



HAL
open science

Domain specific convolutional neural networks for dMRI and M/EEG signal analysis

Sara Sedlar

► **To cite this version:**

Sara Sedlar. Domain specific convolutional neural networks for dMRI and M/EEG signal analysis. Computer Science [cs]. Université Côte d'Azur, 2022. English. NNT : . tel-03946862v1

HAL Id: tel-03946862

<https://theses.hal.science/tel-03946862v1>

Submitted on 19 Jan 2023 (v1), last revised 30 Mar 2023 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



THÈSE DE DOCTORAT

Réseaux de neurones convolutifs adaptés au domaine pour l'analyse des signaux IRMd et M/EEG

Sara SEDLAR

*Centre Inria d'Université Côte d'Azur
Équipe-Projet Athena*

Présentée en vue de l'obtention du grade de docteur en Automatique, traitement du signal et des images d'Université Côte d'Azur, dirigée par Théodore Papadopoulo et co-encadrée par Samuel Deslauriers-Gauthier.
Date de soutenance : 22 Décembre 2022.

Devant le jury composé de :

Directeur de thèse : Théodore Papadopoulo

Directeur de recherche

Centre Inria d'Université Côte d'Azur

Co-encadrant : Samuel Deslauriers-Gauthier

Chargé de recherche

Centre Inria d'Université Côte d'Azur

Rapporteurs: Baba C. Vemuri

Professeur distingué d'université

University of Florida

Michael Tangermann

Professeur associé

Donders Institute, Radboud University

Examineurs: Rachid Deriche

Directeur de recherche

Centre Inria d'Université Côte d'Azur

Guido Gerig

Professeur d'institut

Tandon School of Engineering, New York University

Christian-George Bénar

Directeur de recherche

INSERM, Aix-Marseille Université

Abstract

The analysis of neuroimaging data is essential for the interpretation of the functional or structural characteristics of the human brain. New machine learning algorithms usually require a high amount of data often infeasible to acquire in clinical and practical conditions. This requirement is a consequence of significant data variability arising from numerous factors (various recording procedures, subjects and sessions, presence of high levels of noise). To address this problem, in this thesis, we have investigated and proposed convolutional machine learning models adapted to the properties and well grounded assumptions about the acquired data. Therefore, the models are endowed with valuable knowledge and consequently more efficiently learn to perform certain inferences. In particular, we have studied models for the analysis of non-invasive and in-vivo structural and functional neuroimaging data, namely diffusion Magnetic Resonance Imaging (dMRI) and magneto- and electroencephalography (M/EEG) signals.

Diffusion MRI is a nuclear imaging modality which captures micro-structural properties of the examined tissue. As q-space sampling has been the most widely used high angular resolution diffusion imaging protocol (HARDI) over the last decade, we have studied spherical rotation equivariant convolutional neural networks (CNNs) for dMRI local modeling. As a first contribution, we have proposed a spherical U-net for the estimation of fiber orientation distribution functions (fODFs) with convolutions and non-linearities realized in the spectral and signal domains, respectively. To avoid aliasing, our second contribution proposes a Fourier domain CNN for micro-structure parameter estimation, where non-linearities are defined in the spectral domain.

M/EEG are functional imaging techniques which measure magnetic field strength and electric field potential caused by neural electric activities in the cerebral cortex. Measured signals can be explained by Maxwell's equations with quasi-static approximations. Consequently, we can assume that cortical brain activities spread instantaneously and linearly over the measuring sensors, thus a multivariate M/EEG signal can be represented as a sum of rank-1 multivariate signals corresponding to individual sources in the cortex and noise. Considering this assumption, the second part of the thesis firstly investigates an M/EEG spatial and temporal dictionary learning approach with an L_0 constraint. A second contribution is a CNN classifier with rank-1 spatio-temporal kernels regularized in the spectral domain, where the spatial components of the kernels are represented in terms of spherical harmonics basis, while the temporal components are represented in terms of discrete cosine basis.

Keywords: dMRI local modeling, rotation equivariant CNNs, rank-1 CNN classifier, M/EEG spatio-temporal pattern learning

Résumé

L'analyse des données de neuroimagerie est essentielle pour l'interprétation des caractéristiques fonctionnelles ou structurelles du cerveau humain. Les algorithmes d'apprentissage automatique récents requièrent généralement une grande quantité de données souvent impossibles à acquérir dans des conditions cliniques et pratiques. Une telle exigence est une conséquence de la variabilité importante des données résultant de nombreux facteurs (différentes procédures d'enregistrement, sujets et sessions, présence de niveaux élevés de bruit). Pour résoudre ce problème, dans cette thèse, nous avons étudié et proposé des modèles convolutifs d'apprentissage automatique adaptés aux propriétés et aux hypothèses bien fondées sur les données acquises. Par conséquent, les modèles sont dotés de connaissances précieuses et apprennent plus efficacement à effectuer certaines inférences. En particulier, nous avons étudié des modèles d'analyse des données de neuroimagerie structurelle et fonctionnelle non-invasives et in-vivo pour de l'imagerie par résonance magnétique de diffusion (IRMd) et des signaux de magnéto et d'électro-encéphalographie (M/EEG).

L'IRM de diffusion est une modalité d'imagerie nucléaire qui capture les propriétés microstructurales des tissus examinés. Comme l'échantillonnage de q-space est le protocole d'imagerie de diffusion à haute résolution angulaire (HARDI) le plus largement utilisé au cours de la dernière décennie, nous avons étudié les réseaux de neurones convolutionnels (CNN) sphériques équivariants par rotation pour la modélisation locale de l'IRMd. Comme première contribution, nous avons proposé un U-net sphérique pour l'estimation des fonctions de distribution d'orientation des fibres (fODF) avec des convolutions et des non-linéarités réalisées respectivement dans les domaines spectral et signal. Pour éviter l'aliasing, la deuxième contribution propose un CNN travaillant entièrement dans le domaine spectral – y compris pour les non-linéarités – pour l'estimation des paramètres de microstructure.

La M/EEG est une technique d'imagerie fonctionnelle qui mesure l'intensité du champ magnétique et le potentiel du champ électrique provoqués par les activités électriques neurales dans le cortex cérébral. Les signaux mesurés peuvent être expliqués par les équations de Maxwell avec des approximations quasi-statiques. Par conséquent, nous pouvons supposer que les activités cérébrales corticales se propagent instantanément et linéairement sur les capteurs de mesure, ainsi un signal M/EEG multivarié peut être représenté comme une somme de signaux multivariés de rang 1 correspondant à des sources individuelles dans le cortex et le bruit. Partant de cette hypothèse, la deuxième partie de la thèse étudie une approche d'apprentissage de dictionnaire spatio-temporel M/EEG sous contrainte L_0 . Une deuxième contribution dans cette partie est un classificateur CNN à noyaux spatio-temporels de rang 1 régularisés dans le domaine spectral, où les composantes spatiales et temporelles des noyaux sont représentées respectivement en termes d'éléments de base d'harmoniques sphériques et de base de cosinus discrets.

Mots clés : modélisation locale d'IRMd, CNN équivariant par rotation, classifieur rang-1 CNN, apprentissage spatio-temporel M/EEG

Acknowledgments

Funding

This work was supported by the ERC under the European Union's Horizon 2020 research and innovation program (ERC Advanced Grant agreement No 694665 CoBCoM: Computational Brain Connectivity Mapping).



Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

Contents

List of Acronyms	ix
1 Introduction	1
2 Background	5
2.1 Human brain structure and function	6
2.1.1 Structure and function of neurons	7
2.1.2 Gray matter	8
2.1.3 White matter	10
2.2 Structural and functional brain imaging techniques	14
2.2.1 Diffusion MRI	15
2.2.2 Magneto and electro encephalography	26
2.3 Conclusion	32
3 Diffusion MRI local analysis	35
3.1 dMRI acquired on spheres	36
3.2 dMRI probability density functions	39
3.3 dMRI multi-compartment micro-structure imaging	42
3.4 Deep learning models for spherical signals	44
3.5 Deep learning models in dMRI local modeling	49
3.6 Conclusion	55
4 Spherical U-net for dMRI fiber orientation distribution function estimation	57
4.1 Introduction	58
4.2 Method	58
4.2.1 Estimation of spherical harmonic (SH) coefficients	60
4.2.2 Convolutional layers	61
4.2.3 Rectified linear unit (ReLU) non-linearity	61
4.2.4 Pooling	62
4.2.5 Transposed convolutional layers	62
4.2.6 Loss function	64
4.3 Datasets	64
4.4 Experiments and implementation details	65
4.5 Results	65
4.6 Conclusion	69

5	Fourier domain spherical CNN for dMRI local analysis	71
5.1	Introduction	72
5.2	Theory	73
5.2.1	Convolution (correlation) between S^2 and zonal functions	73
5.2.2	S^2 quadratic function	74
5.2.3	Convolution (correlation) between $SO(3)$ functions	75
5.2.4	$SO(3)$ quadratic function	75
5.2.5	Power spectrum of S^2 and $SO(3)$ functions	76
5.3	Methods	77
5.3.1	Fourier domain convolutional neural network (CNN) with quadratic S^2 non-linearities	78
5.3.2	Fourier domain CNN with quadratic $SO(3)$ non-linearities	78
5.4	Experiments	81
5.4.1	Axon bundle counting experiment	81
5.4.2	Multi-compartment micro-structure estimation	87
5.4.3	Brain tissue segmentation	97
5.5	Conclusion	101
6	MEEG spatial and temporal pattern analysis	103
6.1	MEEG multivariate signal modeling	104
6.2	MEEG inverse problems	106
6.3	State of the art	108
6.3.1	Dictionary learning	108
6.3.2	CNN classification models	113
6.4	Conclusion	117
7	Rank-1 M/EEG waveform and spatial pattern learning with L_0 constraint	119
7.1	Introduction	120
7.2	Method	121
7.2.1	Encoding	122
7.2.2	Decoding	123
7.2.3	Loss and update of the dictionaries	124
7.2.4	Testing	125
7.3	Databases	125
7.4	Implementation details	128
7.5	Results	130
7.6	Conclusion	144
8	Shallow CNN for M/EEG classification	145
8.1	Theory	146
8.2	Method	146
8.2.1	Feature extraction	147
8.2.2	Feature selection and normalization	148

8.2.3	Feature classification	149
8.2.4	Training	149
8.2.5	Validation and test	151
8.3	Experiments	151
8.3.1	Databases	152
8.3.2	Implementation details	153
8.4	Results	157
8.5	Conclusion	162
9	Conclusions and perspectives	163
	Appendices	167
A	S^2 and $SO(3)$ signal related derivations appendix	169
B	Microstructure estimation experiments appendix	181
C	Dictionary learning experiments appendix	193
D	M/EEG classification experiments appendix	217
	Bibliography	229

List of Acronyms

CNS Central Nervous System	6
PNS Peripheral Nervous System	6
GM gray matter	20
WM white matter	20
CSF cerebrospinal fluid	38
AP action potential	26
PSP postsynaptic potential	26
MRI Magnetic Resonance Imaging	14
CT Computed Tomography	15
PET Positron Emission Tomography	15
EM electro-magnetic	5
RF radio frequency	16
DSG diffusion sensitizing gradients	22
dMRI diffusion Magnetic Resonance Imaging	1
PSGE Pulsed Gradient Spin-Echo	24
EEG Electroencephalography	2
MEG Magnetoencephalography	2
fNIRS functional Near Infrared Spectroscopy	14
SPECT Single Photon Emission Computed Tomography	15
DTI Diffusion Tensor Imaging	24
HARDI High Angular Resolution Diffusion Imaging	25
SQUID superconducting quantum interference device	32
SERF spin exchange relaxation-free	32
ZH zonal harmonic	38
SH spherical harmonic	v

RH rotation harmonic	45
PDF probability density function	35
ADC apparent diffusion coefficient	24
DSI diffusion spectrum imaging	25
EAP Ensemble Average Propagator	39
dODF Diffusion Orientation Distribution Function	40
fODF Fiber Orientation Distribution Function	2
BCI brain-computer interfaces	1
CNN convolutional neural network	vi
FCN Fully Connected Network	51
ReLU rectified linear unit	v
ICA independent component analysis	107
HCP Human Connectome Project	2
SNR signal to noise ratio	64
MSE mean square error	64
MAE mean angular error	65
DL deep learning	44
MLP multi layer perceptron	49
CNN convolutional neural network	vi
NODDI neurite orientation dispersion and density imaging	49
SMT spherical mean technique	81
MCSC Multivariate Convolutional Sparse Coding	193

Introduction

The development of neuroimaging techniques over the last and current century has facilitated gathering of the new insights in the structure and function of the central nervous system, mainly in an *in-vivo* and *non-invasive* manner [de Beeck & Nakatani 2019]. Firstly invented structural neuroimaging techniques allowed the analysis of the shape, the distribution, and the volume of different neural tissues [Lenroot & Giedd 2006]. Therefore, they have been used in the diagnosis and characterization of multiple brain diseases, including brain tumors, multiple sclerosis, and traumatic brain injuries [Gordillo *et al.* 2013, Filippi *et al.* 2019, Lindberg *et al.* 2019]. The development of diffusion Magnetic Resonance Imaging (dMRI) enabled structural analysis at a micro-scale by providing valuable information on the orientation of neural micro-structures, principally white matter axon bundles [Le Bihan *et al.* 2006]. This has also opened the door to the research field of structural brain connectivity [Sporns *et al.* 2005]. Functional neuroimaging techniques have been used to represent brain activities [Orrison *et al.* 2017]. Apart from being employed in clinical practice for detection and characterization of brain conditions such as epilepsy and sleep disorders, functional neuroimaging has been widely used in cognitive science, brain-computer interfaces (BCI) and functional connectivity analysis [Kauhanen *et al.* 2006, da Silva 2013]. Besides the independent analysis of the structural and functional properties of the brain, in the last two decades, a field of research has been dedicated to understanding of their relationships [Deriche 2016].

To facilitate and improve the interpretation of the acquired medical data, a broad research area is devoted to development of the models for their analysis [Erickson *et al.* 2017]. New machine learning algorithms, such as deep learning models, usually require a high amount of data (and possibly its annotation) often infeasible to acquire in clinical and practical conditions. This request is a consequence of a high variability of the same imaging modalities between acquisition centers, imaging devices, acquisition protocols, subjects, recording sessions and often, also due to high levels of noise. To account for some of these variabilities, data harmonization [Pezoulas *et al.* 2020] and transfer learning [Cheplygina *et al.* 2019] methods are being investigated.

To exploit the learning capacity of the neural networks, on one side and to account for the data variability and/or low quantity, on the other, in this thesis, we have investigated CNN models adapted to the properties and well grounded assumptions about the acquired data. In this way, the models are endowed with

valuable prior knowledge, before seeing any training data. As a consequence, the models show higher generalization power. In particular, we have investigated the convolutional models for the local analysis of **dMRI** data acquired with q-space sampling protocol [Caruyer *et al.* 2013] and for the analysis of the multivariate Magnetoencephalography (**MEG**) and Electroencephalography (**EEG**) signals. The former take into account real and spherical nature of the **dMRI** signals, their rotation equivariance with respect to the underlying microstructures, antipodal symmetry, and random uniform distribution of the sampling points. M/EEG convolutional models are designed under assumptions that the measured signals can be represented as a sum of rank-1 multivariate signals corresponding to individual brain activities, and noise and that the brain waveforms are of transient and recurrent nature. In addition, to reduce the effects of inter-session and inter-subject variability, a model for M/EEG signal classification which assumes spherical head model has been investigated.

The thesis is organized as follows:

- **Chapter 2.** This chapter contains an overview of the principal structural and functional properties of the human brain. This is followed by a description of biophysical phenomena in neural tissues and medical structural and functional imaging methodologies for their measuring, namely **dMRI**, **EEG**, and **MEG**.
- **Chapter 3.** In Chapter 3, firstly, properties of the **dMRI** signals acquired with q-space sampling schemes are provided. Further, an overview of the state-of-the-art **dMRI** local modeling approaches is given, in particular probability density functions on sphere and biophysically inspired micro-structure multi-compartment models. Following sections include a detailed overview of the most recent deep learning approaches used in the analysis of spherical data and in **dMRI** local modeling.
- **Chapter 4.** Our first contribution is presented in Chapter 4. It introduces spherical U-net for the Fiber Orientation Distribution Function (**fODF**) estimation with details related to the estimation of **SH** coefficients via Gram-Schmidt orthonormalization, convolutions with zonal kernels, pooling layers and transposed convolution layers. The model is positively evaluated on the real Human Connectome Project (**HCP**) and synthetic data generated with the **dmipy** library.
- **Chapter 5.** Our second contribution from the domain of **dMRI** local modeling is given in Chapter 5. It introduces the Fourier domain spherical **CNN** for **dMRI** local parameter estimation. The principal ingredients of this model are quadratic non-linearities realized in the Fourier domain. The model is evaluated on the synthetic data on the problem of the axon bundle count, estimation of the micro-structure parameters and on the brain tissue segmentation.
- **Chapter 6.** In this chapter, first, the modeling of the functional **EEG** and **MEG** signals is presented. After that, a detailed overview of the state-of-the-

art multivariate dictionary learning approaches is provided. This is followed by a description of the classification models used in BCI with a focus on CNN models.

- **Chapter 7.** This chapter contains a contribution in the domain of EEG and MEG analysis, in particular a multivariate rank-1 convolutional dictionary learning approach with an L_0 penalty. The model is thoroughly quantitatively examined on the synthetic data generated with MNE and qualitatively on the real motor task MEG HCP data and on somatosensory MEG data.
- **Chapter 8.** Our second contribution in the domain of EEG and MEG signal analysis is provided in Chapter 8. We have proposed a shallow CNN classifier with rank-1 kernels regularized in the spectral domain, both along spatial and temporal dimensions. The model is evaluated on passive and active BCI classification problems, namely on the EEG mental workload and motor-task MEG HCP data.
- **Chapter 9.** The last chapter contains general conclusions of the presented models and related perspectives.
- **Appendix A.** In Appendix A, we have provided derivations related to the Fourier transform of the real S^2 and $SO(3)$ signals, their convolutions and quadratic functions in the spectral domain. It accompanies chapters related to dMRI local modeling, namely Chapters 3, 4 and 5.
- **Appendix B** In this appendix, we have provided additional information related to the experiments conducted with the Fourier domain spherical CNN and compared methods, presented in Chapter 5.
- **Appendix C** The additional experiment materials related to convolutional dictionary learning, presented in Chapter 7, are provided in Appendix C.
- **Appendix D** The materials related to the experiments performed with the shallow rank-1 CNN and compared methods, presented in Chapter 8, are provided in Appendix D.

Background

Contents

2.1	Human brain structure and function	6
2.1.1	Structure and function of neurons	7
2.1.2	Gray matter	8
2.1.3	White matter	10
2.2	Structural and functional brain imaging techniques	14
2.2.1	Diffusion MRI	15
2.2.1.1	Free and restricted diffusion of water molecules . . .	16
2.2.1.2	Magnetic Resonance Imaging (MRI)	16
2.2.1.3	Diffusion weighted MRI	22
2.2.2	Magneto and electro encephalography	26
2.2.2.1	Neural electrical potentials	26
2.2.2.2	Modeling of electro-magnetic (EM) fields of neural currents in cortex	28
2.2.2.3	Electro-encephalography	31
2.2.2.4	Magneto-encephalography	31
2.3	Conclusion	32

Executive summary

In this chapter, firstly, a brief overview of the functional and structural properties of the human nervous system is provided. It includes the information about the neurons as its essential element and about the neural organizations at a macro-scale, namely the cortical brain lobes and the white matter fiber tracts. Further, an outline of the most prominent functional and structural medical imaging techniques is given, followed by a detailed description of the physical phenomena in the neural tissues and methodologies which allow diffusion Magnetic Resonance Imaging and magneto- and electro-encephalography signal recording.

2.1 Human brain structure and function

Anatomically, the nervous system of vertebrates is composed of the Central Nervous System (CNS) which includes the *brain* and the *spinal cord* and the Peripheral Nervous System (PNS) which is composed of the *nerves* and the *ganglia* outside the CNS. An overview of the principal structural and functional properties of the human's CNS is provided in Figure 2.1 and of the PNS in Figure 2.2. For more details, we refer the reader to [Snell 2010, Johns 2014].

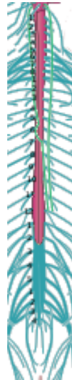
	Cerebrum	Diencephalon	Cerebellum	Brain stem	Spinal cord	
Structure	<ul style="list-style-type: none"> Cerebral cortex Cerebral white matter Limbic structures Basal ganglia Hypophysis 	<ul style="list-style-type: none"> Thalamus Hypothalamus Epithalamus Metathalamus Subthalamus 	<ul style="list-style-type: none"> Cerebellar cortex Cerebellar white matter Cerebellar deep nuclei 	<ul style="list-style-type: none"> Medulla oblongata Pons Midbrain 	 <ul style="list-style-type: none"> Continuation of medulla oblongata 31 segments, each attached to a pair of sensory and a pair of motor nerve roots Segments: 8 cervical, 12 thoracic, 5 lumbar, 5 sacral, 1 coccygeal 	Structure
Function	<ul style="list-style-type: none"> Cognition Emotions Learning, memory Thermoregulation Movement coordination Attention Motivation 	<ul style="list-style-type: none"> Relay point for sensory and motor signals Memory Emotions 	<ul style="list-style-type: none"> Motor control Attention Language Emotion control Balance Posture 	<ul style="list-style-type: none"> Cardiac and respiratory functions Consciousness Circadian clock Motor and sensory nerve supply to face and neck 		

Figure 2.1: An overview of the structural and functional properties of the CNS. Images adapted from: [mid-sagittal plane of the brain](#) and [image credited to William Crochot](#).

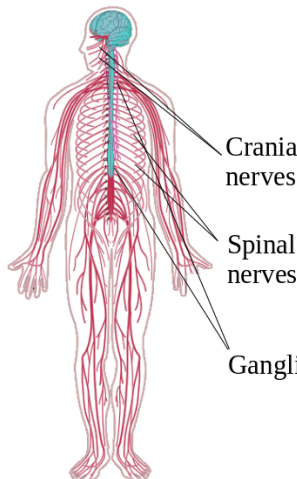
Structure	Function			
	Somatic		Autonomic	
	sensory	motor	sympathetic	parasympathetic
	<i>Relay for sensory inputs</i> <ul style="list-style-type: none"> Sight Hearing Smell Taste Touch Temperature 	<i>Relay for motor outputs</i> <ul style="list-style-type: none"> Speech Eye movements Body movements 	<i>"Fight & flight" response</i> <ul style="list-style-type: none"> Speeds up heart rate and breathing Pupil dilation Adrenalin release 	<i>"Rest & digest" response</i> <ul style="list-style-type: none"> Slows down heart rate and breathing Pupil contraction Stimulates digestive system
	Ganglia	• a type of relay station		

Figure 2.2: An overview of the structural and functional properties of the PNS. Image adapted from: [image credited to William Crochot](#).

2.1.1 Structure and function of neurons

The essential elements of the nervous system are neurons, a majority of which make a part of the brain. On average, an adult human brain contains $\sim 86 \times 10^9$ neurons and $\sim 85 \times 10^9$ non-neural cells [Azevedo *et al.* 2009, Herculano-Houzel 2012]. Typically, a neuron is composed of a soma, dendrites and an axon with multiple terminals. The soma is the metabolic center of a neuron and is responsible for generating proteins necessary for neuron maintenance and functioning. The region of the soma where the axon emerges is called the axon hillock. Dendrites and axons, also referred to as neurites, are projections from the soma responsible for communication and information processing. An illustration of a neuron with its main structures is given in Figure 2.3. Each of the neuron components give rise to a morphological diversity of neurons, thus they can differ in terms of position, shape and size of the soma, length of neurites, number of dendrites and axon terminals, as well as their spatial organization. Crucial electro-physiological properties of neurons are excitability, conductivity and secretion, which enable them to receive and process information and based on the processing outcome, to transmit information further. Given their connections, neurons can be classified as *interneurons* which communicate only with other neurons, *afferent neurons* which convert environmental stimuli into signals, and *efferent neurons* which transmit signals to organs [Peters *et al.* 1976]. In general, signal reception takes place at the level of dendrites. In the case of afferent neurons, dendrites directly or indirectly translate received stimuli into sensory signals. Otherwise, in interneurons and efferent neurons, reception is performed via synapses which are, most commonly, established with dendrites and axons of different neurons. As each synapse has an associated weight, signal processing starts at reception and continues within dendrites. Depending on the spatial distribution of the synaptic inputs, processing at the level of dendrites can be modeled in a linear or non-linear manner [Grienberger *et al.* 2015]. Processed signals are integrated in axon hillock and if the voltage of the resulting signal reaches a high enough amplitude in a short period of time, an action potential is generated. This action potential is transmitted along the axon until its terminals. Some axons are wrapped in a myelin sheath which acts as an insulator and ensures their high conductivity and efficient action potential transmission. In the PNS, the myelin sheath originates from Schwann cells and in the CNS from oligodendroglial cells [Morell & Quarles 1999]. Once the action potential reaches axon terminals, secretion of neurotransmitters enables information transmission to the following neuron or an organ cells in the case of efferent neurons.

In the CNS, the spatial organization of neurons creates tissues that at macroscopic scale appear as the *gray* and *white matter*. Gray matter is composed of cell bodies, dendrites, unmyelinated axons and glial cells [Solomon *et al.* 2014]. White matter contains axons and a much higher concentration of glial cells, a majority of which are oligodendroglial cells which create myelin sheath and give rise to the whitish color of the tissue.

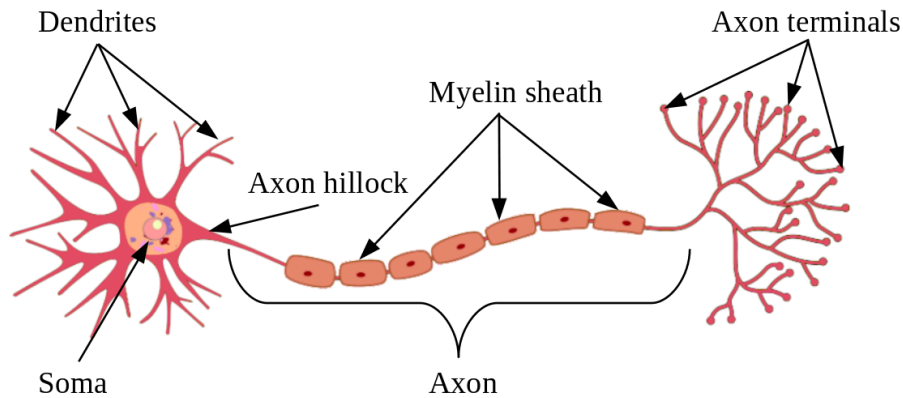


Figure 2.3: Structure of a neuron.
Image adapted from: [structure of neuron](#).

2.1.2 Gray matter

Gray matter tissue constitutes the outer layers of the cerebrum and cerebellum known respectively as the cerebral and cerebellar cortices, but also some of their inner structures such as the basal ganglia and the deep cerebellar nuclei. It is also the principal component of the diencephalon structures and is present in some segments of the brain stem. Further, it constitutes the inner part of the spinal cord also known as the gray column. As in the context of this thesis, we are only interested in the signals emerging from the cerebral cortex, in this section, we focus on its structural and functional properties.

The surface of the cortex is highly wrinkled, where a distinction can be made between tissue bumps known as gyri (singular: gyrus) and grooves known as sulci (singular: sulcus) [Spielman *et al.* 2020]. The cerebral cortex is divided by the longitudinal fissure into the right and left hemispheres. Furthermore, each hemisphere is composed of four lobes, namely frontal, temporal, parietal and occipital lobe.



Courtesy of: Deslauriers-Gauthier S. [nimesh]

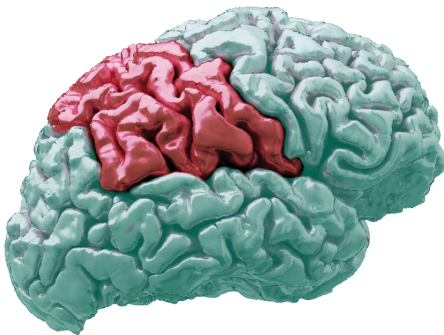
The frontal lobe takes the largest portion of the cerebral cortex. It is separated from the rest of the cortex by the central sulcus (fissure of Rolando) and the lateral sulcus (Sylvian fissure). It contains the precentral, superior frontal, middle frontal, and inferior frontal gyri, separated by precentral, superior frontal and inferior frontal sulci. From the functional point of view, the frontal lobe is often termed as the "action cortex". The precentral gyrus contains the primary motor cortex. The

premotor cortex and supplementary motor area are situated anterior to it. These three regions make the motor cortex and are responsible for planning, control and execution of voluntary movements [Foerster 1936]. The frontal part of the frontal lobe is termed as prefrontal cortex and it participates in higher cognitive functions, such as attention, problem solving, short-term memory, personality expression, etc [Miller *et al.* 2002]. The frontal lobe also includes Broca's area responsible for speech production [Keller *et al.* 2009].



Courtesy of: Deslauriers-Gauthier S. [nimesh]

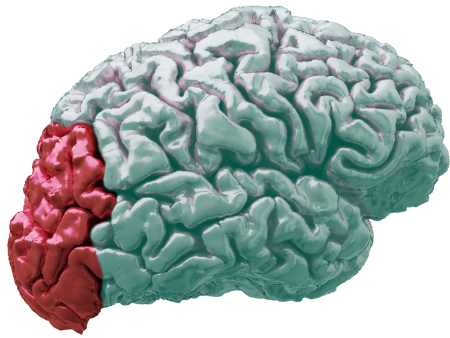
The temporal lobe is separated from the frontal lobe by the lateral sulcus and from the rest of the cortex by an imaginary parietotemporal line. It contains the superior, middle and inferior temporal gyri, separated by superior temporal and inferior temporal sulci. The temporal lobe includes the auditory cortex composed of primary, secondary and tertiary cortices, also referred to as core, belt and parabelt areas, which are responsible for processing of auditory information [Pickles 1998]. A region of the temporal lobe termed as medial temporal lobe, which includes the hippocampus, amygdala and parahippocampal regions is essential in the creation of long-term memory [Eichenbaum *et al.* 1993]. The superior temporal gyrus contains the Wernicke's area which is traditionally associated with understanding of written and spoken language, although some more recent studies indicate that it also participates in speech production [Binder 2015]. Finally, the temporal lobe also include regions which participate in processing of visual information, in particular object recognition [Milner & Goodale 2006].



Courtesy of: Deslauriers-Gauthier S. [nimesh]

The parietal lobe is placed behind the frontal lobe and above the temporal and occipital lobes. From the frontal lobe, it is separated by the central sulcus and from the temporal and occipital lobes by the lateral sulcus, the parieto-occipital sulcus and imaginary borders. It contains the postcentral gyrus, which is situated just after the central sulcus and is followed by the postcentral sulcus. The remaining part of the parietal lobe is the posterior parietal cortex, which is composed of the superior and inferior parietal lobules, separated by the intraparietal sulcus [Vingerhoets 2014]. The postcentral gyrus contains the primary somatosensory cortex, while the secondary somatosensory cortex is situated in the superior bank of the lateral sulcus. Together,

they constitute the somatosensory cortex involved in reception and processing of sensory information [Penfield & Rasmussen 1950]. The superior parietal lobule is involved in attention and visuospatial perception, while the inferior parietal lobule takes part in reading, writing, and solving mathematical operations [Johns 2014].



Courtesy of: Deslauriers-Gauthier S. [nimesh]

The occipital lobe is the smallest lobe and corresponds to the posterior part of the cortex. More precisely, it is separated from the parietal and temporal lobes by the parietooccipital sulcus and the imaginary lateral parietotemporal line. The morphology of this lobe varies most significantly between subjects, but three gyri can be identified, namely the superior, middle and inferior occipital gyri. The occipital lobe contains the primary visual cortex known as the striate cortex and the visual association cortex also

known as extrastriate visual cortex. They are responsible for the processing of visual information, in particular color determination, perception of size, depth and distance, object and face recognition, visuospatial processing, memory formation [Johns 2014, Rehman & Al Khalili 2019].

2.1.3 White matter

White matter tissue is present inside the cerebrum and cerebellum. It is composed of myelinated axons, which are grouped in bundles also called tracts or fibers. These tracts make links between distant gray matter regions. It is also present in the structures of diencephalon and the brain stem and surrounds the gray matter in the spinal cord. As in the context of this thesis, we are only interested in the cerebral white matter, in this section, we focus on its structural and functional properties. White matter tracts can be classified into three groups, namely projection, association and commissural fibers.

The projection tracts connect the cerebral cortex with the other structures of the CNS. Traditionally, they are classified into efferent (brain output) and afferent (brain input). The most prominent *efferent projection tracts* are the corticospinal, corticobulbar and corticopontine fibers. The *corticospinal fibers* primarily emerge from the motor cortex, but some originate from the somatosensory cortex as well. The axons terminate either by connections to motor neurons or to interneurons of the spinal cord. Along this path, they pass through the brain stem, where they form medullary pyramids. At the exit of the medullary pyramids, a larger fraction of the fibers decussate and create the lateral corticospinal tract, while the remaining fibers create the anterior corticospinal tract. The principal function of the corticospinal tract is to transmit the signals responsible for voluntary movements and sensory-

driven reflexes, but they are also involved in modulation of the sensory information. The *corticobulbar fibers* originate in the primary motor cortex, in particular from the regions above the lateral fissure. By passing through the corona radiata and the internal capsule, they end in the medullary pyramids also called bulbar. Corticobulbar fibers transmit motor signals, directly or via interneurons, to the cranial nerves which innervate muscles of the face, mastication, tongue, pharynx, larynx, etc. The *corticopontine fibers* emerge from all the regions of the cerebral cortex, but the largest number of fibers comes from the frontal lobe. They end in the pontine nuclei, just at the entrance to cerebellum. Corticopontine fibers establish communication between the cerebral and cerebellar cortices and are involved in the coordination of voluntary movements [Rea 2015]. Illustrations of the corticospinal, corticobulbar and corticopontine fibers are provided in Figure 2.4.

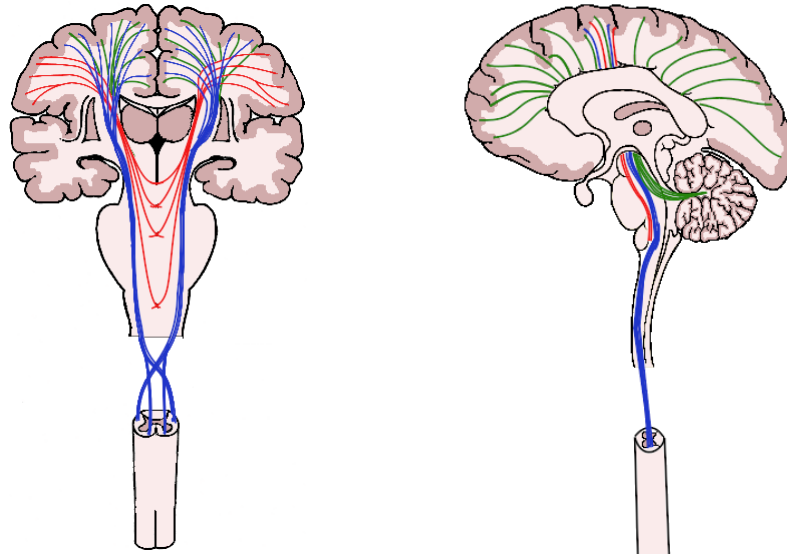


Figure 2.4: Illustration of the corticospinal (blue), corticobulbar (red) and corticopontine (green) fibers in coronal (left) and sagittal (right) views. Images adapted from: [The projection tracts \(coronal\)](#) and [the projection tracts \(sagittal\)](#).

The *afferent projection tracts* transmit information from the subcortical CNS structures to the cortex. Some examples of well recognized afferent projection tracts are the optic and acoustic radiations which make part of the optic and auditory pathways. The optic pathways start with the optic nerves originating in the retina. The nerves meet and partially decussate in the optic chiasm, creating the optic tracts which terminate in the lateral geniculate nucleus, located in the thalamus [Mehra & Moshirfar 2021]. The remaining pathways correspond to the *optic radiations* which connect the thalamus and the visual cortex. The auditory pathways start with the cochlear nerves originating in the cochleas. They pass and partially decussate in brain stem, creating tracts termed lateral lemnisci [Peterson *et al.* 2018]. The lateral lemnisci terminate in the medial geniculate nuclei, located in the thalamus. The remaining pathways correspond to the *acoustic radiations* which connect

the thalamus and the auditory cortex. Illustrations of the optic and auditory pathways are illustrated in Figure 2.5.

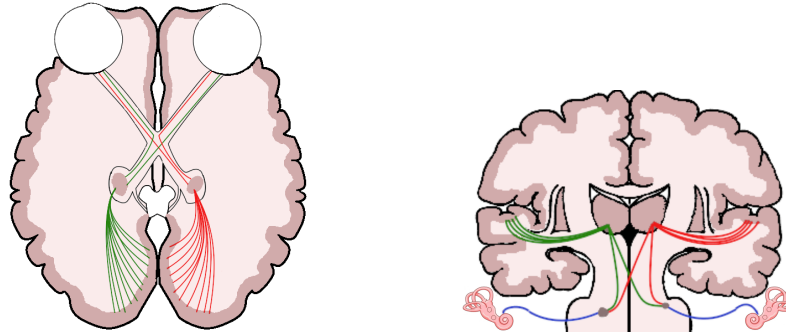


Figure 2.5: Illustrations of the optic (left) and the auditory (right) pathways. Images adapted from: [Optic pathways](#), [auditory pathways](#) and [cochlea](#).

The association tracts form the intrahemispherical connections. They can be classified into short and long tracts. Short fibers, situated closely beneath gray matter, make connections between adjacent gyri. Long tracts connect more distant regions of the cortex. Some of the most prominent long association fibers are the cingulum, the superior and inferior longitudinal fasciculi, the uncinate fasciculus, the vertical occipital fasciculus, the inferior fronto-occipital fasciculus, the arcuate fasciculus, etc. The *cingulum* connects the frontal, parietal and medial temporal regions, and the subcortical nuclei to the cingulate cortex, situated in the medial part of the cerebrum [Bubb *et al.* 2018]. The *superior longitudinal fasciculus* makes connection between the temporoparietal junction area and the parietal lobe on the one side and the frontal lobe on the other side [Wang *et al.* 2016]. It is involved in signal transmission related to language, attention, memory, and emotions. The *uncinate fasciculus* connects the anterior temporal lobe with the inferior region of the frontal lobe [Von Der Heide *et al.* 2013]. It is considered to be involved in some aspects of episodic memory, language and emotional processing [Von Der Heide *et al.* 2013]. The *vertical occipital fasciculus* connects the dorsolateral and ventrolateral visual cortices and is important in signal transmission related to visual and cognitive functions [Yeatman *et al.* 2014]. The *inferior fronto-occipital fasciculus* originates in the frontal lobe and terminates in the regions of occipital cortex, temporo-basal areas, and superior parietal lobe [Wu *et al.* 2016b]. It is associated with language processing and goal-oriented behavior [Conner *et al.* 2018]. The *inferior longitudinal fasciculus* arises from the occipital and temporal-occipital areas and terminates in the inferior region of the temporal lobe. It is involved in a wide range of brain functions, such as object recognition, reading, lexical and semantic processing, emotions and visual processing [Herbet *et al.* 2018]. The *arcuate fasciculus* is historically defined as a fiber connecting two language-related areas, namely the Wernicke's and Broca's area. More precisely, a recent study showed that the fibers arise from the ventrolateral frontal cortex and via the parietal cortex reach the middle and inferior

temporal lobe [Eichert *et al.* 2019]. Illustrations of the short and long association fibers are provided in Figure 2.6.

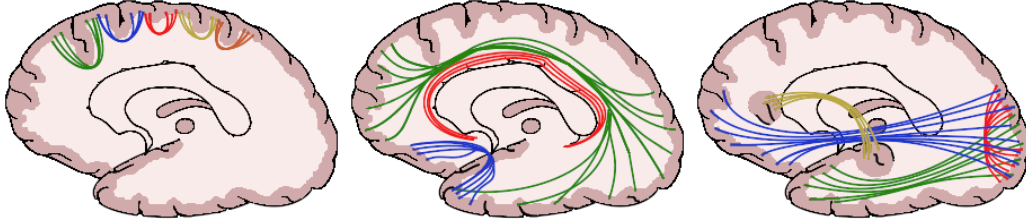


Figure 2.6: Illustrations of the short (left) and long (middle, right) association fibers. The long fibers (middle) include: the *cingulum* (red), the *superior longitudinal fasciculus* (green), the *uncinate fasciculus* (blue). The long fibers (right) include: the *vertical occipital fasciculus* (red), the *inferior fronto-occipital fasciculus* (green), the *inferior longitudinal fasciculus* (blue), the *arcuate fasciculus* (yellow). Images adapted from: [the association tracts](#) (sagittal).

The commissural tracts form interhemispherical connections. The most important commissural fibers are the corpus callosum, the hippocampal commissure, and the anterior and posterior commissures. The *corpus callosum* is the largest commissural tract situated beneath the cerebral cortex and above the thalamus. It is composed of four parts, namely the rostrum, the genu, the body, and the splenium. The rostrum connects the orbital regions of the frontal lobes. The genu connects the medial and lateral regions of the frontal lobe. The body contains fibers which make part of the corona radiata and connect the temporal and occipital lobes. The splenium connects the occipital lobes. The corpus callosum is responsible for signal transmission related to sensory, motor, and high-level cognitive functions. The *anterior commissure* is situated anteriorly with respect to the corpus callosum. It connects the olfactory, amygdaloid and temporal regions [Fenlon *et al.* 2021]. Although still not completely understood, some studies have shown that the anterior commissure is involved in olfactory functions, memory and visual processing [Fenlon *et al.* 2021]. The *posterior commissure* is a small bundle of axons, posterior to the corpus callosum, which connects the structures of epithalamus. It is considered to be involved in signal transmission between language processing centers [Standring 2020]. The *hippocampal commissure*, also known as commissure of the fornix, makes connection between hippocampus [Standring 2020].

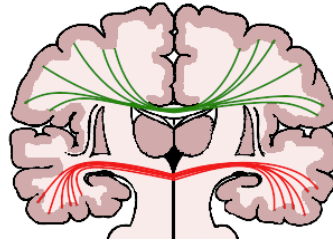


Figure 2.7: Illustration of the principal commissural tracts: *corpus callosum* (green) and the *anterior commissure* (red).

Image adapted from: [the commissural tracts \(coronal\)](#).

2.2 Structural and functional brain imaging techniques

Neuroimaging refers to the creation of images which reflect the structural and/or the functional characteristics of the examined part of the nervous system, via utilization of certain imaging techniques. Apart from the characteristics they reflect, these techniques can be differentiated along multiple axes, such as spatial and temporal resolution, contrast, signal to noise ratio, required acquisition time, portability and price of acquisition devices, invasivity, patient-friendly assessments, etc. An overview of the well developed and commonly used techniques in brain imaging is given below.

