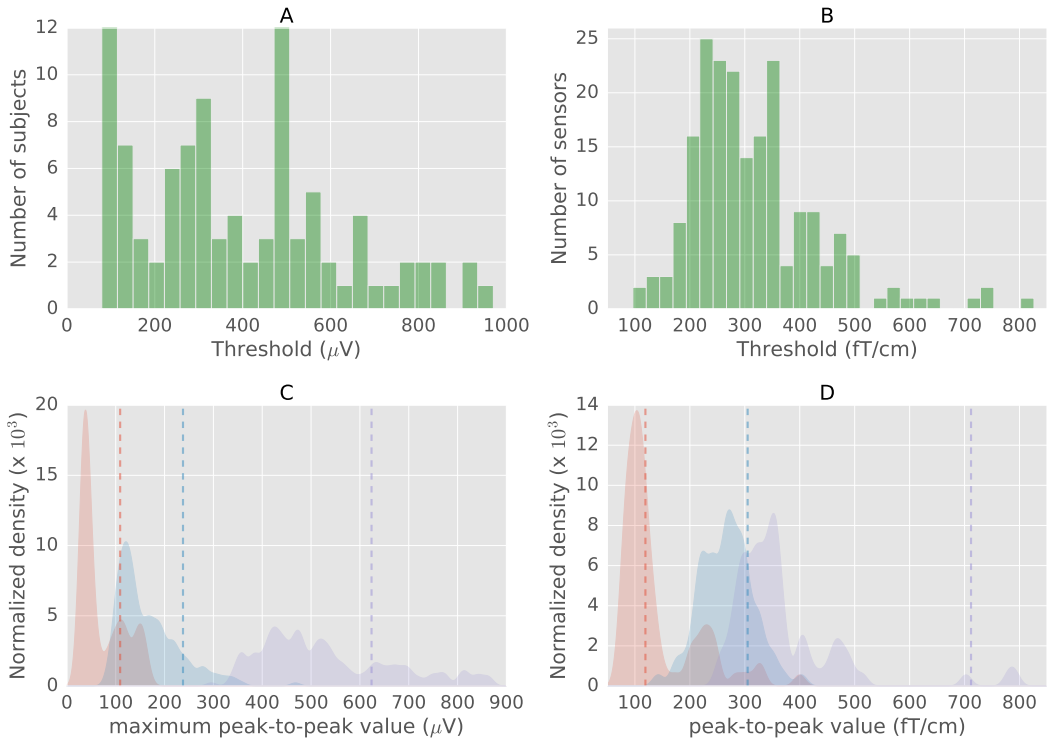


Figure 5.4: A. Histogram of thresholds for subjects in the EEGBCI dataset with *autoreject (global)* B. Histogram of sensor-specific thresholds in gradiometers for the MNE sample dataset (Section 5.4). C. Normalized kernel density plots of maximum peak-to-peak value across sensors for three subjects in the EEGBCI data. Vertical dashed lines indicate estimated thresholds. Density plots and thresholds corresponding to the same subject are the same color. D. Normalized Kernel Density plots of peak-to-peak values for three MEG sensors in the MNE sample dataset. The threshold indeed has to be different depending on the data (subject and sensor).

First, let us convince ourselves that the peak-to-peak thresholds indeed need to be learned.

[4] http://mne-tools.github.io/mne-hcp/

In Figure 5.4A, we show a histogram of the thresholds learned on subjects in the EEGBCI dataset using *autoreject (global)*. This figure shows that thresholds vary a lot across subjects. One could argue that this is due to variance in the estimation process. To rule out such a possibility, we plotted the distribution of maximum peak-to-peak thresholds as kernel density plots in Figure 5.4C for three different subjects. We can see that these distributions are indeed subject dependent, which is why a different threshold must be learned for each subject. In fact, if we were to use a constant threshold of $150\mu V$, in 17% of the subjects, all the trials would be dropped in one of the two conditions. Of course, from Figure 5.4A, we can now observe that $150\mu V$ is not really a good threshold to choose for many subjects.

We show here the maximum peak-to-peak amplitude per sensor because this is what decides if a trial should be dropped or not in the case of *autoreject (global)*. Note that, if instead, we examined the distribution of peak-to-peak amplitudes across all sensors and trials, we would see a quasi-normal distribution. When all the sensors are taken together, a "smoothing" effect is observed in the distribution. This is a consequence of the central limit theorem due to which the sum of independent random variables tends towards a normal distribution. This also explains why we cannot learn a global threshold using all the peak-to-peak amplitudes across trials and sensors.

With the *autoreject (local)* approach, a threshold is estimated for each sensor separately. The histogram of thresholds for the MNE sample dataset is plotted in Figure 5.4B. It shows that the threshold varies even across homogeneous MEG sensors. Figure 5.4D shows the distribution of peak-to-peak thresholds for three different MEG sensors. This graph confirms actual sensor-level differences in amplitude distributions, which was also previously reported in the literature (Junghöfer et al., 2000). With this work, we go one step further by learning automatically the thresholds in a data-driven way rather than asking users to mark them interactively.

5.4.2 Visual quality check

The average response plotted in a single graph, better known as "butterfly plots", constitutes a natural way to visually assess the performance of the algorithm for three different datasets – MNE sample data, HCP MEG data, and EEG faces data. In Figure 5.5, the subplots in the left column show the evoked response with the bad sensors marked in red. Right subplots, show data after applying the *autoreject (local)* algorithm, with the repaired bad sensors in red. The algorithm works for different acquisition modalities – MEG and EEG, and even when multiple sensors are bad. A careful look at the results, show that *autoreject (local)* does not completely remove eyeblinks in the data as some of the blinks are time-locked to the evoked response. We will later discuss (Section 5.5) the possible solutions of applying ICA-based artifact correction in combination with *autoreject (local)*.

5.4.3 Quantification of performance and comparison with state-of-the-art

We now compare these algorithms to *autoreject (local)* using the data quality metric defined in Equation (5.9) (See Section 5.4.4 for explanation as to why l_∞ norm is a better

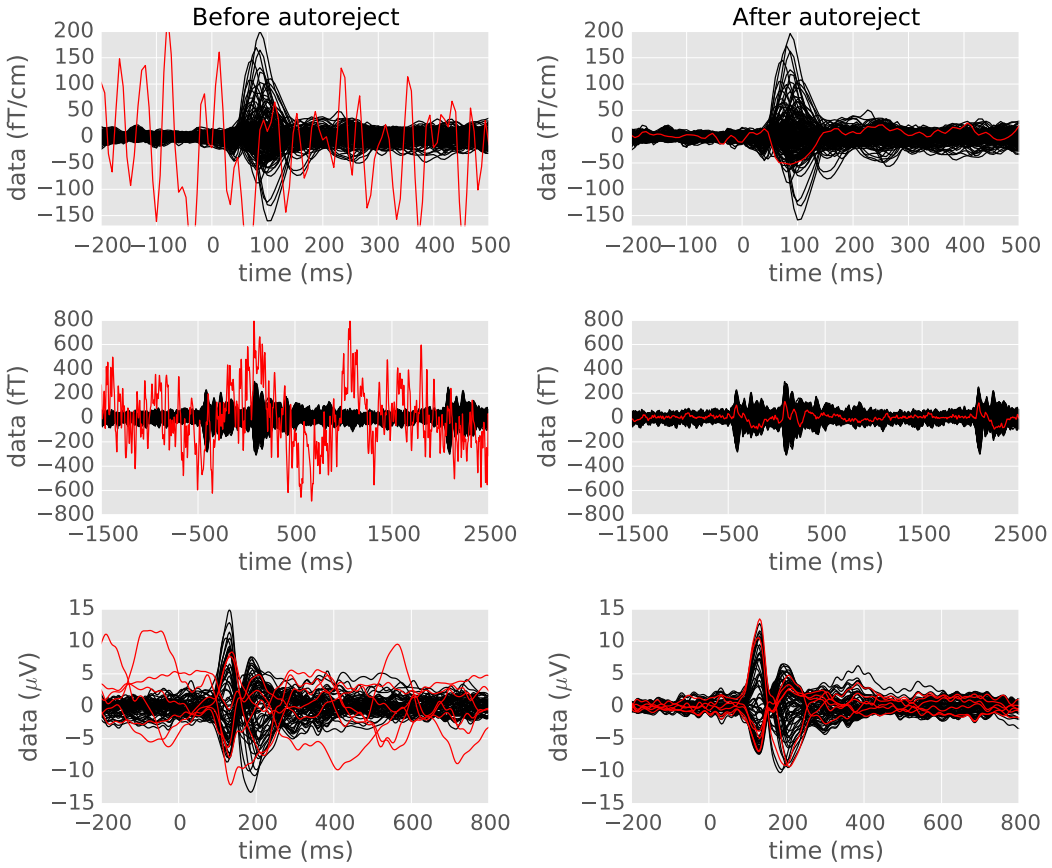


Figure 5.5: The evoked response (average of data across trials) on three different datasets before and after applying *autoreject* — the MNE sample data, the HCP data and the EEG faces data. Each sensor is a line on the plots. On the left, manually annotated bad sensors are shown in red. The algorithm finds the bad sensors automatically and repairs them for the relevant trials. Note that it can even fix multiple sensors at a time and works for different modalities of data acquisition.

choice than l_2 norm). We are interested not only in how the algorithms perform on average but at the level of individual subjects. To detail single subject performance, we present the data quality as scatter plots where each axis corresponds to the performance of a method. Figure 5.6, contains results on the HCP MEG data. We can observe from the top-left subplot of the figure that *autoreject (local)* does indeed improve the data quality in comparison to the *no rejection* approach. In Figure 5.6B, *autoreject (local)* is compared against SNS. The SNS algorithm focuses on removing noise isolated on single sensors. Its results can be affected by the presence of multiple bad sensors and globally bad trials. This explains why *autoreject (local)* outperforms SNS is this setting. In Figure 5.6C, we compare against FASTER. Even though *autoreject (local)* is slightly worse than FASTER for a few subjects, FASTER is clearly much worse than *autoreject (local)* for at least 3 subjects, and *autoreject (local)* yields therefore less errors on average. Finally, Figure 5.6D shows comparison to RANSAC. In the PREP implementation, this algorithm is not fully data-driven in the classic sense of RANSAC. This is due to the fact that the inlier model is not learned but rather derived from the physics of the interpolation. It is therefore an algorithm which is conceptually close to *autoreject*. However, a critical difference is that

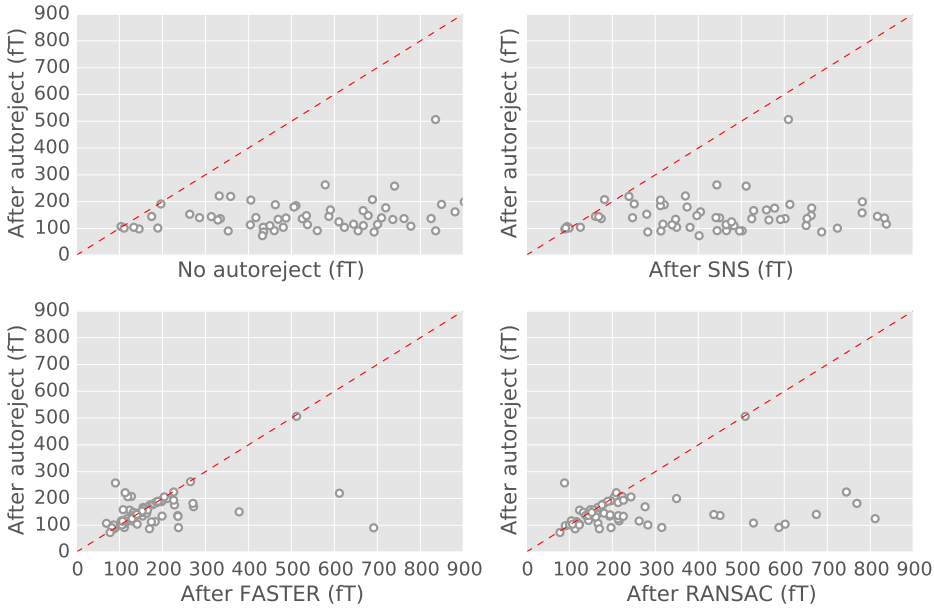


Figure 5.6: Scatter plots for the results with the HCP data. For each method, the $\|\cdot\|_\infty$ norm of the difference between the HCP ground truth and the method is taken. Each circle is a subject. (A) *autoreject (local)* against no rejection, (B) *autoreject (local)* against Sensor Noise Suppression (SNS) (SNS), (C) *autoreject* against FASTER, (D) *autoreject (local)* against RANSAC. Data points below the dotted red line indicate subjects for which *autoreject (local)* outperforms the alternative method.

the parameters of this method still need to be tuned. This can be a problem as these parameters can be suboptimal on some datasets. Some experiments showed that it is for example the case for the EEG faces data, where it is possible to obtain better results by manually tuning the RANSAC parameters, rather than using the values proposed by the original authors.

Figure 5.7 presents scatter plots for the EEG faces data. Here, we restrict our comparison to RANSAC as it is conceptually the closest to *autoreject*. On this data, we apply the algorithms on both the conditions – famous and unfamiliar faces. It should be noted that the ground truth for this data was generated automatically with no additional annotations from human experts. However, a sanity check was performed on the ground truth by visual inspection. Here too, *autoreject* offers good results across all subjects, and even for the subjects for which RANSAC underperforms.

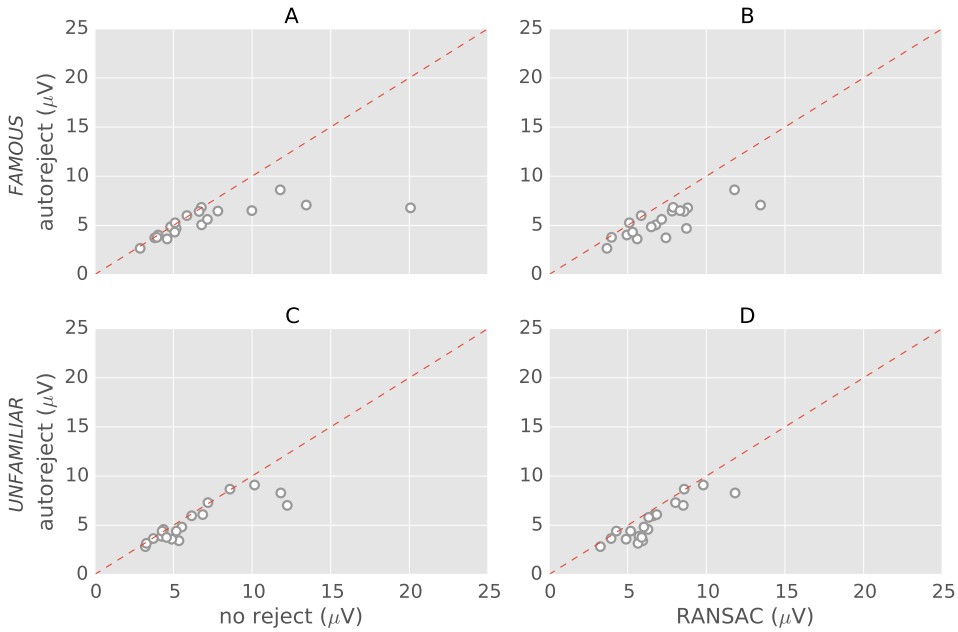


Figure 5.7: Scatter plots for the results with the 19 subjects from Faces dataset. The first row (A) and (B) is for the condition "famous" and the second row (C) and (D) is for the condition "unfamiliar" faces. For each method, the $\|\cdot\|_\infty$ norm of the difference between the ground truth and the estimates is computed. Each circle is a subject. Data points below the dotted red line indicate subjects for which *autoreject (local)* outperforms the alternative method.