Magnetic Resonance Imaging (MRI) uses a strong magnetic field, magnetic field gradients and electro-magnetic radio frequency pulses to interact with nuclei present in the tissues in order to create images. Spatial and temporal organizations of the gradients and the pulses allow acquisition of different MRI modalities. Some of the broadly used structural modalities include conventional T_1 , T_2 and T_2^* weighted images, and diffusion MRI. Examples of MRI modalities which reflect functional properties of the tissues are perfusion weighted images and functional MRI.

EEG is a functional imaging technique which uses electrodes placed on the scalp or intra-cranially to record the electric potential produced by the electric activity of the cerebral cortex. It is characterized by a very high temporal resolution, but a low spatial one in comparison to functional MRI. In addition to its high temporal resolution, another important advantage of the EEG imaging technique is the portability and low cost of its measuring devices.

MEG is a functional imaging technique which measures the magnetic field strength produced by the electric activity of the cerebral cortex. Acquisition is achieved with magnetometers placed on the scalp or in its proximity. As EEG, it is characterized by a high temporal resolution. The spatial resolution is in general higher than with EEG, but lower than that of functional MRI.

Functional Near Infrared Spectroscopy (fNIRS) is a functional imaging technique which uses near-infrared light to capture the haemodynamic activity in the cortex which appears as a consequence of a neural activity (the same physical phenomena is measured by functional MRI). Measuring is achieved using light emitters

and detectors placed on the scalp. Its temporal resolution is better than in functional MRI, but lower than with EEG and MEG. Localization of active regions is more accurate than with EEG and MEG, mostly because fNIRS is only able to measure activities that are close to the cortical surfaces. As for EEG, fNIRS devices can be portable.

Computed Tomography (CT) uses X-ray sources and detectors to measure X-ray attenuation along multiple angles. The obtained measurements are combined using computerized algorithms which perform a tomographic reconstruction to obtain the final images. Conventional CT scans are used for anatomical imaging, whereas CT perfusion imaging is a functional modality which uses contrast agents to quantify blood perfusion in the brain. Compared to MRI, CT scans can have higher spatial resolution and lower acquisition times. MRI however provides better contrast between soft tissues.

Positron Emission Tomography (PET) uses radiotracers which emit positrons which when colliding with electrons emit gamma rays measurable by detectors placed around the examined region. Similarly to CT, a computerized tomographic reconstruction is applied on the measured signals to obtain the final scan. In brain imaging, PET scans are used to measure the blood flow associated with neural activity. Compared to MRI, both spatial and temporal resolutions of PET scans are lower.

Single Photon Emission Computed Tomography (SPECT) uses radiotracers which directly emit gamma rays measurable by detectors placed around the examined regions. As in the previously mentioned tomography imaging techniques, images are computed using computerized tomographic reconstruction algorithms. As with PET, it is a functional imaging technique which measures the blood flow whose increase is correlated with an increase of neural activity. Compared to PET, in general, its spatial and temporal resolutions are lower, as well as the price of scanner.

As in this thesis, we have proposed models for the analysis of EEG, MEG and dMRI data, a more detailed description of the physical phenomena in the neural tissues and methodologies which allow their recording of the corresponding signals is provided.

2.2.1 Diffusion MRI

dMRI is an MRI imaging modality which captures the structural properties of tissues. In comparison to conventional anatomical MRI scans, such as T_1 and T_2 weighted images, dMRI images provide information about the microstructures of the examined tissue.

2.2.1.1 Free and restricted diffusion of water molecules

Molecular diffusion is a phenomenon which corresponds to a type of particle motion occurring at temperatures higher than the absolute zero. If a particle concentration gradient is present in a substance, diffusion leads to their uniform distribution. This process can be described using *Fick's first law of diffusion* [Fick 1855]

$$\mathbf{J} = -D\nabla C, \quad (2.1)$$

which relates the diffusive flux $\mathbf{J}[\frac{mol}{m^2s}]$ to the gradient of the concentration $C[\frac{mol}{m^3}]$ via the diffusion coefficient $D[\frac{m^2}{s}]$. D is often referred to as diffusivity and depends on temperature, viscosity, particle size and presence of boundaries in the medium. *Fick's second law of diffusion* explains how concentration changes over time due to the diffusion process

$$\frac{\partial C}{\partial t} = \nabla \cdot (D\nabla C), \quad (2.2)$$

where $t[s]$ is time. Even if the distribution of particles within a substance is uniform, microscopic motions of the particles exist if the absolute temperature is higher than the absolute zero, although the net flux \mathbf{J} from Eq. 2.1 through any surface is equal to zero. This type of motion is known as Brownian motion [Brown 1828] as it was firstly described by Robert Brown. Displacement of particles only in the presence of Brownian motion can be described by solving Eq. 2.2, where diffusivity D depends on the properties of the medium. For spherical particles in an isotropic medium, diffusivity can be considered constant and is defined using the *Stokes-Einstein equation* as

$$D = \frac{k_B T}{6\pi\eta r} \quad (2.3)$$

where $k_B[\frac{J}{K}]$ is the Boltzmann constant, $T[K]$ is the absolute temperature, $\eta[\frac{kg}{m \cdot s}]$ is the dynamic viscosity and $r[m]$ is the radius of the particle. In an anisotropic medium, diffusivity can be represented as a symmetric positive-definite tensor

$$D = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{zx} & D_{zy} & D_{zz} \end{bmatrix} \quad (2.4)$$

or for more complex structures of the medium, as a positive function on a sphere $D : S^2 \rightarrow \mathbb{R}^+$. Illustration of the displacement of one particle in the same substance, without and with obstacles is provided in Figure 2.8.

2.2.1.2 Magnetic Resonance Imaging (MRI)

MRI is an imaging technique, based on the property of nuclei of certain atoms to absorb and emit **EM** waves at a specific radio frequency (**RF**). In imaging of the human body, a majority of these atoms are hydrogen atoms from the water molecules, thus a nucleus H^+ corresponds to a proton p^+ . To create an image, the received **EM** waves are averaged over small volumes called *voxels* of the order of

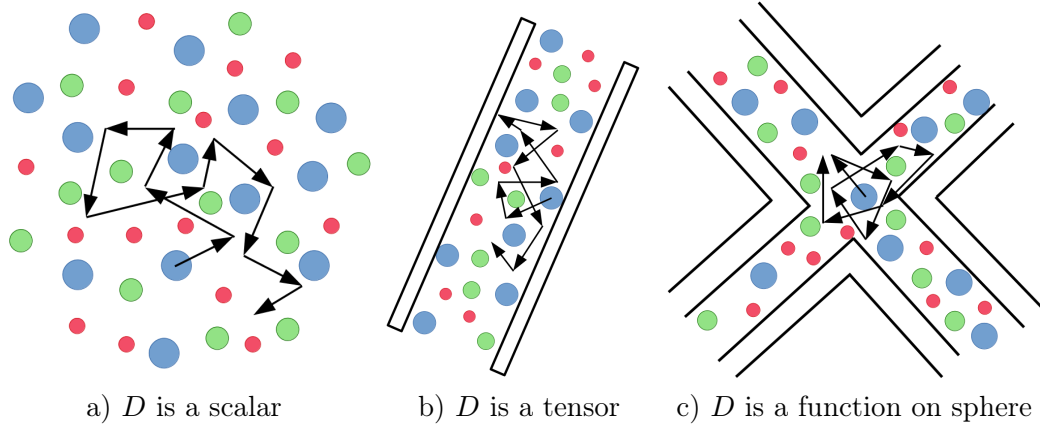


Figure 2.8: Illustration of displacement of one particle in a medium: a) without obstacles, b) in a tube and c) in a tube junction

magnitude $\sim 1mm^3$. One voxel of water, of volume $\sim 1mm^3$, contains 0.67×10^{20} hydrogen protons. This can give us an idea of the amount of protons within one voxel which participate into the EM signal generation for different tissues, bearing in mind that $\sim 73\%$ of the brain and the heart is water, as well as $\sim 31\%$ of the bones [Mitchell *et al.* 1945].

Protons are characterized by their mass, electric charge and spin. When the examined tissue is not exposed to a strong enough external magnetic field, the orientations of the spins of the hydrogen protons are random as illustrated in Figure 2.9 a). In general, the acquisition of an MRI scan requires the utilization of a strong external magnetic field, of three gradient magnetic fields for spatial encoding and of RF EM pulses at the resonance frequency. The external magnetic field is also referred to as the *main magnetic field* $\mathbf{B}_0[T]$. Spatial encoding gradient fields alter \mathbf{B}_0 with a term $\Delta\mathbf{B}_z(x, y, z, t)[T]$ in a way that the EM waves associated to the voxels at different positions have different frequencies and/or times of application. The RF pulses emitted at Larmor frequency enable signal acquisition as it will be further explained.

Once the the main magnetic field \mathbf{B}_0 is activated, spins align with and against it and start to precess at the Larmor or resonance frequency $\omega_0 = \gamma|\mathbf{B}_0|$ around \mathbf{B}_0 which is oriented along the z -axis in Figure 2.9 b). $\gamma[\frac{rad}{s.T}]$ is the gyromagnetic ratio - a constant equal to the ratio of the magnetic moment and the angular momentum of the particle. For the hydrogen proton in a water molecule $\gamma = 267.52 \times 10^6 \frac{rad}{s.T}$. Taking into account the spatial encoding magnetic field gradients, the resonance frequency can be expressed as $\omega_0(x, y, z, t) = \gamma|\mathbf{B}_0 + \Delta\mathbf{B}_z(x, y, z, t)|$. The alignment of the spins is illustrated in Figure 2.9 b). Although both orientations of the spin alignments are possible and are spread between these two orientations, alignments with the external field have a lower energy state. Given this, at each moment, a slightly higher amount of spins aligns with \mathbf{B}_0 . The ratio between the number of

spins aligned with (n_-) and against the external (n_+) field is given by

$$\frac{n_-}{n_+} = e^{\frac{\gamma\hbar|\mathbf{B}_0 + \Delta\mathbf{B}_z(x,y,z,t)|}{k_B T}} \quad (2.5)$$

where \hbar is the Plank constant. The difference between the number of spins at lower and higher energy states gives raise to the *net magnetization*. Although the spins, within one voxel, precess at the same frequency, since they do not precess in phase, the net magnetization in the xy -plane sums up to 0. Thus it exists only along the z -axis and it is denoted with $\mathbf{M}_z[T]$ in Figure 2.9 b), where $|\mathbf{M}_z| = M_0$ is a non-zero net magnetization. \mathbf{M}_z is called the longitudinal component of the magnetization. Assuming the presence of only \mathbf{B}_0 , using Eq. 2.5 one can obtain that for $n_+ = 10^6$ and $|\mathbf{B}_0| = 3T$, $n_- \approx 10^6 + 20$, while for $|\mathbf{B}_0| = 9T$, $n_- \approx 10^6 + 59$. Higher the difference between n_- and n_+ the amplitude of the produced net magnetization is higher ("more protons participate in the contrast creation"), thus the emitted EM waves are less susceptible to noise. This shows why scanners with higher main magnetic field strengths are characterized by a higher signal-to-noise ratio.

If an EM RF pulse $\mathbf{B}_1[T]$ at the Larmor frequency is applied perpendicularly to the main magnetic field \mathbf{B}_0 , spins spiral down to the xy -plane and continue to precess around the z -axis. But now, the precession of the spins are in phase, as depicted in Figure 2.9 c). In this step, the net magnetization is non-zero only in the xy -plane - $|\mathbf{M}_{xy}| = M_0$ and it rotates at the Larmor frequency, while $|\mathbf{M}_z| = 0$. \mathbf{M}_{xy} is called the transverse component of the magnetization.

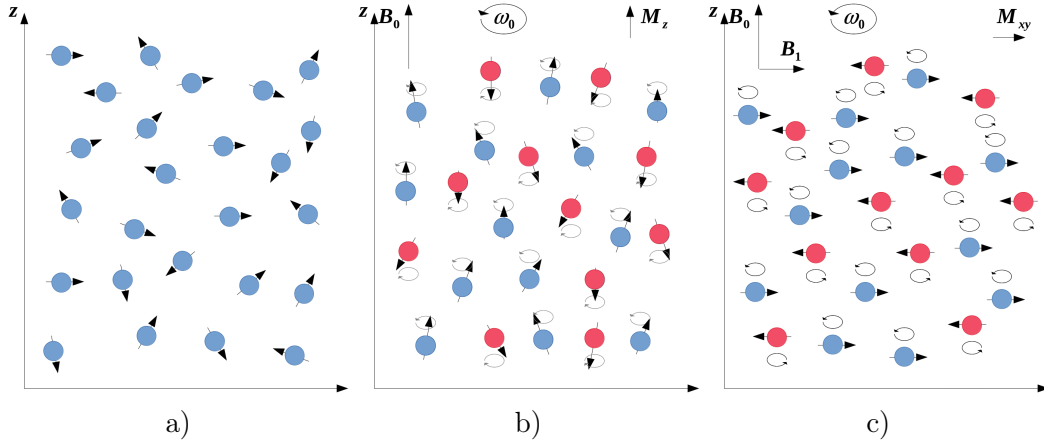


Figure 2.9: Hydrogen proton spins: a) with random orientations when there is no external field, b) aligned with and against the external magnetic field \mathbf{B}_0 , and c) after receiving RF pulse \mathbf{B}_1 at Larmor frequency

Once the RF pulse is turned off, the spins start to emit the received EM energy at the resonance frequency. As a consequence, they start to dephase and re-align with and against the external \mathbf{B}_0 field. This process was firstly described by Felix

Bloch [Bloch 1946] with a set of equations termed as *Bloch equations*

$$\frac{dM_x(t)}{dt} = \gamma(\mathbf{M}(t) \times \mathbf{B}(t))_x - \frac{M_x(t)}{T_2}, \quad (2.6)$$

$$\frac{dM_y(t)}{dt} = \gamma(\mathbf{M}(t) \times \mathbf{B}(t))_y - \frac{M_y(t)}{T_2}, \quad (2.7)$$

$$\frac{dM_z(t)}{dt} = \gamma(\mathbf{M}(t) \times \mathbf{B}(t))_z - \frac{M_z(t) - M_0}{T_1} \quad (2.8)$$

where $\mathbf{B}(t) = (B_x(t), B_y(t), |\mathbf{B}_0 + \Delta\mathbf{B}_z(x, y, z, t)|)$ and $\mathbf{M}(t) = (M_x(t), M_y(t), M_z(t))$. T_1 and T_2 are longitudinal and transverse relaxation times. If the RF pulse is $|\mathbf{B}_1| = 0$, then $\mathbf{B}(t) = (0, 0, |\mathbf{B}_0 + \Delta\mathbf{B}_z(x, y, z, t)|)$ and the Bloch equations can be simplified as

$$\frac{dM_x(t)}{dt} = -\frac{M_x(t)}{T_2} + \gamma B_z(t) M_y(t) = -\frac{M_x(t)}{T_2} + \omega_0(x, y, z, t) M_y(t), \quad (2.9)$$

$$\frac{dM_y(t)}{dt} = -\frac{M_y(t)}{T_2} - \gamma B_z(t) M_x(t) = -\frac{M_y(t)}{T_2} - \omega_0(x, y, z, t) M_x(t), \quad (2.10)$$

$$\frac{dM_z(t)}{dt} = -\frac{M_z(t) - M_0}{T_1}. \quad (2.11)$$

Assuming that $\omega_0(x, y, z, t) = \omega_0(x, y, z)$, by solving equations 2.9 and 2.10, the exponential decay of the magnitude of the transverse magnetization \mathbf{M}_{xy} is defined as

$$|\mathbf{M}_{xy}(t)| = |\mathbf{M}_{xy}(0)| e^{-\frac{t}{T_2}}. \quad (2.12)$$

This is termed as the T_2 relaxation process which is illustrated in Figure 2.10 b). The magnitude of the longitudinal magnetization \mathbf{M}_z recovers exponentially as

$$|\mathbf{M}_z(t)| = M_0 + (|\mathbf{M}_z(0)| - M_0) e^{-\frac{t}{T_1}}. \quad (2.13)$$

This is termed as the T_1 relaxation process which is illustrated in Figure 2.10 a).

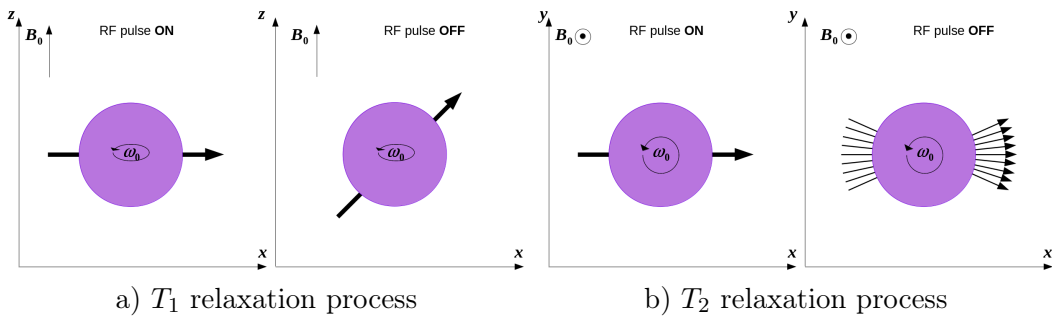


Figure 2.10: Illustration of the longitudinal and transverse net magnetization during the relaxation period. (Note that the axes in T_1 and T_2 are different. For T_1 the main magnetic field is oriented vertically, while for T_2 it points out of the paper plane.)

The T_1 relaxation time describes how quickly the longitudinal component of the net magnetization recovers and is defined as the time necessary to reach $(1 - \frac{1}{e}) \approx 63\%$ of the initial magnitude before the RF pulse - M_0 . The T_1 relaxation occurs due to the energy dissipation via the interactions between H^+ spins at higher energy levels and their environment, leading to a slight increase of temperature. The T_1 relaxation time is approximately 10 times lower in fat than in water.

The T_2 relaxation time describes how quickly the transverse component of the net magnetization decays and it corresponds to the time necessary to reach $\frac{1}{e} \approx 37\%$ of its initial magnitude after the RF pulse - M_0 . The energy dissipation associated to the T_1 relaxation leads to the T_2 relaxation as well. A second cause is the local magnetic fields produced by the nuclei of surrounding atoms, causing the precession frequency to slightly increase or decrease. Local magnetic fields associated with the H^+ spins impact each other as well. The T_2 relaxation times are in general much shorter than the T_1 .

Values of T_1 and T_2 relaxation times in white matter (WM) and gray matter (GM) for scanners with $|B_0| = 1.5T$ and $|B_0| = 3T$ are provided in Table 2.1 [Smith & Webb 2010] and corresponding relaxation curves are illustrated in Figure 2.11.

Table 2.1: Brain white and gray matter tissue T_1 and T_2 relaxation times for $|\mathbf{B}_0| = 1.5T$ and $|\mathbf{B}_0| = 3T$ in ms [Smith & Webb 2010]

Tissue type / Relaxation	$T_1(1.5T)$	$T_1(3T)$	$T_2(1.5T)$	$T_2(3T)$
White matter	790	1100	90	60
Gray matter	920	1600	100	80

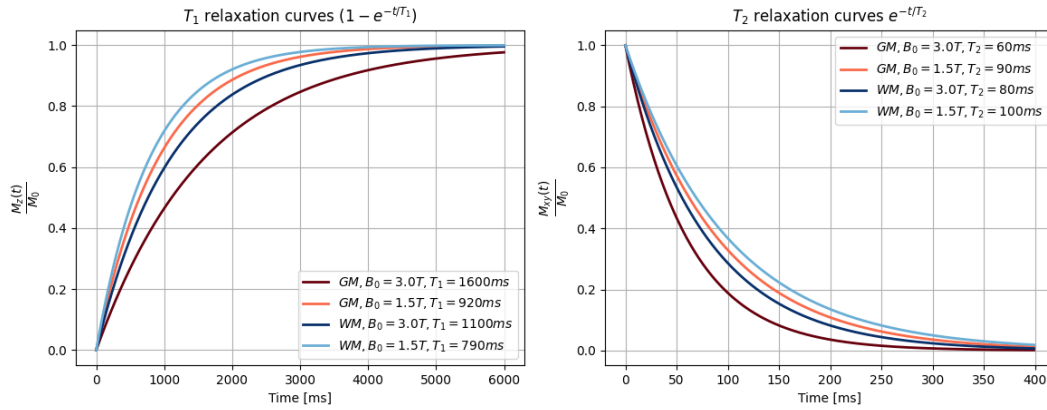


Figure 2.11: The brain white and gray matter tissue T_1 and T_2 relaxation curves corresponding to T_1 and T_2 relaxation times from Table 2.1.

EM signals emitted from excited protons are recorded using RF coils which are placed parallel to the main magnetic field. A rotating magnetic field $\mathbf{M}_{xy}(t)$ produces an oscillating current in the coil whose magnitude is determined using Fourier

transform. On the other hand, the longitudinal component of the magnetization $\mathbf{M}_z(t)$ is very weak compared to the main magnetic field \mathbf{B}_0 and cannot be measured along the z axis, thus it is tipped down by another RF pulse to the transversal plane in order to be measured.

As already mentioned, in addition to the main magnetic field \mathbf{B}_0 which is constantly active, applications of three gradient magnetic fields are used for spatial encoding. They allow us to disentangle signals recorded with the RF coil to signals originating from individual voxels. The gradient along the z axis, denoted as \mathbf{g}_z is used to select the axial slice to be recorded and it is applied at the same time as the \mathbf{B}_1 pulse. Another gradient is applied along the y axis right after the pulse, denoted as \mathbf{g}_y and is also called the *phase encoding* gradient, as it causes that proton spins along the y axis rotate with different phases. After phase encoding, a third gradient \mathbf{g}_x , termed as the *frequency encoding* gradient, is applied along the x axis, causing spins along x to rotate with slightly different frequencies. While this gradient is applied, the EM signal emerging from the entire slice is recorded with the RF coil. With a Fourier transform, we can determine the magnitudes corresponding to different positions along the x axis, however since those magnitudes correspond to the superposition of the signals with the same frequency but different phases, the entire process needs to be repeated multiple times with different amount of phase encoding (amplitude of \mathbf{g}_y) in order to determine magnitudes of the signals emerging from the individual voxels along the y axis. If the number of voxels along the y axis is N_y , then the number of phase encodings with different amplitudes of \mathbf{g}_y must also be N_y . This pulse sequence is called the gradient echo sequence and is illustrated in Figure 2.12 a). Since the main magnetic field \mathbf{B}_0 is not perfectly homogeneous, the existing inhomogeneities cause much faster dephasing of the spins than if only random spin-spin interactions are present. These inhomogeneities are constant in time, so their effect can be reversed using a RF 180° pulses applied at $TE/2$ which flip spins so that all the phase accumulated due to inhomogeneities during the first $TE/2$ period is reversed. Thus the differences due to inhomogeneities sum up to zero with the newly accumulated phase during another $TE/2$ period. This pulse sequence is called the spin echo sequence and is illustrated in Figure 2.12 b).

The T_1 and T_2 weighting of an image is achieved by adjusting the repetition time interval TR and the echo time TE interval. These values are optimized on the longitudinal and transverse relaxation times of the different tissues. Basically, one would like to read an echo signal when the amplitudes of the longitudinal or transversal components differ the most between the tissues. For a T_1 weighting, TR is relatively short and once the RF pulse is applied to flip the longitudinal component to the xy plane, the echo is read shortly after in order to avoid amplitude decrease due to dephasing. Since the recovery of the longitudinal component is long, for T_2 weighting, TR is relatively long, as well as TE . When the longitudinal component is recovered, it is tipped down to the xy plane and a TE period is given to spins to dephase before reading the echo. If the longitudinal component is not recovered only a fraction of spins participate in the evaluation of transversal relaxation.

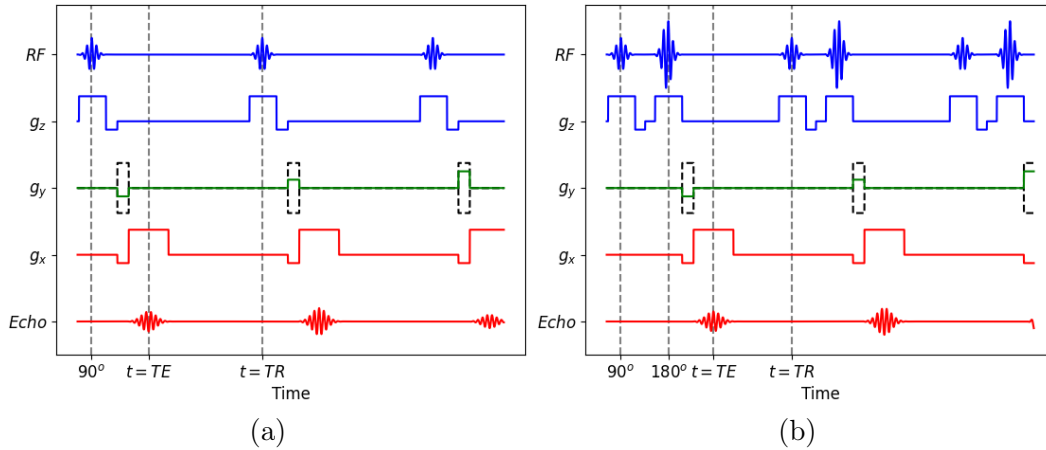


Figure 2.12: Illustration of a gradient echo sequence (a) and a spin echo sequence (b).

2.2.1.3 Diffusion weighted MRI

Diffusion weighting of MRI images is achieved by diffusion sensitizing gradients (DSG)s. A DSG can be created by using gradient fields \mathbf{g}_z , \mathbf{g}_y and \mathbf{g}_x . By adjusting the amplitudes of \mathbf{g}_z , \mathbf{g}_y and \mathbf{g}_x , a DSG can have different orientations. DSGs are combined with the T_2 relaxation process in order to create a contrast. The principal idea behind this is that when spins are tipped down to the transversal plane, a DSG is applied during a short period of time δ along a certain direction. As a consequence, as spins along the DSG direction experience slightly different gradient intensities, they accumulate slightly different phases. Thus, the first DSG is called the phase encoding gradient. After the refocusing RF pulse of 180° is applied and before the echo time, a DSG with the same direction but a reversed amplitude is applied during δ , thus the accumulated phases during the first δ period would be reversed. The second DSG gradient is called the phase decoding gradient. An example of a pulse sequence with diffusion weighting is illustrated in Figure 2.13.

If the displacement of the spins along the DSG is restricted, the second DSG cancels the majority of the dephasing effect of the first DSG. This is illustrated in Figure 2.14. On the other hand, if the displacement of the spins along the DSG is free, spins with initially encoded phases move around, thus when the second DSG is applied, the encoded phases of the spins would not be canceled. This is illustrated in Figure 2.15. Thus, if the diffusion of the water molecules is restricted along the DSG, the amplitude of the transverse component would be high, otherwise if the diffusion is free, due to additional dephasing, the amplitude of the transverse component would be low.

To incorporate the effects of the molecule diffusion, Torrey, defined the *Bloch-Torrey*

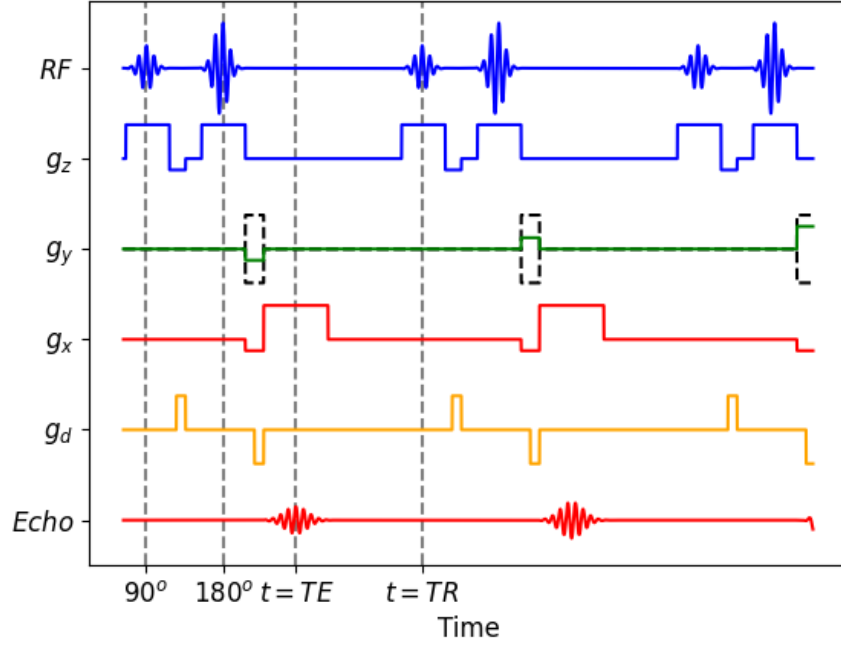


Figure 2.13: Illustration of a spin echo sequence with diffusion weighting.

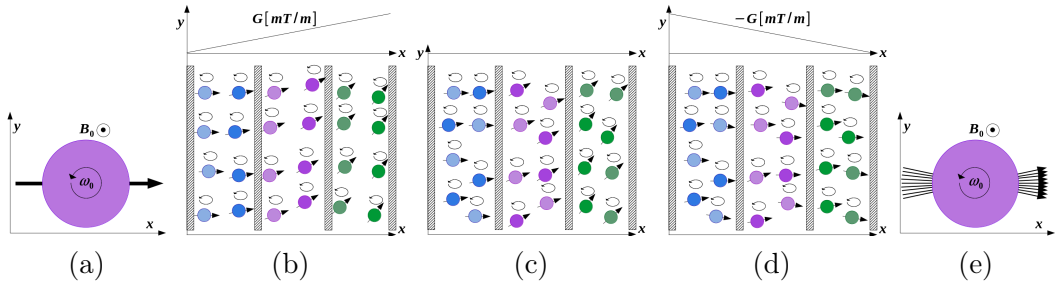


Figure 2.14: Illustration of the spin phases with restricted molecule diffusion. After the spins are tipped down with an RF pulse to the transverse plane (a), after phase encoding with a DSG (b), after a free diffusion period and a refocusing RF pulse of 180° (c), after phase decoding with a reversed DSG (d) and the resulting net magnetization (e).

equations [Torrey 1956] as

$$\frac{dM_x(t)}{dt} = \gamma(\mathbf{M}(t) \times \mathbf{B}(t))_x - \frac{M_x(t)}{T_2} + \nabla \cdot D \nabla (M_x(t) - M_{x0}), \quad (2.14)$$

$$\frac{dM_y(t)}{dt} = \gamma(\mathbf{M}(t) \times \mathbf{B}(t))_y - \frac{M_y(t)}{T_2} + \nabla \cdot D \nabla (M_y(t) - M_{y0}), \quad (2.15)$$

$$\frac{dM_z(t)}{dt} = \gamma(\mathbf{M}(t) \times \mathbf{B}(t))_z - \frac{M_z(t) - M_0}{T_1} + \nabla \cdot D \nabla (M_z(t) - M_{z0}) \quad (2.16)$$

where D is the diffusion coefficient and M_{x0} , M_{y0} and M_{z0} are the x , y and z components of the equilibrium magnetization. Attenuation of the amplitude of the

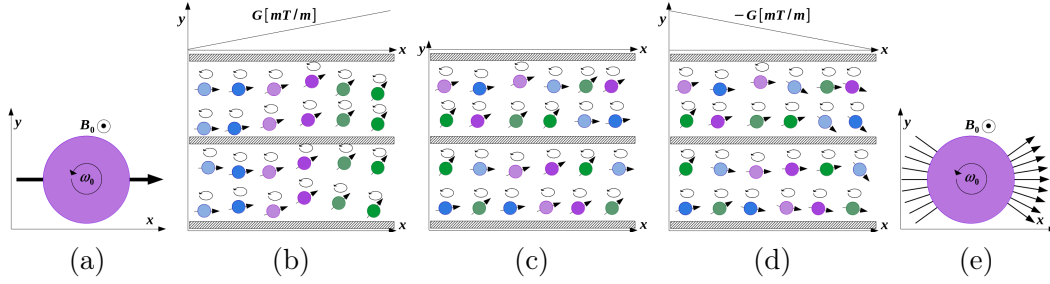


Figure 2.15: Illustration of the spin phases with free molecule diffusion. After the spins are tipped down with an RF pulse to the transverse plane (a), after phase encoding with a DSG (b), after a free diffusion period and a refocusing RF pulse of 180° (c), after phase decoding with a reversed DSG (d) and the resulting net magnetization (e).

transverse component of the magnetic field $M_{xy}(t)$ described by Eqs. 2.14 and 2.15, due to the diffusion process and for the Pulsed Gradient Spin-Echo (PSGE) sequence, is defined by the Stejskal-Tanner equation [Stejskal & Tanner 1965]

as

$$\frac{A(TE)}{A(0)} = e^{-D\gamma^2 G^2 (\Delta - \frac{\delta}{3}) \delta^2} \quad (2.17)$$

where $A(0)$ is the amplitude of $M_{xy}(0)$, when the 90° RF pulse is applied and $A(TE)$ is the amplitude of $M_{xy}(TE)$, when the signal is being recorded. G is the amplitude of the DSG \mathbf{G} . Δ is the interval between encoding and decoding DSG and δ is their duration. $b = \gamma^2 G^2 (\Delta - \frac{\delta}{3}) \delta^2$ is the b -value which describes diffusion weighting of the signal. Phase encoding and decoding DSGs are characterized by direction, strength, shape, duration and temporal spacing which all together constitute a high dimensional acquisition space termed as q -space [Callaghan *et al.* 1988]. A point of the q -space for the PSGE sequence is defined as $\mathbf{q} = \frac{\gamma \mathbf{G} \delta}{2\pi}$.

Starting from a single point q -space sampling via PSGE [Stejskal & Tanner 1965], a number of more advanced q -space sampling schemes have been developed [Descoteaux *et al.* 2014]. The first diffusion weighted MRI scans were acquired with a sampling protocol containing three differently oriented and noncollinear pairs of DSGs as introduced in [Le Bihan *et al.* 1986]. This imaging protocol allowed differentiation of the intravoxel incoherent motions between healthy and pathological tissues via apparent diffusion coefficient (ADC) [Le Bihan *et al.* 1986]. As diffusion of the water molecules in neural tissues is not uniform along all directions, in [Basser *et al.* 1994], an imaging protocol termed Diffusion Tensor Imaging (DTI), comprising acquisition over seven noncollinear q -space points for different gradient strengths, has been proposed. DTI allowed the estimation of the effective diffusion tensors capable to quantify anisotropic diffusion of the water molecules [Basser *et al.* 1994]. Being able to estimate the principal direction of the water molecule diffusion enabled tracking of the white matter pathways, a process known as tractography [Basser *et al.* 2000]. Since the white matter might contain multiple

axon bundle populations, such as crossing, kissing and fanning axon bundles, more advanced High Angular Resolution Diffusion Imaging (HARDI) protocols have been proposed [Descoteaux *et al.* 2014]. Some of the most prominent HARDI protocols are diffusion spectrum imaging (DSI) [Wedeen *et al.* 2000], single [Jones 2010] and multi shell *q-space* sampling schemes [Ye *et al.* 2012, Caruyer *et al.* 2013]. They enabled utilization of more insightful mathematical tools and the estimation of the dMRI 3D probability density functions, which have led to the development of more accurate tractography algorithms.

2.2.2 Magneto and electro encephalography

EEG and MEG are functional neuroimaging techniques which measure electric field potential and magnetic field strength produced by the neural electrical activities occurring in the pyramidal neurons which constitute more than 80% of the cerebral cortex.

2.2.2.1 Neural electrical potentials

The principal task of neurons is the processing of the input signals that might come from other neurons or from external stimuli and the transmission of the signals to other neurons or muscular cells that are supposed to perform certain actions. In the context of EEG and MEG, we are interested in the activities of the neurons that communicate between each other, also called *interneurons*, and are situated in the cerebral cortex. During this communication, two principal types of electric potentials are generated at the level of neurons, and in particular at the level of their membranes, namely action potential (AP)s and postsynaptic potential (PSP)s. These potentials are generated by the exchange of ions through the membrane of the neurons. The ions include positively charged ions such as sodium (Na^+), potassium (K^+), calcium (Ca^{2+}) and negatively charged ions such as chloride (Cl^-) and some proteins (A^-).

When a neuron is in a resting state, the concentration of K^+ and A^- ions is higher in the intracellular space, while the concentration of Na^+ , Ca^{2+} and Cl^- is higher in the extracellular space. This results in a difference between potentials between the interior and exterior of the neuron of approximately $-70mV$, which varies depending on the neuron type. The membrane contains ion channels and ion pumps, which enable passive and active displacements of the ions through the membrane. An illustration of ion distribution when a neuron is in resting state is depicted in Figure 2.16.

When a neuron receives stimuli via dendrites, they are integrated in the axon hillock and if the resulting stimulus is strong enough in a short period of time, it provokes an AP, also called *spike*, which travels along the axon. Firstly, the stimulus provokes voltage gated sodium channels to open, thus the Na^+ ions enter the cell and raise the membrane potential, a process called *depolarization*. At the end of the depolarization, the voltage gated sodium channels start to close and the voltage gated potassium channels start to open causing the K^+ ions to pass to the extracellular space. The increase of K^+ concentration in the extracellular space leads to a decrease of the membrane potential also termed as *repolarization* which terminates with *hyperpolarization*, meaning that the membrane potential reaches values lower than before the stimulus. When the hyperpolarization is reached, the voltage gated potassium channels close. This is followed by a *refractory period* when the intra- and extra-cellular concentrations of Na^+ and K^+ ions return to their resting state distributions. This entire process repeats along the axon, thus the AP travels down the axon until it reaches the axon terminals.

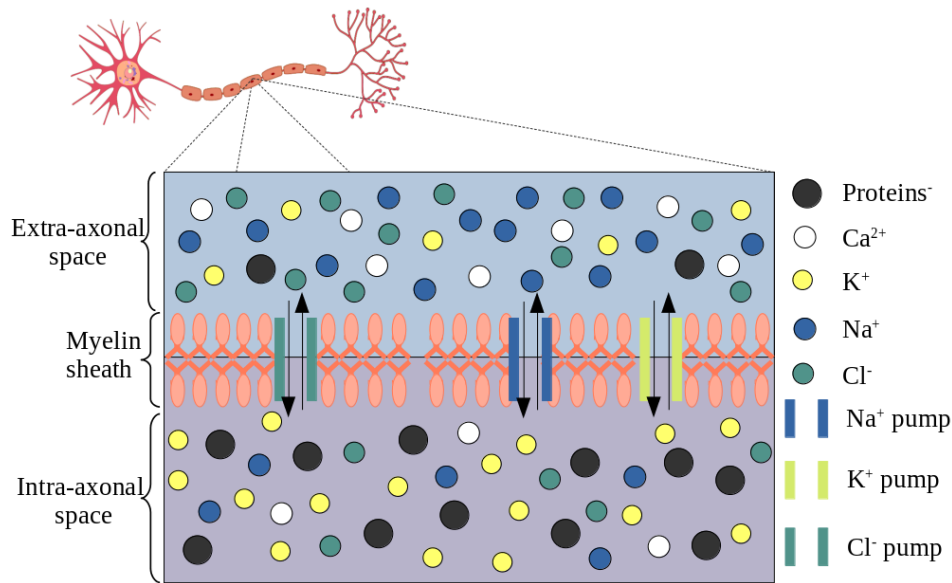


Figure 2.16: Illustration of ion distribution in intra- and extra-axonal spaces during resting state.

Neurotransmitters, situated in small vesicles in axon terminals, are crucial for the generation of PSP. Once the AP reaches the axon terminals, depolarization of its membrane causes opening of voltage gated calcium channels, causing a rush of Ca^{2+} ions into the intracellular space. These ions provoke the release of neurotransmitters from vesicles into the synaptic cleft - the extracellular space between presynaptic axon terminals and postsynaptic dendrites. The released neurotransmitters attach to receptor proteins situated at the membrane of the postsynaptic dendrites, causing certain ion channels to open or close. If sodium channels are opened, this causes an influx of Na^+ ions into the intracellular space leading to a membrane depolarization. This type of postsynaptic potential is called *excitatory*. On the other hand, if potassium channels are opened, K^+ ions pass from intra- to extra-cellular space causing membrane hyperpolarization. This type of PSP is called *inhibitory*.

While the APs are often referred to as *all-or-none*, PSPs are *graded potentials*. All-or-none principle refers to the fact that no matter how strong or long a stimulus is (yet above activation threshold), the amplitude of the AP is the same. On the other hand, graded potentials can have different amplitudes depending on temporal and spatial distances of individual potentials. If there are multiple APs arriving to the axon terminals shortly one after the other, the PSPs sum up at the postsynaptic membrane. Similar effect occurs if the synapses where the PSPs are generated are spatially close. Another important differences between an AP and a PSP are in their durations and amplitudes. Whereas, the amplitude of an AP traveling along an axon can be considered constant and is in the range of $20 - 40mV$, the amplitude of a PSP decreases with time and distance is in the range of $1 - 4mV$. APs are very short, approximately $1ms$, while the duration of the PSPs is around tens of ms .

These differences between APs and PSPs lead to different mathematical modeling of the two. An AP is modeled with an electric quadrupole whose EM field decreases with $\frac{1}{r^3}$, while a PSP is modeled with an electric dipole whose EM field decreases with $\frac{1}{r^2}$.

2.2.2.2 Modeling of EM fields of neural currents in cortex

Even though the amplitude of the APs is significantly higher than that of the PSPs, due to short duration, random orientation and fast decay with distance of EM fields, their electric potential and magnetic field strength outside of head are considered non-measurable by standard EEG and MEG devices. On the other hand, PSPs in pyramidal cells, if occurring synchronously in a large population of cells, can be recorded.

Pyramidal cells are the most common type of neural cells in the cerebral cortex. They are characterized by apical dendrites whose direction can be considered perpendicular to the surface of the cortex. Thus PSP potentials generated in these dendrites can be modeled with current dipoles with the same direction.

A current dipole can be seen as an electric current which is characterized by its position \mathbf{p} , and orientation and magnitude represented by its moment $\mathbf{q} = I d\boldsymbol{\theta}$ with units $[A \cdot m]$, where I is the current intensity and $d\boldsymbol{\theta}$ is an infinitesimal short vector between the current sink and source. The dipole current density at position \mathbf{p} can be written as

$$\mathbf{J}^{\mathbf{P}}(\mathbf{r}) = \mathbf{q}\delta(\mathbf{r} - \mathbf{p}) \quad (2.18)$$

where $\delta(\mathbf{r})$ is the Dirac delta function. Electric field lines of the current dipole start at a source and finish in a sink, while magnetic field lines correspond to concentric circles around $d\boldsymbol{\theta}$. The electric and magnetic field lines are illustrated in Figure 2.17.