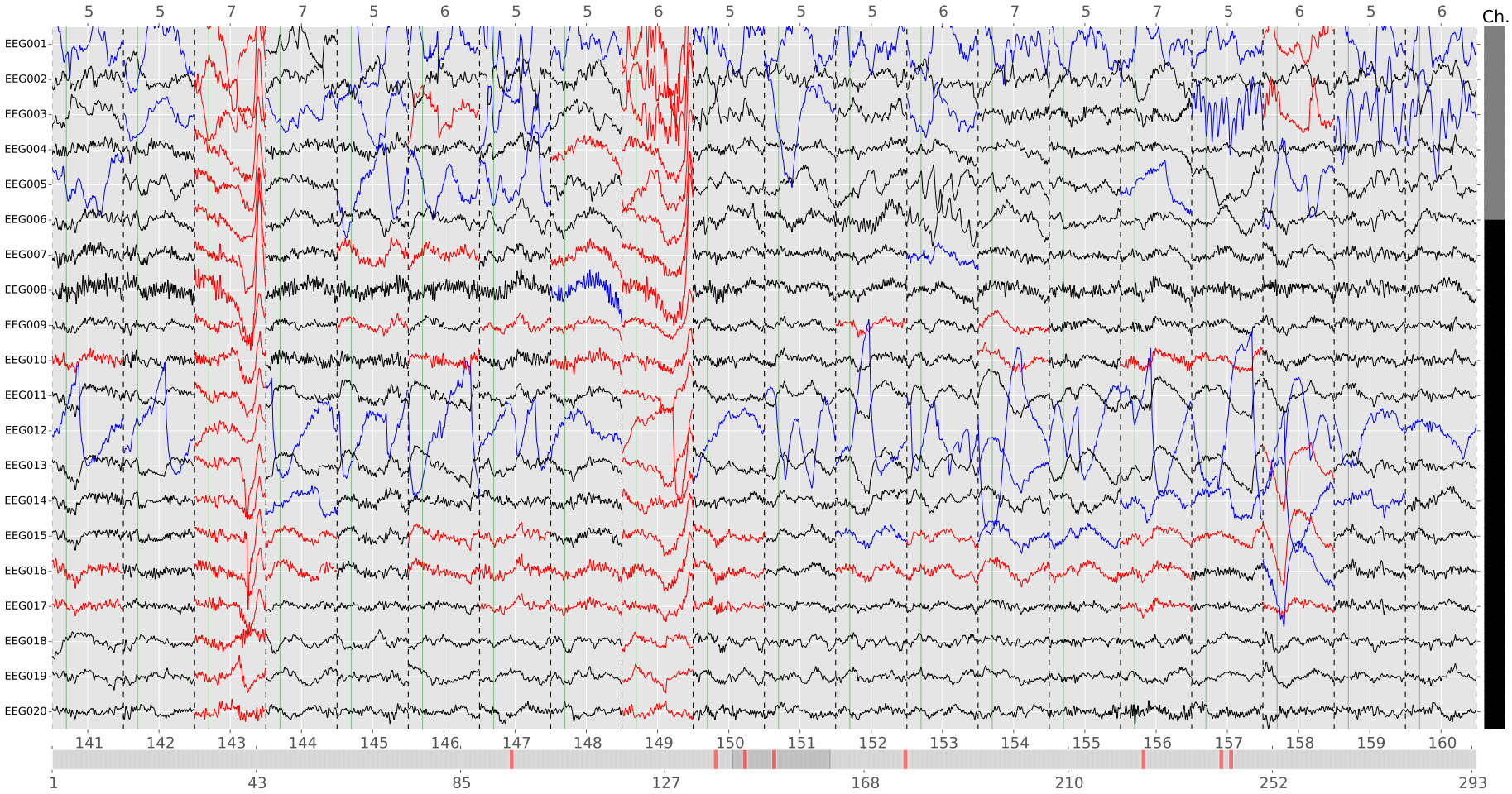


Figure 5.8: An example diagnostic plot from an interactive viewer with *autoreject (local)*. The data plotted here is subject 16 for the condition 'famous' in the EEG faces data. Each row is a different sensor. The trials are concatenated along the x axis with dotted vertical lines separating consecutive trials. Each trial is numbered at the bottom and its corresponding trigger code is at the top. The horizontal scroll bar at the bottom allows browsing trials and the vertical scroll bar on the right is for browsing sensors. A trial which is marked as bad is shown in red on the horizontal scroll bar and the corresponding column for the trial is also red. A data segment in a good trial is either i) Good (in black) ii) Bad and interpolated (blue), or iii) Bad but not interpolated (in red). Note that the worst sensors in a trial are typically interpolated.

5.4.4 ℓ_2 vs ℓ_∞ norm

Why not use ℓ_2 norm instead of ℓ_∞ norm to report the quantitative results in Figure 5.6 or Figure 5.7? The reason is that the ℓ_2 norm will average across the sensors. If one sensor is badly corrupted, then this would not be obvious with the ℓ_2 norm because the average in the ℓ_2 norm computation conceals the isolated problematic sensors with large artifacts. However, as the ℓ_∞ norm captures the worst sensor, it can be used to visualize pathological cases where even one sensor is corrupted. In Figure 5.9, we reproduce Figure 5.6 using the ℓ_2 norm instead of ℓ_∞. We can observe that, although the pattern remains the same, it is much less clear where one method outperforms the other. Even where *autoreject* isn't performing as well, it is not visible due to the averaging

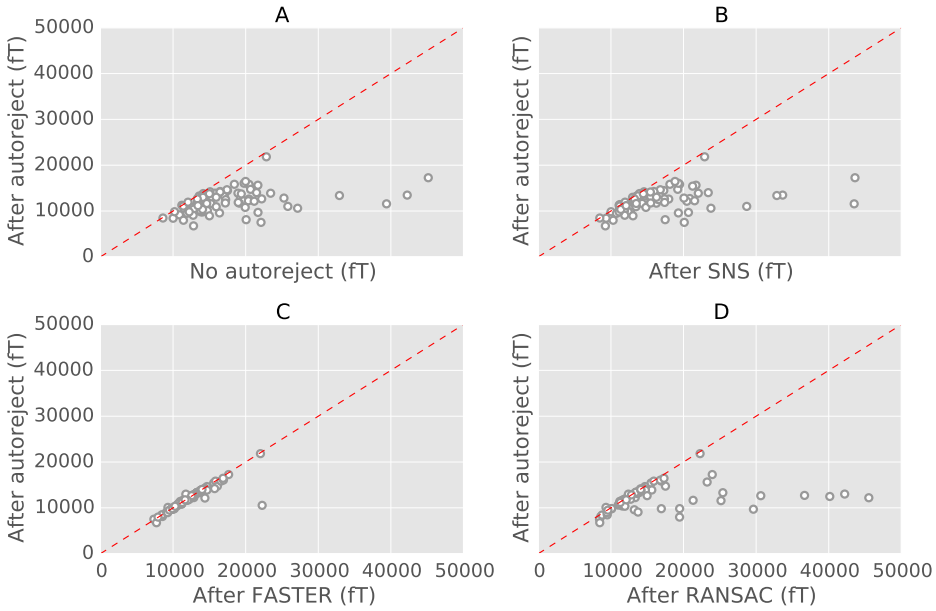


Figure 5.9: Scatter plots for the results with the HCP data. This figure uses the same data as in Figure 5.6 from the main text, but with $\|\cdot\|_2$ norm instead of the $\|\cdot\|_\infty$ norm for computing the difference between the HCP ground truth and the method. As before, each circle is a subject. (A) *autoreject (local)* against no rejection, (B) *autoreject (local)* against Sensor Noise Suppression (SNS) (SNS), (C) *autoreject* against FASTER, (D) *autoreject (local)* against RANSAC. Data points below the dotted red line indicate subjects for which *autoreject (local)* outperforms the alternative method.

5.5 Discussion

In this study, we have presented a novel artifact rejection algorithm called *autoreject* and assessed its performance on multiple datasets showing comparisons with other state-of-the-art methods.

We have shown that learning peak-to-peak rejection thresholds subject-wise is justified as the distribution of this statistic indeed varies considerably across subjects. We have shown qualitatively that *autoreject* yielded clean physiological event related field (ERF) and event related potentials (ERP) by correcting or rejecting contaminated data segments. Finally,

we have shown quantitavely that *autoreject* yields results closer to the ground truth for more subjects than the algorithms presented in Section 5.3.2. We now further discuss the conceptual similarities and differences of our approach to the alternative methods. We also discuss the interaction between *autoreject* and some other steps in the M/EEG analysis pipelines.

5.5.1 Autoreject vs. competing methods

We believe the key advantage of *autoreject (local)* over the other methods consists in combining data-driven parameter tuning with deterministic and physics-driven data interpolation. This interpolation promotes spatial smoothness of the electric potential on the scalp for EEG, and in the case of MEG, explicitly takes into account the well-understood Maxwell's equations. To recapitulate, the sensor-level thresholds mark outlier segments across trials at the level of individual sensors, following a data augmentation step which exploits the full array of sensors. As trials are seen as independent observations, the thresholds can be therefore learned from the data using cross-validation. The cross-validation is stratified so that each fold contains roughly an equal proportion of the original and augmented trials. At repair time, bad segments are replaced with interpolated data from the good sensors. Of course, this is problematic if the sensor locations are not readily available. Fortunately, it turns out that the sensor positions from standard montages are often good enough for reliable interpolation.

In contrast to *autoreject (local)*, SNS is a purely statistical method that does not take into account the physics of sensor locations for repairing the data. In SNS, the sensors are considered in a leave-one-sensor-out protocol. For each sensor, a "clean" subspace is defined from the principal components of the remaining sensors. The data from this sensor is then projected onto the "clean" subspace. As we have seen in Section 5.4 (Figure 5.6), this does not work satisfactorily, presumably because the SNS method makes strong assumptions regarding the orthogonality of the noise and "clean" subspace. The ensuing projection may not improve, and even deteriorate the signal in some cases. The consequence of this is what we observe empirically in Figure 5.6. Applying SNS will also be problematic when multiple sensors are corrupted simultaneously. However, this is less of a problem in the HCP MEG data that we analyzed.

On the other hand, the FASTER method derives its rejection decisions from multiple signal characteristics. It uses criteria such as between-sensor correlation, variance and power spectrum, by considering their univariate Gaussian statistics with thresholds fixed to a z-score of 3. This default threshold appears to be satisfying as they work on a vast majority of subjects. However, the fact that it does not work as well on certain subjects can limit its adoption for large scale studies. Here, the adaptive nature of threshold detection performed by *autoreject* seems to be a clear advantage.

The RANSAC algorithm also performs adaptive outlier detection, but across sensors rather than trials. While *autoreject (local)* operates on segmented data such as trials time-locked to the stimuli, RANSAC was designed for continuous data without any segmentation. In fact, one could readily obtain bad sensor per trial (as illustrated in Figure 5.2) even with RANSAC. However, the authors of the paper did not validate their method on continuous data, and hence, such a modification would require additional work. Although in the case of MEG data, this is not very crucial, this can in fact be critical for EEG data analysis.

Remember, that in EEG, one often has to deal with locally bad sensors. And in this context, it is noteworthy that none of the other methods we have discussed so far provides an explicit treatment for single trial analysis in the presence of locally bad sensors. Our comparison to the RANSAC algorithm seems to suggest that the RANSAC algorithm is indeed sensitive to the parameter settings. Even though the default settings appear to work reasonably well for the EEG data (Figure 5.7), they are not so optimal for the HCP MEG data (Figure 5.6).

It is perhaps worth emphasizing that using cross-validation implies that the trials with artifacts are independent. If this assumption is violated and if artifacts are phase-locked between the training and validation sets, *i.e.* occur for all trials at the same time relative to trial onsets, then this can interfere with the estimation procedure in *autoreject*. Another caveat to be noted is that if the data contains more than ρ^* (the maximum number of sensors that can be interpolated) bad sensors, and if the trial is not dropped, the data in the remaining bad sensors can still spread to other sensors if one were to use spatial filters such as SSP. Finally, *autoreject* considers only peak-to-peak thresholds for detecting bad sensors. Of course, the user must still mark low-amplitude flat sensors using another threshold; however, a simple threshold would suffice here as such sensors are usually consistently flat. Regardless of the method that researchers choose to adopt, diagnostic plots and automated reports (Engemann et al., 2015a) are an essential element to assess and better understand possible failures of automatic procedures. In this regard, transparency of the method in question is important. In the case of our *autoreject (local)* implementation, we offer the possibility for the user to inspect the bad segments marked by the automated algorithm and correct it if necessary. An example of such a plot is shown in Figure 5.8. Automating the detection of bad sensors and trials has the benefit of avoiding any unintentional biases that might be introduced if the experimenter were to mark the segments manually. In this sense, diagnostic visualization should supplement the analysis by ensuring accountability in the case of unexpected results.

5.5.2 Autoreject in the context of ICA, SSP and SSS

It is now important to place these results in the broader context of electrophysiological data analysis. Regarding the correction of specific artifacts such as electrooculogram (EOG) artifacts, *autoreject (local)* does indeed remove or interpolate some of the trials affected by eye blinks. This is because most eye blinks are not time-locked to the trial onsets and therefore get detected in the cross-validation procedure. However, the weaker eye blinks, particularly those smaller in magnitude than the evoked response, are not always removed. Also, the idea of rejection is to remove extreme values which are supposed to be rare events. This is why our empirical observation suggests that *autoreject (local)* is not enough in the presence of too frequent eye blinks, but also not enough to fully get rid of the smallest EOG artifacts.

This is where ICA (Vigário, 1997) and Signal Space Projection (SSP) (Uusitalo and Ilmoniemi, 1997) can naturally supplement *autoreject*. These methods are usually employed to extract and subsequently project out signal subspaces governed by physiological artifacts such as muscular, cardiac and ocular artifacts. However the estimation of these subspaces can be easily corrupted by other even more dramatic environmental or device-related artifacts. This is commonly prevented by band-pass filtering the signals and excluding

high-amplitude artifacts during the estimation of the subspaces. Both ICA and SSP (particularly if it is estimated from the data rather than an empty room recording) are highly sensitive to observations with high variance. Even though they involve estimating spatial filters that do not incorporate any notion of time, artifacts very localized in time will very likely have a considerable impact on the estimation procedure. This leads us to recommend removing globally bad sensors and employing appropriate rejection thresholds to exclude time segments with strong artifacts.

The success of applying *autoreject* to any electrophysiological data hinges critically on its ability to isolate artifacts local in time which cannot necessarily be identified by a prototypical spatial signature. However, the spatial interpolation employed by *autoreject* may not be able to repair sensors which are clustered together. In this case, the software package that implements the spatial interpolation should warn the user if the error due to the interpolation is likely to be high. Such a cluster of bad sensors can be expected in the case of physiological artifacts, such as muscular, cardiac or ocular artifacts. To take care of such artifacts with prototypical spatial patterns, ICA is certainly a powerful method, yet manual identification of artifactual components remains today done primarily manually.

If the context of data processing supports estimation of ICA and signal space projection (SSP) on segmented data, we would recommend to perform it after applying *autoreject*, benefiting from its automated bad sensor and bad trial handling. MEG signals usually contain a strong contribution from environmental electromagnetic fields. Therefore, interference suppression of MEG data is often needed, utilizing hardware and software based approaches (see, *e.g.* Parkkonen (2010) for details). In principle, spatial interpolation of bad sensor signals may not work very well unless the environmental interference has been removed. In the present study, the MNE sample data was recorded in a very well shielded room and did not need separate interference suppression, while the interference in the 4D/BTi data was removed by utilizing the reference channels. Spatial filtering approaches, such as SSP or SSS, may however produce a "chicken and egg" dilemma – whether to apply SSP/SSS or autoreject first - which can be solved using an iterative procedure as suggested by the PREP pipeline (Bigdely-Shamlo et al., 2015). That is, first run autoreject only for detection of bad channels but without interpolation. This is followed by an SSS run excluding the bad channels detected by autoreject. Finally, autoreject can be applied on the data free of environmental interference.

5.5.3 Source localization with artifact rejection

Obviously, artifact-free data benefits almost any analysis that is subsequently performed and the M/EEG inverse problem is no exception. Such benefits not only concern the quality of source estimates but also the choice of source-localization methods, as some of these methods require modification when certain artifact rejection strategies are employed. As *autoreject* amounts to automating a common, pre-existing and early processing step it does not require changes for source-level analyses. For example, evoked responses obtained using *autoreject (local)* can be readily analyzed with various source localization methods such as beamformer methods (Dalal et al., 2008; Groß et al., 2001), or cortically-constrained Minimum Norm Estimates with ℓ_2 penalty (Uutela et al., 1999), and noise-normalized schemes, such as dSPM (Dale et al., 2000b) and sLORETA (Pascual-Marqui et al.,

2002).