Relations between the electric and magnetic fields and the current density are explained via Maxwell's equations, summarized in Table 2.2, where $\mathbf{E}[\frac{V}{m}]$ is the electric field, $\mathbf{B}[T]$ is the magnetic field, $\rho[\frac{C}{m^3}]$ is the charge density, $\mathbf{J}[\frac{A}{m^2}]$ is the current density, $\epsilon_0 = 8.85 \cdot 10^{-12} \frac{1}{kg \cdot m^3}$ is the vacuum permittivity and $\mu_0 = 4\pi \cdot 10^{-7} \frac{mkg}{s^2 A^2}$ is the vacuum permeability. $d\mathbf{r}$ is an infinitesimal volume element, $d\mathbf{s}$ and $d\mathbf{l}$ are infinitesimal vector elements of surface and contour.

From the Maxwell's equations, the charge conservation law can be derived as

$$\int_{\partial\Omega} \mathbf{J} \cdot d\mathbf{s} = - \int_{\Omega} \frac{\partial\rho}{\partial t} d\mathbf{r} \quad (2.19)$$

stating that change over time of the charge density is proportional to the flux of current density through the surface around that volume.

Due to the maximal frequency of the brain waves, but also permittivity and conductivity of brain tissues and head, time derivatives in Ampere's circuital law can be neglected [Hämäläinen *et al.* 1993]. This omitting of time derivatives is called *magneto-quasistatic* assumption [Hämäläinen *et al.* 1993]. Taking into account head

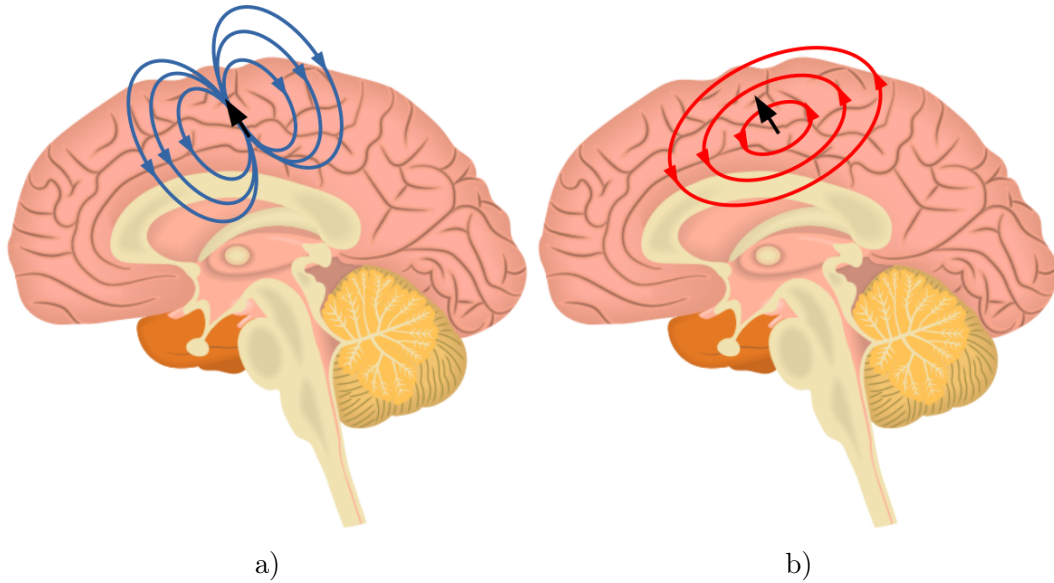


Figure 2.17: Illustrations of the dipole's electric field lines a) and magnetic field lines b).

Images adapted from :Mid-sagittal plane of the brain

Table 2.2: Integral formulae of the Maxwell's equations.

	Integral formulae	Meaning
Gauss's law	$\int_{\partial\Omega} \mathbf{E} \cdot d\mathbf{s} = \int_{\Omega} \frac{\rho}{\epsilon_0} d\mathbf{r}$	The flux of the electric field through any closed surface is proportional to the electric charge within the volume enclosed by this surface.
Gauss's law for magnetism	$\int_{\partial\Omega} \mathbf{B} \cdot d\mathbf{s} = 0$	The flux of the magnetic field through any surface is 0, meaning that the magnetic field is solenoidal vector field.
Faraday's law	$\int_{\partial S} \mathbf{E} \cdot d\mathbf{l} = \int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{s}$	The electromotive force in a contour around a surface is proportional to the change over time of the magnetic field flux through the surface.
Ampere's circuital law	$\int_{\partial S} \mathbf{B} \cdot d\mathbf{l} = \mu_0 \int_S (\mathbf{J} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t}) \cdot d\mathbf{s}$	The magnetic field line integral along a contour around a surface is proportional to total current passing through the surface.

dimensions, as well, leads to the *electro-quasistatic* assumption, where the time derivative in Faraday's law is also neglected [Hämäläinen *et al.* 1993]. With the

quasistatic approximations the Maxwell's equations can be written as in Table 2.3.

Table 2.3: Integral formulae of the quasistatic Maxwell's equations [Hämäläinen *et al.* 1993].

	Integral formulae
Gauss's law	$\int_{\partial\Omega} \mathbf{E} \cdot d\mathbf{s} = \int_{\Omega} \frac{\rho}{\epsilon_0} d\mathbf{r}$
Gauss's law for magnetism	$\int_{\partial\Omega} \mathbf{B} \cdot d\mathbf{s} = 0$
Faraday's law	$\int_{\partial S} \mathbf{E} \cdot d\mathbf{l} = 0$
Ampere's circuital law	$\int_{\partial S} \mathbf{B} \cdot d\mathbf{l} = \mu_0 \int_S \mathbf{J} ds$

A consequence of magneto-quasistatic assumption is that $\int_{\partial\Omega} \mathbf{J} \cdot d\mathbf{s} = 0$, meaning that the dependence of the electric field from the magnetic field can be neglected (from the Faraday's law in particular). On the other hand, the electro-quasistatic assumption neglects only the dependence of the magnetic field on the time varying electric field, while the impact of the electrostatic field which causes Ohmic currents cannot be neglected.

Due to the electro-quasistatic assumption, the electric field can be expressed as the gradient of a scalar function V also known as electrostatic potential as $\mathbf{E} = -\nabla V$. Since current dipoles associated to PSPs, also referred to as *primary currents* with current density \mathbf{J}^p , produce an electric field \mathbf{E} , this electric field produces *Ohmic currents* with current density $\sigma\mathbf{E} = -\sigma\nabla V$ where $\sigma[\frac{1}{\Omega \cdot m}]$ is the tissue conductivity. This means that the total current density is

$$\mathbf{J} = -\sigma\nabla V + \mathbf{J}^p. \quad (2.20)$$

Since we are interested only in the electric field potential generated by PSPs, using the quasistatic charge conservation law, we obtain a relation between the electric potential and the primary currents as

$$\nabla \cdot (\sigma\nabla V) = \nabla \cdot \mathbf{J}^p \quad (2.21)$$

which is a Poisson equation.

As given in Table 2.3, the magnetic field under magneto-quasistatic assumption is $\nabla \times \mathbf{B} = \mu_0\mathbf{J}$, thus $\nabla \times \nabla \times \mathbf{B} = \mu_0\nabla \times \mathbf{J}$ and $\Delta\mathbf{B} = -\mu_0\nabla \times \mathbf{J}$, where a solution is given by the Biot-Savart law as

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \mathbf{J}(\mathbf{r}') \times \frac{\mathbf{r} - \mathbf{r}'}{\|\mathbf{r} - \mathbf{r}'\|^3} d\mathbf{r}' = \frac{\mu_0}{4\pi} \int (\mathbf{J}^p(\mathbf{r}') - \sigma\nabla V(\mathbf{r}')) \times \frac{\mathbf{r} - \mathbf{r}'}{\|\mathbf{r} - \mathbf{r}'\|^3} d\mathbf{r}'. \quad (2.22)$$

From the equation describing the Biot-Savart law, we can see that the magnetic field depends both on the primary PSP and the secondary Ohmic currents.

Complexity of the solutions of the Poisson and Bio-Savart equations depends on the modeling of conductivity σ . The simplest model assumes that conductivity is

constant over all tissues. However, although different tissue conductivities impact both fields, this is more prominent in the case of the electric field due to the low conductivity of the skull. To address this, a model which represents tissues as layers with constant conductivities is proposed. The most advanced model so far assumes that the tissue conductivities are anisotropic and that they can be represented as tensors estimated using **dMRI**.

2.2.2.3 Electro-encephalography

Electro-encephalography (**EEG**) refers to the measuring of the previously described electric potentials arising from the cerebral cortex. Usually, it is performed in a non-invasive manner by placing multiple electrodes on the head, although intra-cranial **EEG** exists too. In order to be measurable on the head, the brain activity must occur synchronously in tens of thousands (≈ 50000) of spatially close pyramidal cells. Such activity in an adult human results in an electric potential in the range of $10 - 100\mu V$ [Aurlien *et al.* 2004]. Distribution of the electrodes over the skull is termed as a *montage*. The two most commonly used types of montage are bipolar and referential. In a bipolar montage, each channel of a multivariate **EEG** signal corresponds to the difference between signals recorded with adjacent electrodes. In a referential montage, from each electrode signal a reference signal is subtracted in order to obtain the final multivariate **EEG** signal. **EEG** signals exhibit very high temporal resolution which can be of the order of the *ms*. On the other hand, the spatial resolution is limited due to the low conductivity of the skull which causes smearing of the electric field. Depending on the number of electrodes, it is of the order of several cm^2 . Apart from the temporal resolution, other advantages of **EEG**, compared to other functional imaging methods, are the low price of the measuring device, its portability and lower storage requirements, and higher robustness to subject motion. In addition to the low spatial resolution, another significant disadvantage of **EEG** is the low signal to noise ratio, where noise comes from the activities of other organs, imperfections of the measuring devices, ambient, electrical sources, etc. Due to the superposition of electric fields ($\nabla \cdot \mathbf{J}^p = 0$ in Eq. 2.21), **EEG** devices have difficulties in recording signals from current dipoles organized into the forms close to solenoidal, whereas magnetic field is measurable [Hämäläinen *et al.* 1993, Grave de Peralta Menendez *et al.* 2000]. An illustration of the electric field lines of dipoles organized into a solenoidal form are illustrated in Figure 2.18 a).

2.2.2.4 Magneto-encephalography

Magneto-encephalography (**MEG**) refers to the measuring of the magnetic field strength arising from the cerebral cortex. This is achieved non-invasively via magnetometers placed at the scalp or slightly above it. As for **EEG**, synchronous activity of tens of thousands of spatially close pyramidal cells is required, so that the magnetic field is detectable by **MEG** device. Amplitudes of the field strength are in the range of $10 - 1000fT$, which is very low compared to the ambient noise of the order

of $10^8 fT$. As a consequence, MEG signals must be recorded in specially magnetically shielded rooms. The most commonly used MEG device is the superconducting quantum interference device (SQUID), which uses magnetometers based on superconducting coils. To achieve superconductivity, coils must be at low temperature. Thus a SQUID device includes a bulky cooling system. In addition, positions of the magnetometers are fixed, thus not well suited to heads of different geometries and sizes. Whereas a standard magnetometer contains a single coil, a special type of magnetometer termed as gradiometer uses multiple coils which allow noise reduction. More recent MEG devices are based on spin exchange relaxation-free (SERF) which use more compact optically pumped magnetometers. As they do not require a cooling system, they can be integrated into a portable helmet. As EEG, MEG signals exhibit very high temporal resolution which can be of order of the *ms*. Since tissue conductivity has a lower impact on the magnetic field, its spatial resolution is higher compared to the electric potential. The higher spatial resolution of the field supports the utilization of a higher number of magnetometers, in the range of 200 – 300. In addition, if MEG signal is recorded in a shielded room, the signal-to-noise ratio of MEG is higher compared to EEG signal. Since a current dipole perpendicular to a magnetometer coil, produces a magnetic field with circular lines parallel to the coil ($B(\mathbf{r}) = 0$ in Eq. 2.22), MEG devices have difficulties in recording signals from radial sources, such as at the top of gyri, whereas electric potential is measurable [Hämäläinen *et al.* 1993, Ahlfors *et al.* 2010]. An illustration of the magnetic field lines of dipoles organized into a radial form at the top of a gyrus are illustrated in Figure 2.18 b).

2.3 Conclusion

In this chapter, we have described the functional and structural properties of the human nervous system, at a micro-scale - the level of neurons and a macro-scale - the level of cortical lobes and white matter fiber tracts. They are presented for better comprehension of the functional and structural medical imaging techniques and their properties. Further, in more detail, we have provided a description of diffusing water molecules in different media and the way dMRI is able to capture structural properties of the examined tissues based on this phenomenon and magnetic properties of the water molecules. In a similar manner, for EEG and MEG techniques, we have firstly described biophysical events leading to PSPs which when occurring in synchronous manner, in the cerebral cortex, provoke measurable electric and magnetic fields, whose potential and strength can be recorded by EEG and MEG devices.

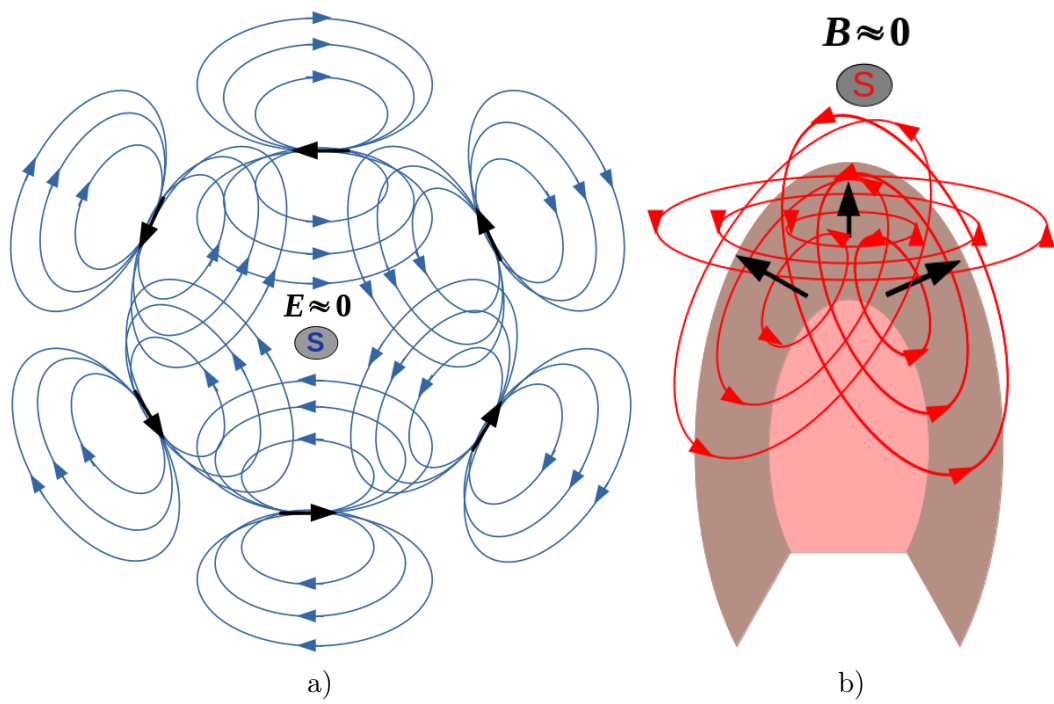


Figure 2.18: Illustrations of the electric field lines of the dipoles organized into a solenoidal form a) and magnetic field lines of the dipoles organized into a radial form at the top of a gyros b). (S denotes measuring sensor.)

Diffusion MRI local analysis

Contents

3.1	dMRI acquired on spheres	36
3.2	dMRI probability density functions	39
3.3	dMRI multi-compartment micro-structure imaging	42
3.4	Deep learning models for spherical signals	44
3.5	Deep learning models in dMRI local modeling	49
3.6	Conclusion	55

Executive summary

In this chapter, we firstly present the properties of the dMRI signals acquired with q-space sampling protocols, namely real and spherical nature, antipodal symmetry and rotation equivariance. Further, we provide an overview of the state-of-the-art dMRI local modeling approaches, which can be categorized into spherical probability density function (PDF)s and biophysically inspired multi-compartment micro-structure models. In the following section, the state-of-the-art deep learning models for the analysis of general spherical signals are presented. The last section contains a detailed description of the deep learning approaches used in local dMRI modeling.

3.1 dMRI acquired on spheres

dMRI signal acquisition with HARDI protocols has enabled the use of more insightful mathematical tools in the challenges, which include local modeling [Descoteaux *et al.* 2014]. The most prominent example is found in the modeling of crossing fibers which was impossible with low angular resolution dMRI signals, such as DTI [Basser *et al.* 1994]. In the last decade, the most commonly used HARDI protocols are single and multi-shell q-space sampling schemes [Jones 2010, Ye *et al.* 2012, Caruyer *et al.* 2013]. The shells correspond to concentric spheres in high-dimensional q-space. In the acquisition protocol proposed by [Caruyer *et al.* 2013], sampling points are randomly uniformly distributed and non-collinear within and between different shells in a way that the optimal angular coverage is achieved as illustrated in Figure 3.1.

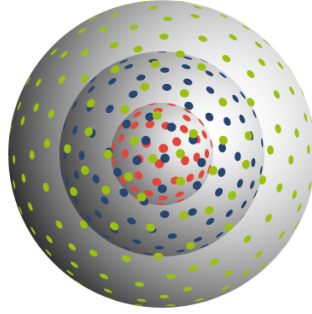


Figure 3.1: Illustration of a q-space sampling points over three shells [Caruyer *et al.* 2013]

Due to the nature of diffusion processes in the neural tissues, noiseless dMRI signals of an arbitrary shell are spherical, antipodally symmetric and real. This means that such a dMRI signal for a single shell, $s : S^2 \rightarrow \mathbb{R}$ can be represented as

$$s(\theta, \phi) = s(\mathbf{r}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{m=l} \hat{s}_{lm} Y_{lm}(\mathbf{r}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{m=l} \hat{s}_{lm} Y_{lm}(\theta, \phi) \quad (3.1)$$

where $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi)$ are colatitude and longitude, $\mathbf{r} \in \mathbb{R}^3$ s.t. $\mathbf{r} = [\sin \theta \cos \phi, \sin \theta \sin \phi, \cos \theta]^T$. \hat{s}_{lm} is a coefficient associated to the real SH basis element of degree l and order m - $Y_{lm} : S^2 \rightarrow \mathbb{R}$. By definition the SH basis are complex, but since we are dealing with the real dMRI signals, we have used a real SH basis, which can be defined using corresponding unitary matrices [Homeier & Steinborn 1996]. Definition of the complex and real SH bases is provided in Appendix A. Given the antipodal symmetry of the signal s , $s(\mathbf{r}) = s(-\mathbf{r})$, only antipodally symmetric SH basis elements are used, which are the elements of even degree l . dMRI signals are rotationally equivariant to the examined tissue structures which can have arbitrary 3D orientations. A function $f : S^2 \rightarrow \mathbb{R}$ is rotationally *equivariant* if the following holds

$$Q(f(\mathbf{r})) = f(Q\mathbf{r}) \quad (3.2)$$

where $\mathbf{r} \in \mathbb{R}^3$ and $Q \in SO(3)$ is a 3D rotation matrix. Another property of interest is rotation *invariance* which is a special case of rotation *equivariance*. A function $f : S^2 \rightarrow \mathbb{R}$ is rotationally *invariant* if the following holds

$$f(\mathbf{r}) = f(Q\mathbf{r}). \quad (3.3)$$

In reality, acquired signals are discrete and affected by noise. Noise which affects dMRI signals is non-additive and of Rician distribution. Due to discretization, they can be represented only with a finite number of SH basis elements. Given this, Eq. 4.1 becomes an approximation

$$s(\mathbf{r}_n) \approx \sum_{l=0}^B \sum_{m=-l}^{m=l} \hat{s}_{lm} Y_{lm}(\mathbf{r}_n) \quad (3.4)$$

where $\{\mathbf{r}_n\}_{n=1}^N$ is a discrete set of N points distributed over one shell, $\mathbf{r}_n \in \mathbb{R}^3$ s.t. $\|\mathbf{r}_n\|_2 = 1$ and B is the signal's bandwidth. This can be written in a matrix-vector notation as

$$\mathbf{s} \approx Y \hat{\mathbf{s}} \quad (3.5)$$

where $\mathbf{s} \in \mathbb{R}^N$ contains the discrete dMRI signal for one shell. $Y \in \mathbb{R}^{N \times N_B}$ is a matrix containing discrete SH basis elements in columns and $\hat{\mathbf{s}} \in \mathbb{R}^{N_B}$ is a vector containing the corresponding SH coefficients. $N_B = \frac{(B+1)(B+2)}{2}$ is the number of SH basis elements of even degrees.

Estimation of dMRI spherical harmonic coefficients

For more efficient processing and an insightful analysis of dMRI signals, it is often of interest to transform it to the Fourier/spectral domain. For signals acquired on a sphere, the Fourier basis is also called SH basis. A challenge in the computation of SH coefficients comes from the fact that there is no discretization process on a sphere which preserves the orthogonality of the SH basis. In analogy to the Nyquist-Shannon sampling theorem for band-limited signals acquired in Euclidean space, a number of sampling theorems for spherical signals have been proposed [Kowsky 1986, Driscoll & Healy 1994, McEwen & Wiaux 2011]. These theorems define sampling grids on spheres which guarantee that all the information from a band-limited spherical signal is preserved. Each sampling grid has a corresponding quadrature formula required for the exact computation of SH coefficients.

However, these sampling grids are not well suited to dMRI. They require a much higher number of sampling points (eg. $B(2B + 1) + 1$ at least for [McEwen & Wiaux 2011]), which is not practical from the clinical point of view. In addition, even if this number can be decreased by exploiting antipodal symmetry, the distribution of their points is not appropriate for signals affected by a significant noise as the sampling is in general dense around the poles and sparse close to the equator.

Coming back to Eq. 3.5, to estimate the SH coefficients $\hat{\mathbf{s}}$ from a signal \mathbf{s} , discretized

at a set of uniformly randomly distributed points, as in the q-space sampling, a number of least square based approaches have been proposed. They require at least $N_B = \frac{(B+1)(B+2)}{2}$ sampling points for a signal of bandwidth B . Initially, a least square solution was used by [Alexander *et al.* 2002, Tournier *et al.* 2004, Hess *et al.* 2006] where the SH coefficients are estimated using the Moore-Penrose pseudo-inverse as

$$\hat{\mathbf{s}} \approx Y_{mp}^\dagger \mathbf{s} = (Y^T Y)^{-1} Y^T \mathbf{s}. \quad (3.6)$$

This approach is very sensitive to noise and yields accurate solutions only for a number of points N much higher than the number of SH coefficients N_B ($N \gg N_B$). To address this problem, higher degree SH coefficients were directly apodized in [Tournier *et al.* 2004], while in [Hess *et al.* 2006] least square problem was regularized with a Tikhonov term, yielding the following

$$\hat{\mathbf{s}} \approx Y_{tikh}^\dagger \mathbf{s} = (Y^T Y + \lambda I)^{-1} Y^T \mathbf{s} \quad (3.7)$$

where λ is a regularization weight and I is the identity matrix of size N_B . Since Tikhonov regularization is not well suited for the S^2 basis (as the regularization term penalizes equally SH basis elements of all degrees), a least square solution with Laplace-Beltrami regularization was proposed by [Descoteaux *et al.* 2007] as follows

$$\hat{\mathbf{s}} \approx Y_{lb}^\dagger \mathbf{s} = (Y^T Y + \lambda L)^{-1} Y^T \mathbf{s} \quad (3.8)$$

where λ is a regularization weight and $L \in \mathbb{R}^{N_B \times N_B}$ is the Laplace-Beltrami smoothing matrix.

Convolution between spherical and zonal signals

As dMRI signals generated by individual neural tissue structures such as single axon bundles, gray matter and cerebrospinal fluid (CSF), at the level of a voxel, are usually assumed to be axially symmetric, it is often of interest to filter dMRI signal with a zonal signal (as it will be clear in the following sections). Zonal signals are a special case of axially symmetric signals, where the symmetry takes place around the z axis. They are also a special case of S^2 signals as they change only along the z axis (or along the inclination angle θ). An S^2 signal $z(\theta, \phi) : S^2 \rightarrow \mathbb{R}$ is a zonal signal iff $z(\theta, \phi) = z(\theta, 0) \forall \phi \in [0, 2\pi)$ and $\forall \theta \in [0, \pi)$. It can be represented in terms of SH and zonal harmonic (ZH) basis elements as

$$z(\theta, \phi) = z(\mathbf{r}) = \sum_{l=0}^{\infty} \hat{z}_{l0} Y_{l0}(\mathbf{r}) = \sum_{l=0}^{\infty} \hat{z}_{l0} Y_{l0}(\theta, \phi) = \sum_{l=0}^{\infty} \hat{z}_l \sqrt{\frac{(2l+1)}{4\pi}} P_l(\cos \theta). \quad (3.9)$$

where $P_l(\cos \theta)$ is the Legendre polynomial or the ZH basis element of degree l and \hat{z}_l is the corresponding coefficient, while \hat{z}_{l0} is the corresponding SH coefficient associated to the SH basis elements $Y_{l0}(\mathbf{r})$. Given an \mathbb{L}^2 signal $s : S^2 \rightarrow \mathbb{R}$ and an \mathbb{L}^2 zonal signal $g : S^2 \rightarrow \mathbb{R}$ of bandwidths B , correlation between them is given by

$$[s * g](\mathbf{r}) = \int_{S^2} s(\mathbf{r}') g(R^{-1}(\theta, \phi, 0)\mathbf{r}') d\mathbf{r}' = \sum_{l=0}^B \sqrt{\frac{4\pi}{2l+1}} \hat{g}_l \sum_{m=-l}^l Y_{lm}(\mathbf{r}) \hat{s}_{lm} \quad (3.10)$$

where $\mathbf{r} = [\sin \theta \cos \phi, \sin \theta \sin \phi, \cos \theta]^T$ and $R(\theta, \phi, 0) \in SO(3)$ is rotation matrix associated to \mathbf{r} . \hat{s}_{lm} is the SH coefficient of degree l and order m of the signal $s(\mathbf{r})$. \hat{g}_l is the ZH coefficient of degree l of the function $g(\mathbf{r})$. If $f(\mathbf{r}) = [s * g](\mathbf{r})$, from Eq. 3.10, its SH coefficients are defined as

$$\hat{f}_{lm} = \sqrt{\frac{4\pi}{2l+1}} \hat{g}_l \hat{s}_{lm} \quad \hat{\mathbf{f}}_l = \sqrt{\frac{4\pi}{2l+1}} \hat{g}_l \hat{\mathbf{s}}_l \quad (3.11)$$

where $\hat{\mathbf{s}}_l, \hat{\mathbf{f}}_l \in \mathbb{R}^{2l+1}$ are vectors which contain the SH coefficients of degree l of the signals $s(\mathbf{r})$ and $f(\mathbf{r})$.

3.2 dMRI probability density functions

One way to explain the HARDI dMRI signals is via 3D PDFs. These functions provide information related to the displacement of water molecules via diffusion within white matter axon bundles or orientation of the axon bundles themselves. They are examples of rotation equivariant functions (see Eq. 4.2). These voxel-wise PDFs opened the possibility of tracking white matter pathways as a process called tractography [Basser *et al.* 2000], which has a great use for the analysis of brain structural connectivity [Jbabdi *et al.* 2015].

Ensemble Average Propagator

The Ensemble Average Propagator (EAP) is a PDF which describes the probability of the water molecule displacement via diffusion in 3D space [Callaghan 1993]. If we denote the density of water molecules at position $\mathbf{R}_0 \in \mathbb{R}^3$ and time instant 0 with $\rho(\mathbf{R}_0)$ and the probability of a molecule displacement from \mathbf{R}_0 to position $\mathbf{R}_\Delta \in \mathbb{R}^3$ at time instant Δ with $P(\mathbf{R}_\Delta|\mathbf{R}_0)$, then the attenuation of the dMRI signal can be written as

$$\frac{s(\mathbf{q})}{s_0} = \int_{\mathbb{R}^3} \rho(\mathbf{R}_0) \int_{\mathbb{R}^3} P(\mathbf{R}_\Delta|\mathbf{R}_0) e^{2\pi i \mathbf{q}^T (\mathbf{R}_\Delta - \mathbf{R}_0)} d\mathbf{R}_\Delta d\mathbf{R}_0 = \int_{\mathbb{R}^3} P(\mathbf{R}) e^{2\pi i \mathbf{q}^T \mathbf{R}} d\mathbf{R} \quad (3.12)$$

where $s(\mathbf{q})$ is the dMRI signal measured at point $\mathbf{q} \in \mathbb{R}^3$ of the q -space and s_0 is the no diffusion weighted signal. $P(\mathbf{R})$ is the probability that a molecule is displaced by $\mathbf{R} = \mathbf{R}_\Delta - \mathbf{R}_0$. It is also known as the EAP. Δ is the interval between the encoding and decoding diffusion sensitizing gradients in direction $\frac{\mathbf{q}}{\|\mathbf{q}\|_2}$. \mathbf{q} is computed as

$$\mathbf{q} = \frac{1}{2\pi} \gamma \int_0^\delta \mathbf{G}(t) dt \quad (3.13)$$

where δ is the duration of diffusion sensitizing gradients. Under the narrow pulse assumption $\delta \ll \Delta$, we can assume that the movement of molecules within the

intervals δ can be neglected, and can also consider $\mathbf{G}(t)$ as a constant over that time. Thus $\mathbf{q} = \frac{1}{2\pi}\gamma\mathbf{G}\delta$. In this scenario, since \mathbf{q} has the same intensity and direction at time instants 0 and Δ , thus at \mathbf{R}_0 and \mathbf{R}_Δ , the EAP can be computed as the Fourier transform of the signal attenuation:

$$P(\mathbf{R}) = \int_{\mathbb{R}^3} \frac{s(\mathbf{q})}{s_0} e^{-2\pi i \mathbf{q}^T \mathbf{R}} d\mathbf{q}. \quad (3.14)$$

Units of EAP are $[\frac{1}{m^3}]$.

Diffusion Orientation Distribution Function

The Diffusion Orientation Distribution Function (dODF) is a PDF on the sphere which describes how water molecules diffuse along different directions. It is thus defined as the radial projection of the EAP. Initially, the dODF has been defined in [Tuch 2004] as

$$dODF(\mathbf{r}) = \frac{1}{Z} \int_0^\infty P(R\mathbf{r}) dR \quad (3.15)$$

where $\mathbf{r} \in \mathbb{R}^3$ s.t. $\|\mathbf{r}\|_2 = 1$ refers to the direction of diffusion and $R \in \mathbb{R}$ is its magnitude. Z is a dimensionless constant which ensures that the PDF $dODF(\mathbf{r})$ sums to one. Since the EAP $P(\mathbf{R})$ actually corresponds to the probability that a water molecule initially placed at origin \mathbf{R}_0 is found in an infinitesimal volume $d\mathbf{R}$ at position \mathbf{R}_Δ after time Δ , in [Wedeen *et al.* 2005] a better grounded definition of $dODF(\mathbf{r})$ has been introduced as

$$dODF(\mathbf{r}) = \int_{\mathbb{R}^3} P(R\mathbf{r}) d\mathbf{R} \quad (3.16)$$

which by representing $d\mathbf{R}$ by $R^2 dR d\Omega$ where $d\Omega$ is infinitesimal solid angle element can be written as

$$dODF(\mathbf{r}) = \frac{1}{4\pi} \int_0^\infty \int_{S^2} P(R\mathbf{r}) R^2 d\Omega dR = \int_0^\infty P(R\mathbf{r}) R^2 dR \quad (3.17)$$

where S^2 is the unit sphere.

In [Descoteaux *et al.* 2007], the authors proposed an analytical solution for dODF approximation from dMRI signals acquired on spheres of q-space. The dODF is obtained as the convolution between a zonal function obtained via the Funk-Hecke theorem and the SH coefficients estimated solving the least square problem with a Laplace-Beltrami regularization as in Eq. 3.8. The convolution is defined as

$$dODF(\mathbf{r}) = \sum_{l=0}^B 2\pi P_l(0) \sum_{m=-l}^l \hat{s}_{ml} Y_{ml}(\mathbf{r}) \quad (3.18)$$

where \hat{s}_{ml} are the real SH coefficients of the dMRI attenuation $s(\mathbf{r})/s_0$ and $P_l(0)$ is the Legendre polynomial of degree l evaluated at $\cos\theta = 0$.

Fiber Orientation Distribution Function

The fODF is a spherical PDF which provides information on the orientation and volume fractions of the axon bundles [Tournier *et al.* 2004, Tournier *et al.* 2007, Jeurissen *et al.* 2014]. Whereas the EAP and the dODF are referred to as "model free", the fODF requires the modeling of a response function corresponding to a single axon bundle. Given the single fiber response function r^{sf} , the fODF is computed by the deconvolution of r^{sf} from the dMRI signal. In the first approach for fODF estimation proposed in [Tournier *et al.* 2004], the dMRI signals were modeled as the convolution between the $fODF : S^2 \rightarrow \mathbb{R}^+$ and a zonal single fiber response function $r^{sf}(\theta)$ as

$$s(\mathbf{r}) = [fODF * r^{sf}](\mathbf{r}) \quad (3.19)$$

where the response function $r^{sf}(\theta)$ is obtained from voxels which are determined as the ones that most probably contain single white matter fibers according to certain rotation invariant measures. As these bundles might have an arbitrary orientation, they are firstly rotated to be zonal and averaged in order to obtain $r^{sf}(\theta)$. In the spectral domain, as given in Eq. 3.11, the convolution from Eq. 3.19 corresponds to

$$\hat{s}_l = \sqrt{\frac{4\pi}{2l+1}} \hat{\mathbf{f}}_l \hat{r}_l^{sf} \quad (3.20)$$

where $\hat{s}_l, \hat{\mathbf{f}}_l \in \mathbb{R}^{2l+1}$ are vectors containing the SH coefficients of degree l of the dMRI signal $s(\mathbf{r})$ and $fODF(\mathbf{r})$. $\hat{r}_l^{sf} \in \mathbb{R}$ is the ZH coefficients of degree l of a single fiber response function $r^{sf}(\theta)$. From Eq. 3.20, we can see that the spectral coefficients of the $fODF(\mathbf{r})$ can be simply obtained by deconvolution as

$$\hat{\mathbf{f}}_l = \sqrt{\frac{2l+1}{4\pi}} \hat{s}_l \frac{1}{\hat{r}_l^{sf}}. \quad (3.21)$$

where a least mean square solution from Eq. 3.6 is used to estimate the SH coefficients of the dMRI signals. Since deconvolution from Eq. 3.21 is susceptible to noise and does not take into account the fact that some voxels contain gray matter or CSF tissues, negative spurious peaks might appear in the estimated fODF. To address this problem, an fODF estimation by deconvolution with non-negativity constraint has been proposed in [Tournier *et al.* 2007]. The minimization problem is defined as

$$\hat{\mathbf{f}} = \arg \min_{\hat{\mathbf{f}}} \|C\hat{\mathbf{f}} - \mathbf{s}\|_2^2 \quad \text{s.t.} \quad A\hat{\mathbf{f}} \geq 0 \quad (3.22)$$

where $\hat{\mathbf{f}}$ are the SH coefficients of $fODF(\mathbf{r})$. The matrix C incorporates convolution of the fODF with response function $r^{sf}(\theta)$ in the spectral domain and the transformation of the resulting SH coefficients into the S^2 domain at the same sampling points as of the signal $s(\mathbf{r})$. The matrix A transforms the SH coefficients $\hat{\mathbf{f}}$ into the S^2 domain on a very dense sampling grid in order to impose the positivity constraint. This approach is termed as *single shell single tissue* constraint spherical deconvolution - SSST-CSD. Since it is designed only for single shell signals and

does not take into account the presence of non-white matter tissues in a voxel, it is further extended into the *multi shell multi tissue* constraint spherical deconvolution - MSMT-CSD [Jeurissen *et al.* 2014], which in addition to white matter fODF provides information on gray matter and CSF volume fractions. The MSMT-CSD minimization problem is defined as

$$\begin{bmatrix} \hat{\mathbf{f}}_1 \\ \hat{\mathbf{f}}_2 \\ \vdots \\ \hat{\mathbf{f}}_n \end{bmatrix} = \arg \min_{\begin{bmatrix} \hat{\mathbf{f}}_1 \\ \hat{\mathbf{f}}_2 \\ \vdots \\ \hat{\mathbf{f}}_n \end{bmatrix}} \left\| \begin{bmatrix} C_{1,1} & \dots & C_{1,n} \\ C_{2,1} & \dots & C_{2,n} \\ \vdots & \dots & \vdots \\ C_m & \dots & C_{m,n} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{f}}_1 \\ \hat{\mathbf{f}}_2 \\ \vdots \\ \hat{\mathbf{f}}_n \end{bmatrix} - \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_m \end{bmatrix} \right\|_2^2 \quad \text{s.t.} \quad \begin{bmatrix} A_1 & \dots & 0 \\ 0 & \dots & 0 \\ \vdots & \dots & \vdots \\ 0 & \dots & A_n \end{bmatrix} \begin{bmatrix} \hat{\mathbf{f}}_1 \\ \hat{\mathbf{f}}_2 \\ \vdots \\ \hat{\mathbf{f}}_n \end{bmatrix} \geq 0 \quad (3.23)$$

where m is the number of shells and n is the number of tissues. \mathbf{s}_i is dMRI signal of shell i and $\hat{\mathbf{f}}_j$ are the SH coefficients of the spherical PDF of tissue j . C_{ji} is a matrix which incorporates the convolution of $\hat{\mathbf{f}}_j$ with the response function of tissue j at shell i , $r_i^j(\theta)$, in the spectral domain and the transformation of the resulting SH coefficients into the S^2 domain at the same sampling points as of the signal \mathbf{s}_i . The obtained reconstructed signals are summed over all tissue types j for the shell i in order to fit it to \mathbf{s}_i . The matrix A_j transforms the SH coefficients $\hat{\mathbf{f}}_j$ into the S^2 domain in order to impose the positivity constraint for the spherical PDF of each tissue type. Since the response functions for gray matter and CSF are spherical (have bandwidth 0), A_j does not need to transform these PDFs on a large number of sampling points. For white matter tissue where bandwidth of response function is much higher a high number of sampling points is needed in order to ensure positivity of the fODF.

Both minimization problems from Eq. 3.22 and Eq. 3.23 can be represented as convex quadratic programming problems which can be solved efficiently [Jeurissen *et al.* 2014].

3.3 dMRI multi-compartment micro-structure imaging

Multi-compartment micro-structure (MCMS) imaging refers to biophysically inspired models which explain the dMRI signal as a linear combination of signals coming from different tissue compartments such as intra- and extra-axonal spaces, gray matter, cerebrospinal fluid, tumorous cell, etc. These models can provide information about axonal density and diameter, neurite dispersion, different tissue volume fractions, which are rotationally invariant measures (see Eq. 3.3), which have shown potential in the evaluation of several neurological diseases [Panagiotaki *et al.* 2014, De Santis *et al.* 2017, Schneider *et al.* 2017, Broad *et al.* 2018] and in characterization of early brain development [Jelescu *et al.* 2015, Bastiani *et al.* 2019].

We provide an overview of the most distinct MCMS models.

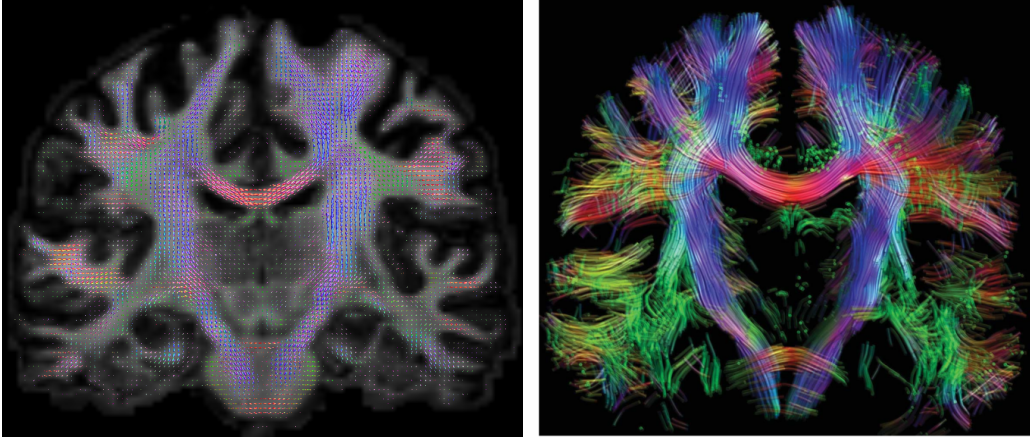


Figure 3.2: Illustration of fODFs and tractography [Tournier *et al.* 2011].

Ball and Stick [Behrens *et al.* 2003, Behrens *et al.* 2007] models the dMRI signal as a linear combination of an isotropic Gaussian (ball) which corresponds to signal generated by extra-axonal water molecule diffusion and N anisotropic diffusion tensors without radial diffusivity (zero radius sticks) for intra-axonal diffusion as

$$s_i = s_0 \left(\nu_0 e^{-b_i d} + \sum_{n=1}^N \nu_n e^{-b_i \mathbf{d} \mathbf{r}_i^T R_n A R_n^T \mathbf{r}_i} \right) \quad (3.24)$$

where s_i is the dMRI signal measured along direction \mathbf{r}_i with a b-value b_i and s_0 is the no diffusion weighted signal. d is diffusivity and $R_n A R_n^T$ is the anisotropic diffusion tensor of the n^{th} fiber. ν_0 and $\{\nu_n\}_{n=1}^N$ are volume fractions of the isotropic and the N fiber compartments.

Composite Hindered And Restricted Model of Diffusion (CHARMED) [Assaf *et al.* 2004, Assaf & Basser 2005] models dMRI generated by white matter tissue as a linear combination of signals generated by hindered and restricted compartments. The former corresponding to between axons diffusion modeled with diffusion tensor and the latter to intra-axonal diffusion modeled with a cylinder as

$$\begin{aligned} s_i &= s_0 \left(\nu_h e^{-4\pi^2(\Delta-\delta/3)\mathbf{q}_i^T D \mathbf{q}_i} + \sum_{n=1}^N \nu_r^n E_h(\mathbf{q}_i, \Delta) \right) \\ &= s_0 \left(\nu_h e^{-4\pi^2(\Delta-\delta/3)\mathbf{q}_i^T D \mathbf{q}_i} + \sum_{n=1}^N \nu_r^n E_h^{\parallel}(\mathbf{q}_i^{n,\parallel}, \Delta) E_h^{\perp}(\mathbf{q}_i^{n,\perp}, \Delta) \right) \end{aligned} \quad (3.25)$$

where D is the effective diffusion tensor. s_i is the dMRI signal measure at point \mathbf{q}_i and s_0 is the no diffusion weighted signal. $\mathbf{q}_i^{n,\parallel}$ and $\mathbf{q}_i^{n,\perp}$ are the parallel and perpendicular components of \mathbf{q}_i with respect to the n^{th} axon bundle. $E_h^{\parallel}(\mathbf{q}_i^{n,\parallel}, \Delta)$ and $E_h^{\perp}(\mathbf{q}_i^{n,\perp}, \Delta)$ are the intra-axonal attenuation factors coming from parallel and perpendicular diffusion within the axon bundle. ν_h and $\{\nu_r^n\}_{n=1}^N$ are the volume fractions of the hindered and the N restricted compartments.