Certain denoising techniques such as SSP (Uusitalo and Ilmoniemi, 1997) or SSS (Taulu et al., 2004) reduce the rank of the data which can be problematic for beamforming techniques (Woolrich et al., 2011b). This needs special attention, and in some software such as MNE, this is handled using a non-square whitening matrix. However, as *autoreject* does not systematically reduce the rank of the data, it does not even require sophisticated handling of the data rank. At the same time, it works seamlessly with noise-normalization, where the estimation of the between-sensor noise covariance depends on the number of trials. To estimate the noise covariance during baseline periods, one computes the covariance of non-averaged data and then, assuming independence of each trial, the covariance gets divided by the number of trials present in the average (Engemann and Gramfort, 2015). Most existing pipelines scale the covariance by an integer number of trials. In contrast, methods such as robust regression (Diedrichsen and Shadmehr, 2005) that preferentially give less weight to noisy trials, require the noise normalization to be modified. Concretely, one would have to estimate an approximate number of trials or estimate the covariance matrix by restricting the computation to a subset of trials. *Autoreject* does not necessitate any such modifications to the source-localization pipeline, and hence, helps reduce the cognitive load of integration with pre-existing tools.

5.6 Conclusion

In summary, we have presented a novel algorithm for automatic data-driven detection and repair of bad segments in single trial M/EEG data. We therefore termed it *autoreject*. We have compared our method to state-of-the-art methods on four different open datasets containing in total more than 200 subjects. Our validation suggests that *autoreject* performs at least as good as diverse alternatives and commonly used procedures while often performing considerably better. This is the consequence of the combination of a data-driven outlier-detection step combined with physics-driven channel repair where all parameters are calibrated using a cross-validation strategy robust to outliers. The insight about the necessity to tune parameters at the level of single sensors and for individual subjects was further consolidated by our analyses of threshold distributions. The empirical variability of optimal thresholds across datasets emphasizes the importance of statistical learning approaches and automatic model selection strategies for preprocessing M/EEG signals. While *autoreject* makes use of black-box minimization strategies such as Bayesian hyperparameter optimization, it is also grounded in the physics underlying the data generation. It is therefore not purely a black-box data-driven approach. It balances the trade-off between accuracy and interpretability. Indeed all *autoreject* parameters have a meaning from a user standpoint and the algorithmic decisions can be explained. Supplemented by efficient diagnostic visualization routines, *autoreject* can be easily integrated in MEG/EEG analysis pipelines, including clinical ones where understanding algorithmic decisions is mandatory for tool adoption.

By offering an automatic and data-driven algorithmic solution to a task mostly so far done manually, *autoreject* reduces the cost of data inspection by experts. By allowing to repair data rather than removing it from the study, it allows saving data which are also costly to acquire. In addition, it removes the experts' bias which are due to specific

training or prior experience, as well as some expectations about the data. It does so by defining a clear set of rules serving as inclusion criteria for M/EEG data, making results more easily reproducible and eventually limiting the risk of false discoveries. Furthermore, as data sharing across centers has become a common practice, *autoreject* addresses the issue of heterogeneous acquisition setups. Indeed, each acquisition set-up has its intrinsic signal qualities, which means that preprocessing parameters can vary significantly between datasets. As opposed to alternative methods, *autoreject* automates the estimation of its parameters.

Chapter 6

Temporal representation learning

"Sparse is better than dense."

—The Zen of Python

Contents

So far, we studied automation in neuroimaging with the objective of enabling scalable data analysis and reproducibility. While reproducibility and large-scale data analysis allow us to consolidate upon existing studies, *per se* they are not tools to uncover new and interesting phenomena. In this chapter, we will explore this dimension of automation using what is known as *representation learning*.

Representations are the building blocks of signal processing. It is quite easy to convince ourselves of this fact, if we simply use a Fast Fourier Transform (FFT) to filter data. When we are using an FFT, we are in effect, decomposing the signal into a sum of sinusoids of varying frequencies. If we are interested in a time-frequency analysis, a common choice of representation for neurosience signals consists in using Morlet wavelets.

Traditionally, the choice of representation has been mainly driven by analytical concern and ease of mathematical manipulation. However, the recent surge of deep learning has ignited an interest in data-driven representations. It is because good representations that compactly capture the properties of the data are essential for efficient and accurate learning systems. In computer vision, for instance, handcrafted features such as SIFT (Lowe, 1999) and GIST descriptors (Oliva and Torralba, 2001), Deformable Parts Model (DPM) (Felzenszwalb et al., 2010), Histogram of Oriented Gradient (HOG) (Dalal and Triggs, 2005) *etc.* had been the norm, before it was realized that unsupervised learning and autoencoders performed much better.

Today, unsupervised learning is used as a first step for a supervised learning task in computer vision. Representation learning, by itself, is perhaps not as interesting, except for diagnostic visualizations in deep learning (Zeiler and Fergus, 2014). Despite this, there has always been an interest in understanding representations in the human brain (visual system particularly), as it was thought that this would help us build better learning systems. One of the pioneers in this area of research is Bruno Olshausen, whose work on dictionary learning (Olshausen and Field, 1996) demonstrated that Gabor patches are indeed fundamental to natural images, similar to the ones that Hubel and Wiesel (Hubel and Wiesel, 1962; Marĉelja, 1980) found in the cat visual cortex, and to what is used in GIST features. Barring this line of studies, the learned representation itself is not considered as meaningful as performance metrics like the prediction score or reconstruction loss. However, in the case of neural signals, we realized that this is not the case and the fidelity of the representation is in itself interesting. Indeed, the shape of the signal is a crucial biomarker in many clinical applications for neuroscience (Cole and Voytek, 2017).

A parallel development in the field of neuroimaging has been the rise in interest for learning prototypical shapes which are shift invariant (Jost et al., 2006; Barthélemy et al., 2013; Brockmeier and Príncipe, 2016; Hitziger et al., 2017). It is motivated by the fact that existing approximations using the Fourier basis often distorts the signal. There is, for example, a debate regarding the type of filters that should be used (See Section 4.3.3 and Widmann et al. (2015); Parks and Burrus (1987); Ifeachor and Jervis (2002); Götz et al. (2015)). Even though some success has been reported with these algorithms in neuroimaging, they are limited in applicability due to their heuristic nature. Remarkably, there has been so far very little cross-pollination of ideas between the computer vision and neuroimaging communities on these sparse coding aspects. Our work is an attempt to bridge this gap. We propose a model which builds upon existing shift-invariant sparse

coding models to be able to handle heavy-tailed noise and artifacts. It assumes positivity of the coefficients to account for the fact that an atom does not change polarity over time.

Our model is a novel probabilistic convolutional sparse coding (CSC) model for learning shift-invariant atoms from unprocessed neural time series data containing potentially severe artifacts. In the core of our model, which we call αCSC, lies a family of heavy-tailed distributions called α-stable distributions. We develop a novel, computationally efficient Monte Carlo expectation-maximization algorithm for inference. The maximization step boils down to a weighted CSC problem, for which we develop a computationally efficient optimization algorithm.

In our work, we rigorously evaluate the computational efficiency of our algorithm against the competing benchmarks. Because the CSC problem is non-convex, the optimization procedure involves nested loops and theoretical analysis often falls short in dealing with the complexity of non-convex functions. The optimization procedure is nested as the problem is convex when one of the variables is fixed: the atoms or the activations. The outer loop alternates between these two variables while the inner loop learns them when the other is fixed. An experimental approach, while challenging, is not completely out of reach. The final result depends on the initialization, and therefore algorithms can be compared only if they are tested for many different random seeds and their results averaged. Our qualitative analysis also goes beyond the narrative of verifying the existence of known waveforms to uncovering more complex structures in the data.

Our results show that the proposed algorithm achieves state-of-the-art convergence speeds. Besides, αCSC is significantly more robust to artifacts when compared to three competing algorithms: it can extract spike bursts, oscillations, and even reveal more subtle phenomena such as cross-frequency coupling when applied to noisy neural time series.

Section 6.1 to Section 6.5 was published in:

- M. Jas, L. Tour, T. Dupré, U. Şimşekli, and A. Gramfort. Learning the morphology of brain signals using alpha-stable convolutional sparse coding. In *Advances in Neural Information Processing Systems (NIPS)*, 2017c

6.1 Introduction

Neural time series data, either non-invasive such as electroencephalography (EEG) or invasive such as electrocorticography (ECoG) and local field potentials (LFPs), are fundamental to modern experimental neuroscience. Such recordings contain a wide variety of 'prototypical signals' that range from beta rhythms (12–30 Hz) in motor imagery tasks and alpha oscillations (8–12 Hz) involved in attention mechanisms, to spindles in sleep studies, and the classical P300 event related potential, a biomarker for surprise. These prototypical waveforms are considered critical in clinical and cognitive research (Cole and Voytek, 2017), thereby motivating the development of computational tools for learning such signals from data.

Despite the underlying complexity in the morphology of neural signals, the majority of the computational tools in the community are based on representing the signals with rather simple, predefined bases, such as the Fourier or wavelet bases (Cohen, 2014). While such bases lead to computationally efficient algorithms, they often fall short at capturing the precise morphology of signal waveforms, as demonstrated by a number of recent studies (Jones, 2016; Mazaheri and Jensen, 2008). An example of such a failure is the disambiguation of the alpha rhythm from the mu rhythm (Hari and Puce, 2017), both of which have a component around 10 Hz but with different morphologies that cannot be captured by Fourier- or wavelet-based representations.

Recently, there have been several attempts for extracting more realistic and precise morphologies directly from unfiltered electrophysiology signals, via dictionary learning approaches (Jost et al., 2006; Brockmeier and Príncipe, 2016; Hitziger et al., 2017; Gips et al., 2017). These methods all aim to extract certain *shift-invariant* prototypical waveforms (called 'atoms' in this context) to better capture the temporal structure of the signals. As opposed to using generic bases that have predefined shapes, such as the Fourier or the wavelet bases, these atoms provide a more meaningful representation of the data and are not restricted to narrow frequency bands.

In this line of research, Jost et al. (2006) proposed the MoTIF algorithm, which uses an iterative strategy based on generalized eigenvalue decompositions, where the atoms are assumed to be orthogonal to each other and learnt one by one in a greedy way. More recently, the 'sliding window matching' (SWM) algorithm (Gips et al., 2017) was proposed for learning time-varying atoms by using a correlation-based approach that aims to identify the recurring patterns. Even though some success has been reported with these algorithms, they have several limitations: SWM uses a slow stochastic search inspired by simulated annealing and MoTIF poorly handles correlated atoms, simultaneously activated, or having varying amplitudes; some cases which often occur in practical applications.

A natural way to cast the problem of learning a dictionary of shift-invariant atoms into an optimization problem is a CSC approach (Grosse et al., 2007). This approach has gained popularity in computer vision (Heide et al., 2015; Wohlberg, 2016; Zeiler et al., 2010; Šorel and Šroubek, 2016; Kavukcuoglu et al., 2010), biomedical imaging (Pachitariu et al., 2013) and audio signal processing (Grosse et al., 2007; Mailhé et al., 2008), due to its ability to obtain compact representations of the signals and to incorporate the temporal structure of the signals via convolution. In the neuroscience context, Barthélemy et al. (2013) used an extension of the K-SVD algorithm using convolutions on EEG

data. In a similar spirit, Brockmeier and Príncipe (2016) used the matching pursuit algorithm combined with a rather heuristic dictionary update, which is similar to the MoTIF algorithm. In a very recent study, Hitziger et al. (2017) proposed the AWL algorithm, which presents a mathematically more principled CSC approach for modeling neural signals. Yet, as opposed to classical CSC approaches, the AWL algorithm imposes additional combinatorial constraints, which limit its scope to certain data that contain spike-like atoms. Also, since these constraints increase the complexity of the optimization problem, the authors had to resort to dataset-specific initializations and many heuristics in their inference procedure.

While the current state-of-the-art CSC methods have a strong potential for modeling neural signals, they might also be limited as they consider an ℓ_2 reconstruction error, which corresponds to assuming an additive Gaussian noise distribution. While this assumption could be reasonable for several signal processing tasks, it turns out to be very restrictive for neural signals, which often contain heavy noise bursts and have low signal-to-noise ratio.

In this study, we aim to address the aforementioned concerns and propose a novel probabilistic CSC model called αCSC, which is better-suited for neural signals. αCSC is based on a family of *heavy-tailed* distributions called α-stable distributions (Samorodnitsky and Taqqu, 1994) whose rich structure covers a broad range of noise distributions. The heavy-tailed nature of the α-stable distributions renders our model robust to impulsive observations. We develop a Monte Carlo expectation maximization (MCEM) algorithm for inference, with a weighted CSC model for the maximization step. We propose efficient optimization strategies that are specifically designed for neural time series. We illustrate the benefits of the proposed approach on both synthetic and real datasets.

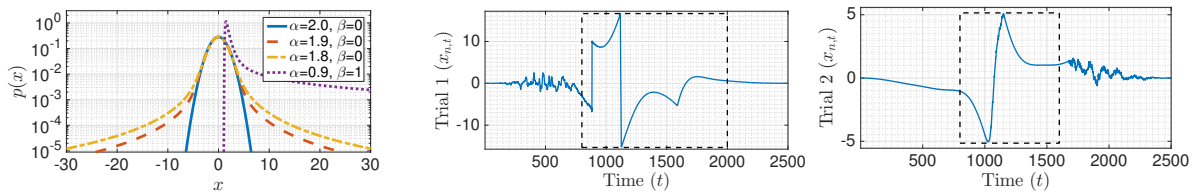
6.2 Preliminaries

Notation: For a vector $v \in \mathbb{R}^n$ we denote the ℓ_p norm by $\|v\|_p = (\sum_i |v_i|^p)^{1/p}$. The convolution of two vectors $v_1 \in \mathbb{R}^N$ and $v_2 \in \mathbb{R}^M$ (as explained in Definition 2.5) is denoted by $v_1 * v_2 \in \mathbb{R}^{N+M-1}$. We denote by x the observed signals, d the temporal atoms, and z the sparse vector of *activations*. The symbols $\mathcal{U}$, $\mathcal{E}$, $\mathcal{N}$, $\mathcal{S}$ denote the univariate uniform, exponential, Gaussian, and α-stable distributions, respectively.

Convolutional sparse coding: The CSC problem formulation adopted in this work follows the Shift Invariant Sparse Coding (SISC) model from Grosse et al. (2007). It is defined as follows:

$$\min_{d,z} \sum_{n=1}^{N} \Big(\frac{1}{2} \|x_n - \sum_{k=1}^{K} d^k * z_n^k\|_2^2 + \lambda \sum_{k=1}^{K} \|z_n^k\|_1 \Big), \quad \text{s.t.} \quad \|d^k\|_2^2 \leq 1 \text{ and } z_n^k \geq 0, \forall n, k \ , \tag{6.1}$$

where $x_n \in \mathbb{R}^T$ denotes one of the N observed segments of signals, also referred to as *trials* in this paper. We denote by T the length of a trial, and K the number of atoms. The aim in this model is to approximate the signals x_n by the convolution of certain *atoms* and their respective *activations*, which are sparse. Here, $d^k \in \mathbb{R}^L$ denotes the kth atom of the *dictionary* $d \equiv \{d^k\}_k$, and $z_n^k \in \mathbb{R}_+^{T-L+1}$ denotes the activation of the kth atom in the nth trial. We denote by $z \equiv \{z_n^k\}_{n,k}$.



(a) (b)

Figure 6.1: (a) PDFs of α-stable distributions. (b) Illustration of two trials from the striatal LFP data, which contain severe artifacts. The artifacts are illustrated with dashed rectangles.

The objective function (6.1) has two terms, an ℓ_2 data fitting term that corresponds to assuming an additive Gaussian noise model, and a regularization term that promotes sparsity with an ℓ_1 norm. The regularization parameter is called $\lambda > 0$. Two constraints are also imposed. First, we ensure that d^k lies within the unit sphere, which prevents the scale ambiguity between d and z. In other words, if one were to scale up d by a certain amount and scale down z by the same amount, it would not change $d * z$, and hence the objective function. Second, a positivity constraint on z is imposed to be able to obtain physically meaningful activations and to avoid sign ambiguities between d and z. This positivity constraint is not present in the original SISC model (Lewicki and Sejnowski, 1999; Grosse et al., 2007).

α-Stable distributions: The α-stable distributions have become increasingly popular in modeling signals that might incur large variations (Kuruoglu, 1999; Mandelbrot, 2013; Şimşekli et al., 2015; Wang et al., 2016; Leglaive et al., 2017) and have a particular importance in statistics since they appear as the limiting distributions in the generalized central limit theorem (Samorodnitsky and Taqqu, 1994). A distribution is said to be stable if a linear combination of two independent random variables with the distribution also has the same distribution upto a location and scale parameter. They are characterized by four parameters: α, β, σ, and μ: (i) $\alpha \in (0, 2]$ is the *characteristic exponent* or the *stability* parameter and determines the tail thickness of the distribution: the distribution will be heavier-tailed as α gets smaller. (ii) $\beta \in [-1, 1]$ is the *skewness* parameter. If $\beta = 0$, the distribution is symmetric. (iii) $\sigma \in (0, \infty)$ is the *scale* parameter and measures the spread of the random variable around its mode (similar to the standard deviation of a Gaussian distribution). Finally, (iv) $\mu \in (-\infty, \infty)$ is the location parameter (for $\alpha > 1$, it is simply the mean).