Neurite Orientation Dispersion and Density Imaging (NODDI) [Zhang *et al.* 2012] models the dMRI signal as a linear combination of three types of compartments. The CSF compartments is modeled with an isotropic Gaussian (ball), while the signals from intra- and extra-neurite spaces are modeled with zero radius cylinders (sticks) distributed respectively to a Watson distribution and an anisotropic Gaussian (zeppelin) whose diffusion tensor corresponds to Watson distributed neurites as

$$s_i = s_0 \left(\nu_{iso} e^{-b_i d_{iso}} + (1 - \nu_{iso}) (\nu_{in} E_{in}(\mathbf{q}_i, d_{\parallel}) + \nu_{en} E_{en}(\mathbf{q}_i, d_{\perp}, d_{\parallel})) \right) \quad (3.26)$$

where s_i is the dMRI signal measured at point \mathbf{q}_i and s_0 is the no diffusion weighted signal. b_i is the b-value corresponding to \mathbf{q}_i . ν_{iso} is the CSF volume fraction and ν_{in} and ν_{en} are the intra and extra-neurite volume fractions with respect to non-isotropic contributions. d_{iso} , d_{\parallel} and d_{\perp} are isotropic, parallel and perpendicular diffusivities. Parallel diffusivities of intra- and extra-neurite compartments are the same, while the perpendicular diffusivity of the extra-neurite compartment is related to parallel diffusivity via the tortuosity model [Szafer *et al.* 1995] as $d_{\perp} = d_{\parallel}(1 - \nu_{in})$. Signal attenuation due to intra and extra-neurite diffusions are defined as

$$E_{in}(\mathbf{q}_i, d_{\parallel}) = \int_{S^2} W(\mathbf{r}, \kappa, \mu) e^{-b_i d_{\parallel} (\mathbf{q}_i^T \mathbf{r})^2} d\mathbf{r} \quad (3.27)$$

and

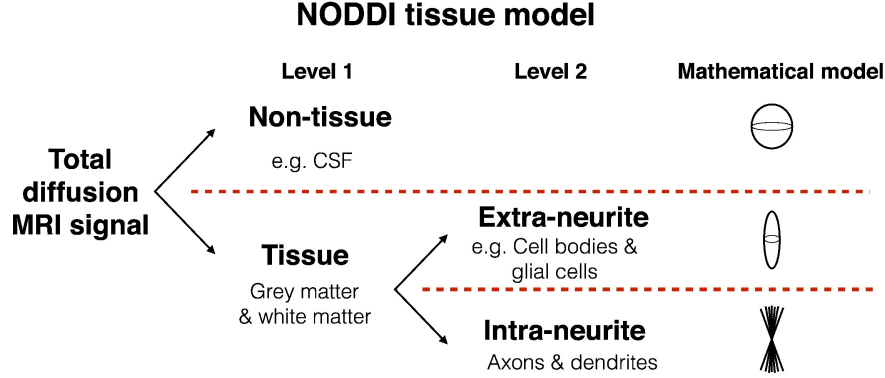
$$E_{en}(\mathbf{q}_i, d_{\perp}, d_{\parallel}) = e^{-b_i \mathbf{q}_i^T D_{en} \mathbf{q}_i} \quad \text{where} \quad D_{en} = \int_{S^2} W(\mathbf{r}, \kappa, \mu) D(\mathbf{r}) d\mathbf{r} \quad (3.28)$$

where $W(\mathbf{r}, \kappa, \mu)$ is the orientation distribution function (axially symmetric), where μ is its orientation and κ determines dispersion around μ . κ is used to define the orientation dispersion index as $OD = \frac{2}{\pi} \arctan(\frac{1}{\kappa})$ whose range is in $[0, 1]$. $D(\mathbf{r})$ is cylindrical diffusion tensor with orientation \mathbf{r} with parallel and perpendicular diffusivities d_{\parallel} and d_{\perp} .

3.4 Deep learning models for spherical signals

A number of 3D rotationally equivariant general purpose deep learning (DL) approaches have been proposed for the analysis of arbitrary S^2 signals. Among the first notable rotationally equivariant neural networks is the $S^2 - CNN$ proposed by [Cohen *et al.* 2018]. The main contribution of their work are the layers with convolutions (correlations) performed in the S^2 and $SO(3)$ spectral domain [Driscoll & Healy 1994, Kostelec & Rockmore 2008], so that the computationally expensive interpolations in the signal space are avoided.

In the first convolutional layer, given an input data sample $f : S^2 \rightarrow \mathbb{C}$ and a trainable kernel $\psi : S^2 \rightarrow \mathbb{C}$ which is sampled at circles around pole (otherwise

Figure 3.3: NODDI compartments [Tariq *et al.* 2016].

is zero), both sampled at a Driscoll-Healy grid [Driscoll & Healy 1994], the SH coefficients $\{\{\hat{f}_l^m\}_{m=-l}^{m=l}\}_{l=0}^B$ and $\{\{\hat{\psi}_l^m\}_{m=-l}^{m=l}\}_{l=0}^B$ are first computed using the corresponding quadrature formulae [Driscoll & Healy 1994]. Convolution (correlation) is performed as follows

$$\begin{aligned}
 g(R) &= [f * \psi^*](R) = \int_{S^2} f(\mathbf{r})\psi^*(R^{-1}\mathbf{r})d\mathbf{r} \\
 &= \sum_{l=0}^B \sum_{m=-l}^l \sum_{n=-l}^l D_l^{mn}(R)\hat{f}_l^m\hat{\psi}_l^{n*} = \sum_{l=0}^B \sum_{m=-l}^l \sum_{n=-l}^l D_l^{mn}(R)\hat{G}_l^{m,n}
 \end{aligned} \tag{3.29}$$

where $R \in SO(3)$ is a rotation matrix. $D_l^{mn}(R)$ is the Wigner-D basis element of degree l and orders m and n which is a Fourier basis element of the $SO(3)$ manifold and $\hat{G}_l^{m,n}$ is the corresponding rotation harmonic (RH) coefficient. As $\hat{G}_l^{m,n} = \hat{f}_l^m\hat{\psi}_l^{n*}$, in matrix-vector notation we can write $\hat{G}_l = \hat{\mathbf{f}}_l\hat{\psi}_l^*$, where $\hat{\mathbf{f}}_l, \hat{\psi}_l \in \mathbb{C}^{(2l+1)}$ are the SH coefficients of degree l of the signal $f(\mathbf{r})$ and kernel $\psi(\mathbf{r})$, respectively. $\hat{G}_l \in \mathbb{C}^{(2l+1) \times (2l+1)}$ are the RH coefficients of degree l of the resulting signal $g(R)$. This is illustrated in Figure 3.4 a). The full derivation of the convolution between two S^2 signals is given in Appendix A. As shown in Eq. 3.29, after convolution in the spectral domain, the signal in the $SO(3)$ domain is obtained as a linear combination of Wigner-D basis elements. Then, the spectral coefficients are projected back onto the equiangular $SO(3)$ sampling grid analogue to the Driscoll-Healy grid used for the discretization of S^2 signals and the Rectified Linear Unit (ReLU) non-linearity is applied. As the convolution of two S^2 signals gives a signal in $SO(3)$ manifold, all layers following the first one perform a convolution between $SO(3)$ signals and kernels after the ReLU non-linearity is applied in the $SO(3)$ domain.

In the i^{th} convolutional layer with $i > 1$, given the input $SO(3)$ signal and kernel $f, \psi : SO(3) \rightarrow \mathbb{C}$, both sampled at equiangular grids which allow the computation of the respective RH coefficients using quadrature formulae denoted as $\{\{\{\hat{F}_l^{m,n}\}_{m=-l}^{m=l}\}_{n=-l}^{n=l}\}_{l=0}^B$ and $\{\{\{\hat{\Psi}_l^{m,n}\}_{m=-l}^{m=l}\}_{n=-l}^{n=l}\}_{l=0}^B$, the convolution (correlation)

is performed as

$$\begin{aligned}
 g(R) &= [f * \psi^*](R) = \int_{SO(3)} f(Q)\psi^*(R^{-1}Q)dQ \\
 &= \sum_{l=0}^B \sum_{m=-l}^l \sum_{n=-l}^l D_l^{mn}(R) \sum_{k=-l}^l \hat{F}_l^{mk} \hat{\Psi}_l^{nk*} = \sum_{l=0}^B \sum_{m=-l}^l \sum_{n=-l}^l D_l^{mn}(R) \hat{G}_l^{mn}
 \end{aligned} \tag{3.30}$$

where $R, Q \in SO(3)$. \hat{F}_l^{pq} and $\hat{\Psi}_l^{pq}$ are the RH coefficients of degree l and orders p and q of the signal $f(R)$ and kernel $\psi(R)$, respectively. $D_l^{pq} : SO(3) \rightarrow \mathbb{C}$ is an element of the Wigner-D matrix of degree l and orders p and q . As $\hat{G}_l^{mn} = \sum_{k=-l}^l \hat{F}_l^{mk} \hat{\Psi}_l^{nk*}$, in matrix notation we can write $\hat{G}_l = \hat{F}_l \hat{\Psi}_l^*$, where $\hat{F}_l, \hat{\Psi}_l, \hat{G}_l \in \mathbb{C}^{(2l+1) \times (2l+1)}$ are the RH coefficients of degree l of $f(R)$, $\psi(R)$ and $g(R)$, respectively. This is illustrated in Figure 3.4 b). Full derivation of the convolution between two $SO(3)$ signals is given in Appendix A. As after the first convolutional layer, ReLU is applied in $SO(3)$ domain.

As in standard Euclidean space CNNs, pooling layers are important as their task is to summarize feature maps by decreasing its resolution (e.g. with max or average pooling). In $S^2 - CNN$, this is achieved simply by discarding the RH coefficients of the highest degree after each ReLU. After the last convolutional layer and non-linearity, only the RH coefficients of degree $l = 0$ are extracted and fed to a chain of fully connected layers whose task is to perform the final inference based on the extracted features, such as regression or classification.

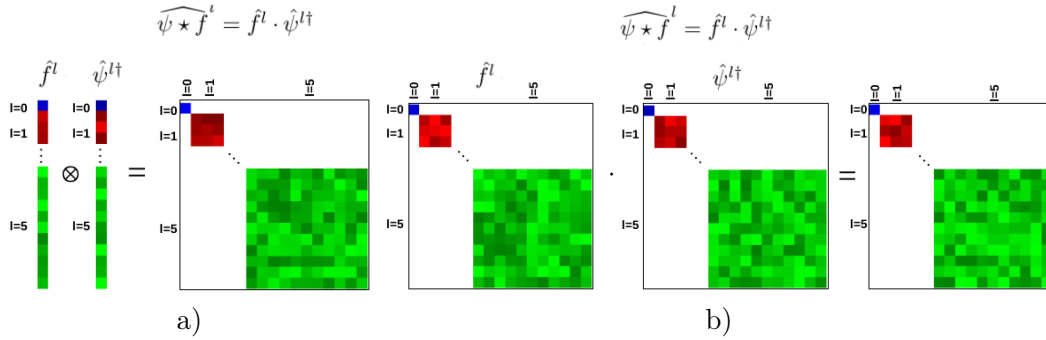


Figure 3.4: Illustration of convolutions in spectral domain between a) two S^2 signals and b) two $SO(3)$ signals.

As the transformations between the $SO(3)$ spectral and signal domains and vice versa are computationally expensive, [Esteves *et al.* 2018] have proposed a spherical CNN model with zonal kernels. In this case the convolution between the S^2 signals and zonal kernels remains in the S^2 space, which is less computationally expensive. The convolution between the input signal and a trainable kernel is illustrated in Figure 3.5 and is performed in the spectral domain as given in Eq. 3.11. As in $S^2 - CNN$ proposed by [Cohen *et al.* 2018], the ReLU non-linearity is applied in the signal domain, that is S^2 in this model and pooling is performed by discarding

the SH coefficients of the highest degree. Finally, as in [Cohen *et al.* 2018], feature maps of degree $l = 0$ are extracted after the last convolutional layer and are fed into a fully connected network.

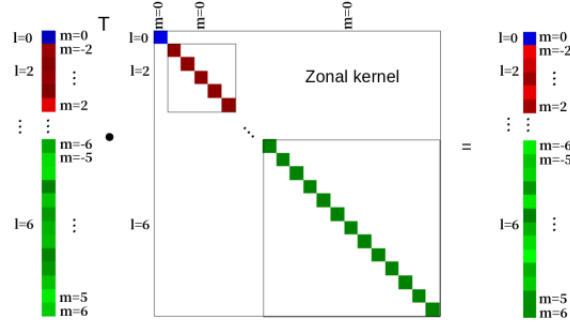


Figure 3.5: Illustration of convolutions in spectral domain between an S^2 signal and a zonal kernel.

An issue that arises from the application of the non-linearity in the signal domain is the appearance of high frequency components, which might introduce aliasing and decrease rotation equivariance of the model. In the work presented by [Kondor *et al.* 2018], a fully Fourier space CNN has been proposed, where rotation invariant Fourier domain non-linearities of quadratic nature have been introduced, thus eliminating the need for conversion from spectral to signal domain and distortions introduced by aliasing. This is achieved by decomposing the tensor product of $SO(3)$ covariant vectors into irreducible fragments (vectors) using the Clebsch-Gordan decomposition. Given an input data sample $f : S^2 \rightarrow \mathbb{C}$ sampled at Driscoll-Healy grid [Driscoll & Healy 1994] or Gauss-Legendre grid, firstly the SH coefficients $\{\{\hat{f}_l^m\}_{m=-l}^{m=l}\}_{l=0}^B$ and $\{\{\hat{\psi}_l^m\}_{m=-l}^{m=l}\}_{l=0}^B$ are computed using corresponding quadrature formulae. We denote with $\mathbf{f}_l \in \mathbb{C}^{(2l+1)}$ vector of the SH coefficients of degree l , here also referred to as the $SO(3)$ covariant vectors or fragments. If there are multiple input channels, we denote with $F_l \in \mathbb{C}^{(2l+1) \times C}$ the matrix which contains the SH coefficients of each of the C channels. The authors proposed a Fourier domain non-linearity achieved via the Clebsch-Gordan decompositions as

$$G_l = \bigsqcup_{|l_1 - l_2| \leq l \leq |l_1 + l_2|} C_{l_1, l_2, l}^T [F_{l_1} \otimes F_{l_2}] \quad (3.31)$$

where $C_{l_1, l_2, l}^T \in \mathbb{R}^{(2l_1+1)(2l_2+1) \times (2l+1)}$ is a sparse matrix containing the Clebsch-Gordan coefficients which are non-zero only for $m_1 + m_2 = m$, where m_1 , m_2 and m are the orders of the SH coefficients in fragments of degrees l_1 , l_2 and l . \bigsqcup refers to concatenation over channels. We can notice that with this type of non-linearity the total number of channels is squared, which is addressed by a covariant linear transformation defined as

$$H_l = G_l W_l \quad (3.32)$$

where $W_l \in \mathbb{C}^{C \times Q}$ where $Q < C$. This can be seen as filtering the channels with different zonal kernels and their sum. Illustration of a single layer of a Clebsch-Gordan network containing a Clebsch-Gordan non-linearity and a linear transform is shown in Figure 3.6. In the final layer, only H_0 are computed and fed into a fully connected network as previously described for models [Cohen *et al.* 2018, Esteves *et al.* 2018].

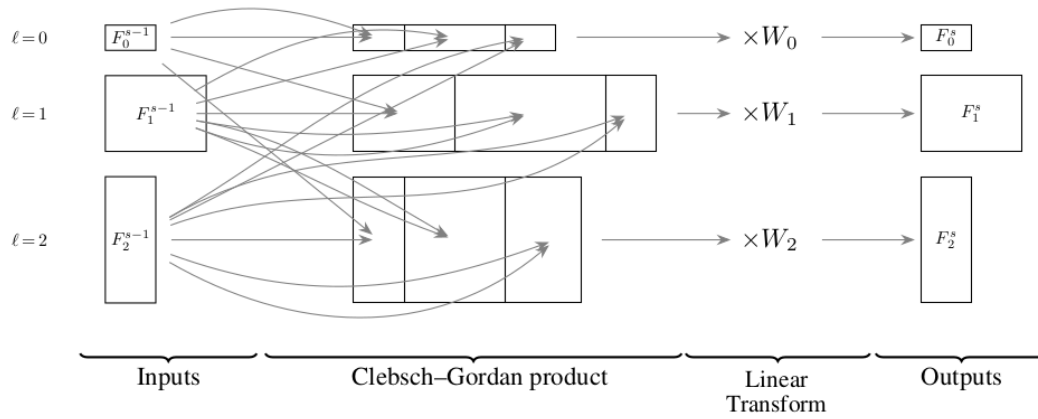


Figure 3.6: Illustration of a single layer of Clebsch-Gordan network [Kondor *et al.* 2018].

3.5 Deep learning models in dMRI local modeling

To address some of the problems in dMRI local modeling, as in other computer vision domains, focus has moved towards data driven approaches, such as DL which have been recognized as a powerful tool to extract information from diffusion MRI signals.

Among the first DL models adapted to address the problem of the estimation of micro-structure parameters from dMRI data acquired with clinically desirable acquisition schemes (containing a low number of sampling points) was the multi layer perceptron (MLP) [Golkov *et al.* 2016]. The model was composed of fully connected layers with trainable weights and biases $\{W_i\}_{i=1}^L$ and $\{\mathbf{b}_i\}_{i=1}^L$, where L is the total number of layers. Each layer maps the input signal \mathbf{s}_{i-1} to the output as $\mathbf{a}_i = g_i(W_i \mathbf{s}_{i-1} + \mathbf{b}_i)$, where g_i is an activation function of the i^{th} layer. Except for g_L which is identity, all previous layers used a ReLU non-linearity. To reduce the effect of overfitting, the authors proposed to use drop-out regularization. The model was successfully evaluated on the problem of diffusion kurtosis imaging and neurite orientation dispersion and density imaging (NODDI) parameter estimation. MLP models have also been investigated in the context of the estimation of rotationally invariant features (RIFs) [Zucchelli *et al.* 2020] from different dMRI signal representations [Zucchelli *et al.* 2021].

In the work of [Ye 2017], an iterative hard thresholding (IHT) algorithm [Blumensath & Davies 2009], used as a solution of sparse reconstruction problem, has been unfolded into a DL approach specifically designed for NODDI parameter estimation. The model was termed as Microstructure Estimation using a Deep Network (MEDN). It is composed of two stages. Its architecture is illustrated in Figure 3.7.

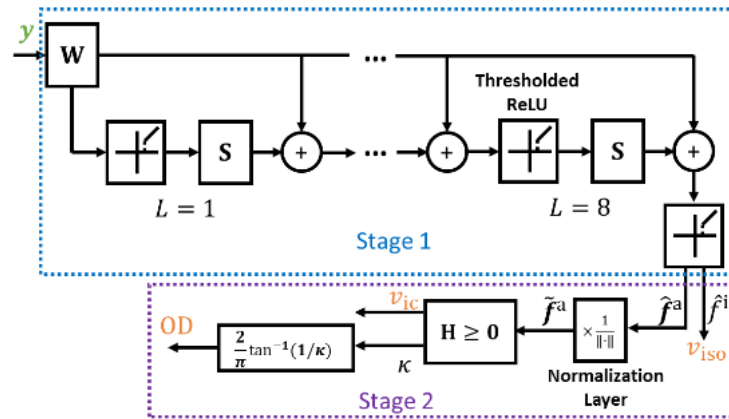


Figure 3.7: Illustration of the MEDN architecture (image source: [Ye 2017]).

In IHT, sparse codes \mathbf{x}^{t+1} are computed as

$$\mathbf{x}^{t+1} = h_\lambda(W\mathbf{y} + S\mathbf{x}^t) \quad (3.33)$$

where \mathbf{y} is the input signal, W is the dictionary of atoms, $S = I - WW^T$, \mathbf{x}^t is the sparse reconstruction at the t^{th} iteration, $\mathbf{x}^0 = 0$ and h_λ is the thresholding operator defined as $h_\lambda(x) = x$ if $x > \lambda$ and $h_\lambda(x) = 0$ otherwise. In MEDN, instead of using a predefined dictionary, the matrices W and S are learned via backpropagation independently. In the second stage, the isotropic volume fraction ν_{iso} corresponds to the last entry of \mathbf{x} , while the previous entries denoted as \mathbf{x}^a correspond to anisotropic compartments. For numerical stability \mathbf{x}^a are firstly normalized as $\tilde{\mathbf{x}}^a = (\mathbf{x}^a + \tau \mathbf{1}) / \|\mathbf{x}^a + \tau \mathbf{1}\|_1$, where $\tau = 10^{-10}$. Finally, the intra-cellular parameter ν_{ic} and the parameter κ associated to the Watson distribution are estimated as $[\nu_{ic}, \kappa]^T = H \tilde{\mathbf{x}}^a$, where H is also a trainable matrix. The orientation dispersion index is obtained as $OD = \frac{2}{\pi} \arctan(\frac{1}{\kappa})$.

The authors of MEDN proposed in [Ye *et al.* 2019] another DL model, inspired by the IHT algorithm, based on a modified long-short term memory (LSTM) units, which is capable to incorporate information from the neighborhood voxels for the estimation of micro-structure parameters. The model is termed as Microstructure Estimation with Sparse Coding Net (MESCNet). It is composed of two stages and its architecture is illustrated in Figure 3.8. Contrarily to MEDN, MESCNet is designed for the estimation of arbitrary micro-structure parameters.

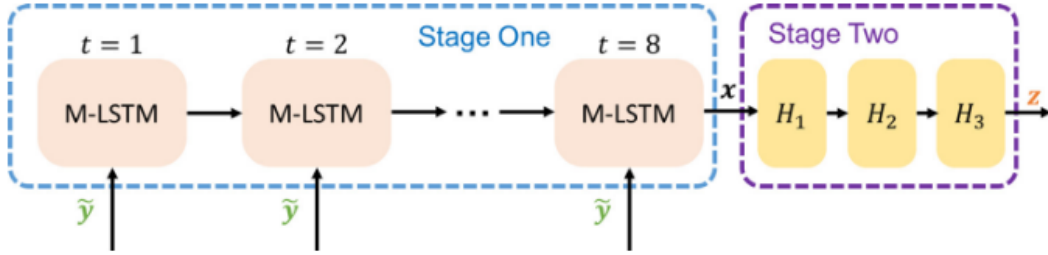


Figure 3.8: Illustration of MESCNet architecture (image source: [Ye *et al.* 2019]).

In the first stage, given the input signals $\mathbf{y} = [\mathbf{y}_1^T, \dots, \mathbf{y}_{nb}^T]^T$, where \mathbf{y}_i is the dMRI signal in voxel i and nb is the total number of voxels in a cubic neighborhood, the estimation of the sparse coefficients \mathbf{x} in the t^{th} layer is given by

$$\mathbf{x}^t = h_\lambda(\mathbf{c}^t) \quad \text{where} \quad \mathbf{c}^t = \mathbf{f}^t \circ \mathbf{c}^{t-1} + \mathbf{i}^t \circ \tilde{\mathbf{c}}^t \quad \text{where} \quad \tilde{\mathbf{c}}^t = W\mathbf{y} + S\mathbf{x}^{t-1} \quad (3.34)$$

where as in MEDN W and S are trainable parameters. \mathbf{f}^t and \mathbf{i}^t are respectively the weighting terms of coefficients from previous layer \mathbf{c}^{t-1} and an intermediate estimate of the coefficients from the current layer $\tilde{\mathbf{c}}^t$. $\mathbf{x}^0 = 0$. \circ refers to element wise multiplication. Comparing the sparse vector estimations in MEDN and in MESCNet, given in equations 3.33 and 3.34, we can see that MESCNet incorporates historical information in the estimate of \mathbf{x}^t . Weights \mathbf{f}^t and \mathbf{i}^t are estimated adaptively as

$$\mathbf{f}^t = \sigma(W_{fx}\mathbf{x}^{t-1} + W_{fy}\mathbf{s}) \quad \text{and} \quad \mathbf{i}^t = \sigma(W_{ix}\mathbf{x}^{t-1} + W_{iy}\mathbf{s}) \quad (3.35)$$

where W_{fx} , W_{fy} , W_{ix} and W_{iy} are trainable matrices. σ is the sigmoid function defined as $\sigma(x) = 1/(1 + e^{-x})$. All together, the structure of the layer used for

the estimation of the coefficients \mathbf{x}^t corresponds to a modified LSTM unit which is illustrated in Figure 3.9 (a). In the second stage, once the sparse codes \mathbf{x} are

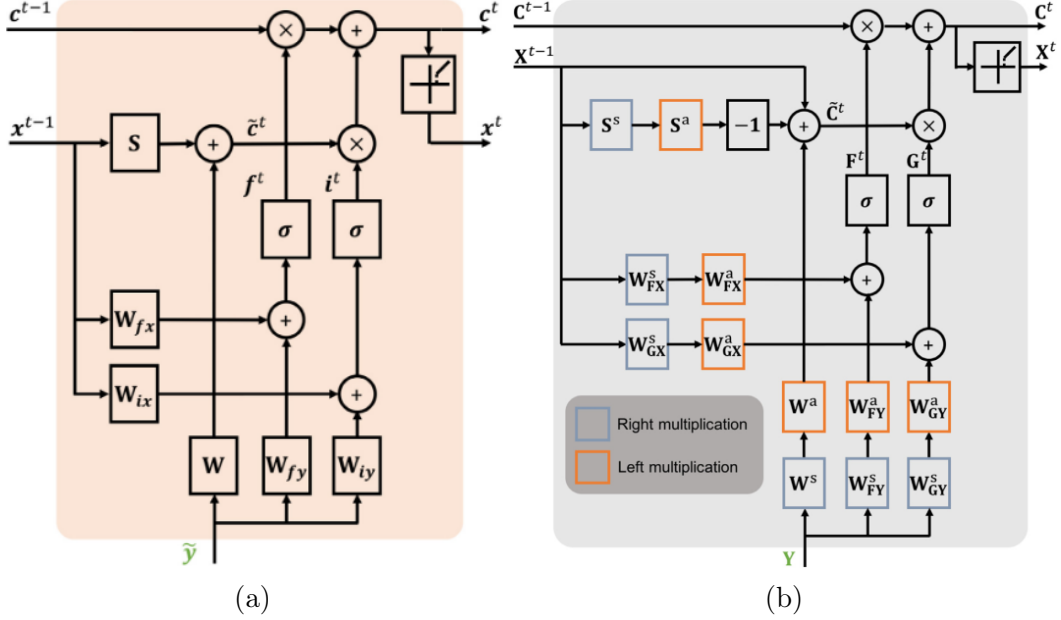


Figure 3.9: Illustration of modified LSTM units (image sources: [Ye *et al.* 2019, Ye *et al.* 2020])

estimated, they are mapped to micro-structure parameters via a Fully Connected Network (FCN), where each layer i has associated weights and biases H_i and \mathbf{b}_i . Given the input \mathbf{a}_{i-1} to a fully connected layer i , the output is estimated as $\mathbf{a}_i = \text{ReLU}(H_i \mathbf{a}_{i-1} + \mathbf{b}_i)$, where $\mathbf{a}_0 = \mathbf{x}$

As the input signal is taken from a neighborhood, the size of the matrices W , W_{fy} and W_{iy} is very large (e.g. assuming 60 points in q-space, neighborhood of size $3 \times 3 \times 3$ and length of sparse codes 300, size of a matrix is the $27 \times 60 \times 300 = 486000$). Training of such a model is computationally and storage-wise demanding, requiring a large amount of training data. To address this problem, in [Ye *et al.* 2020] an improved version of MESCNNet has been proposed, where the weights are separately defined for spatial patterns and q-space patterns. The architecture is also composed of two stages as illustrated in Figure 3.8, but this time with separable weights. In the first stage, given input in matrix form $Y \in \mathbb{R}^{Q \times V}$, where Q is the number of sampling points in q-space and V is the number of voxels in neighborhood, the sparse vectors in the layer t are estimated as

$$X^t = h_\lambda(C^t) \quad \text{s.t.} \quad C^t = F^t \circ C^{t-1} + I^t \circ \tilde{C}^t \quad \text{s.t.} \quad \tilde{C}^t = W^a Y W^s + S^a X^{t-1} S^s \quad (3.36)$$

where W^a , S^a are trainable weights applied along the q-space related (angular) dimension of the input Y and the matrix of sparse code X^{t-1} , while W^s , and S^s weights along the neighborhood related (spatial) dimension. Similarly, weighting

factors F^t and I^t are given by

$$F^t = \sigma(W_{fx}^a X^{t-1} W_{fx}^s + W_{fy}^a Y W_{fy}^s) \quad \text{and} \quad I^t = \sigma(W_{ix}^a X^{t-1} W_{ix}^s + W_{iy}^a Y W_{iy}^s) \quad (3.37)$$

where the pairs W_{fx}^a, W_{fy}^a and W_{ix}^a, W_{iy}^a are trainable weights applied along the q-space related dimension and W_{fx}^s, W_{fy}^s and W_{ix}^s, W_{iy}^s along the neighbourhood dimension. They are used together to estimate the weighting factors of the coefficients C^{t-1} and the intermediate estimate of the coefficients from the current layer \hat{C}^t . This modified LSTM unit with separable weight is illustrated in Figure 3.9 (b). Once the sparse codes in form of a matrix X are estimated, they are mapped to micro-structure parameters via a set of fully connected layers containing separable filters. Each layer i contains a pair of weights W_i^a and W_i^s and bias terms B_i^a and B_i^s . For an input X_i to the i^{th} layer, coefficients are estimated as $A_i = ReLU((W_i^a A_{i-1} + B_i^a)W_i^s + B_i^s)$, where $A_0 = X$. This version of MESCNet, termed as MESCNetSepDict, also has a possibility to provide output for multiple voxels at once. All presented models MLP, MEDN, MESCNet, MESCNetSepDict does not take into account any property of the dMRI signals, such as antipodal symmetry or spherical nature.

One of the first DL models adjusted to the specific properties of dMRI data was proposed in [Banerjee *et al.* 2019]. It is composed of homogeneous CNN (HCNN) designed for signals living in Riemannian homogeneous spaces which extract *intra-voxel* features and 2D planar CNN which extract *inter-voxel* features. The model is termed dMRI-CNN and its architecture is illustrated in Figure 3.10. In the first convolutional layer of HCNN, correlation is performed between the dMRI signal \mathbf{s}^1 and a filter $\mathbf{s}^1, \mathbf{w}_i^1 : S^2 \times \mathbb{R}^+ \rightarrow \mathbb{R}$ which are represented in the SHORE basis [Özarlan *et al.* 2013, Fick *et al.* 2016]. i refers to the ordinal number of the filter. It is denoted by the \mathcal{M} -Corr layer in Figure 3.10. Since $(\mathbf{s}^1 * \mathbf{w}_i^1) : SO(3) \times \mathbb{R}^* \rightarrow \mathbb{R}$, the following convolutional layers contain correlation between $\mathbf{s}_i^l, \mathbf{w}_i^{lj} : SO(3) \times \mathbb{R}^* \rightarrow \mathbb{R}$, where \mathbf{w}_i^{lj} is the trainable filter of layer l ($l > 1$), for the input channel i , contributing to the output channel j . These layers are denoted by the G -Corr layers in Figure 3.10. After each convolutional layer a ReLU non-linearity is applied. Once the features are extracted for each voxel independently, a 2D CNN network is used to extract spatial patterns between them. This model was applied to the problem of classification of dMRI scans into Parkinson disease patients and control group subjects. Application of DL approaches on dMRI data has been investigated for the evaluation of other neurological diseases, as well. In [Minaee *et al.* 2018], a convolutional auto-encoder has been applied on dMRI metrics (e.g. fractional anisotropy; axial, mean and radial kurtosis; white matter integrity metrics), to extract spatial patterns from 3D patches relevant for identification of mild traumatic brain injury features. Furthermore, in [Müller *et al.* 2021] a rotation and translation equivariant network has been developed and applied to the problem of multiple sclerosis lesion segmentation from dMRI data.

DL models have been also investigated for the estimation of voxel-wise PDFs, such as fODFs. In [Lin *et al.* 2019], a 3D CNN applied on the SH coefficients of dMRI

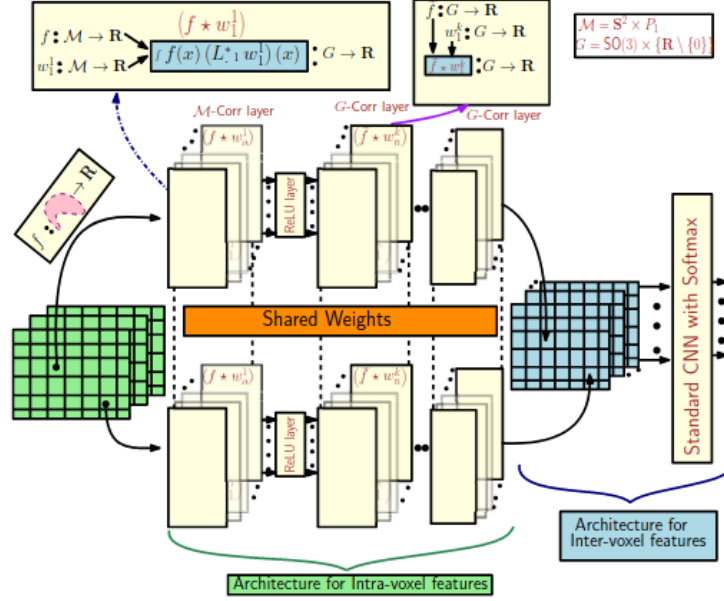


Figure 3.10: Illustration of the architecture of dMRI-CNN (image source: [Banerjee *et al.* 2019]).

signals has been proposed for fODF estimation. The architecture of the model is illustrated in Figure 3.11. As input, it takes the dMRI SH coefficients estimated using Moore-Penrose pseudo-inverse, over multiple shells and over a neighbourhood of size $3 \times 3 \times 3$. Denoting by \hat{s}_i , the SH coefficients of shell i , the input vector corresponding to one voxel is obtained by simple concatenation as $\hat{\mathbf{s}} = [\hat{s}_1^T, \dots, \hat{s}_{N_{sh}}^T]^T$, where N_{sh} is the number of shells. Each entry of $\hat{\mathbf{s}}$ is treated as one input channel (analogue to R,G or B channels of color images). This input is processed by two convolutional layers with kernels of size $2 \times 2 \times 2$, which are followed by three fully connected layers as illustrated in Figure 3.11. After each convolutional or fully connected layer, apart from the last one, a ReLU non-linearity is applied.

Although the model proposed in [Lin *et al.* 2019] achieves competitive results, it does not take into account the properties of the dMRI data. Thus, it requires a higher number of parameters and consequently a higher number of training data. In [Elaldi *et al.* 2021], an unsupervised rotation equivariant U-net with graph convolutions has been proposed for fODF estimation. The architecture of the model is illustrated in Figure 3.12. This model takes as input single- or multi-shell dMRI signals which are transformed to the spectral domain and then re-projected to the S^2 space onto a hierarchical Healpix sampling grid [Gorski *et al.* 2005]. Graph convolution of one such signal $\tilde{\mathbf{s}}$ with a filter \mathbf{w} is defined as

$$\tilde{\mathbf{s}} * \mathbf{w} = \sum_{p=0}^P w_p L^p \tilde{\mathbf{s}} \quad (3.38)$$

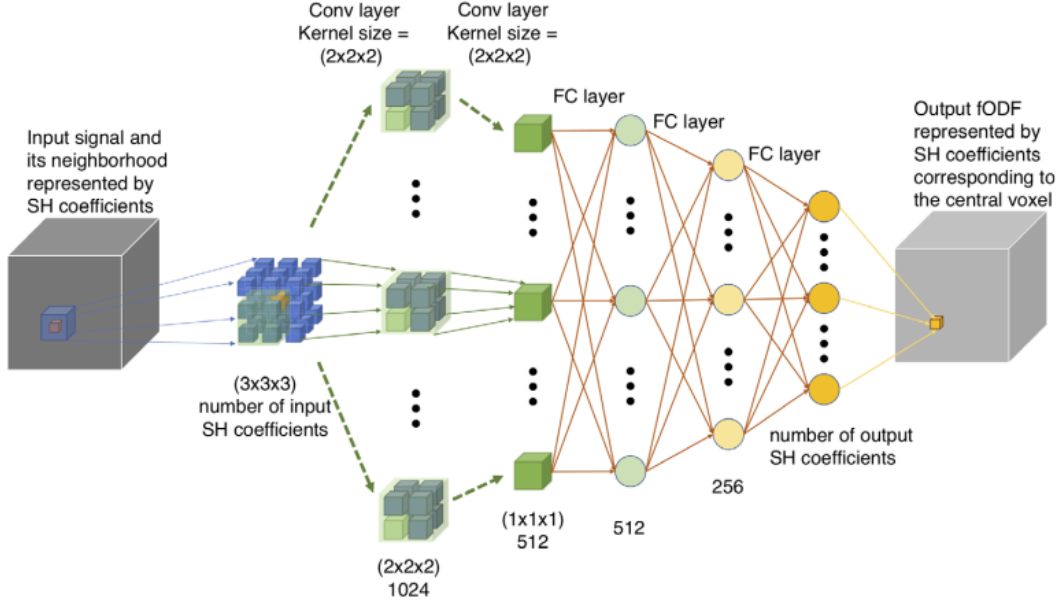


Figure 3.11: Illustration of the architecture of 3DCNN for fODF estimation (image source: [Lin *et al.* 2019]).

where w_p is p^{th} entry of \mathbf{w} . L is graph Laplacian defined as $L = D - A$ with D being degree and A adjacency matrix. The degree matrix of the graph is diagonal, with an i^{th} diagonal entry equal to $\sum_j w_{ij}$, where $w_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{\rho}}$ if $i \neq j$ and $w_{ii} = 0$. \mathbf{x}_k are coordinates of k^{th} vertex and ρ is average distance between two vertices. Entries i, j of the adjacency matrix A are 1 if there is an edge between the vertices and 0 otherwise. In both, the contracting and expanding parts of the U-net, convolutions are followed by ReLU non-linearities and batch normalization, except for the last layer, where a Soft plus activation was used for the multi-shell case and ReLU for the single-shell case. Due to hierarchical nature of the sampling grid, pooling in contracting and up-sampling in expansive parts can be performed without a need for re-sampling. The loss function is defined as

$$\mathcal{L} = \sum_{n=1}^N \|\mathbf{s}_n - \mathbf{f}_n * \mathbf{r}\|_2^2 + \lambda \sum_{i=1}^I \log\left(1 + \frac{f_n^i}{2\sigma_c^2}\right) + \|\mathbf{f}_n \circ \mathbf{m}_n\|_2^2 \quad (3.39)$$

where \mathbf{s}_n and \mathbf{f}_n are the n^{th} dMRI sample and the estimated fODF, respectively. \mathbf{r} is the response function precomputed with the *mrtrix* library [Tournier *et al.* 2019]. \mathbf{m}_n is a mask whose entries are 1 for negative entries of \mathbf{f}_n and 0 otherwise. Constants λ and σ_c control the sparsity of the estimated fODFs.

Apart from the before mentioned applications, DL approaches have also been used for dMRI data synchronization over different sites [Ning *et al.* 2018], segmentation of brain tissues [Zhang *et al.* 2021], signal enhancement [Aggarwal *et al.* 2019] and reconstruction [Hong *et al.* 2019], etc.

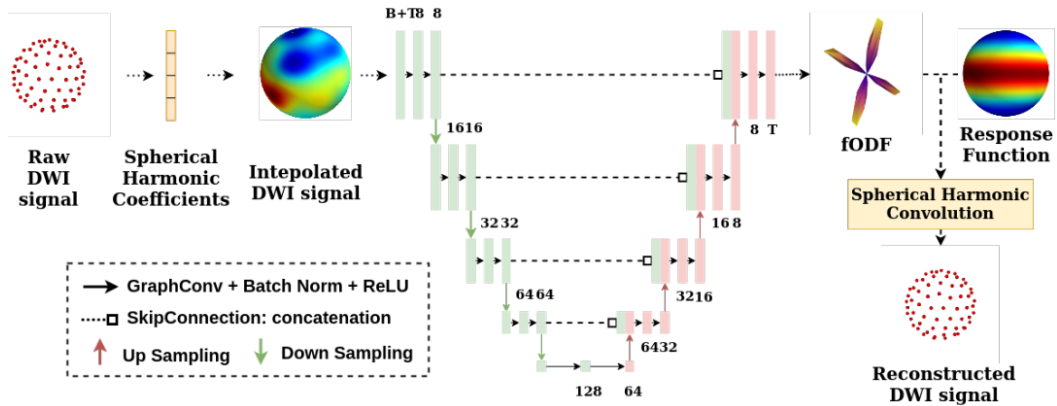


Figure 3.12: Illustration of the architecture of the rotationally equivariant U-net for fODF estimation (image source: [Elaldi *et al.* 2021])

3.6 Conclusion

In this chapter, we have firstly presented the properties of the dMRI signals acquired with q-space sampling protocols, namely their real and spherical nature, antipodal symmetry and rotation equivariance with respect to the underlying tissue structures. Due to the spherical nature and the rotation equivariance, representation of the dMRI signals in the SH basis is often used in their analysis, thus we have provided an overview of the most relevant methods for the SH coefficient estimation. As the dMRI signals associated with individual axon bundles can be considered axially symmetric, we have provided definition of the convolution with zonal filters which is used in the estimation of certain dMRI related PDF functions. Further, we have described the most relevant PDF functions, namely the EAP, dODF and fODF and the most prominent biophysically inspired multi-compartment models for dMRI local modeling. As we are interested in the analysis of the spherical signals, we have provided an overview of the recent rotationally equivariant DL models used for arbitrary spherical signals. Finally, in the last part of the chapter, the most relevant DL approaches used in dMRI local modeling are described in details. In the following two chapters, we will present our contributions in dMRI local analysis, concretely, rotation equivariant models for the fODF estimation and microstructure parameter estimation.

Spherical U-net for dMRI fiber orientation distribution function estimation

Contents

4.1	Introduction	58
4.2	Method	58
4.2.1	Estimation of SH coefficients	60
4.2.2	Convolutional layers	61
4.2.3	ReLU non-linearity	61
4.2.4	Pooling	62
4.2.5	Transposed convolutional layers	62
4.2.6	Loss function	64
4.3	Datasets	64
4.4	Experiments and implementation details	65
4.5	Results	65
4.6	Conclusion	69

Executive summary

This chapter contains our first contribution in dMRI local modeling, namely a spherical U-net for fODF estimation. Firstly, we have presented SH coefficient estimation via the Gram-Schmidt orthonormalization process with an analysis of its orthogonality properties. Further, we provide details related to the architecture of spherical U-net and its main building blocks, namely convolutional and transposed convolutional layers with zonal trainable kernels realized in the spectral domain, non-linear activations ReLU applied in the signal domain and pooling layer realized in the spectral domain. The model is compared with a deep learning 3DCNN approach and a traditional multi-shell multi-tissue constrained spherical deconvolution on the real HCP data and synthetic dMRI signals, both resampled to the reduced grids which are more clinically desirable.

4.1 Introduction

U-net is a type of CNN initially designed for segmentation of biomedical images in [Ronneberger *et al.* 2015]. In contrast to firstly introduced CNNs which have a contracting architecture [O’Shea & Nash 2015], a U-net architecture is composed of a contracting and an expanding parts, which allow it to produce a high resolution outputs, instead of pixel wise (low resolution). It is a type of fully convolutional network [Long *et al.* 2015]. Whereas contracting part of the U-net enables learning of relevant features at different scales, expanding part which contains upsampling operations, instead of pooling, enables propagation of contextual information from the layers of lower to the layers of higher bandwidth. High resolution compared to the pixel-wise segmentation adds a regularization, as the loss is computed over larger areas, not just one pixel, thus the model requires less training samples. At the same time, it is faster.

In the context of spherical signal analysis, a spherical U-net has been proposed for saliency detection in 360° videos in [Zhang *et al.* 2018]. In this model, convolutions between a spherical signal and kernel are realized in the signal domain by stretching and rotating the kernel to match with locations of sampling points of the signal. In the domain of medical imaging, a spherical U-net has been proposed for the analysis of cortical surfaces in [Zhao *et al.* 2019]. In their work, instead of kernel stretching, for each vertex direct neighbors are extracted from the signal and rotated around the vertex. This is followed by a simple inner product with a kernel, representing a convolution in the signal domain. A recent work, presented in more details in Chapter 3, used a spherical U-net trained in an unsupervised manner for the estimation of the fODFs [Elaldi *et al.* 2021].

In this chapter, we present a supervised voxel-wise spherical U-net for the problem of fODF estimation from dMRI data sampled at multiple spheres (shells). The model is tailored to the properties of the dMRI signals, namely its real nature, the uniform distribution of sampling points, the rotation equivariance with respect to the underlying tissues, and the antipodal and axial symmetry of the signals generated by individual fibers. Contrary to the models proposed in [Zhang *et al.* 2018, Zhao *et al.* 2019], our U-net contains convolutional layers where the convolutions are performed in the spectral domain.

4.2 Method

The architecture of our spherical U-net model is illustrated in Figure 4.1.

As input the model takes multi-shell dMRI data of one voxel or a small 3D neighborhood that in total results in $N_{sh} \times N_{nb}^3$ channels, where N_{sh} is the number of shells and N_{nb} is the neighborhood size. Taking into account a small neighborhood rather than a single voxel as input allows incorporation of the spatial information, in addition to the angular information extracted from the q-space. Although in the models proposed in [Ronneberger *et al.* 2015, Zhang *et al.* 2018, Zhao *et al.* 2019] output is of the same resolution as input, for multi-shell dMRI data it is reasonable

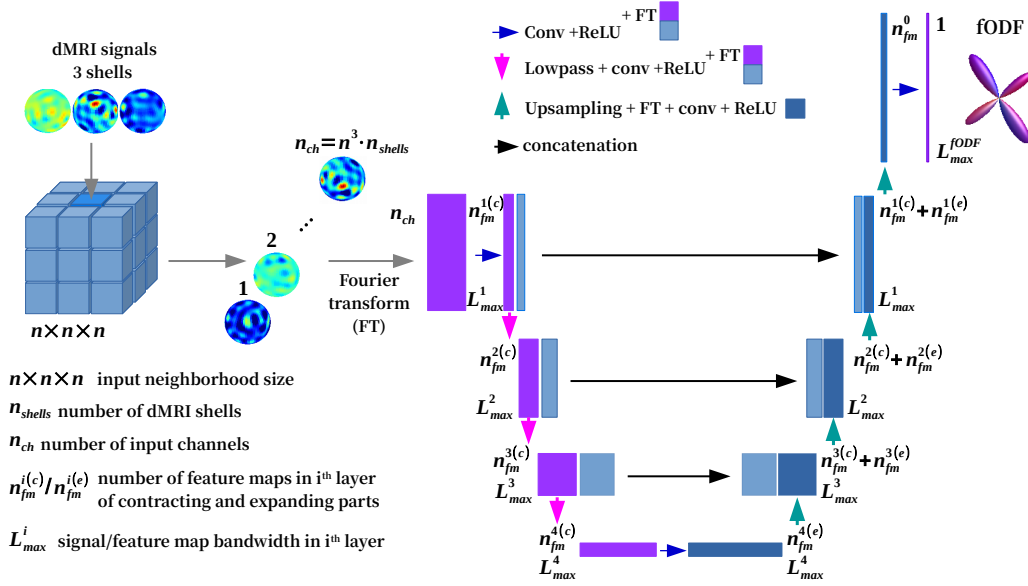


Figure 4.1: Illustration of a spherical U-net architecture with corresponding convolutional operations in contracting and expanding parts.

to assume that fODFs of higher resolution can be estimated. This is explained by the fact that multi-shell dMRI signals are sampled over noncollinear points between shells, distributed over continuous q-space.

As the standard U-net, our model is composed of contracting and expanding parts. The main operations are convolutions, pooling and transposed convolutions. Due to assumed axial symmetry of the signals emerging from individual axon bundles and antipodal symmetry of dMRI signals, convolutional kernels in our model are zonal and antipodally symmetric. Each convolutional layer of the contracting part takes as input the SH coefficients of multi-channel signal and performs convolution with zonal kernels also represented in the SH basis. Resulting SH coefficients are transformed to S^2 space onto a q-space sampling grid [Caruyer *et al.* 2013] where ReLU non-linearity is applied. The S^2 signals obtained after ReLU are forwarded to the parallel layer of the expanding part, while their low-passed SH coefficients are passed to the convolutional layer below. Low-pass filtering corresponds to simple discarding of the SH coefficients of the highest degree as in [Cohen *et al.* 2018, Esteves *et al.* 2018] which corresponds to the operation of pooling. Each transposed convolutional layer, of expanding part, takes as input S^2 domain signals from the layer below and inserts zero samples among existing samples. Following this, the signals are transformed to the spectral domain and convolution with zonal kernels is performed. The obtained SH coefficients are transformed to the S^2 domain where the ReLU non-linearity is applied. The resulting signals in the S^2 domain are concatenated with parallel signals of the contracting part and they serve as input to the next transposed convolutional layer. The last layer in the expanding part, only performs convolution

with one convolutional kernel and as output gives the fODF SH coefficients.

4.2.1 Estimation of SH coefficients

To estimate the SH coefficients of the input and intermediate S^2 signals, the SH basis Y are inverted using Gram-Schmidt orthonormalization process. The inverted basis are denoted with Y_{gs}^\dagger . If y_i and y_j^{gs} correspond to i^{th} columns of Y and $Y_{gs}^\dagger{}^T$, respectively, y_i^{gs} are determined as

$$y_i^{gs} = y_i - \sum_{j=0}^{i-1} \frac{\langle y_i, y_j^{gs} \rangle}{\langle y_j^{gs}, y_j^{gs} \rangle} y_j^{gs}, \quad y_i^{gs} = \frac{y_i}{\|y_i^{gs}\|_2} \quad (4.1)$$

where $y_0^{gs} = y_0$. The SH basis elements in the matrix Y are ordered so that column 0 corresponds to the basis element of degree $l = 0$ and order $m = 0$, following columns are the basis elements of degree $l = 2$ and orders $m = \{-2, -1, 0, 1, 2\}$, etc. Since aliasing affects more the SH coefficients of a higher degree l , it is convenient to start the orthonormalization process with a basis of a lower degree, as we know that they are determined by a lower number of sampling points. On the other hand, in order to avoid a bias due to basis element ordering, the Gram-Schmidt process is repeated N_{it} times, each time randomly shuffling the order of the basis elements of the same degree, which are at the end averaged. Finally, for an input signal $s : S^2 \rightarrow \mathbb{R}$, SH coefficients \hat{s} are estimated as

$$\hat{s} \approx Y_{gs}^\dagger s. \quad (4.2)$$

In Figure 4.2 and 4.3 orthogonality properties of bases inverted with different approaches, presented in Chapter 3, are depicted for 30 uniformly randomly distributed points and the antipodally symmetric basis of bandwidth 6. The approaches we have compared are the Moore-Penrose pseudo inverse (mp), least square with Tikhonov regularization (tikh) with the regularization constants $\lambda \in \{1, 0.1\}$, least square with Laplace-Beltrami regularization (lb) with the regularization constants $\lambda \in \{0.001, 0.0001\}$, and the approach with the Gram-Schmidt orthonormalization (gs) for different number of repetitions $N_{it} \in \{1000, 1\}$. Orthogonality with respect to the basis Y , illustrated in Figure 4.2 indicates how accurately the SH coefficients can be estimated if there is no noise. In this scenario, we can see that Moore-Penrose yields the exact solution, least square with Tikhonov regularization penalizes equally SH basis elements of all degrees, while least square with Laplace-Beltrami penalizes more the SH coefficients of the highest degree, as well as the approach with Gram-Schmidt orthonormalization process. Orthogonality of the inverted SH basis with itself, as illustrated in Figure 4.3 indicates their robustness to noise and aliasing. The illustrations show that the Moore-Penrose and the least square with Tikhonov regularization (for $\lambda = 0.1$) are very sensitive to the noise. Least square with the Laplace Beltrami regularization and the Gram-Schmidt orthonormalization process averaged over 1000 iterations perform stronger regularization of the SH coefficients of the highest degree and therefore are more robust with respect to the noise and aliasing.

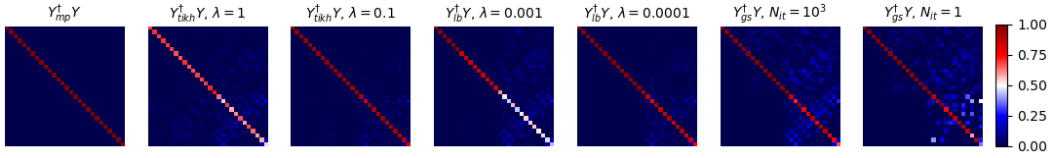


Figure 4.2: Orthogonality between the SH basis Y and inverted SH basis for 30 randomly uniformly distributed points (28 SH basis elements in total).

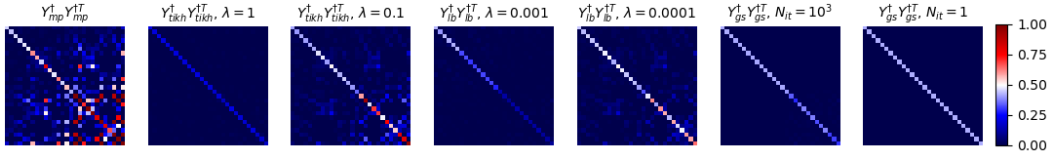


Figure 4.3: Orthogonality of the inverted SH basis for 30 randomly uniformly distributed points (28 SH basis elements in total).

4.2.2 Convolutional layers

Input SH coefficients to a convolutional layer are denoted as $\{\{\hat{\mathbf{s}}_l^i\}_{l=0}^L\}_{i=1}^I$, where l is the SH degree, i refers to channel (shell), L is the bandwidth and I is the total number of input channels. A convolutional zonal kernel is denoted as $\{\{\{\hat{w}_l^{i,j}\}_{l=0}^L\}_{i=1}^I\}_{j=1}^J$, where i, j indicate the input and output channels, respectively and I, J their total number. The convolution between the input $\{\{\hat{\mathbf{s}}_l^i\}_{l=0}^L\}_{i=1}^I$ and the trainable zonal kernel $\{\{\{\hat{w}_l^{i,j}\}_{l=0}^L\}_{i=1}^I\}_{j=1}^J$, is defined as

$$\hat{\mathbf{g}}_l^j = \sum_{i=1}^I \hat{\mathbf{s}}_l^i \hat{w}_l^{i,j} \quad \text{for } l \in \{0, 2, \dots, L\} \quad \text{and} \quad j \in \{1, 2, \dots, J\}. \quad (4.3)$$

By transforming a zonal kernel into a diagonal matrix, convolution in the spectral domain can be illustrated as in Figure 4.4.

4.2.3 ReLU non-linearity

After a convolutional layer, the obtained SH coefficients $\{\hat{\mathbf{g}}_j^J\}_{j=1}^J$ are transformed to S^2 domain as $\mathbf{g}^j = Y \hat{\mathbf{g}}^j$. The ReLU non-linearity is performed as

$$\mathbf{a}^j = \text{ReLU}(\mathbf{g}^j + b_j) \quad (4.4)$$

where b_j is a bias term associated with the channel j . We note that the thresholding of the signal with ReLU might introduce sharp signal transitions between neighboring points, which cannot be represented with given bandwidth. Thus, the ReLU non-linearity can cause the aliasing. When the SH coefficients $\{\hat{\mathbf{g}}_j^J\}_{j=1}^J$ are transformed to S^2 domain, in order to minimize the effect of the aliasing it is better to project the coefficients to $\{\mathbf{g}_j^J\}_{j=1}^J$ sampled at a higher number of sampling

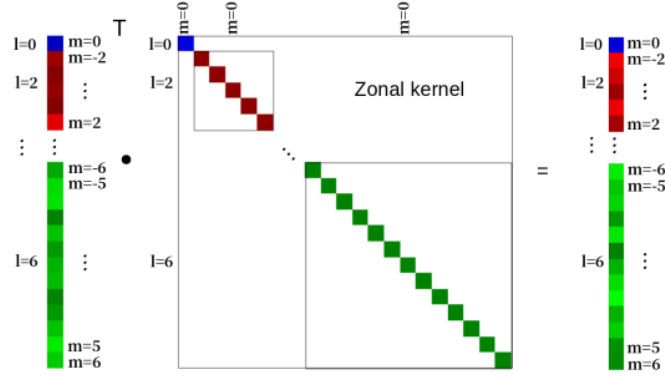


Figure 4.4: Illustration of convolution between S^2 signal and a zonal kernel in spectral domain.

points. (This is simply a consequence of the fact that the SH coefficient estimation is more accurate for a higher number of sampling points.) The minimal number of the sampling points we have used is $\frac{(L+1)(L+2)}{2}$ as it corresponds to the number of the SH basis elements for the bandwidth L .

4.2.4 Pooling

After the non-linearity is applied, pooling is performed in the spectral domain. Obtained $\{\mathbf{a}^j\}_{j=1}^J$ signals are transformed to spectral domain as $\hat{\mathbf{a}}^j = [Y_{gs}^\dagger]_{(L-2)} \mathbf{a}^j$, where $[Y_{gs}^\dagger]_{(L-2)}$ contains the inverted SH basis of the highest degree $(L - 2)$. This can be seen as low pass filtering. In planar CNNs, one way to perform pooling is by averaging values of a small neighborhood as illustrated in Figure 4.5. Similarly, performed in the spectral domain, pooling corresponds to discarding of the SH coefficients of the highest degree as illustrated in Figure 4.6.

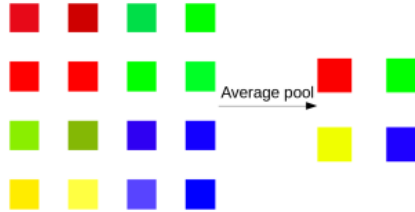


Figure 4.5: Illustration of average pooling in planar CNNs

4.2.5 Transposed convolutional layers

Given input S^2 signals $\{\mathbf{s}_i\}_{i=1}^I$ to a transposed convolutional layer, firstly, by insertion of zero samples we obtain the $\{\mathbf{q}_i\}_{i=1}^I$ signals. If the signals $\{\mathbf{s}_i\}_{i=1}^I$ have bandwidth L , the number of inserted zeros increases the number of sampling points

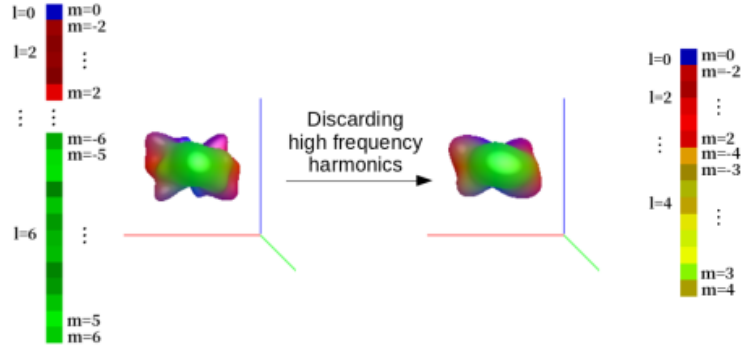


Figure 4.6: Illustration of spectral domain pooling in spherical CNNs

which corresponds to the bandwidth $(L + 2)$ (e.g. from $\frac{(L+1)(L+2)}{2}$ to $\frac{(L+3)(L+4)}{2}$). This is followed by the estimation of the SH coefficients $\hat{\mathbf{q}}^i = [Y_{gs}^\dagger]_{(L+2)} \mathbf{q}^i$ and convolution with kernels $\{\{\{\hat{w}_l^{i,j}\}_{l=0}^L\}_{i=1}^I\}_{j=1}^J$, as defined in equation 4.3. For a comparison, illustrations of a transposed convolution in a planar CNN and in our model are given respectively in Figures 4.7 and 4.8. The obtained SH coefficients are transformed to the S^2 domain, where bias terms are added and ReLU nonlinearities are applied. The resulting signals are concatenated with the signals of the same bandwidth, from the parallel layer in the contracting part of the U-net and serve as input to the following transposed convolution layer.

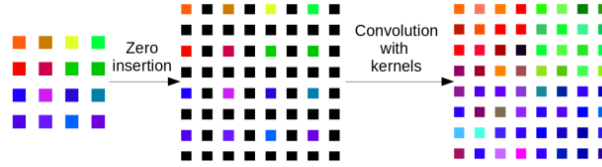


Figure 4.7: Illustration of transposed convolution in planar CNNs.

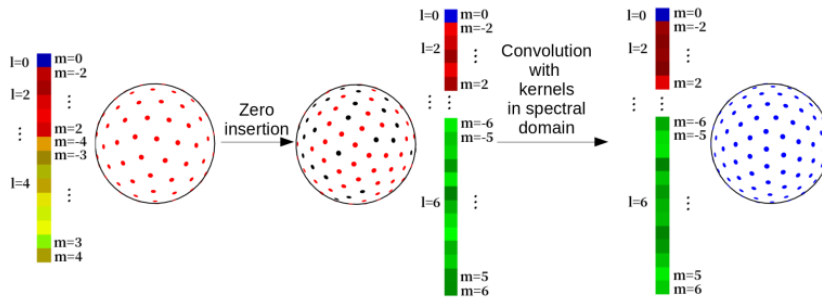


Figure 4.8: Illustration of transposed convolution in our spherical U-net.

4.2.6 Loss function

The loss function is defined as mean square error (MSE) between the SH coefficients of gold-standard fODFs and the estimated fODFs as

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^N (\mathbf{fODF}_n - \mathbf{fODF}_n^e)^2 \quad (4.5)$$

where \mathbf{fODF}_n and \mathbf{fODF}_n^e are the gold standard and estimated SH coefficients of the fODF of the n^{th} sample, respectively. N is the number of samples in a batch.

4.3 Datasets

We have used in our experiments real data from the HCP [Van Essen *et al.* 2013] (referred to as *Real dataset*) and synthetic data generated from the same real HCP scans following the procedure described in [Wilkins *et al.* 2015]. The *Real data* was acquired on Siemens 3T Skyra system with 100 mT /m gradient, over three shells with b-values of 1000, 2000 and 3000 s/mm^2 , each with 90 gradient directions and 18 b = 0 images at resolution $1.25 \times 1.25 \times 1.25 mm^3$. To generate the synthetic data, firstly, up to three fiber orientations and corresponding volume fractions were estimated per voxel using the *bedpostx* tool from the *FSL* library [Smith *et al.* 2004]. These parameters were then used to generate synthetic data using the multi-fiber ball and stick model [Behrens *et al.* 2007] as in [Wilkins *et al.* 2015] for each shell independently. In the generation process, the free diffusivity coefficients are set to $\{0.68, 0.96, 2.25\} \cdot 10^{-3} s/mm^2$ for the white matter, gray matter and cerebrospinal fluid, respectively while the single-fiber tensor's eigenvalues are set to $\{\lambda_1, \lambda_2, \lambda_3\} = \{1.7, 0.17, 0.17\} \cdot 10^{-3} s/mm^2$ [Wilkins *et al.* 2015]. To simulate more realistic dMRI data, a Rician noise with a signal to noise ratio (SNR) of 18db was added to the synthesized data. In addition, in order to investigate the robustness of the compared methods, one synthetic dataset is generated with the constant diffusion single-fiber tensor eigenvalues (*Synthetic dataset 1*) as in [Wilkins *et al.* 2015] and another one with the eigenvalues taken from the uniform distribution around these values (values taken from the range of $\pm 10\%$) (*Synthetic dataset 2*). Experiments were conducted on *Real dataset*, *Synthetic dataset 1* and *Synthetic dataset 2* with downsampled acquisition schemes. To select relevant white matter voxels, we have used brain tissue segmentation computed from *T1w* images using the *FAST* algorithm [Zhang *et al.* 2001] implemented in the *mrtrix* library [Tournier *et al.* 2019]. In the experiments, where comparing models take into account neighborhood information, white matter masks are extended using the 3D binary dilation operator. Gold standard fODFs, of SH degree 8, were estimated using the multi-shell multi-tissue constrained spherical deconvolution (MSMT-CSD) approach [Jeurissen *et al.* 2014] on signals acquired on full sampling scheme using the *mrtrix* library [Tournier *et al.* 2019]. In the case of synthetic data, the fODFs were estimated on the noiseless data. We have used 50 subjects in total, 30 for training, 10 for validation and 10 for testing.

Table 4.1: Sizes of the trainable parameters of the *3DCNNs* and *S²U-nets* (MB) for N_p sampling points.

Model / N_p	20	30	40	60	90	120
<i>3DCNN</i>	18.12	18.12	18.12	18.96	20.18	20.18
<i>S²U-net^{1×1×1}</i>	15.65	15.65	15.65	19.30	20.52	20.52
<i>S²U-net_s^{3×3×3}</i>	3.99	3.99	3.99	4.89	5.17	5.17
<i>S²U-net^{3×3×3}</i>	15.80	15.80	15.80	19.42	20.60	20.60

4.4 Experiments and implementation details

To evaluate our method on data similar to those used in clinical practice, experiments have been performed on data with significantly reduced number of sampling points N_p (20, 30, 40, 60, 90 and 120 in total for the three shells). We compared our approach with another deep learning model - 3DCNN [Lin *et al.* 2019] and with MSMT-CSD [Jeurissen *et al.* 2014]. To investigate importance of the neighbourhood information, our model is trained with single voxel multi-shell signals (termed as *S²U-net^{1×1×1}*) and with multi-shell signals from a neighbourhood of size $N_{nb} = 3$ (termed as *S²U-net^{3×3×3}*), which is also the case with the 3DCNN model. In addition, to investigate the potential of our model, we have trained one model with a significantly lower number of trainable parameters - termed as *S²U-net_s^{3×3×3}*. Sizes of the trainable parameters of the deep learning networks are given in Table 4.1. All DL approaches are implemented with the *tensorflow* library [Abadi *et al.* 2015]. Models are trained over 100 epochs. In each epoch, 3 dMRI scans are randomly selected from the 30 training samples. For all three models, the loss function is defined as MSE between the estimated and gold standard fODFs represented in the spectral domain as given in Eq. 4.5. The initial learning rate is 0.001 and after 50 epochs it is reduced to 0.0001. Model weights updates are computed using the Adam optimization algorithm [Kingma & Ba 2014].

4.5 Results

The results are compared quantitatively in terms of the MSE over all white matter voxels and the mean angular error (MAE) for single fiber voxels and voxels containing two crossing fibers. To compute peaks of the estimated and gold standard fODFs, we have used the *mrtrix* library [Tournier *et al.* 2019] and the threshold of 0.1 of the highest peak is used to eliminate spurious fibers. Thus, the MAE does not take into account the voxels where the number of peaks differs from the number of peaks in the gold standard. In Figure 4.9, we can see that our model *S²U-net^{3×3×3}* achieves a lower MSE compared to the other models on both real and synthetic datasets. This difference is especially significant when compared to the models that do not use neighbourhood information. This performance drop of single voxel based models is expected when the number of sampling points over

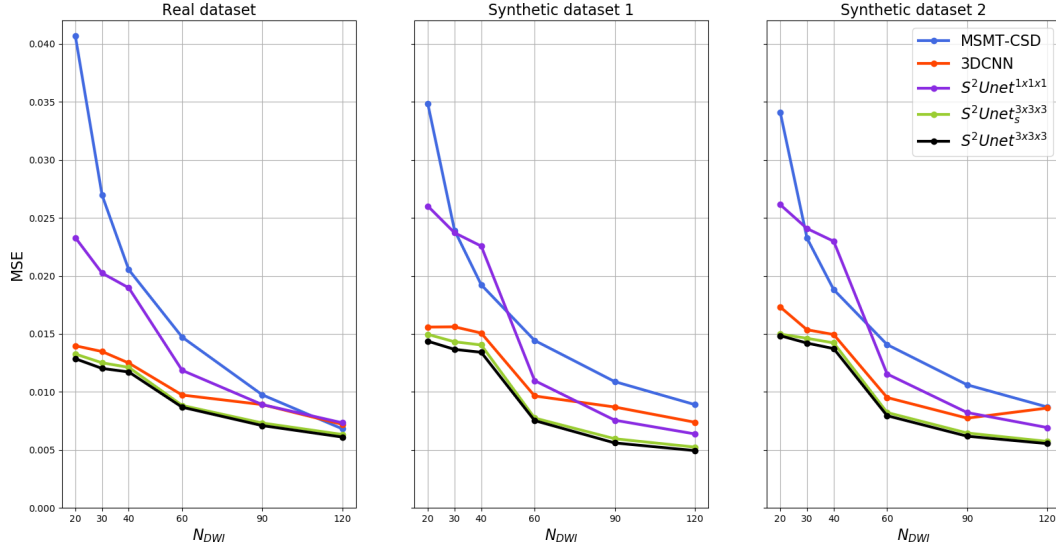


Figure 4.9: Comparison of the MSE averaged over 10 testing subjects for the *real HCP dataset*, *Synthetic dataset 1* and *Synthetic dataset 2* for different numbers of sampling points.

three shells (like 20, 30, 40) is lower than the number of SH coefficients of fODFs (which is 45 for bandwidth 8). We can also notice that almost equal performance to $S^2U-net^{3 \times 3 \times 3}$ can be achieved with the more compact model $S^2U-net_s^{3 \times 3 \times 3}$. In Figure 4.10, we can see that for the single fiber voxels and the real dataset, the MAE of the models $S^2U-net^{3 \times 3 \times 3}$ and $S^2U-net_s^{3 \times 3 \times 3}$ is almost equal to the one achieved with MSMT-CSD. However, these results are a consequence of the fact that MSMT-CSD often produces large spurious peaks when the number of sampling points is reduced, as illustrated in Figures 4.11 and 4.12, which means that they will not be taken into account if the gold standard contains a different number of peaks. The results obtained on synthetic data indicate that our approach is more robust to noise, as the gold standard is estimated on noiseless data. As depicted in Figure 4.10, $S^2U-net^{3 \times 3 \times 3}$ and $S^2U-net_s^{3 \times 3 \times 3}$ achieve a lower MAE in the voxels with crossing fibers. Qualitative comparison of MSMT-CSD, 3DCNN and $S^2U-net^{3 \times 3 \times 3}$ is provided in Figure 4.11 and 4.12 for 60 sampling points. Figure 4.11 compares the fODFs of the gold standard and compare the methods with *Real Dataset*, *Synthetic Dataset 1* and *Synthetic Dataset 2*. A similar comparison is depicted in Figure 4.12, where the estimated fODFs are overlaid over the gold standard peaks. We can notice that MSMT-CSD compared to 3DCNN and $S^2U-net^{3 \times 3 \times 3}$ is more prone to produce spurious fibers, while these DL approaches are more likely to omit some. The 3DCNN model tends to estimate more smoothed fODFs and/or lobes with lower amplitude compared to our approach.

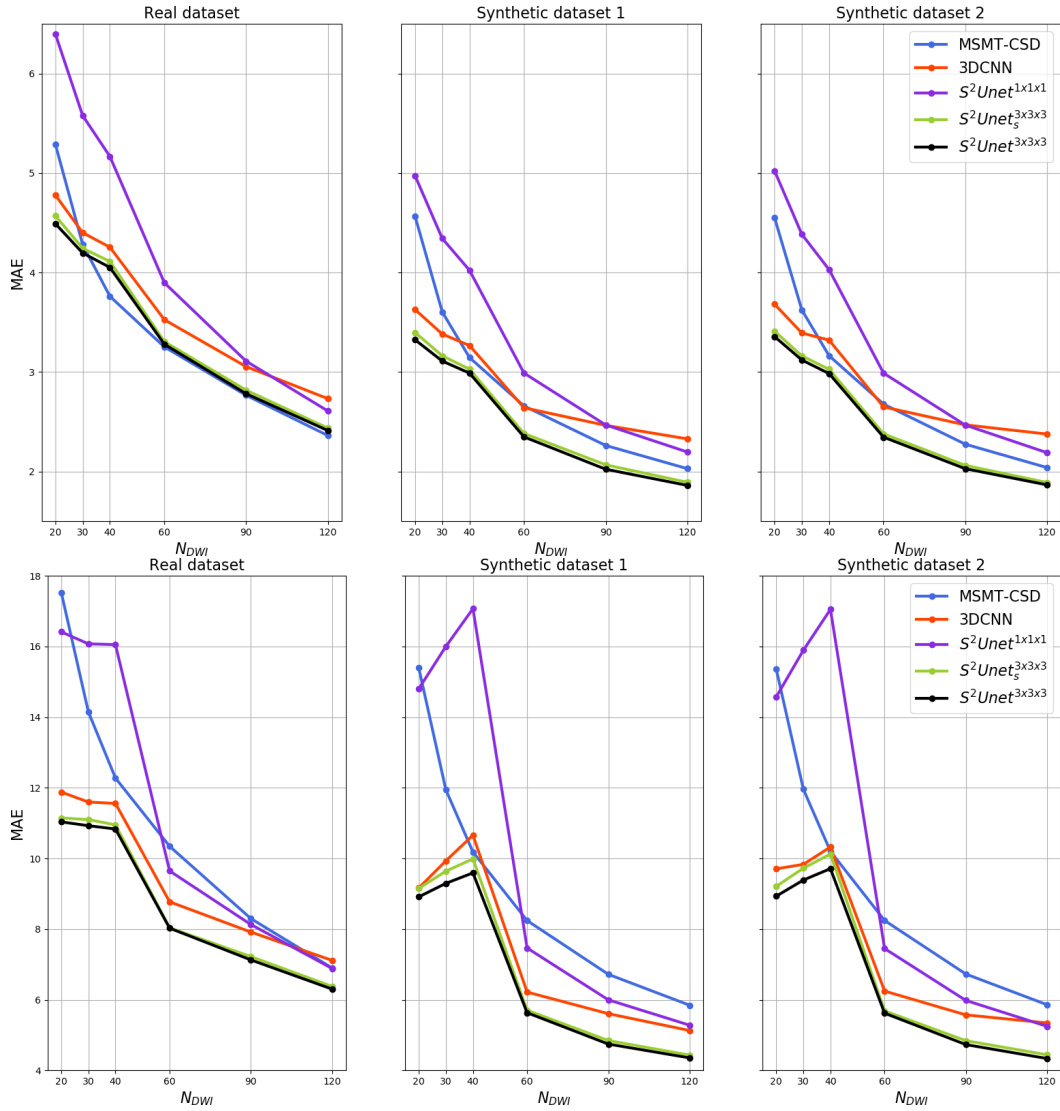


Figure 4.10: Comparison of the MAE averaged over 10 testing subjects for *real HCP dataset*, *Synthetic dataset 1* and *Synthetic dataset 2* for different numbers of sampling points for voxels containing single fibers (upper three sub-figures) and voxels containing two crossing fibers (lower three sub-figures)

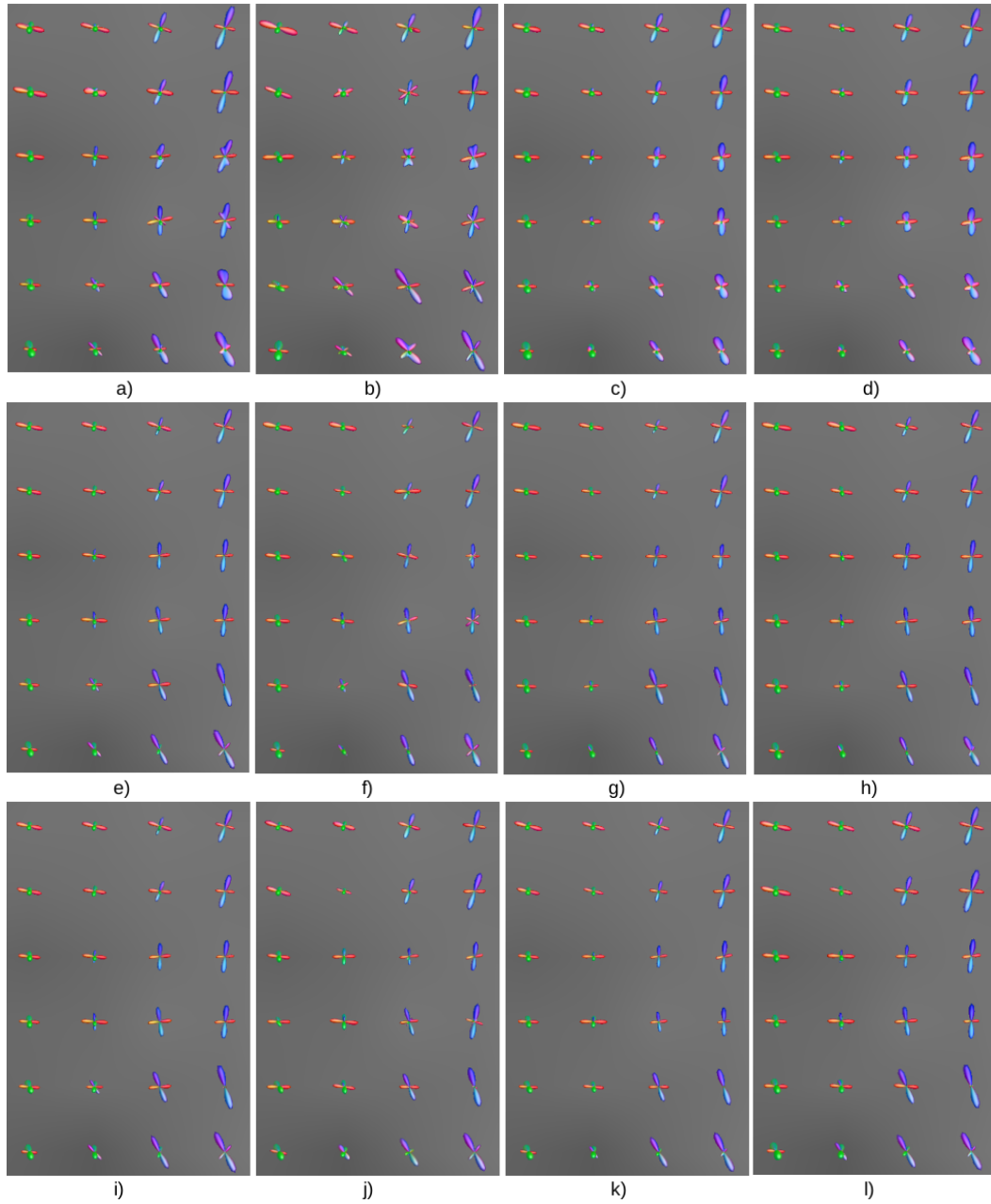


Figure 4.11: Illustration of the fODF gold standard and estimates obtained using MSMT-CSD, 3DCNN and $S^2U-net^{3 \times 3 \times 3}$ with angular resolution decreased to 60 points in total for the three shells. Sub-figures a), e) and i) correspond to the gold standard fODFs for *real HCP dataset*, *Synthetic dataset 1* and *Synthetic dataset 2*, respectively. Sub-figures b), f) and j) correspond to the fODF estimates obtained using MSMT-CSD; sub-figures c), g) and k) using 3DCNN and sub-figures d), h) and l) correspond to the fODF estimation with $S^2U-net^{3 \times 3 \times 3}$

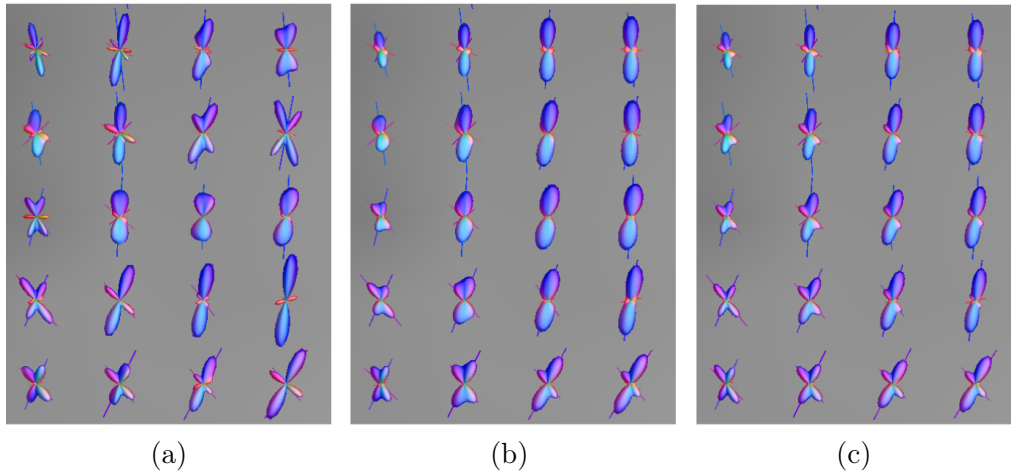


Figure 4.12: Comparison of the fODFs estimated with MSMT-CSD (a), 3DCNN (b) and our $S^2U\text{-net}^{3\times 3\times 3}$ (c), overlaid over gold standard fiber peaks

4.6 Conclusion

In this chapter, we have described a spherical U-net model adjusted to the properties of dMRI data, namely the real and spherical nature of the signals, their antipodal symmetry, the random distribution of the sampling points and the axial symmetry of signals coming from individual fibers. We have demonstrated that the proposed spherical U-net is suitable for a high resolution inference such as the estimation of the fODFs from dMRI data acquired on schemes which contain a lower number of sampling points, which is required in clinical practice. It is also shown that the model is capable of successfully incorporating neighboring information to boost the model’s performance, yielding significant improvements for low numbers of sampling points (≤ 40) when compared to the single voxel based models. The results are compared on the real and synthetic datasets. The results obtained on synthetic data indicate a better robustness of our model with respect to the noise.

This work was used in the Diffusion Simulated Connectivity (DiSCo) Challenge associated with the CDMRI MICCAI workshop 2021 where our team was ranked as fifth. In the DiSCo challenge the model was trained on single voxel data generated by the *dmipy* library [Fick *et al.* 2019], showing that our model has a high generalization power. In addition, correlation between ground truth and validation connectivity matrices obtained with MSMT-CSD fODFs was in the range 87 – 90% and with the fODFs obtained with spherical U-net in the range 94 – 97%.

Fourier domain spherical CNN for dMRI local analysis

Contents

5.1	Introduction	72
5.2	Theory	73
5.2.1	Convolution (correlation) between S^2 and zonal functions	73
5.2.2	S^2 quadratic function	74
5.2.3	Convolution (correlation) between $SO(3)$ functions	75
5.2.4	$SO(3)$ quadratic function	75
5.2.5	Power spectrum of S^2 and $SO(3)$ functions	76
5.3	Methods	77
5.3.1	Fourier domain CNN with quadratic S^2 non-linearities	78
5.3.2	Fourier domain CNN with quadratic $SO(3)$ non-linearities	78
5.4	Experiments	81
5.4.1	Axon bundle counting experiment	81
5.4.2	Multi-compartment micro-structure estimation	87
5.4.3	Brain tissue segmentation	97
5.5	Conclusion	101

Executive summary

In this chapter, we have investigated rotation equivariant CNNs with quadratic non-linearities realized in the spectral domain for local analysis of dMRI data. The spectral domain non-linearities are introduced to avoid often computationally expensive conversions from the spectral to the signal domain in order to apply non-linearities such as ReLU and to avoid the aliasing that such non-linearities generate. First, in Section 1.2, we introduce the mathematical grounds necessary for understanding and defining the Fourier domain CNNs, which are presented in the following Section 1.3. The models are evaluated in Section 1.4 on the problem of axon bundle counting on synthetic data, and on the real HCP dMRI data on the problems of micro-structure parameter estimation and brain tissue segmentation.

5.1 Introduction

Although the data acquired on spheres have been present over the last several decades in different scientific areas such as astronomy, meteorology, satellite imaging, point cloud applications, medical imaging, etc, it was only recently that neural network models, properly taking into account their spherical nature have been introduced for their analysis. Some of the most relevant rotation equivariant CNN models for arbitrary spherical signals are presented in Chapter 3 [Cohen *et al.* 2018, Esteves *et al.* 2018, Kondor *et al.* 2018]. From the point of view of dMRI data acquired with q-space sampling schemes, the first drawback of these models is that they take as input signals sampled on grids which have associated quadrature formulae for the exact computation of the SH coefficients such as Driscoll-Healy and Gauss-Legendre grids. Furthermore, as already mentioned, models proposed by [Cohen *et al.* 2018, Esteves *et al.* 2018] use signal domain non-linearities. A drawback of the spectral domain non-linearity of quadratic nature, introduced in [Kondor *et al.* 2018], is its quadratic increase of the output channels, consequently requiring a higher number of trainable parameters compared to the other models [Cohen *et al.* 2018, Esteves *et al.* 2018]. The first rotation equivariant CNN adapted to the properties of dMRI data, with signal domain non-linearities, has been introduced in [Banerjee *et al.* 2019], as a part of the model used in Parkinson’s disease classification (detailed description in Chapter 3).

In this chapter, we present the following contributions and findings:

- As in the work introduced in [Sedlar *et al.* 2020], to estimate the SH coefficients of the input dMRI data, we have used the Gram-Schmidt orthonormalization process. Furthermore, for the multi-shell dMRI data, we have introduced denoising layers which exploit the facts that q-space is continuous and that the sampling points are noncollinear within and between shells. The signal from one shell can thus be improved by incorporating information of each point’s direct and antipodal neighbourhood and the information from other shells.
- Secondly, we have introduced channel-wise spectral-domain non-linearities. We have investigated two types of models, one which uses zonal convolutional kernels resulting in S^2 feature maps and a second model which uses S^2 and $SO(3)$ convolutional kernels which result in $SO(3)$ feature maps. Consequently, we have introduced channel-wise S^2 and $SO(3)$ quadratic non-linearities, respectively.
- Finally, as the purpose of the sequence of the rotationally equivariant convolutional layers is to extract rotationally invariant features, contrary to the models [Cohen *et al.* 2018, Esteves *et al.* 2018, Kondor *et al.* 2018, Banerjee *et al.* 2019] which use average value of each of the output channels of the last layer (which corresponds to the spectral harmonic of degree 0), we have introduced degree-wise power spectrum features, which are also ro-

tationally invariant. They are extracted from the model's input and from the channels after each non-linearity.