The probability density function of an α-stable distribution cannot be written in closed-form except for certain special cases; however, the characteristic function can be written as follows:

$$x \sim \mathcal{S}(\alpha, \beta, \sigma, \mu) \iff \mathbb{E}[\exp(i\omega x)] = \exp(-|\sigma\omega|^{\alpha}\left[1 + i\,\mathrm{sign}(\omega)\beta\psi_{\alpha}(\omega)\right] + i\mu\omega) \ ,$$

where $\psi_{\alpha}(\omega) = \log|\omega|$ for $\alpha = 1$, $\psi_{\alpha}(\omega) = \tan(\pi\alpha/2)$ for $\alpha \neq 1$, and $i = \sqrt{-1}$. As an important special case of the α-stable distributions, we obtain the Gaussian distribution when $\alpha = 2$ and $\beta = 0$, *i.e.* $\mathcal{S}(2, 0, \sigma, \mu) = \mathcal{N}(\mu, 2\sigma^2)$. In Fig. 6.1(a), we illustrate the (approximately computed) probability density functions (PDFs) of the α-stable distribution for different values of α and β. The distribution becomes heavier-tailed as we decrease α,

whereas the tails vanish quickly when $\alpha = 2$.

The moments of the α-stable distributions can only be defined up to the order α, i.e. $\mathbb{E}[|x|^p] < \infty$ if and only if $p < \alpha$, which implies the distribution has infinite variance when $\alpha < 2$. Furthermore, despite the fact that the PDFs of α-stable distributions admit no general analytical form, it is straightforward to draw random samples from them (Chambers et al., 1976).

6.3 Alpha-Stable Convolutional Sparse Coding

6.3.1 The Model

From a probabilistic perspective, the CSC problem can be also formulated as a maximum a posteriori (MAP) estimation problem on the following probabilistic generative model:

$$z^k_{n,t} \sim \mathcal{E}(\lambda), \quad x_{n,t}|z, d \sim \mathcal{N}(\hat{x}_{n,t}, 1), \quad \text{where,} \quad \hat{x}_n \triangleq \sum_{k=1}^{K} d^k * z^k_n \ . \tag{6.2}$$

Here, $z^k_{n,t}$ denotes the tth element of z^k_n. We use the same notations for $x_{n,t}$ and $\hat{x}_{n,t}$. It is easy to verify that the MAP estimate for this probabilistic model, *i.e.* $\max_{d,z} \log p(d, z|x)$, is identical to the original optimization problem defined in (6.1)[1].

It has been long known that, due to their light-tailed nature, Gaussian models often fail at handling noisy high amplitude observations or outliers (Huber, 1981). As a result, the 'vanilla' CSC model turns out to be highly sensitive to outliers and impulsive noise that frequently occur in electrophysiological recordings, as illustrated in Fig. 6.1(b). Possible origins of such artifacts are movement, muscle contractions, ocular blinks or electrode contact losses.

In this study, we aim at developing a probabilistic CSC model that would be capable of modeling challenging electrophysiological signals. We propose an extension of the original CSC model defined in (6.2) by replacing the light-tailed Gaussian likelihood (corresponding to the ℓ_2 reconstruction loss in (6.1)) with heavy-tailed α-stable distributions. We define the proposed probabilistic model (αCSC) as follows:

$$z^k_{n,t} \sim \mathcal{E}(\lambda), \quad x_{n,t}|z, d \sim \mathcal{S}(\alpha, 0, 1/\sqrt{2}, \hat{x}_{n,t}) \ , \tag{6.3}$$

where $\mathcal{S}$ denotes the α-stable distribution. While still being able to capture the temporal structure of the observed signals via convolution, the proposed model has a richer structure and would allow large variations and outliers, thanks to the heavy-tailed α-stable distributions. Note that the vanilla CSC defined in (6.2) appears as a special case of αCSC, as the α-stable distribution coincides with the Gaussian distribution when $\alpha = 2$.

[1] Note that the positivity constraint on the activations is equivalent to an exponential prior for the regularization term rather than the more common Laplacian prior.

6.3.2 Maximum A-Posteriori Inference

Given the observed signals x, we are interested in the MAP estimates, defined as follows:

$$(d^\star, z^\star) = \operatorname*{argmax}_{d,z} \sum_{n,t} \Big(\log p(x_{n,t}|d,z) + \sum_k \log p(z^k_{n,t}) \Big). \tag{6.4}$$

As opposed to the Gaussian case, unfortunately, this optimization problem is not amenable to classical optimization tools, since the PDF of the α-stable distributions does not admit an analytical expression. As a remedy, we use the product property of the symmetric α-stable densities (Samorodnitsky and Taqqu, 1994; Godsill and Kuruoglu, 1999) and re-express the αCSC model as conditionally Gaussian. It leads to:

$$z^k_{n,t} \sim \mathcal{E}(\lambda), \quad \phi_{n,t} \sim \mathcal{S}\Big(\frac{\alpha}{2}, 1, 2(\cos\frac{\pi\alpha}{4})^{2/\alpha}, 0\Big), \quad x_{n,t}|z,d,\phi \sim \mathcal{N}\Big(\hat{x}_{n,t}, \frac{1}{2}\phi_{n,t}\Big) \ , \tag{6.5}$$

where ϕ is called the *impulse* variable that is drawn from a *positive* α-stable distribution (i.e. $\beta = 1$), whose PDF is illustrated in Fig. 6.1(a). It can be shown that both formulations of the αCSC model are identical by marginalizing the joint distribution $p(x, d, z, \phi)$ over ϕ (Samorodnitsky and Taqqu, 1994, Proposition 1.3.1).

The impulsive structure of the αCSC model becomes more prominent in this formulation: the variances of the Gaussian observations are modulated by stable random variables with infinite variance, where the impulsiveness depends on the value of α. It is also worth noting that when $\alpha = 2$, $\phi_{n,t}$ becomes deterministic and we can again verify that αCSC coincides with the vanilla CSC.

The conditionally Gaussian structure has a crucial practical implication: if the impulse variable ϕ were to be known, then the MAP estimation problem over d and z in this model would turn into a 'weighted' CSC problem, which is a much easier task compared to the original problem. In order to be able to exploit this property, we propose an expectation maximization (EM) algorithm, which iteratively maximizes a lower bound of the log-posterior $\log p(d, z|x)$, and algorithmically boils down to computing the following steps in an iterative manner:

$$\text{E-Step:} \quad \mathcal{B}^{(i)}(d,z) = \mathbb{E}\left[\log p(x,\phi,z|d)\right]_{p(\phi|x,z^{(i)},d^{(i)})}, \tag{6.6}$$

$$\text{M-Step:} \quad (d^{(i+1)}, z^{(i+1)}) = \operatorname{argmax}_{d,z} \mathcal{B}^{(i)}(d,z). \tag{6.7}$$

where $\mathbb{E}[f(x)]_{q(x)}$ denotes the expectation of a function f under the distribution q, i denotes the iterations, and $\mathcal{B}^{(i)}$ is a lower bound to $\log p(d, z|x)$ and it is tight at the current iterates $z^{(i)}$, $d^{(i)}$.

The E-Step: In the first step of our algorithm, we need to compute the EM lower bound $\mathcal{B}$ that has the following form:

$$\mathcal{B}^{(i)}(d,z) =^+ -\sum_{n=1}^{N} \Big(\|\sqrt{w_n^{(i)}} \odot (x_n - \sum_{k=1}^{K} d^k * z_n^k)\|_2^2 + \lambda \sum_{k=1}^{K} \|z_n^k\|_1 \Big), \tag{6.8}$$

where $=^+$ denotes equality up to additive constants, $\odot$ denotes the Hadamard (element-wise) product, and the square-root operator is also defined element-wise. Here, $w_n^{(i)} \in \mathbb{R}_+^T$

are the *weights* that are defined as follows: $w_{n,t}^{(i)} \triangleq \mathbb{E}\left[1/\phi_{n,t}\right]_{p(\phi|x,z^{(i)},d^{(i)})}$. As the variables $\phi_{n,t}$ are expected to be large when $\hat{x}_{n,t}$ cannot explain the observation $x_{n,t}$ – typically due to a corruption or a high noise – the weights will accordingly suppress the importance of the particular point $x_{n,t}$. Therefore, the overall approach will be more robust to corrupted data than the Gaussian models where all weights would be deterministic and equal to 0.5.

Algorithm 2 α-stable Convolutional Sparse Coding

Require: Regularization: $\lambda \in \mathbb{R}_+$, Num. atoms: K, Atom length: L, Num. iterations: I , J, M

for $i = 1$ to I **do**
  / E-step: */*
  for $j = 1$ to J **do**
    Draw $\phi_{n,t}^{(i,j)}$ via MCMC (6.9)
  end for
  $w_{n,t}^{(i)} \approx (1/J) \sum_{j=1}^{J} 1/\phi_{n,t}^{(i,j)}$
  / M-step: */*
  for $m = 1$ to M **do**
    $z^{(i)}$ = L-BFGS-B on (6.10)
    $d^{(i)}$ = L-BFGS-B on the dual of (6.11)
  end for
end for
return $w^{(I)}$, $d^{(I)}$, $z^{(I)}$

Unfortunately, the weights $w^{(i)}$ cannot be computed analytically, therefore we need to resort to approximate methods. In this study, we develop a Markov chain Monte Carlo (MCMC) method to approximately compute the weights, where we approximate the intractable expectations with a finite sample average, given as follows: $w_{n,t}^{(i)} \approx (1/J) \sum_{j=1}^{J} 1/\phi_{n,t}^{(i,j)}$, where $\phi_{n,t}^{(i,j)}$ are some samples that are ideally drawn from the posterior distribution $p(\phi|x, z^{(i)}, d^{(i)})$. Unfortunately, directly drawing samples from the posterior distribution of ϕ is not tractable either, and therefore, we develop a *Metropolis-Hastings* algorithm (Chib and Greenberg, 1995), that asymptotically generates samples from the *target* distribution $p(\phi|\cdot)$ in two steps. In the j-th iteration of this algorithm, we first draw a random sample for each n and t from the prior distribution (cf. (6.5)), *i.e.*, $\phi'_{n,t} \sim p(\phi_{n,t})$. We then compute an acceptance probability for each $\phi'_{n,t}$ that is defined as follows:

$$\text{acc}(\phi_{n,t}^{(i,j)} \to \phi'_{n,t}) \triangleq \min\left\{1, p(x_{n,t}|d^{(i)}, z^{(i)}, \phi'_{n,t})/p(x_{n,t}|d^{(i)}, z^{(i)}, \phi_{n,t}^{(i,j)})\right\} \tag{6.9}$$

where j denotes the iteration number of the MCMC algorithm. Finally, we draw a uniform random number $u_{n,t} \sim \mathcal{U}([0, 1])$ for each n and t. If $u_{n,t} < \text{acc}(\phi_{n,t}^{(i)} \to \phi'_{n,t})$, we accept the sample and set $\phi_{n,t}^{(i+1)} = \phi'_{n,t}$; otherwise we reject the sample and set $\phi_{n,t}^{(i+1)} = \phi_{n,t}^{(i)}$. This procedure forms a Markov chain that leaves the target distribution $p(\phi|\cdot)$ invariant, where under mild ergodicity conditions, it can be shown that the finite-sample averages converge to their true values when J goes to infinity (Liu, 2008). More detailed explanation of this procedure is given in Section 6.3.3.

The M-Step: Given the weights w_n that are estimated during the E-step, the objective of the M-step (6.7) is to solve a weighted CSC problem, which is much easier when compared to our original problem. This objective function is not jointly convex in d and z, yet it is convex if one fix either d or z. Here, similarly to the vanilla CSC approaches (Gips et al., 2017; Grosse et al., 2007), we develop a *block coordinate descent* strategy, where we solve the problem in (6.7) for either d or z, by keeping respectively z and d fixed. We first focus on solving the problem for z while keeping d fixed, given as follows:

$$\min_{z} \sum_{n=1}^{N} \left(\|\sqrt{w_n} \odot (x_n - \sum_{k=1}^{K} D^k \bar{z}_n^k)\|_2^2 + \lambda \sum_{k} \|z_n^k\|_1 \right) \quad \text{s.t. } z_n^k \geq 0, \forall n, k \ . \tag{6.10}$$

Here, we expressed the convolution of d^k and z_n^k as the inner product of the zero-padded activations $\bar{z}_n^k \triangleq [(z_n^k)^\top, 0 \cdots 0]^\top \in \mathbb{R}_+^T$, with a Toeplitz matrix $D^k \in \mathbb{R}^{T \times T}$, that is constructed from d^k. The matrices D^k are never constructed in practice, and all operations are carried out using convolutions. This problem can be solved by various constrained optimization algorithms. Here, we choose the quasi-Newton L-BFGS-B algorithm Byrd et al. (1995) with a box constraint: $0 \leq z_{n,t}^k \leq \infty$. This approach only requires the simple computation of the gradient of the objective function with respect to z (*cf.* Section 6.3.4). Note that, since trials are mutually independent, we can solve this problem for each z_n in parallel.

We then solve the problem for the atoms d while keeping z fixed. This optimization problem turns out to be a constrained weighted least-squares problem. In the non-weighted case, this problem can be solved either in the time domain or in the Fourier domain (Grosse et al., 2007; Heide et al., 2015; Wohlberg, 2016). The Fourier transform simplifies the convolutions that appear in least-squares problem, but it also induces several difficulties, such as that the atom d_k have to be in a finite support L, an important issue ignored in the seminal work of Grosse et al. (2007) and addressed with an alternating direction method of multipliers (ADMM) solver in Heide et al. (2015); Wohlberg (2016). In the weighted case, it is not clear how to solve this problem in the Fourier domain. We thus perform all the computations in the time domain.

Following the traditional filter identification approach (Moulines et al., 1995), we need to embed the one-dimensional signals z_n^k into a matrix of delayed signals $Z_n^k \in \mathbb{R}^{T \times L}$, where $(Z_n^k)_{i,j} = z_{n,i+j-L+1}^k$ if $L - 1 \leq i + j < T$ and 0 elsewhere. Equation (6.1) then becomes:

$$\min_{d} \sum_{n=1}^{N} \|\sqrt{w_n} \odot (x_n - \sum_{k=1}^{K} Z_n^k d^k)\|_2^2, \quad \text{s.t. } \|d^k\|_2^2 \leq 1 \ . \tag{6.11}$$

Due to the constraint, we must resort to an iterative approach. The options are to use (accelerated) projected gradient methods such as FISTA (Beck and Teboulle, 2009) applied to (6.11), or to solve a dual problem as done in Grosse et al. (2007). The dual is also a smooth constraint problem yet with a simpler positivity box constraint (*cf.* Section 6.3.4). The dual can therefore be optimized with L-BFGS-B. Using such a quasi-Newton solver turned out to be more efficient than any accelerated first order method in either the primal or the dual (*cf.* benchmarks in Section 6.4.1).

Our entire EM approach can be summarized in the Algorithm 2. Note that during the alternating minimization, thanks to convexity we can warm start the d update and the

z update using the solution from the previous update. This significantly speeds up the convergence of the L-BFGS-B algorithm, particularly in the later iterations of the overall algorithm. We will not describe the E step and the M steps in more detail.

6.3.3 Details of the E-Step

Computing the weights that are required in the M-step requires us to compute the expectation of $\frac{1}{\phi_{n,t}}$ under the posterior distribution $p(\phi_{n,t}|x,d,z)$, which is not analytically available.

Monte Carlo methods are numerical techniques that can be used to approximately compute the expectations of the form:

$$\mathbb{E}[f(\phi_{n,t})] = \int f(\phi_{n,t})\pi(\phi_{n,t})d\phi_{n,t} \approx \frac{1}{J}\sum_{j=1}^{J} f(\phi_{n,t}^{(j)}) \tag{6.12}$$

where $\phi_{n,t}^{(j)}$ are some samples drawn from $\pi(\phi_{n,t}) \triangleq p(\phi_{n,t}|x,d,z)$ and $f(\phi) = 1/\phi$ in our case. However, in our case, sampling directly from $\pi(\phi_{n,t})$ is also unfortunately intractable.

MCMC methods generate samples from the target distribution $\pi(\phi_{n,t})$ by forming a Markov chain, whose stationary distribution is $\pi(\phi_{n,t})$, so that $\pi(\phi_{n,t}) = \int \mathcal{T}(\phi_{n,t}|\phi'_{n,t})p(\phi'_{n,t})d\phi'_{n,t}$, where $\mathcal{T}$ denotes the transition kernel of the Markov chain.

In this study, we develop a Metropolis-Hastings (MH) algorithm, that implicitly forms a transition kernel. The MH algorithm generates samples from a target distribution $\pi(\phi_{n,t})$ in two steps. First, it generates a random sample $\phi'_{n,t}$ from a *proposal* distribution $\phi'_{n,t} \sim q(\phi'_{n,t}|\phi_{n,t}^{(j)})$, then computes an acceptance probability $\text{acc}(\phi_{n,t}^{(j)} \to \phi'_{n,t})$ and draws a uniform random number $u \sim \mathcal{U}([0,1])$. If $u < \text{acc}(\phi_{n,t}^{(j)} \to \phi'_{n,t})$, it accepts the sample and sets $\phi_{n,t}^{(j+1)} = \phi'_{n,t}$; otherwise it rejects the sample and sets $\phi_{n,t}^{(j+1)} = \phi_{n,t}^{(j)}$. The acceptance probability is given as follows

$$\begin{aligned}\text{acc}(\phi_{n,t} \to \phi'_{n,t}) &= \min\Big\{1, \frac{q(\phi_{n,t}|\phi'_{n,t})\pi(\phi'_{n,t})}{q(\phi'_{n,t}|\phi_{n,t})\pi(\phi_{n,t})}\Big\} \\ &= \min\Big\{1, \frac{q(\phi_{n,t}|\phi'_{n,t})p(x_{n,t}|\phi'_{n,t},d,z)p(\phi'_{n,t})}{q(\phi'_{n,t}|\phi_{n,t})p(x_{n,t}|\phi_{n,t},d,z)p(\phi_{n,t})}\Big\}\end{aligned} \tag{6.13}$$

where the last equality is obtained by applying the Bayes rule on π.

The acceptance probability requires the prior distribution of ϕ to be evaluated. Unfortunately, this is intractable in our case since this prior distribution is chosen to be a positive α-stable distribution whose PDF does not have an analytical form. As a remedy, we choose the prior distribution of $\phi_{n,t}$ as the proposal distribution, such $q(\phi_{n,t}|\phi'_{n,t}) = p(\phi_{n,t})$. This enables us to simplify the acceptance probability. Accordingly, for each $\phi_{n,t}$, we have the following acceptance probability:

$$\text{acc}(\phi_{n,t}^{(i,j)} \to \phi'_{n,t}) \triangleq \min\Big\{1, \exp(\log\phi_{n,t}^{(i,j)} - \log\phi'_{n,t})/2 + (x_{n,t} - \hat{x}_{n,t}^{(i)})^2(1/\phi_{n,t}^{(i,j)} - 1/\phi'_{n,t})\Big\}. \tag{6.14}$$

Thanks to the simplification, this probability is tractable and can be easily computed.

6.3.4 Details of the M-Step

Solving for the activations: In the M-step, we optimize (6.10) to find the activations $z_n^{(i)}$ of each trial n independently. To keep the notation simple, we will drop the index for the iteration number i of the EM algorithm.

First, this equation can be rewritten by concatenating the Toeplitz matrices for the K atoms into a big matrix $D = [D^1, D^2, ..., D^K] \in \mathbb{R}^{T \times KT}$ and the activations for different atoms into a single vector $\bar{z}_n = [(\bar{z}_n^1)^\top, (\bar{z}_n^2)^\top, ..., (\bar{z}_n^K)^\top]^\top \in \mathbb{R}_+^{KT}$ where $(\cdot)^\top$ denotes the transposition operation. Recall that $\bar{z}_n^k$ is a zero-padded version of z_n^k. This leads to a simpler formulation and the objective function $\mathcal{L}(d, z)$:

$$\mathcal{L}(d, z) = \sum_{n=1}^{N} \frac{1}{2} \|\sqrt{w_n} \odot (x_n - D\bar{z}_n)\|_2^2 + \lambda \mathbb{1}^\top \bar{z}_n \ , \tag{6.15}$$

where $\mathbb{1} \in \mathbb{R}^{KT}$ is a vector of ones.

The derivative w.r.t. z_n now reads:

$$\frac{\partial \mathcal{L}(d, z)}{\partial \bar{z}_n} = D^\top (w_n \odot (x_n - D\bar{z}_n)) + \lambda \mathbb{1}^\top \ . \tag{6.16}$$

In practice, this big matrix D is never assembled and all operations are carried out using convolutions. Note also that we do not update the zeros from the padding in $\bar{z}_n^k$. Now that we have the gradient, the activations can be estimated using a efficient quasi-Newton solver such as L-BFGS-B, taking into account the box posititivy constraint $0 \leq z_n \leq \infty$.

For each trial, one iteration costs $\mathcal{O}(LKT)$.

Solving for the atoms: In the M-step, we optimize (6.11) to find the atoms d^k. As when solving for the activations z_n, we can remove the summation over the atoms by concatenating the delayed matrices into $Z_n = [Z_n^1, Z_n^2, \ldots, Z_n^K] \in \mathbb{R}^{T \times KL}$ and $d = [(d^1)^\top, (d^2)^\top, ..., (d^K)^\top]^\top \in \mathbb{R}^{KL}$. This leads to the simpler formulation:

$$\min_d \sum_{n=1}^{N} \frac{1}{2} \|\sqrt{w_n} \odot (x_n - Z_n d)\|_2^2, \quad \text{s.t. } \|d^k\|_2^2 \leq 1 \ . \tag{6.17}$$

The Lagrangian of this problem is given by:

$$g(d, \beta) = \sum_{n=1}^{N} \frac{1}{2} \|\sqrt{w_n} \odot (x_n - \sum_{k=1}^{K} Z_n^k d^k)\|_2^2 + \sum_k \beta^k (\|d^k\|_2^2 - 1) \quad \text{s.t. } \beta^k \geq 0 \ , \tag{6.18}$$

where $\beta = (\beta^1, \beta^2, ..., \beta^K)$ are the dual variables. Therefore, the dual problem is:

$$\min_d g(d, \beta) = g(d^*, \beta) \tag{6.19}$$

where d^*, the primal optimal, is given by:

$$d^* = (\sum_{n=1}^{N} Z_n^\top (w_n \odot Z_n) + \bar{\beta})^{-1} \sum_{n=1}^{N} (w_n \odot Z_n)^\top x_n \tag{6.20}$$

with $\bar{\beta} = \text{diag}([\mathbb{1}\beta^1, \mathbb{1}\beta^2, ..., \mathbb{1}\beta^K]) \in \mathbb{R}^{KL}$ with $\mathbb{1} \in \mathbb{R}^L$. The gradient for the dual variable β^k is given by:

$$\frac{\partial g(d^*, \beta)}{\partial \beta^k} = \|d^{*k}\|_2^2 - 1, \tag{6.21}$$

with d^{*k} computed from (6.20). We can solve this iteratively using again L-BFGS-B taking into account the positivity constraint $\beta^k \geq 0$ for all k. What we have described so far solves for all the atoms simultaneously. However, it is also possible to estimate the atoms sequentially one at a time using a block coordinate descent (BCD) approach, as in the work of (Mairal et al., 2010). In each iteration of the BCD algorithm, a residual r_n^k is computed as given by:

$$r_n^k = x_n - \sum_{k' \neq k} Z_n^{k'} d^{k'} \tag{6.22}$$

and correspondingly subproblem 6.17 becomes:

$$\min_{d^k} \sum_{n=1}^{N} \frac{1}{2} \|\sqrt{w_n} \odot (r_n^k - Z_n^k d^k)\|_2^2, \quad \text{s.t. } \|d^k\|_2^2 \leq 1, \quad . \tag{6.23}$$

which is solved in the same way as subproblem 6.17. Now, in the simultaneous case, we construct one linear problem in $\mathcal{O}(L^2K^2TN)$ and one iteration costs $\mathcal{O}(L^3K^3)$. However, in the BCD strategy, we construct K linear problems in $\mathcal{O}(L^2TN)$ and one iteration costs only $\mathcal{O}(L^3)$. Interestingly, when the weights w_n are all identical, we can use the fact that for one atom k, the matrix $\sum_{i=1}^{N} (Z_i^k)^T Z_i^k$ is Toeplitz. In this case, we can construct K linear problems in only $\mathcal{O}(LTN)$ and one iteration costs only $\mathcal{O}(L^2)$.

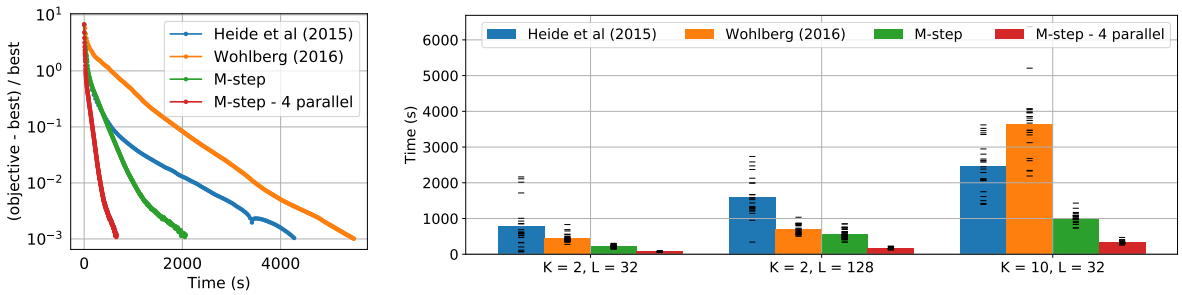
For the benefit of the reader, we summarize the complexity of the M-step in Table 6.1. We note p and q the number of iterations in the L-BFGS-B methods for the activations update and atoms update.

Method	Complexity
Solving activations z	$p \min(L, \log(T))KTN$
Solving atoms d	$L^2K^2TN + qL^3K^3$
Solving atoms d (sequential)	$LKTN + qL^2K$

Table 6.1: Complexity analysis of the M-step, where p and q are the number of iterations in the L-BFGS-B solvers for the activations and atoms updates.

6.4 Experiments

In order to evaluate our approach, we conduct several experiments on both synthetic and real data. First, we show that our proposed optimization scheme for the M-step provides significant improvements in terms of convergence speed over the state-of-the-art CSC methods. Then, we provide empirical evidence that our algorithm is more robust to artifacts and outliers than three competing CSC methods (Jost et al., 2006; Brockmeier and Príncipe, 2016; Wohlberg, 2016). Finally, we consider LFP data, where we illustrate that our algorithm can reveal interesting properties in electrophysiological signals without supervision, even in the presence of severe artifacts. The source code is publicly available at https://alphacsc.github.io/.



(a) $K = 10$, $L = 32$.

(b) Time to reach a relative precision of 0.01.

Figure 6.2: Comparison of state-of-the-art methods with our approach. (a) Convergence plot with the objective function relative to the obtained minimum, as a function of computational time. (b) Time taken to reach a relative precision of 10^{-2}, for different settings of K and L.

Synthetic simulation setup: In our synthetic data experiments, we simulate N trials of length T by first generating K zero mean and unit norm atoms of length L. The activation instants are integers drawn from a uniform distribution in $[\![0, T - L]\!]$. The amplitude of the activations are drawn from a uniform distribution in $[0, 1]$. Atoms are activated only once per trial and are allowed to overlap. The activations are then convolved with the generated atoms and summed up as in (6.1).

6.4.1 M-step performance

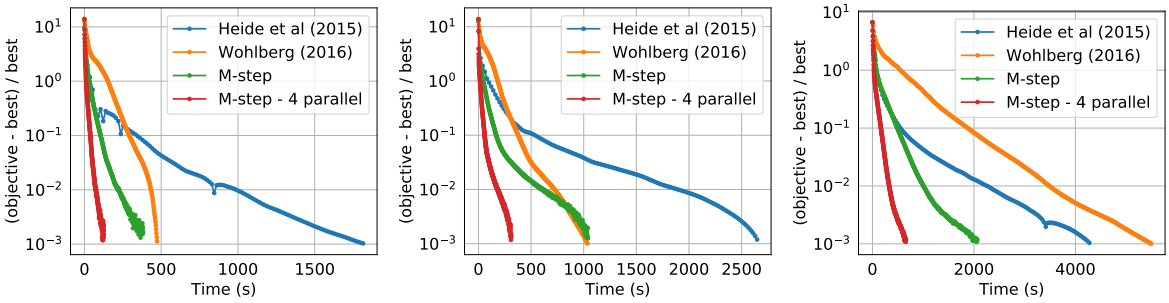
In our first set of synthetic experiments, we illustrate the benefits of our M-step optimization approach over state-of-the-art CSC solvers. We set $N = 100$, $T = 2000$ and $\lambda = 1$, and use different values for K and L. To be comparable, we set $\alpha = 2$ and add Gaussian noise to the synthesized signals, where the standard deviation is set to 0.01. In this setting, we have $w_{n,t} = 1/2$ for all n, t, which reduces the problem to a standard CSC setup. We monitor the convergence of ADMM-based methods by Heide et al. (2015) and Wohlberg (2016) against our M-step algorithm, using both a single-threaded and a parallel version for the z-update. As the problem is non-convex, even if two algorithms start from the same point, they are not guaranteed to reach the same local minimum[2]. Hence, for a fair comparison, we use a multiple restart strategy with averaging across 24 random seeds.

During our experiments we have observed that the ADMM-based methods do not guarantee the feasibility of the iterates. In other words, the norms of the estimated atoms might be greater than 1 during the iterations. To keep the algorithms comparable, when computing the objective value, we project the atoms to the unit ball and scale the activations accordingly. To be strictly comparable, we also imposed a positivity constraint on these algorithms. This is easily done by modifying the soft-thresholding operator to be a rectified linear function. In the benchmarks, all algorithms use a single thread, except "M-step - 4

[2]Note that the M-step can be viewed as a biconvex problem, for which global convergence guarantees can be shown under certain assumptions (Agarwal et al., 2014; Gorski et al., 2007). However, we have observed that it is required to use multiple restarts even for vanilla CSC, implying that these assumptions are not satisfied in this particular problem.

parallel" which uses 4 threads during the z update.

In Fig. 6.2, we illustrate the convergence behaviors of the different methods. Note that the y-axis is the precision relative to the objective value obtained upon convergence. In other words, each curve is relative to its own local minimum (see next paragraph for details). In the right subplot, we show how long it takes for the algorithms to reach a relative precision of 0.01 for different settings (*cf.* next paragraph for more benchmarks). Our method consistently performs better and the difference is even more striking for more challenging setups. This speed improvement on the M-step is crucial for us as this step will be repeatedly executed.



(a) $K = 2$, $L = 32$. (b) $K = 2$, $L = 128$. (c) $K = 10$, $L = 32$.

Figure 6.3: Convergence speed of the relative objective function. The y-axis shows the objective function relative to the obtained minimum for each run: $(f(x) - f(x^*))/f(x^*)$. Each curve is the geometrical mean over 24 different random initializations.

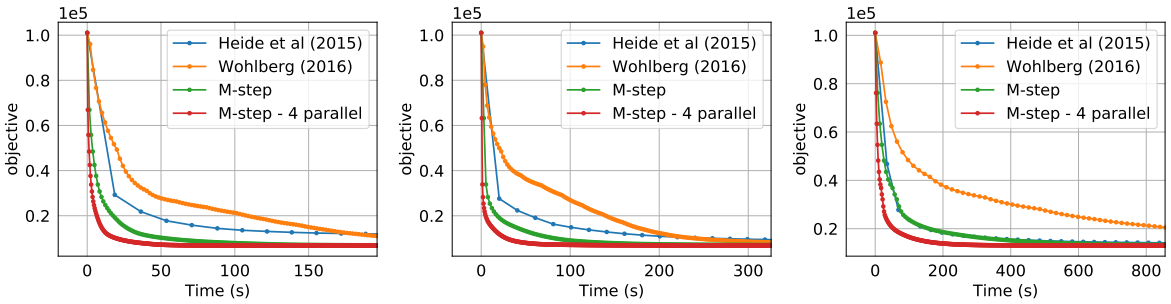
Convergence plots: Here, we compare convergence plots of our algorithm against a number of state-of-art methods. Fig. 6.3 demonstrates on a variety of setups the computational advantage of our quasi-Newton approach to solve the M-step. Note that Fig. 6.2b is in fact a summary of Fig. 6.3. Indeed, we can verify that ADMM methods converge quickly to a modest accuracy, but take much longer to converge to a high accuracy (Boyd et al., 2011).