- In Appendix A, we also provide derivations related to the real SH basis, Wigner-D matrices, convolutions of S^2 and $SO(3)$ signals, and Clebsch-Gordan transformations required to realize quadratic functions of the real S^2 and $SO(3)$ functions. To the best of our knowledge, some of these derivations are not available in the literature, so they can be useful for the researchers in related fields.

5.2 Theory

In this section, we describe the mathematical tools necessary to define Fourier domain rotationally equivariant CNN models with zonal, and with S^2 and $SO(3)$ kernels. Concretely, we provide definitions of convolutions and quadratic non-linearities realized in spectral domain, and rotationally invariant degree-wise power spectrum computed using generalization of the Parseval's theorem.

5.2.1 Convolution (correlation) between S^2 and zonal functions

Although previously introduced, for readability of the section, we briefly repeat definition of correlation between S^2 and zonal functions. Zonal functions are a special case of S^2 ones as they change only along the z axis, thus a correlation between an S^2 and a zonal function is a special case of spherical correlation since the resulting function remains in the S^2 space. Given an \mathbb{L}^2 function $s : S^2 \rightarrow \mathbb{R}$ and an \mathbb{L}^2 zonal function $k : S^2 \rightarrow \mathbb{R}$, where $k(\theta, \phi) = k(\theta)$ for $\theta \in [0, \pi)$, correlation between them is given by

$$[s * k](\mathbf{r}) = \int_{S^2} s(\mathbf{r}') k(R^{-1}(\theta, \phi, 0)\mathbf{r}') d\mathbf{r}' = \sum_{l=0}^B \sqrt{\frac{4\pi}{2l+1}} \hat{k}_l \sum_{m=-l}^l Y_{lm}(\mathbf{r}) \hat{s}_{lm} \quad (5.1)$$

where $\mathbf{r} = [\sin \theta \cos \phi, \sin \theta \sin \phi, \cos \theta]^T$ and $R(\theta, \phi, 0) \in SO(3)$ is a rotation matrix. \hat{s}_{lm} is the SH coefficient of degree l and order m of s . \hat{k}_l is the ZH coefficient of degree l of k . Y_{lm} is the SH real basis element of degree l and order m . If $g(\mathbf{r}) = [s * k](\mathbf{r})$, from Eq. 5.1, the SH coefficients of g are defined as:

$$\hat{g}_{lm} = \sqrt{\frac{4\pi}{2l+1}} \hat{k}_l \hat{s}_{lm}, \quad \hat{\mathbf{g}}_l = \sqrt{\frac{4\pi}{2l+1}} \hat{k}_l \hat{\mathbf{s}}_l, \quad (5.2)$$

where $\hat{\mathbf{s}}_l, \hat{\mathbf{g}}_l \in \mathbb{R}^{2l+1}$ are vectors which contain the SH coefficients of degree l of the functions s and g . Derivations of equations 5.1 and 5.2 are provided in Appendix A. An illustration of the convolution between an S^2 and a zonal function is provided in Figure 5.1.

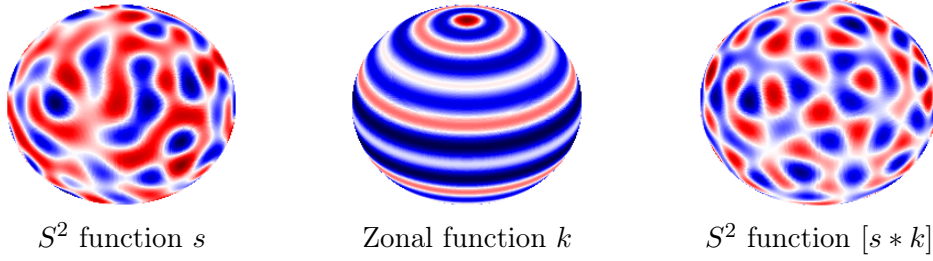


Figure 5.1: Illustration of an S^2 signal $s(\mathbf{r})$ (a), of a zonal kernel $k(\mathbf{r})$ (b) and the S^2 signal $[s * k](\mathbf{r})$ (c). All the signals are of bandwidth 16.

5.2.2 S^2 quadratic function

Given an \mathbb{L}^2 signal $g : S^2 \rightarrow \mathbb{R}$ of bandwidth B_g , $[g \times g](\mathbf{r})$ is defined as

$$[g \times g](\mathbf{r}) = \sum_{l=0}^{2B_g} \sum_{m=-l}^l \hat{h}_{lm} Y_{lm}(\mathbf{r}) \quad (5.3)$$

where

$$\hat{h}_{lm} = \sum_{l'=0}^{B_g} \sum_{l''=0}^{B_g} \sum_{m'=-l'}^{l'} \sum_{m''=-l''}^{l''} \hat{g}_{l'm'} \hat{g}_{l''m''} \sqrt{\frac{(2l'+1)(2l''+1)}{4\pi(2l+1)}} C_{l',m',l'',m''}^{l,m} C_{l',0,l'',0}^{l,0} \quad (5.4)$$

and $C_{l',q',l'',q''}^{l,q} \in \mathbb{R}$ is the Clebsch-Gordan coefficient associated to the real SH basis elements. This can be written in matrix vector notation as

$$\hat{\mathbf{h}}_l = \sum_{l',l''} \sqrt{\frac{(2l'+1)(2l''+1)}{4\pi(2l+1)}} C_{l',0,l'',0}^{l,0} C_{l',l''}^{l,0}{}^T [\hat{\mathbf{g}}_{l'} \otimes \hat{\mathbf{g}}_{l''}] \quad \text{s.t. } |l' - l''| \leq l \leq l' + l'' \quad (5.5)$$

where $C_{l',l''}^{l,0} \in \mathbb{R}^{(2l'+1)(2l''+1) \times (2l+1)}$ is the sparse Clebsch-Gordan matrix whose entries are given with $C_{l',m',l'',m''}^{l,m}$. $\hat{\mathbf{g}}_l, \hat{\mathbf{h}}_l \in \mathbb{R}^{2l+1}$ contain the real SH coefficients of degree l of the functions g and $h = g \times g$. \otimes denotes the Kronecker product of vectors. If the signal g is bandlimited to B_g , h has bandwidth $2B_g$. The definition of the Clebsch-Gordan coefficients associated to the real SH basis elements and the derivation of equations 5.3 and 5.4 are given in Appendix A. In addition to the optimization obtained by operating only on the real SH coefficients, an additional reduction of computational complexity is achieved by noting that $C_{l',l''}^{l,0}{}^T [\hat{\mathbf{g}}_{l'} \otimes \hat{\mathbf{g}}_{l''}] = C_{l'',l'}^{l,0}{}^T [\hat{\mathbf{g}}_{l''} \otimes \hat{\mathbf{g}}_{l'}]$. In the case of an S^2 non-linearity, for $l' = l'' = l$, computational complexity of $C_{l,l}^{l,0}{}^T [\hat{\mathbf{g}}_l \otimes \hat{\mathbf{g}}_l]$ is $\mathcal{O}((2l+1)^3)$.

5.2.3 Convolution (correlation) between $SO(3)$ functions

An S^2 function is a special case of an $SO(3)$ function. Given two \mathbb{L}^2 functions $s, k : S^2 \rightarrow \mathbb{R}$, their correlation is defined as:

$$[s * k](R) = \int_{S^2} s(\mathbf{r})k(R^{-1}\mathbf{r})d\mathbf{r} = \sum_{l=0}^B \sum_{m=-l}^l \sum_{n=-l}^l D_{lmn}(R)\hat{s}_{lm}\hat{k}_{ln} \quad (5.6)$$

where $R = R(\theta, \phi, \psi) \in SO(3)$ is a rotation matrix. \hat{s}_{lq} and \hat{k}_{lq} are the real SH coefficients of degree l and order q of the functions s and k . $D_{lmn} : SO(3) \rightarrow \mathbb{R}$ is an element of the real Wigner-D matrix of degree l and orders m and n . If $g(R) = [s * k](R)$, from Eq. 5.6, its Wigner-D, or here referred to as RH, coefficients are defined as

$$\hat{G}_{lmn} = \hat{s}_{lm}\hat{k}_{ln}, \quad \hat{G}_l = \hat{\mathbf{s}}_l\hat{\mathbf{k}}_l^T, \quad (5.7)$$

where $\hat{\mathbf{s}}_l, \hat{\mathbf{k}}_l \in \mathbb{R}^{2l+1}$ are the vectors which contain the real SH coefficients of degree l of the functions s and k . $\hat{G}_l \in \mathbb{R}^{(2l+1) \times (2l+1)}$ is a the matrix containing the real RH coefficient of degree l of the $SO(3)$ function g .

Given two \mathbb{L}^2 functions $s, k : SO(3) \rightarrow \mathbb{R}$, their is defined as:

$$[s * k](R) = \int_{SO(3)} s(Q)k(R^{-1}Q)dQ = \sum_{l=0}^B \sum_{m=-l}^l \sum_{n=-l}^l D_{lmn}(R) \sum_{k=-l}^l \hat{S}_{lmk}\hat{K}_{lnk} \quad (5.8)$$

where $R, Q \in SO(3)$. \hat{S}_{lpq} and \hat{K}_{lpq} are the real RH coefficients of degree l and orders p and q of the functions s and k . $D_{lpq} : SO(3) \rightarrow \mathbb{R}$ is an element of the real Wigner-D matrix of degree l and orders p and q . If $g(R) = [s * k](R)$, from Eq. 5.8, its RH coefficients are defined as:

$$\hat{G}_{lmn} = \sum_{k=-l}^l \hat{S}_{lmk}\hat{K}_{lnk}, \quad \hat{G}_l = \hat{S}_l\hat{K}_l^T, \quad (5.9)$$

where $\hat{S}_l, \hat{K}_l \in \mathbb{R}^{(2l+1) \times (2l+1)}$ are the matrices which contain the real RH coefficients of degree l of the functions s and k . $\hat{G}_l \in \mathbb{R}^{(2l+1) \times (2l+1)}$ is a matrix containing the real RH coefficient of degree l of the function g . Derivations of equations 5.6, 5.7, 5.8 and 5.9 are provided in Appendix A.

5.2.4 $SO(3)$ quadratic function

Given an \mathbb{L}^2 signal $g : SO(3) \rightarrow \mathbb{R}$ of bandwidth B_g , $[g \times g](R)$ is defined as:

$$[g \times g](R) = \sum_{l=0}^{2B_g} \sum_{m=-l}^l \sum_{n=-l}^l \hat{H}_{lmn}D_{lmn}(R) \quad (5.10)$$

where

$$\hat{H}_{lmn} = \sum_{l'=0}^{B_f} \sum_{l''=0}^{B_g} \sum_{m'=-l'}^{l'} \sum_{n'=-l'}^{l'} \sum_{m''=-l''}^{l''} \sum_{n''=-l''}^{l''} \hat{G}_{l'm'n'} \hat{G}_{l''m''n''} C_{l',m',l'',m''}^{l,m} C_{l',n',l'',n''}^{l,n} \quad (5.11)$$

and $C_{l',q',l'',q''}^{l,q} \in \mathbb{R}$ is the Clebsch-Gordan coefficient associated to the real RH basis elements. Similarly as in Eq. 5.5, this can be written in matrix notation as:

$$\hat{H}_l = \sum_{l',l''} C_{l',l''}^l T [\hat{G}_{l'} \otimes \hat{G}_{l''}] C_{l',l''}^l \quad \text{s.t.} \quad |l' - l''| \leq l \leq l' + l'' \quad (5.12)$$

where $C_{l',l''}^l \in \mathbb{R}^{(2l'+1)(2l''+1) \times (2l+1)}$ is the Clebsch-Gordan matrix as used in Eq. 5.5. $\hat{G}_l, \hat{H}_l \in \mathbb{R}^{(2l+1) \times (2l+1)}$ contain the real RH coefficients of degrees l of the signals g and $h = [g \times g]$. \otimes denotes the Kronecker product of matrices. If the signal g is bandlimited to B_g , h has bandwidth $2B_g$. The derivation of equations 5.11 and 5.12 is given in Appendix A. In addition to the optimization obtained due to the operations on the real RH coefficients, symmetry $C_{l',l''}^l T [\hat{G}_{l'} \otimes \hat{G}_{l''}] C_{l',l''}^l = C_{l'',l'}^l T [\hat{G}_{l''} \otimes \hat{G}_{l'}] C_{l'',l'}^l$, an additional reduction of the computational complexity is obtained as follows. First, we remark that Eq. 5.12 can be written as

$$\hat{H}_l = \sum_{l',l''} C_{l',l''}^l T (\hat{G}_{l'} \otimes I_{2l''+1}) (I_{2l'+1} \otimes \hat{G}_{l''}) C_{l',l''}^l = \sum_{l',l''} \hat{V}_{l',l''}^l [\hat{U}_{l',l''}^l]^T \quad (5.13)$$

s.t. $|l' - l''| \leq l \leq l' + l''$

where the computation of

$$\hat{V}_{l',l''}^l = C_{l',l''}^l T (\hat{G}_{l'} \otimes I_{2l''+1}) \text{ is optimized by } \hat{V}_{l',l''}^l [q, :] = \text{vec}(\tilde{C}_{l',l''}^l [q, :, :]^T \hat{G}_{l'}) \quad (5.14)$$

and

$$\hat{U}_{l',l''}^l = (I_{2l'+1} \otimes \hat{G}_{l''}) C_{l',l''}^l \text{ is optimized by } \hat{U}_{l',l''}^l [q, :] = \text{vec}(\hat{G}_{l''} \tilde{C}_{l',l''}^l [q, :, :]) \quad (5.15)$$

where $q \in \{-l, \dots, 0, \dots, l\}$. I_{2l+1} is the identity matrix of size $(2l+1) \times (2l+1)$. $\tilde{C}_{l',l''}^l \in \mathbb{R}^{(2l+1) \times (2l''+1) \times (2l'+1)}$ is 3D tensor obtained by reshaping the Clebsch-Gordan matrix $C_{l',l''}^l$. If we assume naive matrix and tensor product, for $l' = l'' = l$, replacing $C_{l,l}^l T [\hat{G}_l \otimes \hat{G}_l] C_{l,l}^l$ by the optimized $\hat{V}_{l,l}^l [\hat{U}_{l,l}^l]^T$ expression as given in equations 5.14 and 5.15, reduces the computational complexity from $\mathcal{O}((2l+1)^5 + 2(2l+1)^4)$ to $\mathcal{O}(3(2l+1)^4)$.

5.2.5 Power spectrum of S^2 and $SO(3)$ functions

From the generalization of Parseval's theorem to S^2 and $SO(3)$ functions, given \mathbb{L}^2 functions $g(\mathbf{r}) : S^2 \rightarrow \mathbb{R}$ and $g(R) : SO(3) \rightarrow \mathbb{R}$, the angular and rotation power

spectra corresponding to the spectral degree l are defined as

$$p_l = \sum_{m=-l}^l \hat{g}_{lm}^2, \quad P_l = \frac{8\pi^2}{2l+1} \sum_{m=-l}^l \sum_{n=-l}^l \hat{G}_{lmn}^2 \quad (5.16)$$

where $p_l, P_l \in \mathbb{R}$. \hat{g}_{lm} is the real **SH** coefficient of degree l and order m of the signal $g(\mathbf{r})$ and \hat{G}_{lmn} is the real **RH** coefficient of degree l and orders m and n of the signal $g(R)$.

5.3 Methods

We have investigated two types of Fourier domain rotation equivariant CNNs. One with zonal kernels and S^2 quadratic non-linearities, termed as *Fourier_S^2_zonal* and another one with S^2 and $SO(3)$ kernels and $SO(3)$ quadratic non-linearities, termed as *Fourier_S^2_SO(3)*. Although both types of convolutional layers are rotation equivariant, here we stress the essential differences between them. First, the number of their spectral components of a zonal, an S^2 and an $SO(3)$ kernels of bandwidth L , is $L+1$, $(L+1)^2$ and $(L+1)(4(L+1)^2-1)/3$, respectively. This means that the S^2 and $SO(3)$ kernels have a higher discrimination power. Thus, to make a distinction between two patterns on sphere, one would need to use more zonal kernels than S^2 or $SO(3)$. On the other hand, convolution with zonal kernels is less computationally expensive. In addition, for an S^2 signal input, convolution with a zonal kernel results in S^2 signal, whose quadratic function is much less computationally expensive than the quadratic function of the $SO(3)$ signals.

Architecture of the two models are illustrated in Figures 5.2 and 5.3. As input they take raw multi-shell **dMRI** signals. Since q-space is continuous, signals acquired over different shells are correlated. In addition, since they are sampled at points which are noncollinear within and between shells, they contain a certain amount of supplementary information. To make a use of this and taking into account that **dMRI** signals are positive, we have incorporated into the models a denoising layer composed of a cascade of non-linear layers defined as

$$\mathbf{s}^{(n)} = \text{ReLU}((I + \lambda W_n)\mathbf{s}^{(n-1)}) \quad (5.17)$$

where $\mathbf{s}^{(0)} = [s_0^{sh=1} \dots s_{N_1}^{sh=1}, \dots, s_0^{sh=K} \dots s_{N_K}^{sh=K}]^T$ is a vector that contains concatenated raw **dMRI** signals of K shells, where N_k is the number of points for shell k . Vectors $\{\mathbf{s}^{(n)}\}$ contain denoised **dMRI** signals after application of n denoising steps. I is the identity matrix, $\{W_n\}$ are trainable weights and λ is parameter which ensures that matrices $\{(I + \lambda W_n)\}$ remain close to identity matrix and in this way preserve the spherical nature of the input signal. After the denoising layer, the signals are transformed to the Fourier domain using the real **SH** basis, of even degrees, inverted with Gram-Schmidt orthonormalization process as in [Sedlar *et al.* 2020, Sedlar *et al.* 2021] and as described in Chapter 4. In the context of standard **CNN**, a shell corresponds to a channel. We denote input **SH** coefficients of degree l and of channel k as $\hat{\mathbf{a}}_l^{0,k}$, where $l \in \{0, 2, \dots, L\}$, with L being the input's bandwidth.

5.3.1 Fourier domain CNN with quadratic S^2 non-linearities

In the model with zonal kernels *Fourier_* S^2 *_zonal*, convolutions in the n^{th} convolutional layer are defined as

$$\hat{\mathbf{z}}_l^{n,i} = \sum_j \hat{\mathbf{a}}_l^{n-1,j} \hat{w}_l^{n,j,i} \quad \text{for } l \neq 0, \quad \hat{\mathbf{z}}_0^{n,i} = \sum_j \hat{\mathbf{a}}_0^{n-1,j} \hat{w}_0^{n,j,i} + \hat{b}_0^{n,i} \quad \text{for } l = 0 \quad (5.18)$$

where $\hat{w}_l^{n,j,i}$ is a ZH coefficient of the convolutional kernel in n^{th} layer, corresponding to the input channel j and output channel i , while $\hat{b}_0^{n,i}$ is corresponding bias term. $\hat{\mathbf{a}}_l^{n-1,j}$ and $\hat{\mathbf{z}}_l^{n,i}$ are the vectors containing input and output SH coefficients of degree l for the channels j and i , respectively.

The output or the activation of the n^{th} S^2 nonlinear layer is obtained using Eq. 5.5 as

$$\hat{\mathbf{a}}_l^{n,i} = \sum_{l',l''} \sqrt{\frac{(2l'+1)(2l''+1)}{4\pi(2l+1)}} C_{l',0,l'',0}^{l,0} C_{l',l''}^{l,T} [\hat{\mathbf{z}}_{l'}^{n,i} \otimes \hat{\mathbf{z}}_{l''}^{n,i}] \quad \text{s.t. } |l' - l''| \leq l \leq l' + l'' \quad (5.19)$$

where $C_{l',l''}^l \in \mathbb{R}^{(2l'+1)(2l''+1) \times (2l+1)}$ is the sparse Clebsch-Gordan matrix. $\hat{\mathbf{z}}_l^{n,i}$, $\hat{\mathbf{a}}_l^{n,i}$ are the input and output SH coefficients of degree l of the i^{th} channel. This type of non-linearity is similar to the one proposed in [Kondor *et al.* 2018], with a difference that it is channel wise, and thus it does not lead to quadratic increase of the output channels.

As in [Cohen *et al.* 2018, Esteves *et al.* 2018, Kondor *et al.* 2018], pooling is achieved by discarding high frequency spectral components. Simply, the $\hat{\mathbf{a}}_l^{n,i}$ is computed only for $l < L^n$, where L^n is the output bandwidth of the layer n .

Rotationally invariant power spectrum features are extracted from the input SH coefficients and after each non-linearity. The feature vector is defined as

$$\mathbf{f} = [r_0^{0,1}, \dots, r_0^{0,K}, \dots, r_L^{0,1}, \dots, r_L^{0,K}, \dots, r_0^{n,1}, \dots, r_0^{n,K^n}, \dots, r_{L^n}^{n,1}, \dots, r_{L^n}^{n,K^n}, \dots] \quad (5.20)$$

where K^n refers to the number of output channels of the layer n . $r_l^{n,k}$ is defined using Eq. 5.16 as

$$r_l^{n,k} = \sum_{m=-l}^l [\hat{a}_{lm}^{n,k}]^2. \quad (5.21)$$

Concatenated rotationally invariant power spectrum features are fed into a fully connected network which performs the final inference.

5.3.2 Fourier domain CNN with quadratic $SO(3)$ non-linearities

In the model with S^2 and $SO(3)$ kernels, *Fourier_* S^2 *_* $SO(3)$, convolutions are realized as firstly proposed in [Cohen *et al.* 2018]. Convolution in the 1^{st} convolutional

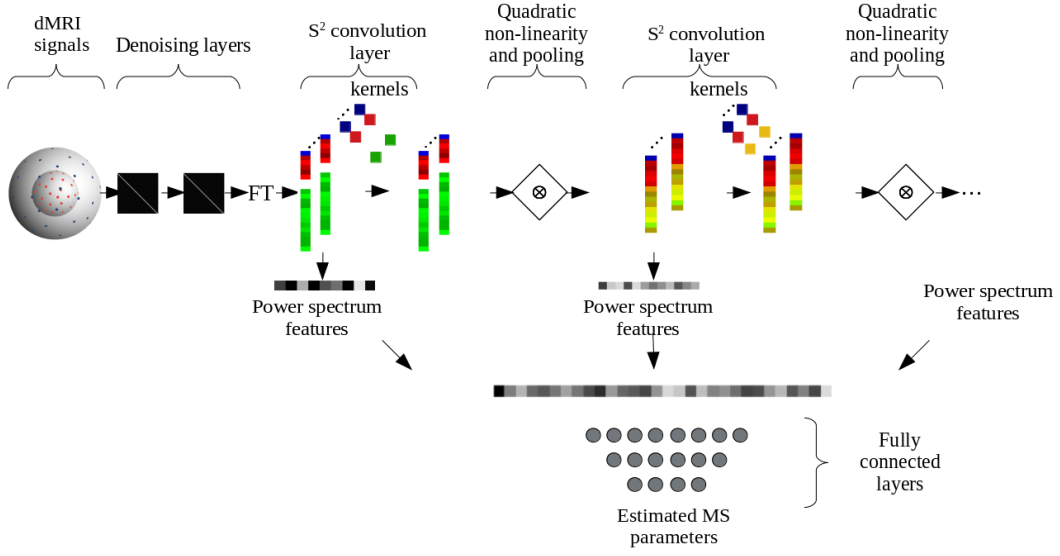


Figure 5.2: Architecture of the proposed model with zonal convolutional kernels and S^2 quadratic non-linearities. The model is termed as *Fourier_S²_zonal*.

layer is defined as:

$$\hat{Z}_l^{1,i} = \sum_j \hat{\mathbf{a}}_l^{0,j} [\hat{\mathbf{w}}_l^{1,j,i}]^T \quad \text{for } l \neq 0, \quad \hat{Z}_0^{1,i} = \sum_j \hat{\mathbf{a}}_0^{0,j} \hat{\mathbf{w}}_0^{1,j,i} + \hat{b}_0^{1,i} \quad \text{for } l = 0, \quad (5.22)$$

where $\hat{\mathbf{w}}_l^{1,j,i}$ are the SH coefficients of the S^2 convolutional kernel in the 1st layer, corresponding to the input channel j and output channel i , while $\hat{b}_0^{1,i}$ is corresponding bias term. $\hat{Z}_l^{1,i}$ is the matrix containing the output RH coefficients of degree l for the channel i .

Since the output of the first and all the following non-linear layers is an $SO(3)$ signal represented in the Fourier domain, convolution in the n^{th} convolutional layer ($n > 1$) is defined as:

$$\hat{Z}_l^{n,i} = \sum_j \hat{A}_l^{n-1,j} [\hat{W}_l^{n,j,i}]^T \quad \text{for } l \neq 0, \quad \hat{Z}_0^{n,i} = \sum_j \hat{A}_0^{n-1,j} \hat{W}_0^{n,j,i} + \hat{B}_0^{n,i} \quad \text{for } l = 0 \quad (5.23)$$

where $\hat{W}_l^{n,j,i}$ are the RH coefficients of the $SO(3)$ convolutional kernel in the n^{th} layer, corresponding to the input channel j and output channel i , while $\hat{B}_0^{n,i}$ is the corresponding bias term. $\hat{A}_l^{n-1,j}$ and $\hat{Z}_l^{n,i}$ are the vectors containing input and output RH coefficients of degree l for the channels j and i , respectively.

The output of the n^{th} $SO(3)$ nonlinear layer is obtained using Eq. 5.12 as:

$$\hat{A}_l^{n,i} = \sum_{l',l''} C_{l',l''}^l{}^T [\hat{Z}_{l'}^{n,i} \otimes \hat{Z}_{l''}^{n,i}] C_{l',l''}^l \quad \text{s.t. } |l' - l''| \leq l \leq l' + l'' \quad (5.24)$$

where $C_{l',l''}^l \in \mathbb{R}^{(2l'+1)(2l''+1) \times (2l+1)}$ is the sparse Clebsch-Gordan matrix. $\hat{Z}_l^{n,i}, \hat{A}_l^{n,i}$ are the input and output RH coefficients of degree l of the i^{th} channel.

In this model as well, pooling is achieved by discarding spectral components of the highest degree, thus $\hat{A}_l^{n,i}$ are computed only for $l < L^n$, with L^n being the output bandwidth of the layer n . Rotationally invariant power spectrum features are extracted from the input SH coefficients and from the RH coefficients after each non-linearity. The feature vector is defined as

$$f = [r_0^{0,1}, \dots, r_0^{0,K}, \dots, r_L^{0,1}, \dots, r_L^{0,K}, \dots, r_0^{n,1}, \dots, R_0^{n,K^n}, \dots, R_{L^n}^{n,1}, \dots, R_{L^n}^{n,K^n}, \dots] \quad (5.25)$$

where K^n refers to the number of output channels of the layer n . $r_l^{n,k}$ is defined as in Eq. 5.21 and $R_l^{n,k}$ according to Eq. 5.16 as:

$$R_l^{n,k} = \sum_{m=-l}^l \sum_{n=-l}^l [\hat{A}_{lmn}^{n,k}]^2, \quad (5.26)$$

where the scaling factor $\frac{8\pi^2}{2l+1}$ is omitted in order to have more balanced magnitudes of the power spectrum features. As in the model with zonal kernels, concatenated rotation invariant power spectrum features are fed into a fully connected network which performs the final inference.

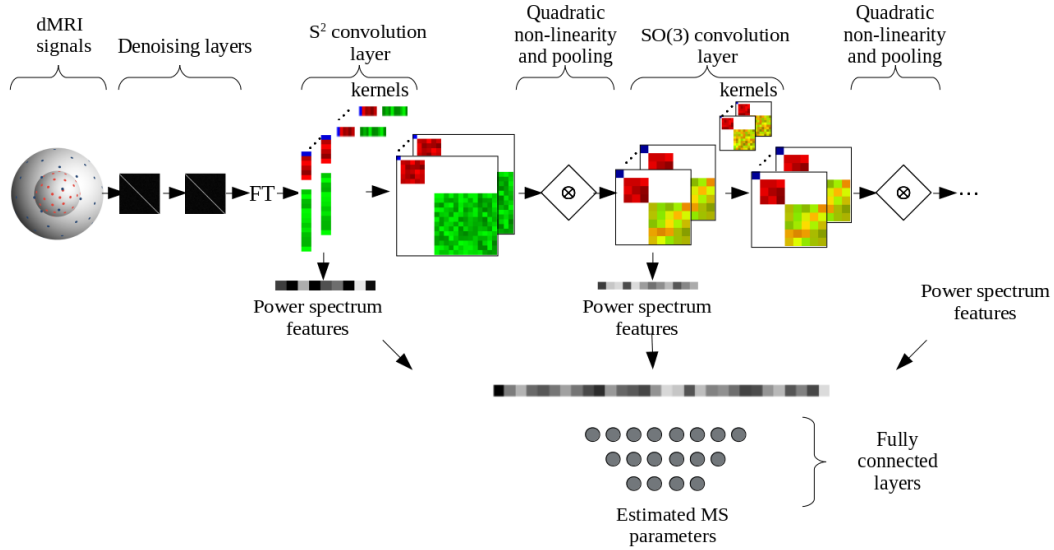


Figure 5.3: Architecture of the proposed model with zonal convolutional kernels and S^2 quadratic non-linearities. The model is termed as *Fourier_S²_SO(3)*.

5.4 Experiments

Firstly, we have compared our model with zonal kernels with a state-of-the-art spherical CNN model, namely S^2CNN proposed by [Cohen *et al.* 2018]. Due to the differences in sampling grids, the models are compared on synthetic dMRI data on the classification problem of axon bundle count. Furthermore, the models are extensively compared with the dMRI state-of-the-art deep learning approaches, namely MLP [Golkov *et al.* 2016], MEDN and MEDN+ [Ye 2017], MescNet [Ye *et al.* 2019] and MescNetSepDict [Ye *et al.* 2020], on the problem of NODDI [Zhang *et al.* 2012] and spherical mean technique (SMT) [Kaden *et al.* 2016] microstructure parameter estimation from dMRI acquired with significantly reduced sampling scheme. Finally, we demonstrate that our model can be successfully used to extract rotation invariant features for brain tissue segmentation, obtaining results comparable to the recently proposed deep learning approach [Zhang *et al.* 2021], while requiring significantly less computational time.

5.4.1 Axon bundle counting experiment

In this experiment, we have compared our *Fourier_S^2_zonal* model with the state-of-the-art S^2CNN [Cohen *et al.* 2018] model on synthetic data on the problem of the axon bundle counting. The experiments highlight the importance of the spectral domain non-linearity used in our model.

Synthetic database

We have generated synthetic dMRI samples distributed over four classes containing zero, one, two or three axon bundles. Data is generated using single fiber white matter, gray matter and CSF response functions and corresponding estimated PDFs of one HCP subject ('100307'). The tissue response functions were estimated using the *mrtrix* command *dwi2response msmt_5tt* and corresponding PDFs with *multi-shell multi-tissue* CSD [Jeurissen *et al.* 2014] with the command *dwi2fod msmt_csd* [Tournier *et al.* 2019]. SH coefficients of response functions for a shell k are noted as $\hat{r}_k^{gm}, \hat{r}_k^{csf} \in \mathbb{R}^1$ and $\hat{r}_k^{sfwm} \in \mathbb{R}^{N_{sh}}$, for gray matter, CSF and single fiber white matter, respectively, where N_{sh} is the number of SH coefficients.

The SH coefficients of synthetic dMRI signals for a shell k are computed as follows:

$$\hat{\mathbf{s}}_k = \nu_{gm} \sqrt{4\pi} \hat{p}^{gm} \hat{r}_k^{gm} + \nu_{csf} \sqrt{4\pi} \hat{p}^{csf} \hat{r}_k^{csf} + \nu_{wm} \sum_{b=1}^{N_b} \nu_{sfwm}^b R_b(\mathbf{c} \odot \hat{\mathbf{p}}_b^{sfwm} \odot \hat{\mathbf{r}}_k^{sfwm}) \quad (5.27)$$

where $\nu_{gm}, \nu_{csf}, \nu_{wm}$ are tissue fractions, ν_{sfwm}^b are axon bundle fractions and $N_b \in \{1, 2, 3\}$ is the number of axon bundles. $\hat{p}^{gm}, \hat{p}^{csf} \in \mathbb{R}$ are the SH coefficients of PDFs of gray matter and CSF (these tissues are modeled as a sphere, thus they have only the SH coefficient of $l = 0$). $\hat{\mathbf{p}}_b^{sfwm} \in \mathbb{R}^{N_{sh}}$ is the fODF of white matter bundle b oriented along z axis. R_b is rotation matrix for bundle b . Vector $\mathbf{c} \in \mathbb{R}^{N_{sh}}$ is

a constant vector $\mathbf{c} = [\sqrt{4\pi}, 0, 0, \sqrt{\frac{4\pi}{2 \cdot 2 + 1}}, \dots, \sqrt{\frac{4\pi}{2 \cdot 4 + 1}}, \dots]$ used in the convolution between response function $\hat{\mathbf{r}}_k^{sfwm}$ and fODF $\hat{\mathbf{p}}_b^{sfwm}$. To simulate white matter samples, we set $\nu_{wm} = 1$ and $\nu_{gm}, \nu_{csf} \sim |\mathcal{N}(0, 0.05)|$, to simulate gray matter $\nu_{gm} = 1$ and $\nu_{wm}, \nu_{csf} \sim |\mathcal{N}(0, 0.05)|$ and to simulate CSF $\nu_{csf} = 1$ and $\nu_{wm}, \nu_{gm} \sim |\mathcal{N}(0, 0.05)|$. Axon bundle fractions are drawn from uniform distribution where minimum ν_{sfwm}^b is 0.2. Realistic PDFs are drawn from random distributions $\hat{p}_k^{gm} \sim \mathcal{N}(\hat{p}_m^{gm}, \hat{p}_{std}^{gm})$, $\hat{p}_k^{csf} \sim \mathcal{N}(\hat{p}_m^{csf}, \hat{p}_{std}^{csf})$, $\hat{\mathbf{p}}_k^{sfwm} \sim \mathcal{N}(\hat{\mathbf{p}}_m^{sfwm}, \hat{\mathbf{p}}_{std}^{sfwm})$. The mean and standard deviation of gray matter and CSF tissue PDFs are computed over corresponding regions determined with five-tissue-type segmentation with FAST algorithm applied on T1w images [Zhang *et al.* 2001]. Single fiber white matter PDFs - fODFs - are selected from brain regions with high fractional anisotropy (> 0.75), they are aligned with z-axis and mean and standard deviation are computed for each zonal harmonic. Rotation of the axon bundle is performed in a way that the minimum angle between bundles is $\frac{\pi}{6}rad$. Bandwidth of generated signals is $L = 8$, thus $N_{sh} = 45$ and they are composed of three shells with b values 1000, 2000, 3000 s/mm^2 . The total number of generated samples is 10^6 , where 0.2×10^6 has been used for training, 0.2×10^6 for validation and 0.6×10^6 for testing. Once the SH coefficients are converted to signal domain they are distorted by a non-additive Rician noise of $SNR = 20$ and afterwards normalized with mean $b = 0$ value and clipped to the range $[0, 1]$. Number of no diffusion weighted signals ($b = 0$) is 18.

To investigate how the models behave with dMRI data with different angular resolutions and to verify their rotation invariance, we have created three datasets (**db 1**, **db 2**, **db 3**). Each of the dataset is generated for two types of grids, Driscoll-Healy grid [Driscoll & Healy 1994] used in the model S^2CNN [Cohen *et al.* 2018] and q-space sampling used in dMRI imaging [Caruyer *et al.* 2013]. In **db1**, SH coefficients of generated samples (degree 8) are projected on 91 and 90 points for Driscoll-Healy and q-space sampling grids, respectively. This corresponds to a bandwidth $L = 4$ for Driscoll-Healy grid. In **db2**, SH coefficients of generated samples (degree 8) are projected on 57 points, which corresponds to $L = 3$ for Driscoll-Healy grid. In **db3**, to investigate rotation invariance of the models, training and validation samples are generated with a restriction on their orientation, while testing samples contain bundles of arbitrary orientation. Concretely, the first bundle is always aligned with z axis, if there are two bundles, the second one is always in $z - x$ plane drawn from uniform distribution $[\frac{\pi}{6}, \frac{\pi}{2}]rad$, if there are three bundles, the third one is rotated for $\theta < \frac{\pi}{2}rad$ and $\phi < \pi rad$, while respecting that the angle with respect to the other two bundles is greater than $\frac{\pi}{6}rad$. Properties of the datasets in terms of number of points with corresponding grid types, and bundle orientations are summarized in Table 5.1. Illustrations of the noiseless fODFs and dMRI for three shells of **db 1** and **db 2** are illustrated in Figure 5.4 and for **db 3** in Figure 5.5.

Table 5.1: Overview of the synthetic databases. Comparing models include S^2CNN [Cohen *et al.* 2018]. Grid type DH refers to Driscoll-Healy [Driscoll & Healy 1994] and Q to multi-shell q-space sampling [Caruyer *et al.* 2013].

<i>Database</i>	<i>db 1</i>		<i>db 2</i>		<i>db 3</i>	
<i>Model</i>	S^2CNN	Our	S^2CNN	Our	S^2CNN	Our
<i>Grid type</i>	DH	Q	DH	Q	DH	Q
<i>No. of points</i>	91	90	57	57	57	57
<i>Bundle orientations</i>	arbitrary		arbitrary		restricted	

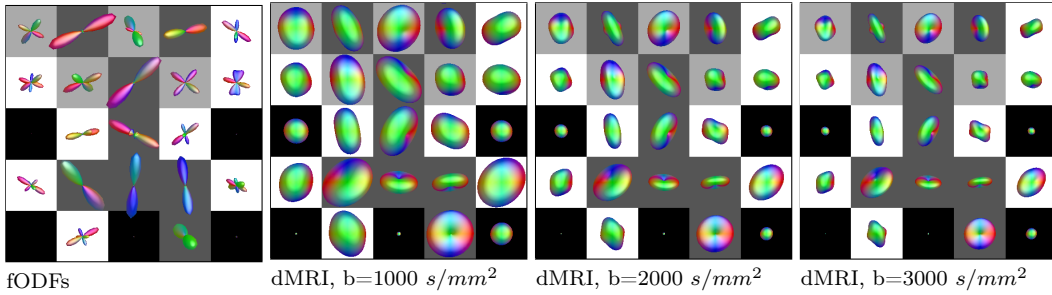


Figure 5.4: Simulated fODFs and dMRI signals with arbitrary orientations of bundles. Background color corresponds to the number of bundles (black-zero bundles, dark gray - one bundle, light gray - two bundle, white - three bundles).

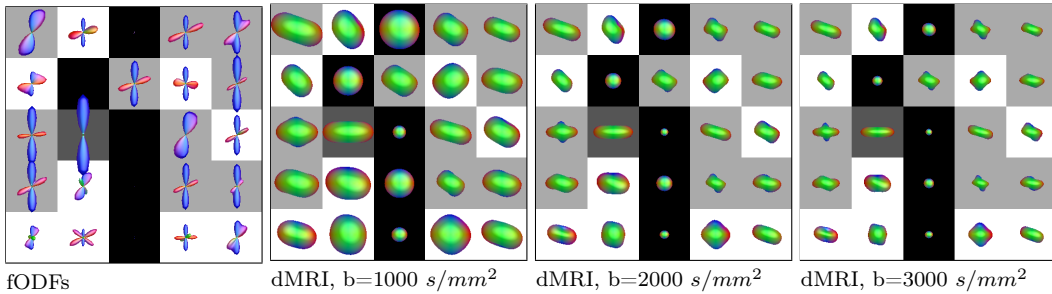


Figure 5.5: Simulated fODFs and dMRI signals with restricted orientations of bundles. Background color corresponds to the number of bundles (black-zero bundles, dark gray - one bundle, light gray - two bundle, white - three bundles).

Implementation details

Our model is implemented with the *tensorflow* library [Abadi *et al.* 2015] and compared to the model S^2CNN implemented with the *torch* [Collobert *et al.* 2002]. These models have been trained over 200 epochs by minimizing categorical cross-entropy loss using an Adam optimizer [Kingma & Ba 2014]. Initial learning rate has been set to 0.001 and batch size to 128. If the difference between validation categorical cross-entropy averaged over two sequential blocks of five epochs is smaller than 10^{-3} , learning rate is reduced by factor of 0.95. For 91 sam-

pling points, S^2CNN has three convolutional layers with input and output bandwidths $(4, 4), (4, 2), (2, 0)$, while for 57 sampling points the bandwidths are $(3, 3), (3, 1), (1, 0)$. For both sampling schemes, containing 90 and 57 points, we have evaluated $Fourier_S^2_zonal$ with three convolutional layers with two different sets of bandwidths, $(8, 4), (4, 2), (2, 0)$ and $(4, 4), (4, 2), (2, 0)$. The number of input and output channels in convolutional layers is $(3, 8), (8, 16), (16, 32)$ and $(3, 16), (16, 32), (32, 64)$, for S^2CNN and $Fourier_S^2_zonal$, respectively, since the number of trainable weights in zonal kernels is much smaller than in S^2 and $SO(3)$ convolutional kernels used in S^2CNN . The extracted rotation invariant features are classified with a fully connected network composed of three layers with output sizes 32, 16, 4. In our models, we have taken into account antipodal symmetry of dMRI signals, thus the convolutional kernels are antipodally symmetric as well. In this experiment, since the number of sampling points is considerably higher than the number of SH basis elements (45 and 15), the model does not contain any denoising layer.

Results

Classification is compared in terms of confusion matrices illustrated in Figures 5.6, 5.7 and 5.8, for *db1*, *db2* and *db3*, respectively. In Figure 5.6, we can notice that the classification accuracy of S^2CNN and $Fourier_S^2_zonal$ are comparable and that both models meet some difficulties in distinguishing between samples containing 2 and 3 axon bundles. This can be a consequence of the lower amplitude of the dMRI signals as the number of bundles increases from 1 to 3, as their volume fractions sum to 1. Figure 5.7 shows that our models keep high classification accuracy even when the number of sampling points is significantly reduced. On the other hand, accuracy of S^2CNN significantly decreases, which might be a consequence of the fact that the model can extract only low frequency information of maximal bandwidth 3. In addition, taking into account antipodal symmetry of the input signals, in S^2CNN , valuable information of the SH coefficients are found only for the degrees 0 and 2. (We denote that for 57 points, with quadrature formulae associated to Driscoll-Healy grid, we cannot compute SH coefficients of a higher degree.) In Figure 5.8, the obtained results highlight the impact of the aliasing introduced by *ReLU* non-linearity applied in the signal domain used in S^2CNN and the benefit of the spectral domain non-linearity used in our models. The S^2CNN is only capable to make a distinction between white and non-white matter samples. For one such inference, a mean of the signal is sufficient (only the SH coefficients of $l = 0$). On the other hand, by comparing the results obtained with *db2* and *db3*, we can also notice that our models preserve a high degree of rotation invariance.

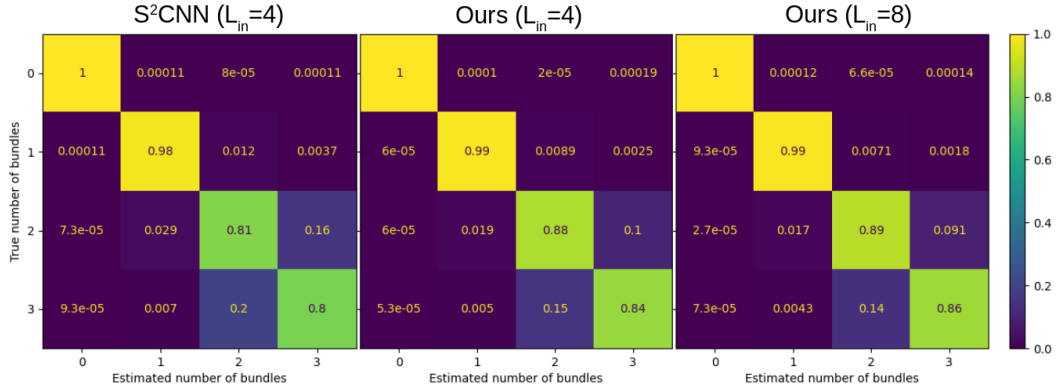


Figure 5.6: Comparison of confusion matrices for the number of axon bundle classification problem, for db1 where axon bundles are arbitrarily oriented in all, train, validation and test subsets, and the number of sampling points is 91 (S^2CNN) and 90 ($Fourier_S^2_zonal$). $SNR = 20$.

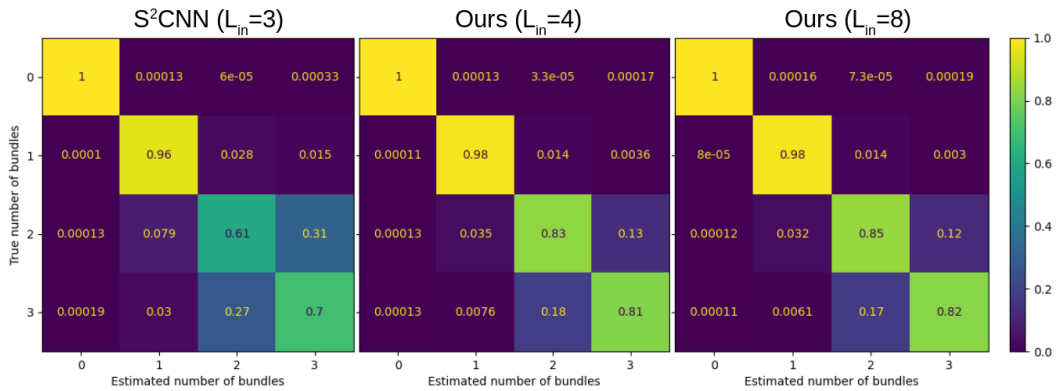


Figure 5.7: Comparison of confusion matrices for the number of axon bundle classification problem, for db2 where axon bundles are arbitrarily oriented in all, train, validation and test subsets, and the number of sampling points is 57. $SNR = 20$.

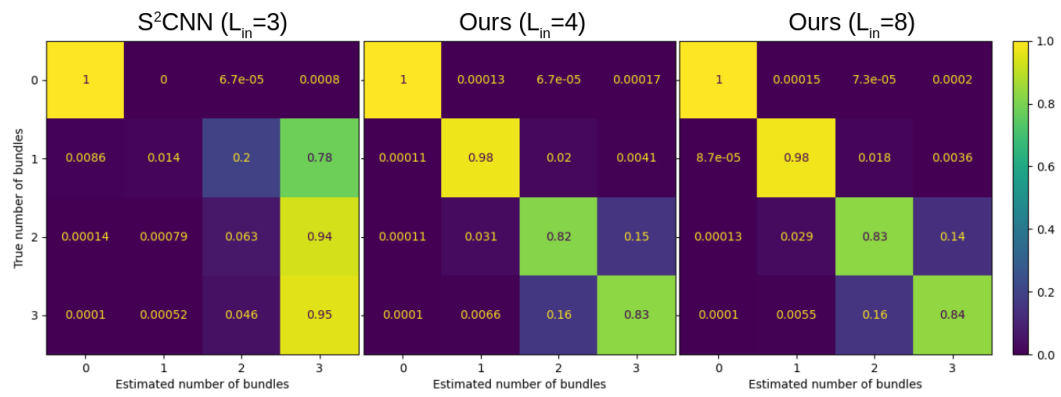


Figure 5.8: Comparison of confusion matrices for axon bundle classification problem, for db3 where the axon bundles orientation is restricted in train and validation subsets, and the number of sampling points is 57. $SNR = 20$.

5.4.2 Multi-compartment micro-structure estimation

In these experiments, we have extensively compared our models *Fourier_S²_SO(3)*, *Fourier_S²_zonal* with the deep learning approaches MLP [Golkov *et al.* 2016], MEDN and MEDN+ [Ye 2017], MescNet [Ye *et al.* 2019] and MescNetSepDict [Ye *et al.* 2020], on the problem of NODDI and SMT microstructure parameter estimation from dMRI signals acquired with a significantly reduced number of q-space sampling points. Concretely, the NODDI parameters include intracellular volume fraction ν_{ic} , isotropic volume fraction ν_{iso} and orientation dispersion indices denoted with *OD*. SMT parameters include extra-neurite fraction ν_{ext} and intrinsic diffusion coefficient λ . In analogy to MEDN+, the FCN+ is designed as the version of FCN which takes as input the signals from a small neighbourhood. For a neighbourhood of size $3 \times 3 \times 3$ the size of input vector is increased by factor 27. In a similar manner, we have created *Fourier_S²_SO(3)+* and *Fourier_S²_zonal+*, which take as input signals from a small neighbourhood treated as different channels.

Real data from HCP and estimation of gold standard

We have used in our experiments a subset of 200 subjects from the Human Connectome Project (HCP) database [Van Essen *et al.* 2013]. We have used 1, 3, 5, 10, 15 or 30 subjects for training, 20 for validation and 150 for the final testing of the algorithm. dMRI scans have been acquired on a Siemens 3T Skyra system with a gradient strength of $100mT/m$. Scans are composed of three shells with b-values of 1000, 2000 and $3000 s/mm^2$, each with 90 gradient directions and 18 $b = 0$ images at resolution $1.25 \times 1.25 \times 1.25 mm^3$. We have used scans that were previously registered to T1w images. As a consequence, although acquired with the same acquisition protocol, after registration, gradient directions and b-values slightly differ from their initial values and between subjects. To select brain region voxels, we have used brain masks provided as a part of HCP dataset, obtained from no diffusion weighted images ($b = 0$) using Otsu thresholding algorithm. Masks are post-processed by excluding voxels with very low mean $b = 0$ value (lower than 100) as they correspond to border voxels with likely erroneous data. dMRI signals are voxel-wise normalized with mean value of $b = 0$ scans and clipped to the range $[0, 1]$. For the estimation of gold standard we have used *brute2fine* optimizer from *dmipy* toolbox applied on dMRI data with full acquisition scheme. Models are compared with dMRI signals acquired over a significantly reduced sampling scheme, containing 30 points over two shells of b-values 1000 and $2000 s/mm^2$.

Implementation details

The models are implemented with the *tensorflow*. They are trained over 300 epochs, where in each epoch 25600 voxels (or windows of size $3 \times 3 \times 3$) are randomly drawn from T training samples, where $T \in \{1, 3, 5, 10, 15, 30\}$. Validation is performed on

25600 voxels randomly drawn from 20 validation subjects. If the difference between validation loss averaged over two sequential blocks of five epochs is smaller than 10^{-6} , the learning rate is reduced by the factor of 0.95. Testing is performed on 150 testing subjects. Models have been trained with batch size of 128 by minimizing mean square error loss using an Adam optimizer [Kingma & Ba 2014].

Results

Results are compared quantitatively in terms of mean absolute error computed over the 150 testing subjects. The mean absolute error and corresponding standard deviations for NODDI parameter estimation, namely ν_{ic} , ν_{iso} and OD , for training on 1, 3, 5, 10, 15, 30 subjects are illustrated in Figure 5.9 for the models which take as input single voxels. Comparison of the models which take as input signals from 3D patches is provided in Figure 5.10. For the single voxel models, we have performed an extensive hyperparameter grid search provided in Appendix B. Figure 5.9 shows that our models *Fourier_S²_zonal* and *Fourier_S²_SO(3)* with the number of trainable parameters $0.0915 \cdot 10^6$ and $0.0789 \cdot 10^6$, respectively give on average similar mean absolute error as MLP with $\sim 0.148 \cdot 10^6$ parameters. Further, we can see that the model *MEDN*, with 0.11×10^6 trainable parameters, which is specifically designed for NODDI parameter estimation yields noticeable higher mean absolute errors for the parameter ν_{iso} regardless of the number of training subjects. More important differences in the mean absolute errors can be observed by comparing the methods which take as input 3D patches, which are compared for the number of training subjects 1, 3 and 5. We can see that our models yield errors slightly higher but comparable with the recently proposed state-of-the-art MESCNetSepDict, with the number of parameters decreased by factors 2.7 and 4.4 for *Fourier_S²_SO(3)+* and *Fourier_S²_zonal+*, respectively. Although the number of parameters is not necessarily proportional to the computational time (for example, the training and testing with MESCNet is more than 8 times faster than with MESCNetSepDict), *Fourier_S²_SO(3)+* is approximately 6 times faster and *Fourier_S²_zonal+* 12 times. As for the single voxel methods, for the 3D patch based methods we can also notice that the model specifically designed for NODDI parameters *MEDN+* yields the highest mean absolute errors over all three parameters ν_{ic} , ν_{iso} and OD . Figures 5.11, 5.12 and 5.13 show qualitative comparison of NODDI parameters estimated with single-voxel and 3D patch based models, trained on one subject. We can see that single-voxel based models tend to underestimate values of ν_{ic} and ν_{iso} in the white matter regions more prominently than 3D patch based models. *MEDN* and *MEDN+* are characterized by the overestimation of OD parameter, especially noticeable in the corpus callosum.

Similarly as for NODDI parameters, *MLP* designed for single voxel inputs gives comparable results to our models on the problem of *SMT* parameter estimation as depicted in Figure 5.14. Compared with 3D patch based models, *Fourier_S²_SO(3)+* and *Fourier_S²_zonal+* models exhibit lower mean absolute values for λ SMT parameter compared to MLP+, MescNet and MescNet-

SepDict, but higher for ν_{ext} in comparison with MescNetSepDict as given in Figure 5.15. Qualitative comparisons of SMT parameter estimation for models trained on one subject are illustrated in Figures 5.16 and 5.17. The comparison shows that the single voxel models highly overestimate ν_{ext} in certain voxels of white matter in comparison with 3D patch based models. Qualitative comparison of the λ parameter estimation shows that our models *Fourier_S2_SO(3)+* and *Fourier_S2_zonal+* yield lower errors in the frontal brain regions compared to other models.

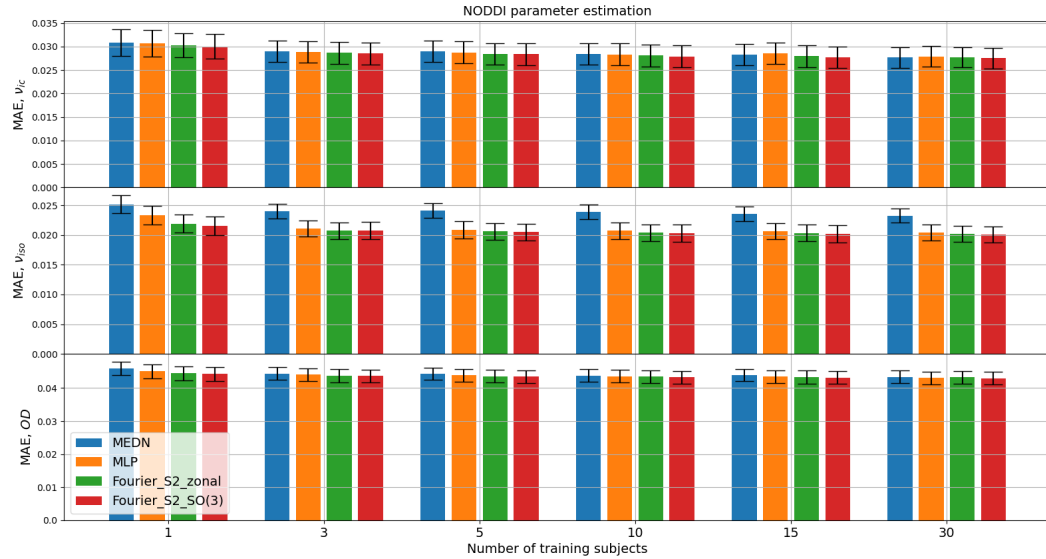


Figure 5.9: Comparison of the mean absolute errors for NODDI ν_{ic} , ν_{iso} and OD parameter estimation for different number of training subjects for single voxel models.

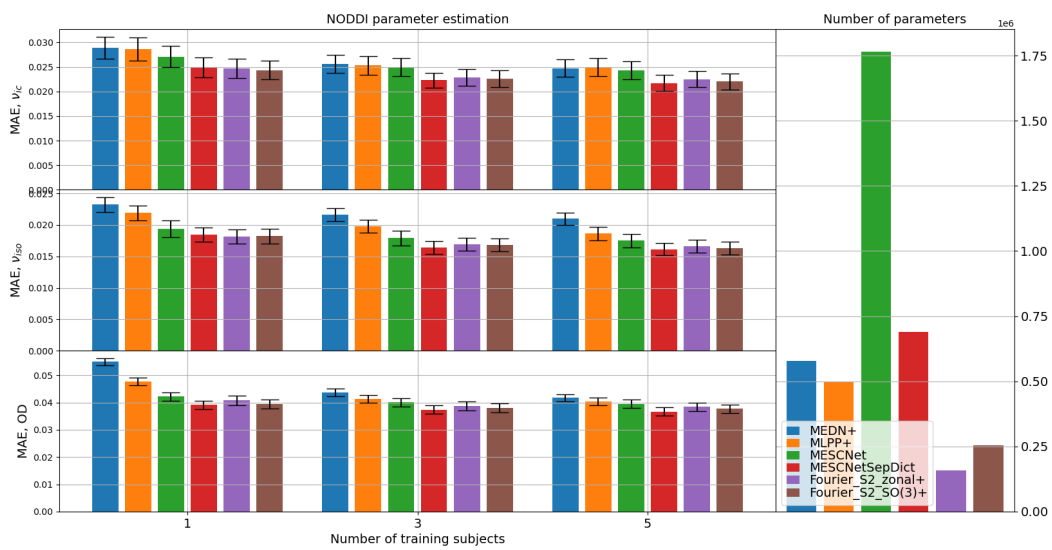


Figure 5.10: Comparison of the mean absolute errors for NODDI ν_{ic} , ν_{iso} parameter estimation for different number of training subjects for 3D patch based models. *MescNetSepDict for 3 subjects: testing performed on 49 subjects, due to memory issues

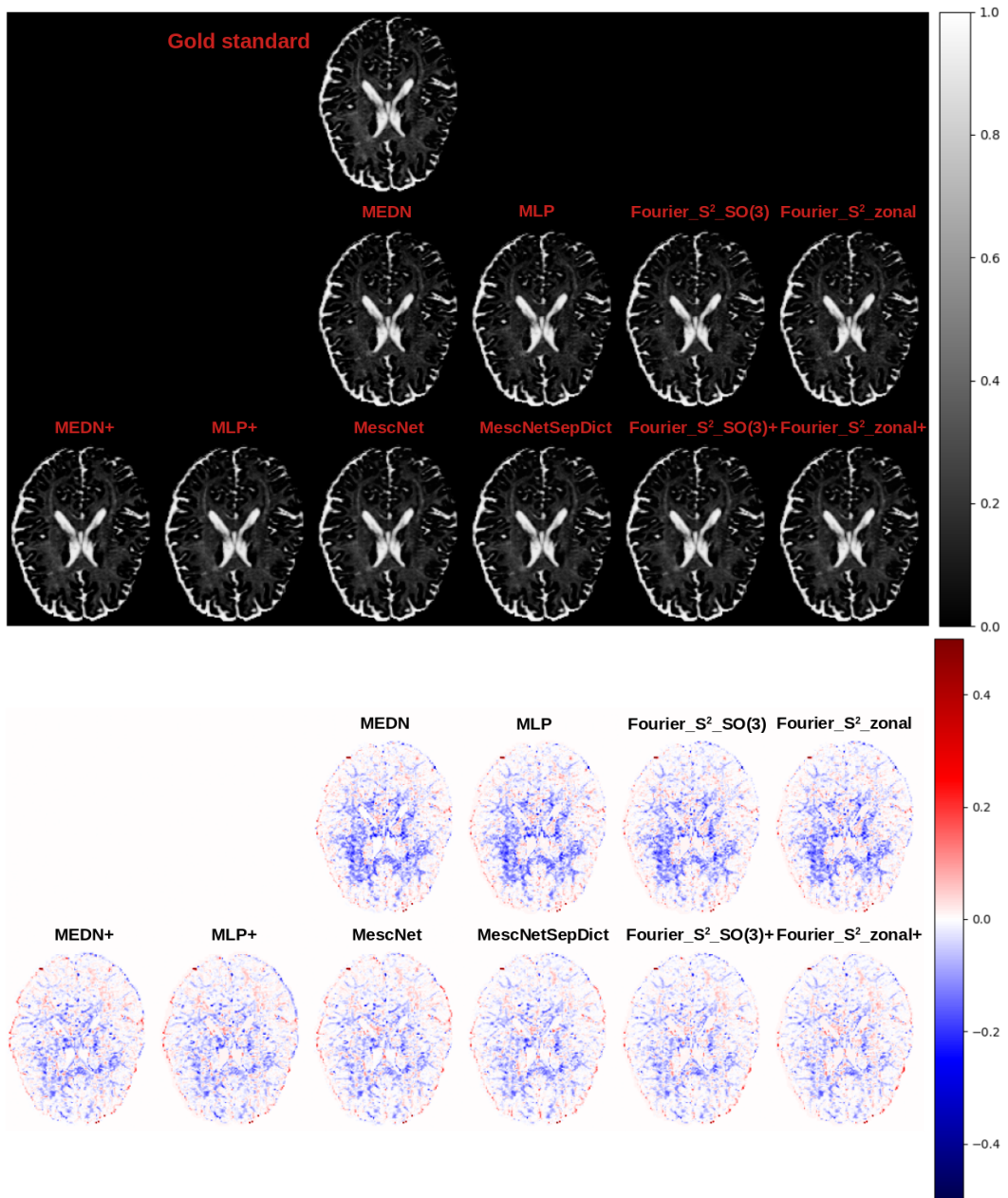


Figure 5.11: Qualitative comparison of NODDI ν_{ic} parameter estimation and difference between the estimated and gold standard values. Training performed on one subject.

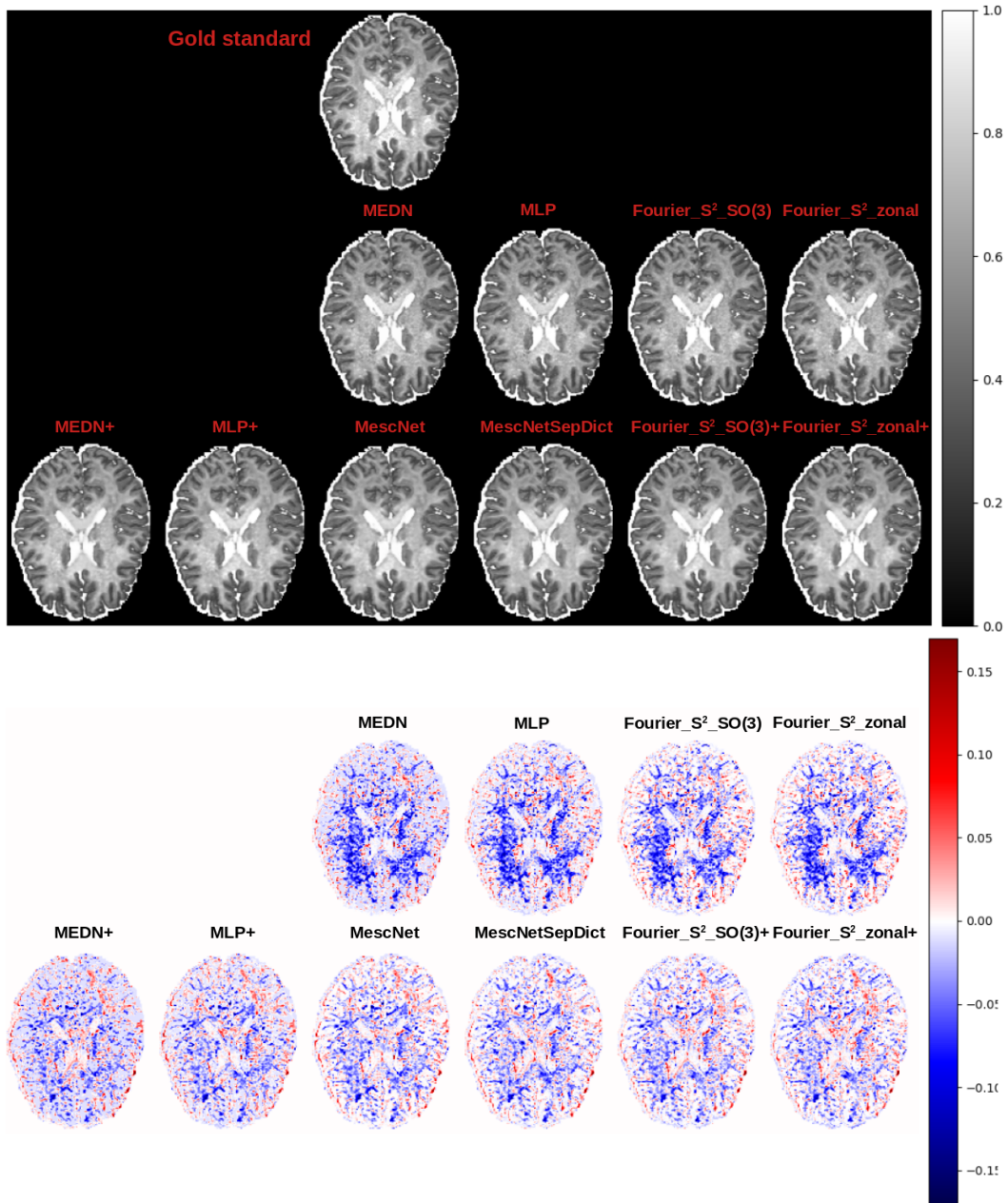


Figure 5.12: Qualitative comparison of NODDI ν_{iso} parameter estimation and difference between the estimated and gold standard values. Training performed on one subject.

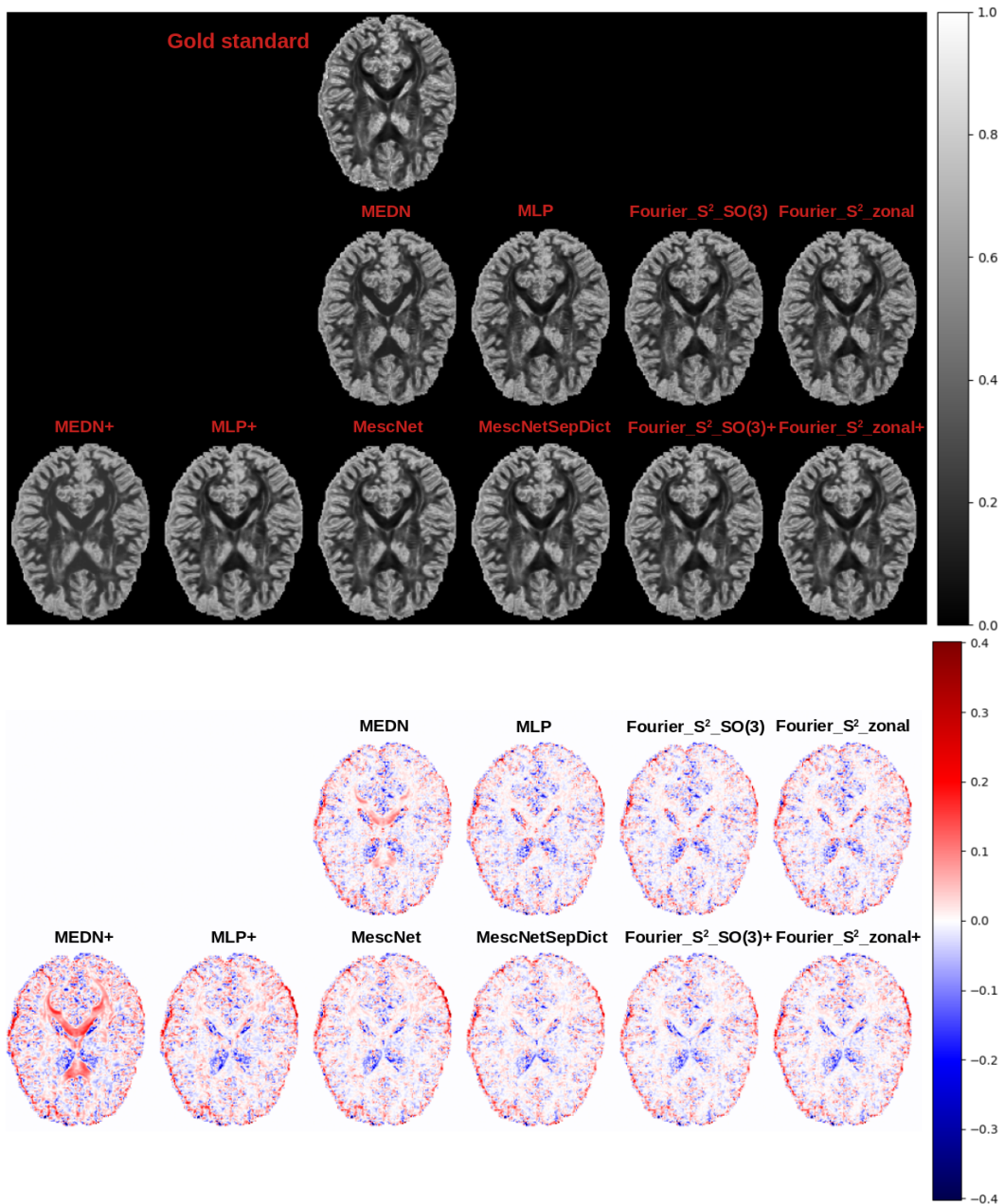


Figure 5.13: Qualitative comparison of NODDI OD parameter estimation and difference between the estimated and gold standard values. Training performed on one subject.

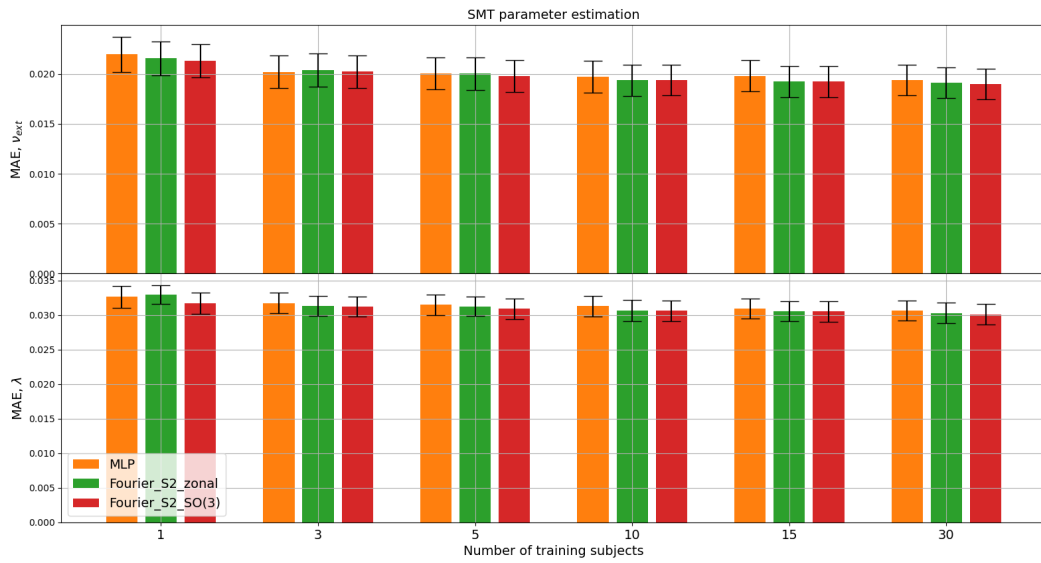


Figure 5.14: Comparison of the mean absolute errors for SMT ν_{ext} and λ parameter estimation for single voxel models. Intrinsic diffusion coefficients λ are normalized to the range of $[0, 1]$.

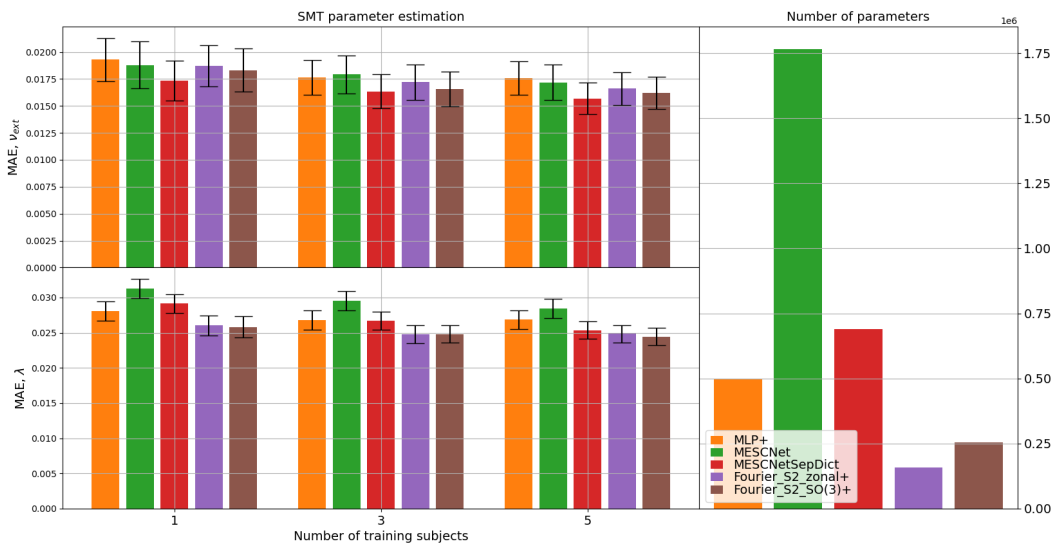


Figure 5.15: Comparison of the mean absolute errors for SMT ν_{ext} and λ parameter estimation for 3D patch based model. Intrinsic diffusion coefficients λ are normalized to the range of $[0, 1]$.

*MescNet for 5 subjects: testing performed on 93 subjects, due to memory issues.

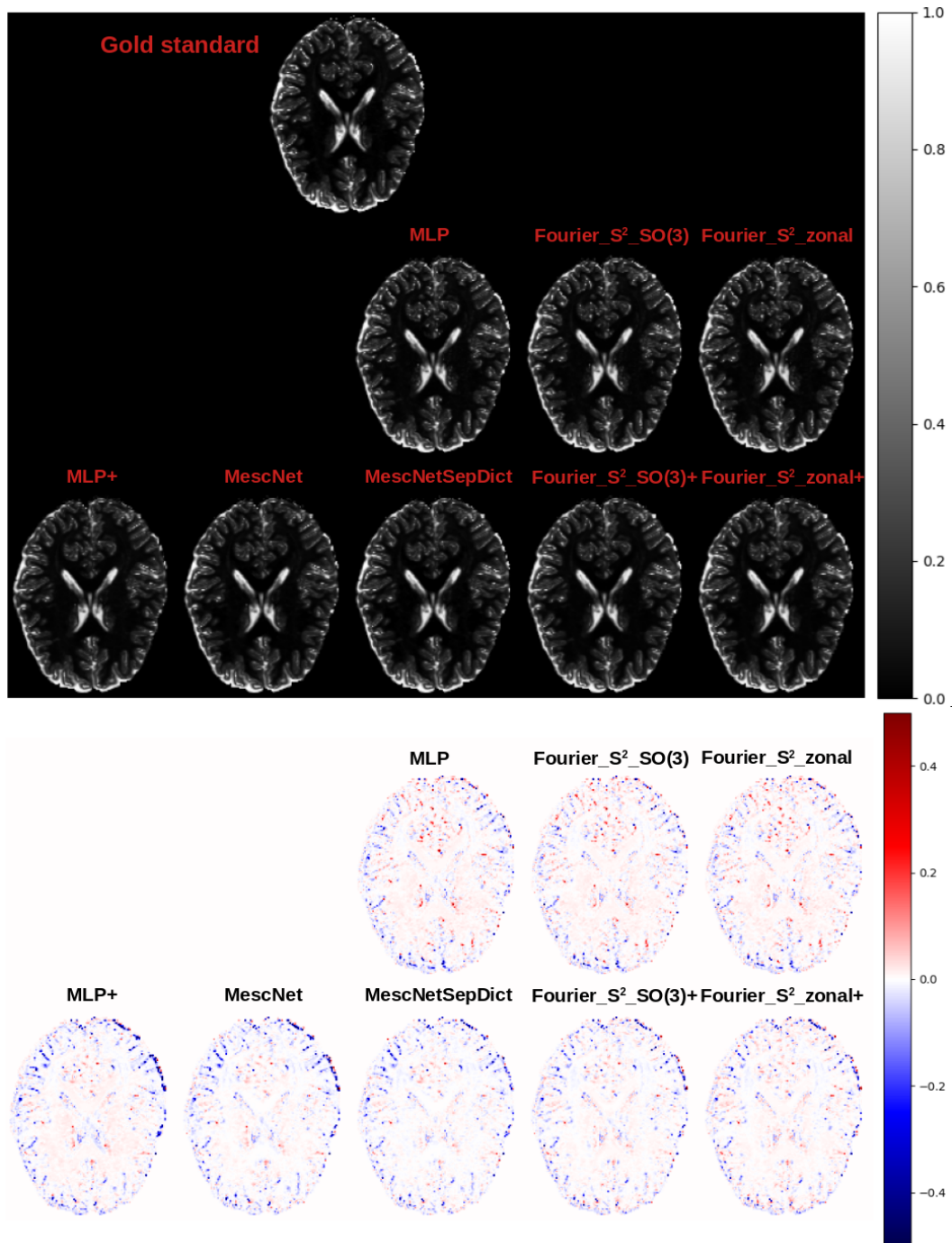


Figure 5.16: Qualitative comparison of SMT ν_{ext} parameter estimation and difference between estimated and gold standard values. Training performed on one subject.

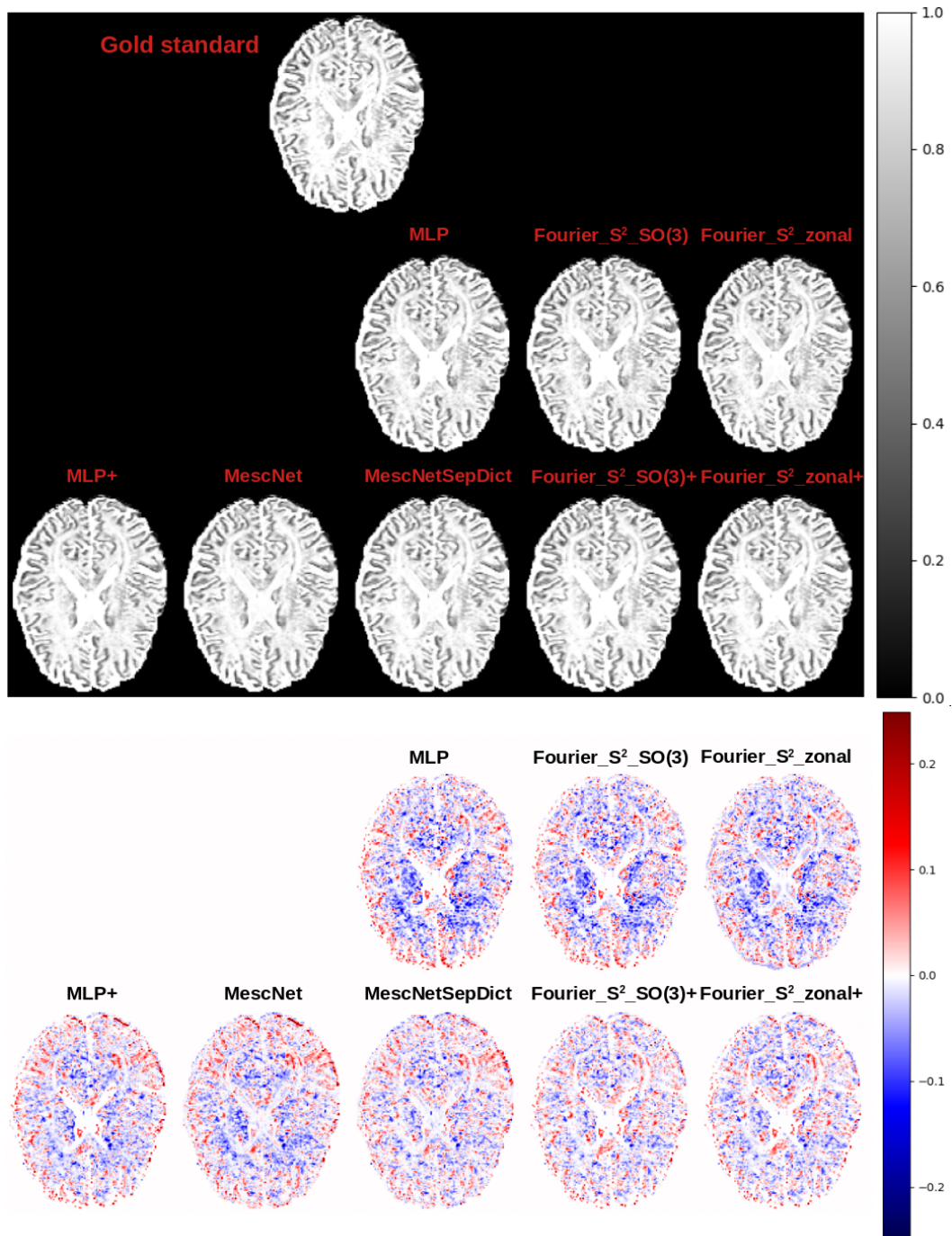


Figure 5.17: Qualitative comparison of SMT intrinsic diffusion coefficients λ normalized to the range $[0, 1]$ and difference between estimated and gold standard values. Training performed on one subject.

5.4.3 Brain tissue segmentation

In this experiment, we demonstrate that our approach can be used in combination with 3D planar CNN for the problem of the brain tissue segmentation.

Real data from HCP and the estimation of gold standard

We have used the same subset of 200 subjects from the HCP database [Van Essen *et al.* 2013] as used in the experiments for microstructure parameter estimation. The preprocessing and normalization of the signals is performed in the same way. A gold standard has been estimated using the FAST algorithm [Zhang *et al.* 2001] applied on T1w images of resolution $1.25 \times 1.25 \times 1.25 \text{ mm}^3$ implemented in the *mrtrix* library [Tournier *et al.* 2019]. It segments tissue into cortical gray matter, sub-cortical gray matter, white matter, CSF and pathological tissue. Since, we have used data from healthy subjects only and since we merged cortical and sub-cortical gray matter classes, only three tissue classes have been considered, namely gray matter, white matter and CSF. We have conducted experiments with the number of training subjects 1, 30 and 70, on full HCP acquisition scheme containing 90 points per each of the three shells and on a reduced sampling scheme containing 60 points per each of the three shells. The number of validation subjects is 20 and the number of testing subjects is 110.

Implementation details

The model is composed of $Fourier_S^2_SO(3)$ which is applied voxel-wise to extract features and 3D planar CNN which takes as input the 3D patches of the extracted features. This enables the integration of 3D spatial information into the segmentation process. For a 3D patch of size $n \times n \times n$, depending on the number of convolutional layers and kernel sizes, the output will be $m \times m \times m$ where $m < n$. Although, n can be chosen such that $m = 1$ (voxel-wise), training a model with $m > 1$ provides regularization of the training process. During the testing phase, extracted features of the entire scan are fed into the CNN model. We have compared $Fourier_S^2_SO(3)$ and MLP [Golkov *et al.* 2016] models for feature extraction followed by a CNN of the same structure. We named these models with $Fourier_S^2_SO(3) + CNN$ and $MLP + CNN$. Both models $Fourier_S^2_SO(3) + CNN$ and $MLP + CNN$ are implemented in *tensorflow* [Abadi *et al.* 2015]. The CNN is composed of three convolutional layers with kernels of size 3. During the training, the spatial sizes of the input 3D patches are $15 \times 15 \times 15$ and of the output $9 \times 9 \times 9$. Given that each voxel contains high dimensional dMRI data acquired over three shells, models' training with 3D patches of size $15 \times 15 \times 15$ might be computationally demanding in terms of GPU RAM since the backpropagation algorithm requires keeping intermediate feature maps and gradients. On the other hand, integrating spatial information of broader context is important, especially for the segmentation of the tissues close to a tissue border.

Since the output patch is of size $9 \times 9 \times 9$, which means that the loss is averaged over 9^3 samples and for an efficient usage of RAM, 3D patch-wise batch size is only 1. To augment training data in a computationally efficient manner, extracted patches of features are axially mirrored, which efficiently increases batch size to 2. In each epoch, 3D patches are randomly extracted from training subjects and validation is performed on 3D patches randomly extracted from validation subjects. Half of the training patches has been selected from the border regions of tissues. The border regions are determined by selecting voxels with tissue class probabilities provided by FAST higher than a threshold 0.9. Models have been trained over 200 epochs by minimizing categorical cross-entropy loss using an Adam optimizer [Kingma & Ba 2014]. Initial learning rate has been set to 0.001. If the difference between validation categorical cross-entropy averaged over two sequential blocks of five epochs is smaller than 10^{-4} , the learning rate is reduced by the factor of 0.95. Once the models are trained, testing is very computationally efficient. It is composed of feature extraction step which is performed voxel-wise with batches of size 128, and a segmentation with 3D CNN which takes as input entire scan of the extracted features and its axially mirrored version. Both *MLP* and *Fourier_S²_SO(3)* extract 64 features. *MLP* is composed of 6 layers of output sizes 128, 128, 128, 256, 128, 64. *Fourier_S²_SO(3)* is composed of three convolutional layers of the input and output bandwidths (8, 6), (6, 4), (4, 2) and the input and output number of channels (3, 2), (2, 4), (4, 8), and three fully connected layers of the output sizes 256, 128, 64. The total number of parameters in *MLP + CNN* is 0.212×10^6 and 0.201×10^6 for 90 and 60 points per shell, respectively. The total number of parameters in *Fourier_S²_SO(3) + CNN* is 0.131×10^6 for both sampling schemes, as the input to the models are the SH coefficients of bandwidth 8. Since the number of sampling points is considerably higher than the number of SH basis elements (45), the model does not contain denoising layer.