Next, in Fig. 6.4, we show more traditional convergence plots. In contrast to Fig. 6.2 or 6.3 where the relative objective function is shown, here we plot the absolute value of the objective function. We can now verify that each of the methods have indeed converged to their respective local minimum. Of course, owing to the non-convex nature of the problem, they do not necessarily converge to the same local minimum.

Comparison of solver for the activations subproblem: Finally, we compare convergence plots of our algorithm using different solvers for the z-update: ISTA, FISTA, and L-BFGS-B. The rationale for choosing a quasi-Newton solver for the z-update becomes clear in Fig. 6.5 as the L-BFGS-B solver turns out to be computationally advantageous on a variety of setups.

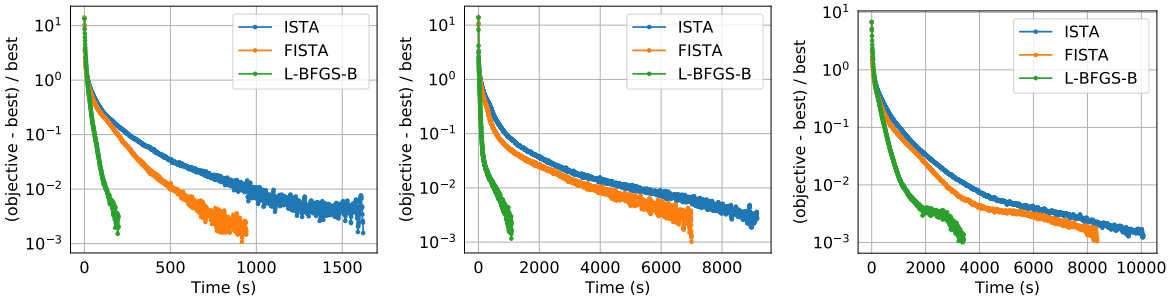
6.4.2 Robustness to corrupted data

In our second synthetic data experiment, we illustrate the robustness of αCSC in the presence of corrupted observations. In order to simulate the likely presence of high amplitude artifacts, one way would be to directly simulate the generative model in (6.3).



(a) $K = 2$, $L = 32$. (b) $K = 2$, $L = 128$. (c) $K = 10$, $L = 32$.

Figure 6.4: Convergence of the objective function as a function of time. The y-axis shows the absolute objective function $f(x)$. Each curve is the mean over 24 different random initializations.



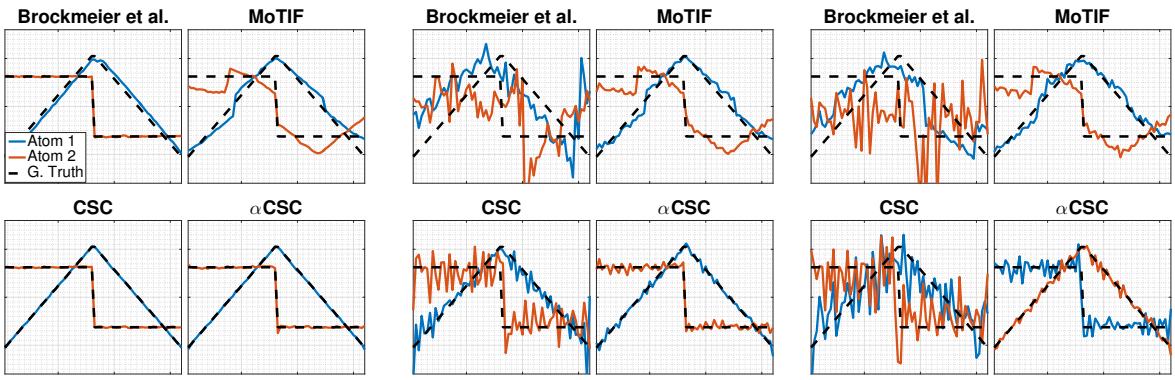
(a) $K = 2$, $L = 32$. (b) $K = 2$, $L = 128$. (c) $K = 10$, $L = 32$.

Figure 6.5: Convergence speed of the relative objective function. The y-axis shows the objective function relative to the obtained minimum for each run: $(f(x) - f(x^*))/f(x^*)$. Each curve is the geometrical mean over 24 different random initializations.

However, this would give us an unfair advantage, since αCSC is specifically designed for such data. Here, we take an alternative approach, where we corrupt a randomly chosen fraction of the trials (10% or 20%) with strong Gaussian noise of standard deviation 0.1, *i.e.* one order of magnitude higher than in a regular trial. We used a regularization parameter of $\lambda = 0.1$. In these experiments, by CSC we refer to αCSC with $\alpha = 2$, that resembles using only the M-step of our algorithm with deterministic weights $w_{n,t} = 1/2$ for all n, t. We used a simpler setup where we set $N = 100$, $T = 512$, and $L = 64$. We used $K = 2$ atoms, as shown in dashed lines in Fig. 6.6.

For αCSC, we set the number of outer iterations to $I = 5$, the number of iterations of the M-step to $M = 50$, and the number of iterations of the MCMC algorithm to $J = 10$. We discard the first 5 samples of the MCMC algorithm as burn-in. To enable a fair comparison, we run the standard CSC algorithm for $I \times M$ iterations, i.e. the *total* number of M-step iterations in αCSC. We also compared αCSC against competing state-of-art methods previously applied to neural time series: Brockmeier and Príncipe (2016) and MoTIF (Jost et al., 2006). Starting from multiple random initializations, the estimated atoms with the smallest ℓ_2 distance with the true atoms are shown in Fig. 6.6.

In the artifact-free scenario, all algorithms perform equally well, except for MoTIF that suffers from the presence of activations with varying amplitudes. This is because it aligns



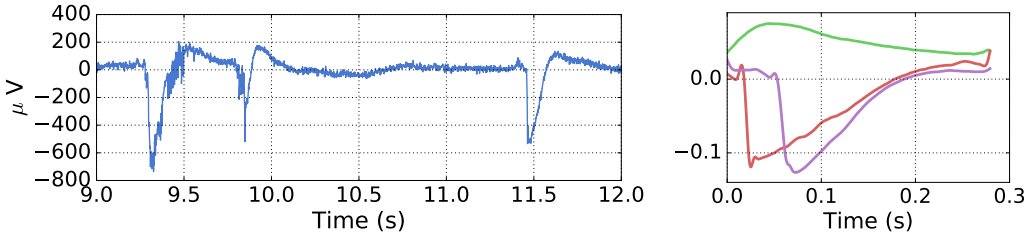
(a) No corruption. (b) 10% corruption. (c) 20% corruption

Figure 6.6: Simulation to compare state-of-the-art methods against αCSC.

the data using correlations before performing the eigenvalue decomposition, without taking into account the strength of activations in each trial. The performance of Brockmeier and Príncipe (2016) and CSC degrades as the level of corruption increases. On the other hand, αCSC is clearly more robust to the increasing level of corruption and recovers reasonable atoms even when 20% of the trials are corrupted.

6.4.3 Results on LFP data

In our last set of experiments, we consider real neural data from two different datasets. We first applied αCSC on an LFP dataset previously used in Hitziger et al. (2017) and containing epileptiform spikes as shown in Fig. 6.7(a). The data was recorded in the rat cortex, and is free of artifact. Therefore, we used the standard CSC with our optimization scheme, (i.e. αCSC with $\alpha = 2$). As a standard preprocessing procedure, we applied a high-pass filter at 1 Hz in order to remove drifts in the signal, and then applied a tapered cosine window to down-weight the samples near the edges. We set $\lambda = 6$, $N = 300$, $T = 2500$, $L = 350$, and $K = 3$. The recovered atoms by our algorithm are shown in Fig. 6.7(b). We can observe that the estimated atoms resemble the spikes in Fig. 6.7(a). These results show that, without using any heuristics, our approach can recover similar atoms to the ones reported in Hitziger et al. (2017), even though it does not make any assumptions on the shapes of the waveforms, or initializes the atoms with template spikes in order to ease the optimization.



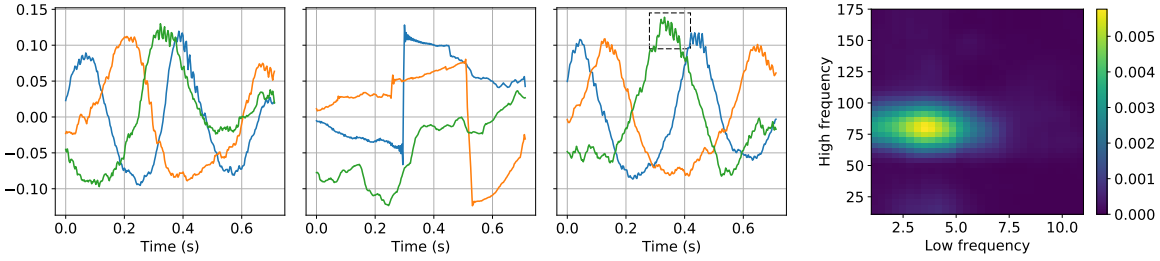
(a) LFP spike data from Hitziger et al. (2017) (b) Estimated atoms

Figure 6.7: Atoms learnt by αCSC on LFP data containing epileptiform spikes with $\alpha = 2$.

The second dataset is an LFP channel in a rodent striatum from Dallérac et al. (2017). We segmented the data into 70 trials of length 2500 samples, windowed each trial with a tapered cosine function, and detrended the data with a high-pass filter at 1 Hz. We set $\lambda = 10$, initialized the weights w_n to the inverse of the variance of the trial x_n. Atoms are in all experiments initialized with Gaussian white noise.

As opposed to the first LFP dataset, this dataset contains strong artifacts, as shown in Fig. 6.1(b). In order to be able to illustrate the potential of CSC on this data, we first *manually* identified and removed the trials that were corrupted by artifacts. In Fig. 6.8(a), we illustrate the estimated atoms with CSC on the manually-cleaned data. We observe that the estimated atoms correspond to canonical waveforms found in the signal. In particular, the high frequency oscillations around 80 Hz are modulated in amplitude by the low-frequency oscillation around 3 Hz, a phenomenon known as cross-frequency coupling (CFC) (Jensen and Colgin, 2007). We can observe this by computing a comodulogram (Tort et al., 2010) on the entire signal (Fig. 6.8(b)). This measures the correlation between the amplitude of the high frequency band and the phase of the low frequency band.

Even though CSC is able to provide these excellent results on the cleaned data set, its performance heavily relies on the manual removal of the artifacts. Finally, we repeated the previous experiment on the full data, without removing the artifacts and compared CSC with αCSC, where we set $\alpha = 1.2$. The results are shown in the middle and the right sub-figures of Fig. 6.8(a). It can be observed that in the presence of strong artifacts, CSC is not able to recover the atoms anymore. On the contrary, we observe that αCSC can still recover atoms as observed in the artifact-free regime. In particular, the cross-frequency coupling phenomenon is still visible.



(a) Atoms learnt by: CSC (clean data), CSC (full data), αCSC (full data)

(b) Comodulogram.

Figure 6.8: (a) Three atoms learnt from a rodent striatal LFP channel, using CSC on cleaned data, and both CSC and αCSC on the full data. The atoms capture the cross-frequency coupling of the data (dashed rectangle). (b) Comodulogram presents the cross-frequency coupling intensity computed between pairs of frequency bands on the entire cleaned signal, following Tort et al. (2010).

6.5 Conclusion

We address the present need in the neuroscience community to better capture the complex morphology of brain waves. Our approach is based on a probabilistic formulation of a CSC model. We propose an inference strategy based on MCEM to deal efficiently with

heavy tailed noise and take into account the polarity of neural activations with a positivity constraint. Our problem formulation allows the use of fast quasi-Newton methods that outperform previously proposed state-of-the-art ADMM-based algorithms, even when not making use of our parallel implementation. Results on LFP data demonstrate that such algorithms can be robust to the presence of transient artifacts in data and reveal insights on neural time-series without supervision.

Chapter 7

Conclusion

"The purpose of computation is insight, not numbers."

—Richard Hamming

Methods research in neuroimaging is a marriage between computer science and neuroscience. It is a collaboration between two complementary disciplines – the aim is to bring to the table computation tools which can help scientists make new discoveries. Certain aspects of this interdisciplinary subfield is of course to incrementally develop existing tools: for example, those that can help achieve a better prediction score, or a better localization accuracy in estimating neural sources. However, an orthogonal but equally important aspect of methods research is to develop tools which allow fundamentally new ways to interact with the data. This thesis is an attempt to advance this goal by developing tools for automated analysis in electrophysiology.

It has now become evident to us that in order to achieve the goal of reproducible research, large public datasets are the key and automated methods to analyze them are indispensable. While every neuroscientist's ambition is to generate new insights and push the frontiers of our knowledge of the brain, this is often not possible due to the weak effect sizes which cannot be uncovered in small datasets. When the null hypothesis cannot be rejected, it is a common practice to start fishing for significant results by testing multiple hypotheses and reporting the most favourable ones. This has resulted in a body of literature where a large fraction of the results lie on shaky grounds.

In this thesis, we developed a new specification known as the Brain Imaging Data Structure (BIDS), which facilitates data sharing between neuroscientists by promoting common standards for storing measurement related metadata. We also provided an overview of the challenges in reproducible data analysis with respect to magnetoencephalography (MEG)/electroencephalography (EEG) data. As contributors to the MNE software package, we felt particularly well positioned to address the software related challenges: complex pipelines, software versions, random initialization *etc.,* and standardized recommendations for each stage of these pipelines. We did this by reanalyzing a group study on Faces dataset (Wakeman and Henson, 2015). To ensure reproducible results, the entire analysis was scripted and the plots generated automatically using the `sphinx_gallery` package[1].

In order to even further push the goal of reproducibility via automation, we developed two

[1] https://sphinx-gallery.github.io

new methods for analyzing electrophysiological data. The first method, called *autoreject*, aims to streamline the removal of data segments containing artifacts which is a basic preprocessing step in almost every analysis chain. We develop an efficient method which uses a parameter search method known as Bayesian optimization. Our approach was able to facilitate re-analysis of the Human Connectome Project (HCP) data for benchmarking. Our second method, known as *alphacsc* enables mining neural time series for new oscillatory structures. Not only that, it is a tool to estimate more accurate waveform shapes than what is possible using traditional Fourier analysis. We demonstrated in our work that it was able to discover nested oscillations from the data.

These technologies can still be considered to be in their infancy in many respects. Just as source localization methods in MEG/EEG have evolved from dipole-based models to distributed methods to more sophisticated models implementing structured sparsity, these new methods are likely to undergo an evolutionary process of incremental improvements. If we consider the example of convolutional sparse coding (CSC), our model based on alpha-stable distributions extended the computer vision models to be able to handle heavy-tailed distributions that is characteristic in neural data. Obviously, this is not the end of the road. Tuning hyperparameters in CSC models is still notoriously difficult, but it is not impossible if there is an supervised task at the end of the pipeline. Multiscale dictionaries might be critical for brain signals considering that the oscillations can have varying support. As the problem is non-convex, smarter initialization strategies such as those based on Markov chain Monte Carlo (MCMC) could lead to more accurate estimates (Bachem et al., 2016). It will also soon be necessary to build streaming CSC algorithms based on stochastic approximations to deal with larger datasets.

Unlike computer vision or natural language processing, high-risk industries such as healthcare require transparent algorithms. It is no longer sufficient to be able to merely achieve higher prediction accuracy. In our work on *autoreject* and *alphacsc*, we leverage such public data to develop algorithms which are easy to interpret and diagnose. *Autoreject* identifies the data segments to be removed based on a single parameter which is easy to understand and is automatically tuned. In the same way, *alphacsc* mines the prototypical waveforms directly so as to replace indirect measures for unearthing phenomena of interest.

In this thesis, I outlined a strategy for reproducible research in the future: public datasets with large sample sizes and automation. However, the focus of some of my work was limited to automation on the scale of single subjects. Even though this does enable us to analyse large datasets, it can be sometimes limiting as it does not allow us to pool data across subjects so as to discover more subtle effects. As we enter an era of fast-paced science, such data-driven tools will become indispensable. While a lot of methods research has been focussed on improving the signal-to-noise ratio in each dataset, this may turn out to be not as important when dealing with larger datasets. Looking ahead, we will increasingly prefer large datasets which are not perfectly denoised rather than a smaller perfectly denoised dataset. New tools will need to be developed in order to enable clinicians to rapidly probe the brain so as to identify signals and structures of interest, quantify uncertainties along with the accuracy scores, perform quality control, and interactively visualize their data.

Bibliography

D. J. Acunzo, G. MacKenzie, and M. C. van Rossum. Systematic biases in early ERP and ERF components as a result of high-pass filtering. *Journal of Neuroscience Methods*, 209(1):212–218, 2012.

A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon. Learning sparsely used overcomplete dictionaries. In *Conference on Learning Theory*, pages 123–137, 2014.

x. Alzheimer's Association et al. 2016 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 12(4):459–509, 2016.

R. H. Baayen, D. J. Davidson, and D. M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412, 2008.

F. Bach, R. Jenatton, J. Mairal, G. Obozinski, et al. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

O. Bachem, M. Lucic, H. Hassani, and A. Krause. Fast and provably good seedings for k-means. In *Advances in Neural Information Processing Systems*, pages 55–63, 2016.

S. Baillet. Magnetoencephalography for brain electrophysiology and imaging. *Nature Neuroscience*, 20(3):327–339, 03 2017a. URL http://dx.doi.org/10.1038/nn.4504.

S. Baillet. Magnetoencephalography for brain electrophysiology and imaging. *Nature neuroscience*, 20(3):327, 2017b.

S. Baillet, K. Friston, and R. Oostenveld. Academic software applications for electromagnetic brain mapping using meg and eeg. *Computational intelligence and neuroscience*, 2011, 2011.

M. Baker. 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604):452, 2016.

A. Barachant, A. Andreev, and M. Congedo. The Riemannian Potato: an automatic and adaptive artifact detection method for online experiments using Riemannian geometry. In *TOBI Workshop lV*, pages 19–20, 2013.

Q. Barthélemy, C. Gouy-Pailler, Y. Isaac, A. Souloumiac, A. Larue, and J. I. Mars. Multivariate temporal dictionary learning for EEG. *Journal of Neuroscience Methods*, 215(1):19–28, 2013.

A. Basirat, S. Dehaene, and G. Dehaene-Lambertz. A hierarchy of cortical responses to sequence violations in three-month-old infants. *Cognition*, 132(2):137–150, 2014.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The arcade learning environment: An evaluation platform for general agents. 2013.

C. M. Bennett. Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction. 2009.

J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, pages 2546–2554, 2011.

N. Bigdely-Shamlo, K. Kreutz-Delgado, K. Robbins, M. Miyakoshi, M. Westerfield, T. Bel-Bahar, C. Kothe, J. Hsi, and S. Makeig. Hierarchical event descriptor (HED) tags for analysis of event-related EEG studies. In *Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–4. IEEE, 2013.

N. Bigdely-Shamlo, T. Mullen, C. Kothe, K.-M. Su, and K. Robbins. The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformtics*, 9, 2015.

N. Bigdely-Shamlo, S. Makeig, and K. A. Robbins. Preparing laboratory and real-world eeg data for large-scale analysis: a containerized approach. *Frontiers in neuroinformatics*, 10, 2016.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

A. J. Brockmeier and J. C. Príncipe. Learning recurrent waveforms within EEGs. *IEEE Transactions on Biomedical Engineering*, 63(1):43–54, 2016.

K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013.

R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

J. Carp. On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments. *Frontiers in neuroscience*, 6:149, Jan. 2012a. ISSN 1662-453X. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3468892&tool=pmcentrez&rendertype=abstract.

J. Carp. The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, 63(1):289 – 300, 2012b. ISSN 1053-8119. URL http://www.sciencedirect.com/science/article/pii/S1053811912007057.

G. C. Cawley and N. L. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11 (Jul):2079–2107, 2010.

J. M. Chambers, C. L. Mallows, and B. W. Stuck. A method for simulating stable random variables. *Journal of the american statistical association*, 71(354):340–344, 1976.

S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

M. X. Cohen. *Analyzing neural time series data: Theory and practice*. MIT Press, 2014. ISBN 9780262319560.

S. R. Cole and B. Voytek. Brain oscillations and the importance of waveform shape. *Trends Cogn. Sci.*, 2017.

O. S. Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

D. Cruse, S. Chennu, C. Chatelle, T. A. Bekinschtein, D. Fernández-Espejo, J. D. Pickard, S. Laureys, and A. M. Owen. Bedside detection of awareness in the vegetative state: a cohort study. *The Lancet*, 378(9809):2088–2094, 2012.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

S. S. Dalal, A. G. Guggisberg, E. Edwards, K. Sekihara, A. M. Findlay, R. T. Canolty, M. S. Berger, R. T. Knight, N. M. Barbaro, H. E. Kirsch, et al. Five-dimensional neuroimaging: localization of the time–frequency dynamics of cortical activity. *NeuroImage*, 40(4): 1686–1700, 2008.

S. S. Dalal, J. M. Zumer, A. G. Guggisberg, M. Trumpis, D. D. E. Wong, K. Sekihara, and S. S. Nagarajan. MEG/EEG source reconstruction, statistical evaluation, and visualization with NUTMEG. *Computational Intelligence and Neuroscience*, 2011, 2011.

A. Dale, B. Fischl, and M. Sereno. Cortical surface-based analysis I: Segmentation and surface reconstruction. *NeuroImage*, 9:179–194, 1999.

A. Dale, A. Liu, B. Fischl, and R. Buckner. Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, 26: 55–67, 2000a.

A. M. Dale, A. K. Liu, B. R. Fischl, R. L. Buckner, J. W. Belliveau, J. D. Lewine, and E. Halgren. Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, 26(1):55–67, 2000b.

A. M. Dale, A. K. Liu, B. R. Fischl, R. L. Buckner, B. J. W., J. D. Lewine, and E. Halgrena. Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, 26(1):55–67, 2000c.

G. Dallérac, M. Graupner, J. Knippenberg, R. C. R. Martinez, T. F. Tavares, L. Tallot, N. El Massioui, A. Verschueren, S. Höhn, J. Bertolus, et al. Updating temporal expectancy of an aversive event engages striatal plasticity under amygdala control. *Nature Communications*, 8:13920, 2017.

J. Dammers, M. Schiek, F. Boers, C. Silex, M. Zvyagintsev, U. Pietrzyk, and K. Mathiak. Integration of amplitude and phase statistics for complete artifact removal in independent components of neuromagnetic recordings. *IEEE Transactions on Biomedical Engineering*, 55(10):2353–2362, 2008.

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on pure and applied mathematics*, 57(11):1413–1457, 2004.

J. De Brabanter, K. Pelckmans, J. Suykens, J. Vandewalle, and B. De Moor. Robust cross-validation score functions with application to weighted least squares support vector machine function estimation. Technical report, K.U. Leuven, 2003.

A. De Cheveigné and J. Simon. Sensor noise suppression. *Journal of Neuroscience Methods*, 168(1):195–202, 2008.

A. Delorme and S. Makeig. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21, 2004.

A. Delorme, T. Mullen, C. Kothe, Z. A. Acar, N. Bigdely-Shamlo, A. Vankov, and S. Makeig. EEGLAB, SIFT, NFT, BCILAB, and ERICA: new tools for advanced EEG processing. *Intell. Neuroscience*, 2011:10:10–10:10, Jan. 2011. ISSN 1687-5265.

A. Delorme, M. Miyakoshi, T.-P. Jung, and S. Makeig. Grand average ERP-image plotting and statistics: A method for comparing variability in event-related single-trial eeg activities across subjects and conditions. *Journal of Neuroscience Methods*, 250 (Supplement C):3 – 6, 2015. ISSN 0165-0270.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

J. Diedrichsen and R. Shadmehr. Detecting and adjusting for artifacts in fMRI time series data. *NeuroImage*, 27(3):624–634, 2005.

P. Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

O. J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

D. Duvenaud, J. Lloyd, R. Grosse, J. Tenenbaum, and Z. Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. *arXiv preprint:1302.4922*, 2013.

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006.

D. Engemann, F. Raimondo, J. King, M. Jas, A. Gramfort, S. Dehaene, L. Naccache, and J. Sitt. Automated measurement and prediction of consciousness in vegetative state and minimally conscious patients. In *Workshop on Statistics, Machine Learning and Neuroscience at the International Conference on Machine Learning (ICML)*, Lille, July 2015a.

D. Engemann, F. Raimondo, J.-R. King, M. Jas, A. Gramfort, S. Dehaene, L. Naccache, and J. Sitt. Automated measurement and prediction of consciousness in vegetative and minimally conscious patients. In *ICML Workshop on Statistics, Machine Learning and Neuroscience (Stamlins 2015)*, 2015b.

D. Engemann, D. Strohmeier, E. Larson, and A. Gramfort. Mind the noise covariance when localizing brain sources with m/eeg. In *Pattern Recognition in NeuroImaging (PRNI), 2015 International Workshop on*, pages 9–12. IEEE, 2015c.

D. A. Engemann and A. Gramfort. Automated model selection in covariance estimation and spatial whitening of MEG and EEG signals. *NeuroImage*, 108:328–342, 2015.

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, et al. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79, 2010.

B. Fischl, M. Sereno, and A. Dale. Cortical surface-based analysis II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9:195–207, 1999.

M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

A. Gelman. Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*, 48(3):432–435, 2006.

C. R. Genovese, N. A. Lazar, and T. Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878, 2002.

F. Gibson, P. G. Overton, T. V. Smulders, S. R. Schultz, S. J. Eglen, C. D. Ingram, S. Panzeri, P. Bream, M. Whittington, E. Sernagor, et al. Minimum information about a neuroscience investigation (mini): electrophysiology. 2009.

B. Gips, A. Bahramisharif, E. Lowet, M. Roberts, P. de Weerd, O. Jensen, and J. van der Eerden. Discovering recurring patterns in electrophysiological recordings. *Journal of Neuroscience Methods*, 275:66–79, 2017.

T. Glatard, L. B. Lewis, R. Ferreira da Silva, R. Adalat, N. Beck, C. Lepage, P. Rioux, M.-E. Rousseau, T. Sherif, E. Deelman, N. Khalili-Mahani, and A. C. Evans. Reproducibility of neuroimaging analyses across operating systems. *Frontiers in Neuroinformatics*, 9:12, 2015. ISSN 1662-5196. doi: 10.3389/fninf.2015.00012. URL https://www.frontiersin.org/article/10.3389/fninf.2015.00012.

S. Godsill and E. Kuruoglu. Bayesian inference for time series with heavy-tailed symmetric α-stable noise processes. *Proc. Applications of heavy tailed distributions in economics, eng. and stat.*, 1999.

A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. Ivanov, et al. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

A. M. Goldfine, J. C. Bardin, Q. Noirhomme, J. J. Fins, N. D. Schiff, and J. D. Victor. Reanalysis of "bedside detection of awareness in the vegetative state: a cohort study.". *Lancet*, 381(9863):289, 2013.

K. J. Gorgolewski and R. A. Poldrack. A practical guide for improving transparency and reproducibility in neuroimaging research. *PLoS biology*, 14(7):e1002506, 2016.

K. J. Gorgolewski, G. Varoquaux, G. Rivera, Y. Schwarz, S. S. Ghosh, C. Maumet, V. V. Sochat, T. E. Nichols, R. A. Poldrack, J.-B. Poline, et al. Neurovault. org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in neuroinformatics*, 9, 2015.

K. J. Gorgolewski, T. Auer, V. D. Calhoun, R. C. Craddock, S. Das, E. P. Duff, G. Flandin, S. S. Ghosh, T. Glatard, Y. O. Halchenko, et al. The Brain Imaging Data Structure: a standard for organizing and describing outputs of neuroimaging experiments. *bioRxiv*, page 034561, 2016.

J. Gorski, F. Pfeuffer, and K. Klamroth. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research*, 66 (3):373–407, 2007.

T. Götz, T. Milde, G. Curio, S. Debener, T. Lehmann, L. Leistritz, O. Witte, H. Witte, and J. Haueisen. Primary somatosensory contextual modulation is encoded by oscillation frequency change. *Clinical Neurophysiology*, 126(9):1769 – 1779, 2015. ISSN 1388-2457.

A. Gramfort, R. Keriven, and M. Clerc. Graph-based variability estimation in single-trial event-related neural responses. *IEEE Transactions on Biomedical Engineering*, 57(5): 1051–1061, 2010.

A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, et al. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7, 2013a.

A. Gramfort, D. Strohmeier, J. Haueisen, M. Hämäläinen, and M. Kowalski. Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations. *NeuroImage*, 70(0):410 – 422, 2013b. ISSN 1053-8119.

A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, et al. MNE software for processing MEG and EEG data. *NeuroImage*, 86(0):446 – 460, 2014. ISSN 1053-8119. doi: http://dx.doi.org/10.1016/j.neuroimage.2013.10.027.

J. Grewe, T. Wachtler, and J. Benda. A bottom-up approach to data annotation in neurophysiology. *Frontiers in neuroinformatics*, 5, 2011.

K. Grill-Spector, N. Knouf, and N. Kanwisher. The fusiform face area subserves face perception, not generic within-category identification. *Nature neuroscience*, 7(5):555–562, 2004.

K. Grill-Spector, K. S. Weiner, K. Kay, and J. Gomez. The functional neuroanatomy of human face perception. *Annual Review of Vision Science*, 3(1), 2017.

E. H. Gronenschild, P. Habets, H. I. Jacobs, R. Mengelers, N. Rozendaal, J. Van Os, and M. Marcelis. The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. *PloS one*, 7(6):e38234, 2012.

J. Gross, J. Kujala, M. Hämäläinen, and L. Timmermann. Dynamic imaging of coherent sources: studying neural interactions in the human brain. *Proceedings of the National Academy of Sciences*, 98(2):694–699, Jan 2001.

J. Groß, J. Kujala, M. Hämäläinen, L. Timmermann, A. Schnitzler, and R. Salmelin. Dynamic imaging of coherent sources: studying neural interactions in the human brain. *Proceedings of the National Academy of Sciences*, 98(2):694–699, 2001.

J. Gross, S. Baillet, G. Barnes, R. Henson, A. Hillebrand, O. Jensen, K. Jerbi, V. Litvak, B. Maess, R. Oostenveld, L. Parkkonen, J. Taylor, V. van Wassenhove, M. Wibral, and J. Schoffelen. Good practice for conducting and reporting MEG research. *NeuroImage*, 65(15):349–363, January 2013a.

J. Gross, S. Baillet, G. R. Barnes, R. N. Henson, A. Hillebrand, O. Jensen, K. Jerbi, V. Litvak, B. Maess, R. Oostenveld, et al. Good practice for conducting and reporting meg research. *Neuroimage*, 65:349–363, 2013b.

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'07, pages 149–158. AUAI Press, 2007. ISBN 0-9749039-3-0.

R. Grosse, R. Raina, H. Kwong, and A. Y. Ng. Shift-Invariant Sparse Coding for Audio Classification. *arXiv preprint arXiv:1206.5241*, 2012.

A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

M. Hämäläinen and R. Ilmoniemi. Interpreting magnetic fields of the brain: minimum norm estimates. Technical Report TKK-F-A559, Helsinki University of Technology, 1984.

M. Hämäläinen and R. Ilmoniemi. Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & Biological Engineering & Computing*, 32(1):35–42, 1994.

R. Hari and A. Puce. *MEG-EEG Primer*. Oxford University Press, 2017.

S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87:96–110, 2014.

O. Hauk, D. G. Wakeman, and R. Henson. Comparison of noise-normalized minimum norm estimates for MEG analysis using multiple resolution metrics. *Neuroimage*, 54(3): 1966–1974, 2011.

E. C. Hayden. The automated lab. *Nature*, 516(7529):131, 2014.

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.

S. Hitziger, M. Clerc, S. Saillet, C. Benar, and T. Papadopoulo. Adaptive Waveform Learning: A Framework for Modeling Variability in Neurophysiological Signals. *IEEE Transactions on Signal Processing*, 2017.

C.-W. Hsu, C.-C. Chang, C.-J. Lin, et al. A practical guide to support vector classification. 2003.

D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

P. J. Huber. *Robust Statistics*. Wiley, 1981.

A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634, 1999.

E. C. Ifeachor and B. W. Jervis. *Digital signal processing: a practical approach*. Pearson Education, 2002.

J. P. Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005a.

J. P. A. Ioannidis. Why most published research findings are false. *PLOS Medicine*, 2 (8), 08 2005b. doi: 10.1371/journal.pmed.0020124. URL https://doi.org/10.1371/journal.pmed.0020124.

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

L. Jardine. A Point of View: Crowd-sourcing comets. *BBC News*, 2013. URL http://www.bbc.com/news/magazine-21802843.

M. Jas, D. Engemann, F. Raimondo, Y. Bekhti, and A. Gramfort. Automated rejection and repair of bad trials in MEG/EEG. In *6th International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 2016.

M. Jas, D. A. Engemann, Y. Bekhti, F. Raimondo, and A. Gramfort. Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, 159:417–429, 2017a.

M. Jas, E. Larson, D. A. Engemann, J. Leppakangas, S. Taulu, M. Hamalainen, and A. Gramfort. MEG/EEG group study with MNE: recommendations, quality assessments and best practices. *bioRxiv*, 2017b. doi: 10.1101/240044.

M. Jas, L. Tour, T. Dupré, U. Şimşekli, and A. Gramfort. Learning the morphology of brain signals using alpha-stable convolutional sparse coding. In *Advances in Neural Information Processing Systems (NIPS)*, 2017c.

O. Jensen and L. L. Colgin. Cross-frequency coupling between neuronal oscillations. *Trends in cognitive sciences*, 11(7):267–269, 2007.

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, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 5, pages V–V. IEEE, 2006.

T.-P. Jung, C. Humphries, T.-W. Lee, S. Makeig, M. J. McKeown, V. Iragui, and T. J. Sejnowski. Extended ICA removes artifacts from electroencephalographic recordings. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 894–900. MIT Press, 1998.

M. Junghöfer, T. Elbert, D. M. Tucker, and B. Rockstroh. Statistical control of artifacts in dense array EEG/MEG studies. *Psychophysiology*, 37(04):523–532, 2000.

K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, and Y. Cun. Learning convolutional feature hierarchies for visual recognition. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1090–1098, 2010.

S. Khan and D. Cohen. Note: Magnetic noise from the inner wall of a magnetically shielded room. *Review of Scientific Instruments*, 84(5):056101, 2013.

J. King and S. Dehaene. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences*, 18(4):203–210, 2014.

N. Kriegeskorte, M. Mur, and P. Bandettini. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2, 2008.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

E. E. Kuruoglu. *Signal processing in α-stable noise environments: a least Lp-norm approach*. PhD thesis, University of Cambridge, 1999.

E. Larson and A. K. C. Lee. The cortical dynamics underlying effective switching of auditory spatial attention. *NeuroImage*, 64(Supplement C):365 – 370, 2013. ISSN 1053-8119.

L. J. Larson-Prior, R. Oostenveld, S. Della Penna, G. Michalareas, F. Prior, A. Babajani-Feremi, J.-M. Schoffelen, L. Marzetti, F. de Pasquale, F. Di Pompeo, et al. Adding dynamics to the Human Connectome Project with MEG. *NeuroImage*, 80:190–201, 2013.

S. Leglaive, U. Şimşekli, A. Liutkus, R. Badeau, and G. Richard. Alpha-stable multichannel audio source separation. In *Acoustics, Speech and Signal Processing, ICASSP*, pages 576–580, 2017.

D. Leung. Cross-validation in nonparametric regression with outliers. *Annals of Statistics*, pages 2291–2310, 2005.

M. S. Lewicki and T. J. Sejnowski. Coding time-varying signals using sparse, shift-invariant representations. 1999.

F.-H. Lin, T. Witzel, S. P. Ahlfors, S. M. Stufflebeam, J. W. Belliveau, and M. S. Hämäläinen. Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates. *Neuroimage*, 31(1):160–171, 2006.

V. Litvak, J. Mattout, S. Kiebel, C. Phillips, R. Henson, J. Kilner, G. Barnes, R. Oostenveld, J. Daunizeau, G. Flandin, et al. EEG and MEG data analysis in SPM8. *Computational intelligence and neuroscience*, 2011, 2011.

J. Liu. *Monte Carlo strategies in scientific computing*. Springer, 2008.

D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

B. Maess, E. Schröger, and A. Widmann. High-pass filters and baseline correction in M/EEG analysis-continued discussion. *Journal of neuroscience methods*, 266:171, 2016.

B. Mailhé, S. Lesage, R. Gribonval, F. Bimbot, and P. Vandergheynst. Shift-invariant dictionary learning for sparse representations: extending K-SVD. In *16th Eur. Signal Process. Conf.*, pages 1–5. IEEE, 2008.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual International Conference on Machine Learning (ICML)*, pages 689–696. ACM, 2009a.

J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach. Supervised dictionary learning. In *Advances in neural information processing systems*, pages 1033–1040, 2009b.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(Jan):19–60, 2010.

S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

B. B. Mandelbrot. *Fractals and scaling in finance: Discontinuity, concentration, risk. Selecta volume E*. Springer Science & Business Media, 2013.

S. Marĉelja. Mathematical description of the responses of simple cortical cells. *JOSA*, 70 (11):1297–1300, 1980.

E. Maris and R. Oostenveld. Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1):177 – 190, 2007. ISSN 0165-0270.

A. Mazaheri and O. Jensen. Asymmetric amplitude modulations of brain oscillations generate slow evoked responses. *The Journal of Neuroscience*, 28(31):7781–7787, 2008.

J. Mosher, P. Lewis, and R. Leahy. Multiple dipole modeling and localization from spatio-temporal MEG data. *IEEE Transactions on Biomedical Engineering*, 39(6): 541–553, 1992.

J. Mosher, R. Leahy, and P. Lewis. EEG and MEG: Forward solutions for inverse methods. *IEEE Transactions on Biomedical Engineering*, 46(3):245–259, 1999.

E. Moulines, P. Duhamel, J.-F. Cardoso, and S. Mayrargue. Subspace methods for the blind identification of multichannel FIR filters. *IEEE Transactions on signal processing*, 43(2):516–525, 1995.

Y. Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2). volume 27, pages 372–376, 1983.

G. Niso, J. Moreau, E. Bock, F. Tadel, R. Oostenveld, J. Schoffelen, A. Gramfort, V. Litvak, K. Gorgolewski, and S. Baillet. An meg extension to bids: Brain imaging data structure - a solution to organize, describe and share neuroimaging data. In *20th International conference on biomagnetism (BIOMAG 2016)*, 2016a.

G. Niso, C. Rogers, J. T. Moreau, L.-Y. Chen, C. Madjar, S. Das, E. Bock, F. Tadel, A. C. Evans, P. Jolicoeur, et al. OMEGA: the open MEG archive. *NeuroImage*, 124: 1182–1187, 2016b.

G. Niso, K. J. Gorgolewski, E. Bock, T. L. Brooks, G. Flandin, A. Gramfort, R. N. Henson, M. Jas, V. Litvak, J. T. Moreau, et al. MEG-BIDS, the brain imaging data structure extended to magnetoencephalography. *Scientific data*, 5:180110, 2018.

H. Nolan, R. Whelan, and R. Reilly. FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection. *Journal of Neuroscience Methods*, 192(1):152–162, 2010.

A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

W. Ollier, T. Sprosen, and T. Peakman. UK Biobank: from concept to reality. 2005.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607, 1996.

R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011, 2011.

M. Pachitariu, A. M. Packer, N. Pettit, H. Dalgleish, M. Hausser, and M. Sahani. Extracting regions of interest from biological images with convolutional sparse block coding. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1745–1753, 2013.

N. Parikh, S. Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

L. Parkkonen. Instrumentation and data preprocessing. In P. Hansen, M. Kringelbach, and R. Salmelin, editors, *MEG: an introduction to methods*. Oxford University Press, New York, 2010.

T. W. Parks and C. S. Burrus. *Digital filter design*. Wiley-Interscience, 1987.

R. D. Pascual-Marqui et al. Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. *Methods Find Exp Clin Pharmacol*, 24(Suppl D):5–12, 2002.

D. B. Paul and J. M. Baker. The design for the wall street journal-based csr corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 357–362. Association for Computational Linguistics, 1992.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

D. B. Percival and A. T. Walden. *Spectral analysis for physical applications*. Cambridge University Press, 1993.

F. Perrin, J. Pernier, O. Bertrand, and J. Echallier. Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, 72 (2):184–187, 1989.

R. A. Poldrack and K. J. Gorgolewski. OpenfMRI: open sharing of task fMRI data. *NeuroImage*, 144:259–261, 2017.

R. A. Poldrack, D. M. Barch, J. P. Mitchell, T. D. Wager, A. D. Wagner, J. T. Devlin, C. Cumba, O. Koyejo, and M. P. Milham. Toward open sharing of task-based fmri data: the openfmri project. *Frontiers in neuroinformatics*, 7, 2013.

P. Ramkumar, M. Jas, S. Pannasch, R. Hari, and L. Parkkonen. Feature-specific information processing precedes concerted activation in human visual cortex. *Journal of Neuroscience*, 33(18):7691–7699, 2013.

C. E. Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

C. E. Rasmussen and C. K. Williams. Gaussian processes for machine learning. 2006. *The MIT Press, Cambridge, MA, USA*, 38:715–719, 2006.

B. Rivet, A. Souloumiac, V. Attina, and G. Gibert. xDAWN algorithm to enhance evoked potentials: application to brain–computer interface. *IEEE Transactions on Biomedical Engineering*, 56(8):2035–2043, 2009.

R. Rosenthal. The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3):638, 1979.

G. A. Rousselet. Does filtering preclude us from studying ERP time-courses? *Frontiers in psychology*, 3:131, 2012.

G. Samorodnitsky and M. S. Taqqu. *Stable non-Gaussian random processes: stochastic models with infinite variance*, volume 1. CRC press, 1994.

G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw. BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004.

M. Scherg and D. Von Cramon. Two bilateral sources of the late AEP as identified by a spatio-temporal dipole model. *Electroencephalography and Clinical Neurophysiology*, 62 (1):32–44, Jan. 1985. ISSN 0013-4694.

A. Schurger, S. Marti, and S. Dehaene. Reducing multi-sensor data to a single time course that reveals experimental effects. *BMC neuroscience*, 14(1):122, 2013.

K. Sekihara, M. Sahani, and S. S. Nagarajan. Localization bias and spatial resolution of adaptive and non-adaptive spatial filters for MEG source reconstruction. *NeuroImage*, 25(4):1056–67, May 2005. ISSN 1053-8119. URL http://www.ncbi.nlm.nih.gov/pubmed/15850724.

J. P. Simmons, L. D. Nelson, and U. Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366, 2011.

U. Şimşekli, A. Liutkus, and A. T. Cemgil. Alpha-stable matrix factorization. *IEEE Signal Processing Letters*, 22(12):2289–2293, 2015.

D. Slepian. Prolate spheroidal wave functions, Fourier analysis, and uncertainty: The discrete case. *Bell System Technical Journal*, 57(5):1371–1430, 1978.

N. Smith and M. Kutas. Regression-based estimation of ERP waveforms: I. the rERP framework. *Psychophysiology*, 52(2):157–168, 2015a.

N. Smith and M. Kutas. Regression-based estimation of ERP waveforms: Ii. nonlinear effects, overlap correction, and practical considerations. *Psychophysiology*, 52(2):169–181, 2015b.

S. Smith and T. Nichols. Statistical Challenges in "Big Data" Human Neuroimaging. *Neuron*, 97(2):263–268, 2017.

S. M. Smith and T. E. Nichols. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, 44(1):83–98, 2009.

J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2951–2959, 2012.

M. Šorel and F. Šroubek. Fast convolutional sparse coding using matrix inversion lemma. *Digital Signal Processing*, 2016.

A. Stoewer, J. Benda, C. Garbers, C. J. Kellner, A. Sobolev, T. Wachtler, and J. Grewe. Single-file solution for storing neuroscience data and metadata. *Frontiers in Neuroinformatics*, (77), 2013. ISSN 1662-5196. doi: 10.3389/conf.fninf.2013.09.00077. URL http://www.frontiersin.org/neuroinformatics/10.3389/conf.fninf.2013.09.00077/full.

D. Szucs and J. Ioannidis. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol*, 15(3):e2000797, 2017.

F. Tadel, S. Baillet, J. C. Mosher, D. Pantazis, and R. M. Leahy. Brainstorm: a user-friendly application for MEG/EEG analysis. *Computational intelligence and neuroscience*, 2011:8, 2011.

D. Tanner, K. Morgan-Short, and S. J. Luck. How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology*, 52(8):997–1009, 2015.

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, 2006.

S. Taulu, M. Kajola, and J. Simola. Suppression of interference and artifacts by the signal space separation method. *Brain Topography*, 16(4):269–275, 2004.

J. R. Taylor, N. Williams, R. Cusack, T. Auer, M. A. Shafto, M. Dixon, L. K. Tyler, R. 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*, 2015.

J. L. Teeters, K. Godfrey, R. Young, C. Dang, C. Friedsam, B. Wark, H. Asari, S. Peron, N. Li, A. Peyrache, et al. Neurodata without borders: creating a common data format for neurophysiology. *Neuron*, 88(4):629–634, 2015.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

A. B. Tort, R. Komorowski, H. Eichenbaum, and N. Kopell. Measuring phase-amplitude coupling between neuronal oscillations of different frequencies. *Journal of Neurophysiology*, 104(2):1195–1210, 2010.

M. Uusitalo and R. Ilmoniemi. Signal-space projection method for separating MEG or EEG into components. *Medical & Biological Engineering & Computing*, 35(2):135–140, 1997.

K. Uutela, M. Hämäläinen, and E. Somersalo. Visualization of magnetoencephalographic data using minimum current estimates. *NeuroImage*, 10(2):173–180, 1999.

D. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. Curtiss, et al. The Human Connectome Project: a data acquisition perspective. *NeuroImage*, 62(4):2222–2231, 2012.

D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium, et al. The WU-Minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

B. Van Veen, W. van Drongelen, M. Yuchtman, and A. Suzuki. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Transactions on Biomedical Engineering*, 44(9):867–80, Sept. 1997. ISSN 0018-9294. URL http://www.ncbi.nlm.nih.gov/pubmed/9282479.

G. Varoquaux, P. R. Raamana, D. A. Engemann, A. Hoyos-Idrobo, Y. Schwartz, and B. Thirion. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*, 145:166–179, 2017.

R. Vigário. Extraction of ocular artefacts from EEG using independent component analysis. *Electroencephalography and Clinical Neurophysiology* , 103(3):395–404, 1997.

R. Vigário, J. Sarela, V. Jousmiki, M. Hamalainen, and E. Oja. Independent component approach to the analysis of EEG and MEG recordings. *IEEE Transactions on Biomedical Engineering*, 47(5):589–593, 2000.

F. C. Viola, J. Thorne, B. Edmonds, T. Schneider, T. Eichele, and S. Debener. Semi-automatic identification of independent components representing EEG artifact. *Clinical Neurophysiology*, 120(5):868–877, 2009.

D. Wakeman and R. Henson. A multi-subject, multi-modal human neuroimaging dataset. *Scientific Data*, 2, 2015.

Y. Wang, Y. Qi, Y. Wang, Z. Lei, X. Zheng, and G. Pan. Delving into α-stable distribution in noise suppression for seizure detection from scalp EEG. *Journal of Neural Engineering*, 13(5):056009, 2016.

C. J. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

P. Welch. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, Jun 1967. ISSN 0018-9278. doi: 10.1109/TAU.1967.1161901.

A. Widmann and E. Schröger. Filter effects and filter artifacts in the analysis of electrophysiological data. *Frontiers in psychology*, 3:233, 2012.

A. Widmann, E. Schröger, and B. Maess. Digital filter design for electrophysiological data–a practical approach. *Journal of Neuroscience Methods*, 250:34–46, 2015.

B. Wohlberg. Efficient algorithms for convolutional sparse representations. *Image Processing, IEEE Transactions on*, 25(1):301–315, 2016.

M. Woolrich, L. Hunt, A. Groves, and G. Barnes. MEG beamforming using Bayesian PCA for adaptive data covariance matrix regularization. *NeuroImage*, 57(4):1466–79, Aug. 2011a. ISSN 1095-9572. URL http://www.ncbi.nlm.nih.gov/pubmed/21620977.

M. Woolrich, L. Hunt, A. Groves, and G. Barnes. MEG beamforming using Bayesian PCA for adaptive data covariance matrix regularization. *NeuroImage*, 57(4):1466–1479, 2011b.

S. Wright and J. Nocedal. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.

T. Yarkoni, R. A. Poldrack, T. E. Nichols, D. C. Van Essen, and T. D. Wager. Large-scale automated synthesis of human functional neuroimaging data. *Nature methods*, 8(8):665, 2011.

A. Yendiki, K. Koldewyn, S. Kakunoori, N. Kanwisher, and B. Fischl. Spurious group differences due to head motion in a diffusion MRI study. *NeuroImage*, 88:79–90, 2014.

M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. Deconvolutional networks. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2528–2535. IEEE, 2010.

Z. g. Zheng and Y. Yang. Cross-validation and median criterion. *Statistica Sinica*, pages 907–921, 1998.