Results

The results are compared in terms of Dice scores and are given in Tables 5.2 and 5.3 for 90 and 60 sampling points per shell. According to Dice scores, the difference in performance between the two models is negligible except when the number of training subjects is one. On the other hand, qualitative comparison of the segmentations illustrated in Figure 5.18 highlights some differences. The comparison is provided for the experiments with one training subject and 90 sampling points per shell (1*t*, 90*p*) and 30 training subjects and 60 points per shell (30*t*, 60*p*). First, by comparing slices in axial view, we can notice that *MLP + CNN* misclassifies several voxels of CSF situated in ventricles into white matter voxels. This is especially prominent for the model trained with one subject. Secondly, illustrations in coronal plane show that *Fourier_S²_SO(3) + CNN* gives better segmentation of gray matter in the region of the left lateral fissure. In the sagittal plane, we can notice some differences in the region of cerebellum and below it, where *MLP + CNN* trained on one subject misclassifies CSF as white matter region. Finally, we remark that

the Dice scores obtained with 70 training subjects and 90 points per shell for both models are comparable with the recently proposed deep learning approach which uses three 2D U-nets applied on combination of mean-kurtosis curve, diffusion kurtosis and diffusion tensor parameters [Zhang *et al.* 2021] also trained on 70 HCP subjects. Whereas the model proposed in [Zhang *et al.* 2021] takes $\sim 20min$ for the segmentation of one scan, $Fourier_S^2_SO(3) + CNN$ requires $\sim 1min$ and $MLP + CNN \sim 15s$.

Table 5.2: Dice scores for brain tissue segmentation obtained with $MLP + CNN$ and $Fourier_S^2_SO(3) + CNN$ for 90 points per shell and 1, 30 and 70 subjects.

Model Tissue	Gray matter	Cerebrospinal fluid	White matter
MLP (1)	0.859 \pm 0.017	0.805 \pm 0.023	0.885 \pm 0.018
Ours (1)	0.871 \pm0.015	0.804 \pm 0.022	0.903 \pm0.015
MLP (30)	0.896 \pm 0.010	0.835 \pm 0.019	0.922 \pm 0.010
Ours (30)	0.903 \pm 0.009	0.840 \pm 0.019	0.930 \pm 0.009
MLP (70)	0.900 \pm 0.008	0.836 \pm 0.018	0.927 \pm 0.009
Ours (70)	0.905 \pm 0.008	0.843 \pm 0.018	0.931 \pm 0.009

Table 5.3: Dice scores for brain tissue segmentation obtained with $MLP + CNN$ and $Fourier_S^2_SO(3) + CNN$ for 60 points per shell and 30 and 70 subjects.

Model Tissue	Gray matter	Cerebrospinal fluid	White matter
MLP (30)	0.896 \pm 0.009	0.834 \pm 0.019	0.923 \pm 0.010
Ours (30)	0.904 \pm 0.009	0.838 \pm 0.019	0.930 \pm 0.010
MLP (70)	0.899 \pm 0.008	0.837 \pm 0.019	0.926 \pm 0.009
Ours (70)	0.906 \pm 0.008	0.843 \pm 0.018	0.932 \pm 0.008

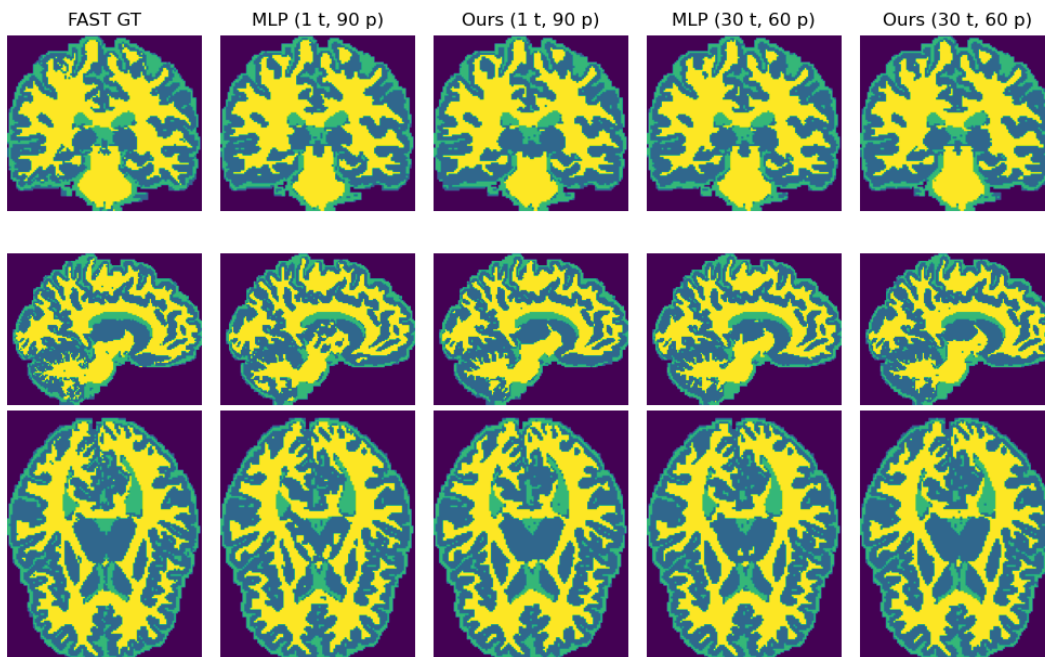


Figure 5.18: Qualitative comparison of brain tissue segmentation into white matter, gray matter and cerebrospinal fluid with $MLP + CNN$ and $Fourier_S^2_SO(3) + CNN$ for one training subject and 90 points per shell ($1t, 90p$) and for 30 training subjects and 60 points per shell ($30t, 60p$).

5.5 Conclusion

In this chapter, we have presented convolutional models adjusted to the spherical and real nature of dMRI signals for its local analysis. Apart from SH coefficient estimation and eventual denoising layers, all other operations in the models are rotation equivariant. We have proposed non-linearities of quadratic nature and degree-wise power spectrum as rotation invariant feature vectors. In the experiment with synthetic data we have demonstrated robustness and rotation invariance of our models with respect to the aliasing and noise. In the extensive comparison with the other deep learning approaches for microstructure parameter estimation, we have shown that the model can achieve state-of-the-art performance with significantly lower number of parameters and with often reduced computational time. Finally, in the last experiments, we have demonstrated that our model can be used to extract voxel-wise rotation equivariant features that can be used for computationally efficient brain tissue segmentation.

Non-linearities of quadratic nature in deep learning are not common due to the fact that they are not bounded. Given a lower computational complexity of convolutions with zonal kernels and of S^2 quadratic non-linearity compared to $SO(3)$ convolutions and non-linearities, in the future work we will investigate how some standard deep learning non-linearities such as sigmoid $\frac{1}{1+e^x}$ and hyperbolic tangent $\frac{e^x - e^{-x}}{e^x + e^{-x}}$ can be approximated via Taylor series in spectral domain.

MEEG spatial and temporal pattern analysis

Contents

6.1	MEEG multivariate signal modeling	104
6.2	MEEG inverse problems	106
6.3	State of the art	108
6.3.1	Dictionary learning	108
6.3.2	CNN classification models	113
6.4	Conclusion	117

Executive summary

In this chapter, we firstly describe modeling of multivariate EEG and MEG signals as sum of rank-1 multivariate signals corresponding to individual brain sources and noise, where temporal courses of the brain activities are modeled as convolution of activation signals and characteristic temporal waveforms. Further, we provide an overview of the most important inverse problems in EEG and MEG signal analysis. Whereas in the section state of the art, we provide a more detailed description of the most prominent multivariate convolutional dictionary learning approaches and CNN models used for EEG and MEG signal classification in BCI is provided.

6.1 MEEG multivariate signal modeling

As the brain is responsible for the functioning of other human's organs, processing of sensory inputs, performing cognitive and motor tasks, controlling emotions, etc, numerous activities are always present in a brain. Each of these activities can be described by cortical regions they arise from and their temporal courses. Magnetic field strength and electric potential, as direct measures of the brain activities, recorded at scalp (or slightly above it) by M/EEG devices can be described with the Maxwell's equations with quasi-static approximations [Sarvas 1987]. As a consequence, we can assume that the cortical brain activities spread instantaneously and linearly over measuring sensors. In order to be measurable by M/EEG devices, neural activity must occur synchronously in a group of pyramidal neural cells in the cortex which counts tens of thousands cells. A common way to model current density present in these groups of cells is via *equivalent electric dipoles* [Hämäläinen *et al.* 1993], often referred to as sources. Since the orientation and position of each source can be considered fixed, the spread of source signal over measuring sensors is fixed as well and can be represented with a vector of weights, also called topographic map. Each weight describes how much a source contributes to the measured signal and depends on the relative orientation of the source with respect to the sensor, their distance and the presence of different amounts of tissues (bones, gray and white matter, cerebrospinal fluid) along the path between the source and the sensor. These weights allow the construction of a so called *leadfield* matrix L and allow the computation of the measured signals as

$$X = LS + \mathcal{N} \quad (6.1)$$

where $L \in \mathbb{R}^{N \times Q}$, with N being the number of sensors and Q the number of sources. Thus, q^{th} column of L describes how q^{th} source signal spreads spatially over sensors. $X \in \mathbb{R}^{N \times T}$ contains measured multivariate signal over T time instants and each row of $S \in \mathbb{R}^{Q \times T}$ represents a source signal over T time instants. \mathcal{N} is an additive noise which includes noise coming from measuring devices, from the environment and from the subject itself. The estimation of a leadfield matrix belongs to the M/EEG forward model problems. A common point in the estimation of MEG and EEG forward models is modeling of the head and brain shapes. However, whereas magnetic permeability can be considered constant over tissues, electric conductivities of different tissue types must be taken into account. The simplest model is the spherical head model, which assumes concentric spheres. Each sphere corresponds to one tissue and has a specific conductivity [Hämäläinen *et al.* 1993, Vatta *et al.* 2010]. More advanced head models require utilization of anatomical and/or structural information usually extracted from MRI data. This allows them to take into account finer head and brain tissue geometries and even to model anisotropic conductivities [Hämäläinen *et al.* 1993, Vatta *et al.* 2010, Ziegler *et al.* 2014].

Assuming K active sources, with $K \leq Q$ and often $K \ll Q$, the measured multi-

variate signal X from Eq. 6.1 can be written as

$$X = \sum_{k=1}^K \mathbf{u}_k \cdot \mathbf{s}_k^T + \mathcal{N} \quad (6.2)$$

where $\mathbf{s}_k \in \mathbb{R}^T$ is the source signal and $\mathbf{u}_k \in \mathbb{R}^N$ its topographic map which corresponds to one column of the leadfield matrix L . Thus, we can notice that a multivariate signal associated to one source k can be represented as a rank-1 matrix $\mathbf{u}_k \cdot \mathbf{s}_k^T$.

Source signals are traditionally classified according to the frequency band they span. They can reveal information related to the organism restoration, cognitive processes and certain brain disorders. *Infra-low waves* (<0.5Hz) or slow cortical potentials are the least investigated ones and are in general considered to be important in the dynamic organization of neural networks at a large scale and the modulation of higher frequency waves [Vanhatalo *et al.* 2004, Fox & Raichle 2007, Grooms *et al.* 2017, Watson 2018]. *Delta waves* (0.5 to 4 Hz) are high energy waves dominant in deep sleep, playing an important role in the stimulation of restoration processes. Delta waves might also be prominent in certain brain disorders such as the attention deficit hyperactivity disorder [Kamida *et al.* 2016] and traumatic brain injuries [Dunkley *et al.* 2015]. *Theta waves* (4 to 8 Hz) are occurring during shallow sleep and meditation. Also several studies have shown increased power in theta range during working memory load and processing [Schacter 1977, Grunwald *et al.* 1999]. *Alpha waves* (8 to 12 Hz) are dominant in the occipital lobe during relaxation with closed eyes when not much information is processed. *Mu waves* occur in the same frequency range as alpha waves but in the sensorimotor cortex and are indicators that the motor system is idling. Once a part of the body is moved or imagined to be moved, the power of these waves decreases which is a phenomenon used in the BCI [Pineda *et al.* 2000, Krusienski *et al.* 2007]. *Beta waves* (12 to 30 Hz) are related to active thinking, problem solving and concentration. Low frequency beta waves are considered to be related to idling and focusing, medium ones to high engagement in mental activity and high frequency beta waves to complex thoughts, high anxiety and excitement. *Gamma waves* (30 to 100 Hz) are related to high level cognitive functioning and are responsible for information processing from different brain regions.

Recent studies have shown that in certain frequency bands, brain waveforms are rather of transient and recurrent nature [van Ede *et al.* 2018]. This is also the case in the active BCI, where the brain waveforms are evoked by external sensor stimuli, with a difference that recurrence is approximately determined based on the repetition of the stimuli. Under the assumption that waveforms of interest are of transient and recurrent nature, Eq. 6.2 can be written as

$$X = \sum_{k=1}^K \mathbf{u}_k \cdot (\mathbf{z}_k * \mathbf{v}_k)^T + \mathcal{N} \quad (6.3)$$

where $\mathbf{v}_k \in \mathbb{R}^\tau$ is a waveform associated to the source k and $\mathbf{z}_k \in \mathbb{R}^{T+\tau-1}$ is a sparse vector with Diracs indicating instants of the activation of the waveform k . τ is duration of the waveforms \mathbf{v}_k .

6.2 MEEG inverse problems

In general, the analysis of EEG and MEG signals can be seen as a joint or an independent analysis of spatial and temporal components of the measured signals, in order to make an inference about the underlying neural activities. Depending on the inference one would like to make, we can distinguish between multiple areas of interest in the domain of EEG and MEG signal analysis, which are not necessarily completely independent of each other. Some of them are inverse problems, source separation, dictionary learning, classification and regression problems, functional brain network analysis, etc.

Inverse problems in functional brain imaging usually refer to the estimation of the distribution, orientation, and intensity of neural activity sources in the cerebral cortex, given the measured signals. Characterization of the sources is important for identification of the cortical regions which are employed while a subject is executing certain functions such as cognitive and motor tasks, or processing of sensory inputs [Bowyer *et al.* 2020], but also in the evaluation of certain neurological disorders [Asadzadeh *et al.* 2020]. Since there is an infinite number of source organizations, including silent ones, and the number of measuring sensors is limited, the inverse MEG and EEG problems are underdetermined. This ill-posedness is addressed via multiple assumptions about the source space. A common assumption is that the relevant sources are situated in the cerebral cortex with orientations perpendicular to the cortex surface [Hämäläinen *et al.* 1993]. Furthermore, assuming a discrete source space, it can be constrained by limiting the number of possible active sources, modeled with equivalent current dipoles [Mosher *et al.* 1992, Mosher & Leahy 1998], while in the case of distributed current sources, minimum norm or smoothness constraints are imposed on the solution [Hämäläinen & Ilmoniemi 1994, Pascual-Marqui *et al.* 1994]. Recent studies have shown that a regularization of the MEG and EEG inverse problems can also be achieved by incorporating information from structural imaging modalities such as dMRI [Belaoucha *et al.* 2015, Kojčić *et al.* 2021].

Source separation refers to disentangling of time courses originating from multiple sources given the measured mixed signals. Mathematically, it is also a class of inverse problems, but with a focus on the temporal aspect of the brain signals, rather than spatial. Source separation is often used as a preprocessing step for artifact removal and denoising [Zou *et al.* 2019, Roy & Shukla 2019], but also for the extraction of event-related responses [Lee *et al.* 2006, Metsomaa *et al.* 2016]. Separating source signals can also facilitate source localization [Zhukov *et al.* 2000]. To address ill-posedness in the source separation problem, assumptions are made

on statistical properties of the source signals. In a broadly used method for source separation - independent component analysis (ICA) the assumption is that the values of each source signal have non-Gaussian distribution and that they are statistically independent [Hyvärinen & Oja 2000]. Under these constraints, the solutions can be estimated by maximizing measures of non-gaussianity such as kurtosis and negentropy, by minimizing mutual information or by the estimation of maximum likelihood [Hyvärinen & Oja 2000].

Dictionary learning is closely related to the source separation and corresponds to the estimation of atoms which constitute a dictionary and allow sparse representation of the measured signals, assuming the presence of recurrent waveforms in the source signals. In addition to being able to separate source signals, dictionary learning frameworks which exhibit translation invariance allow identification of the time instants when the waveforms constituting source signals appear, also referred to as waveform activations. Analysis of such waveforms and their occurrences over time has potential in the evaluation of disorders such as epilepsy and cognitive impairments [Abreu *et al.* 2019], but also used in the extraction of event-related signals [Barthélemy *et al.* 2013, Hamner *et al.* 2011]. In general, dictionary learning is achieved by alternating between update of the dictionary atoms and the update of the corresponding activations [Barthélemy *et al.* 2013, Hitziger *et al.* 2017, La Tour *et al.* 2018]. The difference between objectives of source separation and translation invariant dictionary learning approaches is depicted in Figure 6.1.

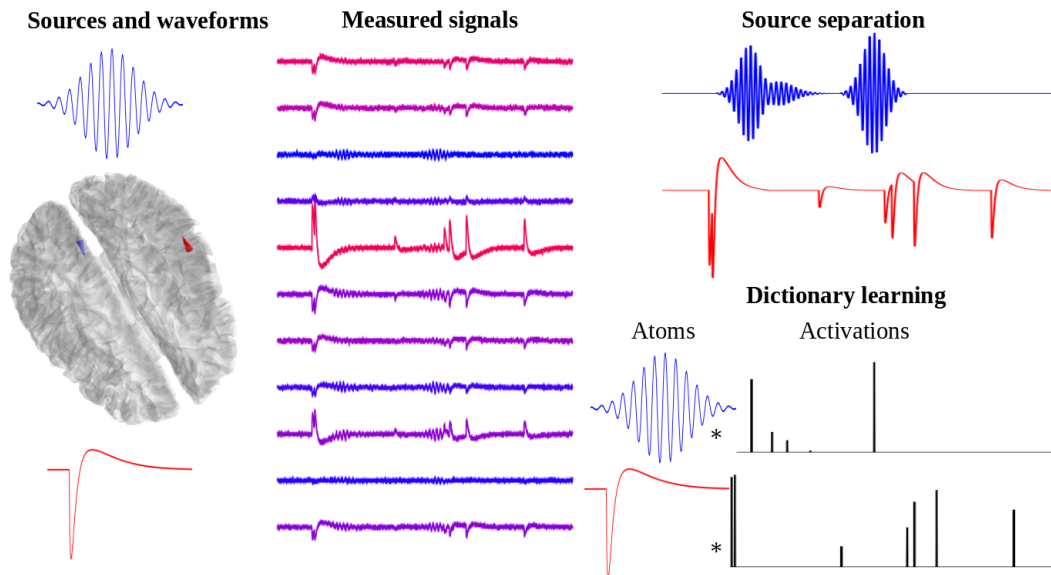


Figure 6.1: Illustrations of objectives of source separation and translation invariant dictionary learning approaches, when two sources with distinct waveforms are active.

Classification and regression models aim to associate a label or a quantity to neural activities given the recorded signals. These models are particularly important in active and passive BCI systems [Lotte *et al.* 2007, Lotte & Roy 2019]. In the context of active BCI, classifiers are necessary in the process of translation of relevant brain activity into a command given to a computer [Allison *et al.* 2007]. More recent, passive BCI systems use classifiers or regression models to assess mental workload, emotional state, drowsiness and alertness of the users [Zander & Kothe 2011, Aricò *et al.* 2018]. Classification and regression problems are addressed by machine learning algorithms trained in a supervised manner. They can be applied directly on raw or preprocessed EEG and MEG signals, but also on extracted features. Recently, a detailed review of the classifiers used in BCI, categorized into adaptive, matrix and tensor, transfer and deep learning, and miscellaneous classifiers has been provided in [Lotte *et al.* 2018]. Although, significantly less studies have addressed the regression problems [Antelis *et al.* 2013, Wu *et al.* 2016a], the majority of the classification models can be simply transformed into regression ones.

Functional brain network analysis aims to understand relationship between activities occurring in different regions of the cortex. Analysis of such networks provides additional insights in highly complex neural activities, while the examined subject is performing cognitive or motor tasks, responding to some sensory stimuli or simply being in a resting state. MEG functional brain networks have been used to identify connectivity markers related to Alzheimer’s and Parkinson’s diseases [Stam 2010] and multiple sclerosis [Nauta *et al.* 2021]. They have also shown an importance in assessment and monitoring of functional reorganization of the brain after a surgery [Wang *et al.* 2010, Pittau & Vulliemoz 2015]. A functional brain network can be represented as a graph composed of nodes which correspond to measuring sensors or their projections to small regions of the cortex. Functional connectivity measures represent the edges between the nodes, which can be undirected such as correlation, phase coherence, mutual information or directed as lagged correlation, transfer entropy, Granger causality [de Vico Fallani *et al.* 2014].

6.3 State of the art

In the context of this thesis, we provide a detailed overview of the dictionary and deep learning approaches which are related or served as an inspiration to our work. Firstly, we provide a description of dictionary learning paradigms, with a focus on the multi-variate dictionary learning. Further, the most prominent deep-learning CNN classifiers used for M/EEG signal classification are presented.

6.3.1 Dictionary learning

Over the last two and half decades, the attention in computer vision community has shifted from Fourier and wavelet analysis towards dictionary learning ap-

proaches. Whereas, a wavelet frame is composed of predefined wavelet functions, dictionary learning aims to estimate a data driven frame, also know as a dictionary. Such dictionaries allow for a sparser representation of data. Thus, they have been initially used for compression and denoising [Kreutz-Delgado *et al.* 2003, Elad & Aharon 2006]. Dictionary learning has been also successfully used in clustering and classification problems, signal reconstruction, etc [Ramirez *et al.* 2010, Sprechmann & Sapiro 2010, Kong & Wang 2012].

In the context of brain wave analysis, the employment of dictionary learning approaches is more recent. This has been motivated by the fact that brain waves of interest are often of a transient and recurrent nature [van Ede *et al.* 2018].

We can distinguish translation-invariant and noninvariant models, and univariate and multivariate models. Given a univariate set of data samples $\{\mathbf{x}_n\}_{n=1}^N$, where N is the number of samples and $\mathbf{x}_n \in \mathbb{R}^T$, with T being the number of sampling points, a *univariate translation-noninvariant* dictionary learning problem can be defined as

$$\operatorname{argmin}_{D, \mathbf{z}_n} \sum_{n=1}^N \|\mathbf{x}_n - D\mathbf{z}_n\|_2^2 \quad \text{s.t.} \quad \mathcal{C}_z(\mathbf{z}_n) \quad \text{and} \quad \mathcal{C}_D(\mathbf{d}_k) \quad (6.4)$$

where $D \in \mathbb{R}^{T \times K}$ is dictionary composed of K atoms $\mathbf{d}_k \in \mathbb{R}^T$ to be estimated, and $\mathbf{z}_n \in \mathbb{R}^K$ is a sparse vector containing coefficients for the sample \mathbf{x}_n [Tošić & Frossard 2011]. \mathcal{C}_z is a constraint which imposes sparsity of the vectors $\{\mathbf{z}_n\}_{n=1}^N$. \mathcal{C}_D is a constraint imposed on the atoms in dictionary. Most commonly, this constraint corresponds to $\|\mathbf{d}_k\|_2 \leq 1$ [Olshausen & Field 1997], alleviating very high amplitudes of the atoms and very low values of the sparse coefficients. Originally, \mathcal{C}_z is defined as $\|\mathbf{z}_n\|_0 \leq \alpha$, however with this penalty, the minimization problem from Eq. 6.4 is not convex and it is NP-hard with respect to \mathbf{z}_n [Tillmann 2014]. Commonly, this minimization problem is addressed by K-singular value decomposition algorithm (K-SVD) [Aharon *et al.* 2006]. Although, this algorithm can end up in a local minima, it has been shown as a sufficiently good solution in practice. In the context of BCI, dictionaries of spatial and temporal EEG patterns have been estimated independently using K-SVD algorithm [Hamner *et al.* 2011]. l_0 penalty is often replaced by l_1 , ensuring convexity of the problem with respect to \mathbf{z}_n , which can be solved by least absolute shrinkage and selection operator (LASSO) method [Tibshirani 1996].

For the analysis of longer brain signals, where waveforms of interest might appear at any time instant, translation-invariant dictionary learning is more suitable. Even if the analysed signals are segmented into epochs, which is a common practice in the analysis of the responses evoked by certain stimuli, the responses might follow the stimuli with different delays. Thus, the models exhibiting translation invariance are better suited for such data. *Univariate translation-invariant* dictionary learning problem can be defined as

$$\operatorname{argmin}_{D, \mathbf{z}_n^k} \sum_{n=1}^N \left\| \mathbf{x}_n - \sum_{k=1}^K \mathbf{z}_n^k * \mathbf{d}_k \right\|_2^2 \quad \text{s.t.} \quad \mathcal{C}_z(\mathbf{z}_n^k) \quad \text{and} \quad \mathcal{C}_D(\mathbf{d}_k) \quad (6.5)$$

where dictionary D is composed of K atoms $\mathbf{d}_k \in \mathbb{R}^\tau$, where $\tau < T$ is length of the atoms [Garcia-Cardona & Wohlberg 2018]. Sparse coefficients $\mathbf{z}_n^k \in \mathbb{R}^{T+\tau-1}$ correspond to the activations of the atom k in the signal \mathbf{x}_n . \mathcal{C}_z is a constraint which imposes sparsity of the activation vectors $\{\{\mathbf{z}_n^k\}_{k=1}^K\}_{n=1}^N$. In the Matching of Time Invariant Features (MoTIF) algorithm, univariate dictionary learning has been achieved independently of the activations and in an iterative manner, where each new atom is estimated under constraint \mathcal{C}_D which imposes that the atom is the most correlated to the data samples, but at the same time the least correlated to the previously estimated atoms [Jost *et al.* 2005]. Once the dictionary is created, sparse coefficients are estimated using Matching Pursuit (MP) algorithm [Mallat & Zhang 1993]. Adaptive Waveform Learning (AWL) is designed for epoched or long EEG recordings, termed with E-AWL and C-AWL, respectively [Hitziger *et al.* 2017]. Dictionary learning is performed by alternating between update of activations and update of dictionary. In addition to translation invariance, AWL can also be dilation invariant. To impose sparsity on the activations, E-AWL model combines l_0 and l_1 regularization terms. The activations are estimated using a modification of the least angle regression shrinkage (LARS) algorithm [Efron *et al.* 2004] termed as LARS-0. This modifications corresponds to an exclusion operator which enforces l_0 sparsity of the l_1 constrained solution. Considering LARS regularization path, at each regularization step, exclusion operator excludes coefficients which corresponds to translation of the atom within a predefined time interval around the epoch center. To reduce computational expenses, in C-AWL, the activations are estimated using the MP algorithm [Mallat & Zhang 1993] with exclusion operator acting within a predefined time interval around any time instant and within an interval of atom dilations. In both versions of AWL, the atoms are constrained to have $\|\mathbf{d}_k\|_2 = 1$ and they are updated via block coordinate descent. Apart from being characterized by waveforms, brain activity can be also described by the brain region from which it arises. Naturally, this has lead to multivariate dictionary learning approaches. Given a multivariate set of data samples $\{X_n\}_{n=1}^N$, where N is the number of samples and $X_n \in \mathbb{R}^{C \times T}$, with C being the number of channels and T the number of sampling points, we can categorized *multivariate translation-invariant* dictionary learning approaches into three groups, illustrated in Figure 6.2:

1. with *multivariate* dictionary and *univariate* activations (Figure 6.2 a))
2. with *univariate* dictionary and *rank-1 multivariate* activations (Figure 6.2 b))
3. with *rank-1 multivariate* dictionary and *univariate* activations (Figure 6.2 c)).

1. Multivariate translation-invariant dictionary learning with multivariate dictionary and univariate activations is defined as

$$\operatorname{argmin}_{D, \mathbf{z}_n^k} \sum_{n=1}^N \left\| X_n - \sum_{k=1}^K \mathbf{z}_n^k * D_k \right\|_2^2 \quad \text{s.t.} \quad \mathcal{C}_z(\mathbf{z}_n^k) \quad \text{and} \quad \mathcal{C}_D(D_k) \quad (6.6)$$

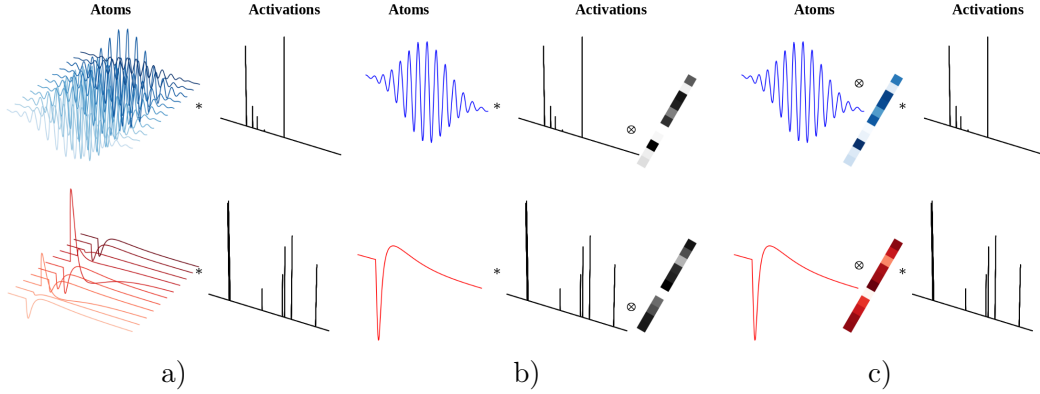


Figure 6.2: Illustration of translation invariant multivariate dictionary learning paradigms: a) with multivariate atoms and univariate activations; b) with univariate atoms and multivariate rank-1 activations; c) with multivariate rank-1 atoms and univariate activations. Each row corresponds to multivariate signal contribution associated to one atom.

where dictionary D is composed of K multivariate atoms $D_k \in \mathbb{R}^{C \times \tau}$, where $\tau < T$ is length of the atoms. Sparse coefficients $\mathbf{z}_n^k \in \mathbb{R}^{T+\tau-1}$ correspond to the activations of the atom k in the signal X_n and convolution between activations and multivariate atom is given by $\text{row}_j[\mathbf{z}_n^k * D_k] = \mathbf{z}_n^k * \text{row}_j[D_k], \forall j \in \{1, \dots, C\}$. In [Barthélemy *et al.* 2012, Barthélemy *et al.* 2013], dictionary learning is achieved by solving Eq. 6.6, where $\mathcal{C}_z(\mathbf{z}_n^k)$ is defined as $|\mathbf{z}_n^k|_0 < P$, with P being maximal number of non-zero entries and $\mathcal{C}_D(D_k)$ is defined as $\|D_k\|_2 = 1$. Proposed multivariate dictionary learning approach is achieved in an *online* manner, by iterating trough the entire dataset and performing estimation of sparse activations and update of atoms for each data sample individually. Approximation of the sparse activation vectors is performed using multivariate orthogonal matching pursuit (M-OMP) developed in [Barthélemy *et al.* 2012] and update of the atoms using stochastic Levenberg–Marquardt second-order gradient descent [Madsen *et al.* 2004].

2. Multivariate translation-invariant dictionary learning with univariate dictionary and rank-1 multivariate activations is defined as

$$\underset{D, \mathbf{z}_n^k, \mathbf{y}_n^k}{\operatorname{argmin}} \sum_{n=1}^N \left\| X_n - \sum_{k=1}^K (\mathbf{y}_n^k \mathbf{z}_n^{kT}) * \mathbf{d}_k \right\|_2^2 \quad \text{s.t.} \quad \mathcal{C}_z(\mathbf{z}_n^k), \mathcal{C}_y(\mathbf{y}_n^k) \quad \text{and} \quad \mathcal{C}_D(\mathbf{d}_k) \quad (6.7)$$

where sparse univariate activations $\mathbf{z}_n^k \in \mathbb{R}^{T+\tau-1}$ correspond to the activations of the atom k and $\mathbf{y}_n^k \in \mathbb{R}^C$ to its spread over channels, for the data sample X_n . Although defined in a slightly different manner, multidimensional jitter-adaptive dictionary learning (JADL), proposed in [Papageorgakis *et al.* 2017], belongs to this group of multivariate translation-invariant methods. The dictionary D composed of the atoms $\{\mathbf{d}_k\}_{k=1}^K$, $\mathbf{d}_k \in \mathbb{R}^T$, is extended to a dictionary D^s by shifting the atoms by small shifts $\delta \in \Delta$, creating a dictionary composed of the atoms $\{\{\mathbf{d}_{k,\delta}\}_{\delta \in \Delta}\}_{k=1}^K$, $\mathbf{d}_{k,\delta} \in \mathbb{R}^T$. With such extension of dictionary, convolution from Eq. 6.7 is replaced

by $\mathbf{a}_n^{k,\delta} \mathbf{d}_{k,\delta}^T$ in [Papageorgakis *et al.* 2017], where $\mathbf{a}_n^{k,\delta} \in \mathbb{R}^C$ performs linear mapping of the atom $\mathbf{d}_{k,\delta}$ to the measuring sensors. To stay in accordance with the notation used in this section, $\mathbf{a}_n^{k,\delta} \mathbf{d}_{k,\delta}^T$ can be written as $\mathbf{y}_n^k \mathbf{z}_n^{k,\delta} \mathbf{d}_{k,\delta}^T$, where $\mathbf{z}_n^{k,\delta} \in \{0, 1\}$. Constraint \mathcal{C}_z is defined as l_0 norm along δ axis as $\|\mathbf{z}_n^k\|_0 \leq 1, \forall k \in \{1, \dots, K\}$ imposing sparse selection of the atom shifts, allowing maximum one shift per atom k . Given a data sample X_n , for each of the k atoms of the original dictionary D , a shift δ_n^k is chosen as the one which gives maximal value of $\|X_n \mathbf{d}_{k,\delta}\|_1$, thus $\mathbf{z}_n^{k,\delta} = 1$ only iff $\delta = \delta_n^k$. Once the shifts are selected, a dictionary D^n containing $\{\mathbf{d}_{k,\delta_n^k}\}_{k=1}^K$ is created. The constraint \mathcal{C}_y is defined as channel-wise l_1 norm along k axis as $\|\mathbf{y}_{n,j}\|_1 \leq \alpha, \forall j \in \{1, \dots, C\}$. For one channel of X_n and given the dictionary D^n and the constraint \mathcal{C}_y , this problem becomes equivalent to the one from Eq. 6.4 when solving with respect to sparse coefficients. In [Papageorgakis *et al.* 2017], it is solved using the LARS algorithm [Efron *et al.* 2004]. Constraint \mathcal{C}_D on the atoms of the dictionary D is $\|\mathbf{d}_k\|_2 = 1$ and they are updated using block coordinate descent, taking account that each dictionary D^n has different atom shifts. Estimation of the activations $\{\{\mathbf{z}_n^{k,\delta}\}_{\delta \in \Delta}\}_{k=1}^K\}_{n=1}^N$, construction of the dictionaries $\{D^n\}_{n=1}^N$ and the estimation of the topographic maps $\{\{\mathbf{y}_n^k\}_{k=1}^K\}_{n=1}^N$, followed by the update of the dictionary D is repeated until convergence.

3. Multivariate translation-invariant dictionary learning with rank-1 multivariate dictionary and univariate activations is defined as

$$\operatorname{argmin}_{U, V, \mathbf{z}_n^k} \sum_{n=1}^N \left\| X_n - \sum_{k=1}^K \mathbf{z}_n^k * (\mathbf{u}_k \mathbf{v}_k^T) \right\|_2^2 \quad \text{s.t.} \quad \mathcal{C}_z(\mathbf{z}_n^k) \quad , \quad \mathcal{C}_V(\mathbf{v}_k) \quad \text{and} \quad \mathcal{C}_U(\mathbf{u}_k) \quad (6.8)$$

where dictionary U and V are composed of K univariate spatial and temporal atoms $\mathbf{u}_k \in \mathbb{R}^C$ and $\mathbf{v}_k \in \mathbb{R}^\tau$, where $\tau < T$ is length of the atoms. Sparse coefficients $\mathbf{z}_n^k \in \mathbb{R}^{T+\tau-1}$ correspond to the activations of the atoms k in the signal X_n and convolution between activations and a rank-1 multivariate atom is given by $\operatorname{row}_j[\mathbf{z}_n^k * (\mathbf{u}_k \mathbf{v}_k^T)] = \operatorname{row}_j[\mathbf{v}_k (\mathbf{z}_n^k * \mathbf{u}_k)^T], \forall j \in \{1, \dots, C\}$. Imposing rank-1 constraint on atoms is motivated by the assumption that the spread of source signals over measuring space is linear and instantaneous, where each possible source has a constant topographic map [Hari & Puce 2017, La Tour *et al.* 2018]. Multivariate convolutional sparse coding (CSC) for dictionary learning with rank-1 constraint imposed on atoms, as given in Eq. 6.8, has been introduced in [La Tour *et al.* 2018]. The constraint \mathcal{C}_z was defined as $\|\mathbf{z}_n^k\|_1 < \alpha$ and $\mathbf{z}_n^k \geq 0$, and constraints \mathcal{C}_V and \mathcal{C}_U as $\|\mathbf{v}_k\|_2 \leq 1$ and $\|\mathbf{u}_k\|_2 \leq 1$. With given constraints, minimization problem from Eq. 6.8 is convex individually with respect to each of the unknowns, $\{\{\mathbf{z}_n^k\}_{k=1}^K\}_{n=1}^N$, $\{\mathbf{v}_k\}_{k=1}^K$ and $\{\mathbf{u}_k\}_{k=1}^K$. The activations are updated using local greedy coordinate descent (LGCD) introduced in [Moreau *et al.* 2018]. Given a data sample X_n , dictionaries U and V , and initialized activations $\{\mathbf{z}_n^k\}_{k=1}^K$, LGCD segments range of coordinates $[1, T-\tau+1]$ into M segments, and updates activation vector corresponding to one pair of atoms k along one coordinate t per segment to its optimal value. The coordinate t and the pair k are selected as ones where the activation value is the furthest from its optimal value. Sequential pass through all segments is repeated

iteratively until convergence. Given $\{X_n\}_{n=1}^N$ and corresponding $\{\{z_n^k\}_{k=1}^K\}_{n=1}^N$, updating dictionaries $\{u_k\}_{k=1}^K$ and $\{v_k\}_{k=1}^K$ can be performed independently using gradient descent. In particular, in [La Tour *et al.* 2018], projected gradient descent with Armijo rule [Nocedal & Wright 2006] has been used, where the Armijo rule governs the amplitude of the updates.

6.3.2 CNN classification models

In the context of M/EEG signal analysis, classification models are essential in the BCI, but they have also been employed in the analysis of epileptic seizures, sleeping disorders, Alzheimer disease, etc. In addition to signal preprocessing, which is common for a majority of M/EEG signal analysis pipelines, process of classification, in general, is composed of multiple steps, namely the feature extraction, their eventual reduction and/or selection and the feature classification.

The feature extraction refers to application of spatial and/or temporal signal processing tools with a goal to extract a pool of possibly relevant features. We can make a distinction between "hand-crafted", connectivity based, and data driven feature extraction. The former group includes power spectral density [Herman *et al.* 2008, Iscan *et al.* 2011], discrete Gabor transform [Kumar *et al.* 2015, Jrad *et al.* 2016], discrete wavelet transform features [Subasi & Gursoy 2010, Bhattacharyya *et al.* 2010], etc. Connectivity based features model strength of connections between brain regions, represented by sensors, via covariance matrices [Barachant *et al.* 2010] or synchrony measures [Wei *et al.* 2007]. Prominent connectivity features are the ones where data is mapped to matrix manifolds such as Hermitian and Grassmann ones which are equipped with Riemannian metrics which are often better suited to BCI than Euclidean space metrics [Barachant *et al.* 2010]. Data driven feature extraction is present in a broad range of unsupervised and supervised paradigms, starting with principal and independent component analysis (PCA and ICA), linear discriminant analysis (LDA), throughout dictionary learning [Zhou *et al.* 2012, Peng *et al.* 2021] and deep learning approaches.

The feature reduction and selection are optional steps in classification process, applied if dimensionality of extracted features is very high. Purpose of this step is to extract the most relevant features and in such a way reduce possibility of classifier overfitting to training samples. Whereas feature reduction transforms feature vector to a space with lower dimensionality, feature selection simply selects a predefined number of features from the given vector. Although used directly for feature extraction, PCA and LDA are linear techniques which have been often used for dimensionality reduction as well [Kołodziej *et al.* 2012, Yu *et al.* 2014].

The feature classification refers to the application of the linear or non-linear classifiers on the extracted features in order to perform the final inference. Among the classifiers applied on the extracted features, broadly used linear ones are LDA and support vector machine (SVM) [Herman *et al.* 2008, Iscan *et al.* 2011, Jrad *et al.* 2016]. Distinct non-linear classifiers are k-nearest

neighbours (k-NN), non-linear Bayesian classifiers, random trees and neural networks [Herman *et al.* 2008, Iscan *et al.* 2011, Bhattacharyya *et al.* 2010, Kumar *et al.* 2015, Jrad *et al.* 2016]. As summarized in the recent review of the BCI models [Lotte *et al.* 2018], we can also identify BCI classifiers which are able to adapt to new data samples termed as adaptive classifiers and ones which allow transfer of their parameters to a domain of another subject or session referred to as transfer learning approaches.

As in another computer vision research fields, over the last two decades, an attention has been drawn to DL approaches in the analysis of M/EEG signals, as well. In general, these models learn to perform the feature extraction, reduction and classification in a joint global training procedure. Given that the brain waveforms of interest can have an arbitrary position over time, CNN, which exhibit translational invariance, have been chosen to address multiple problems.

In analogy to the dictionary learning approaches, we can make a distinction between *univariate* and *multivariate* CNN models. Due to ease of use and portability of the single channel EEG devices, a number of *univariate* CNN models have been investigated in the context of sleep and epilepsy analysis. In [Tsinalis *et al.* 2016] and [Sors *et al.* 2018], classical CNN models have been employed in the studies on single channel EEG sleep scoring. In [Supratak *et al.* 2017], the authors proposed a *DeepSleepNet* model composed of a convolutional module for time-invariant representation learning and a module with bi-directional long short-term-memory (LSTM) units, which is able to learn transitions between the sleep stages. A pyramidal CNN, with a low number of trainable parameters, suitable for lower amount of training data, for classification of single channel EEG signals into normal, ictal and interictal classes has been proposed in [Acharya *et al.* 2018, Ullah *et al.* 2018]. In the context of multivariate CNNs developed for M/EEG signal analysis, we can identify three types of convolutional layers, namely, *standard convolutional layer*, *separable convolutional layers* and *depthwise convolutional layers*. Given a multivariate M/EEG signal $X \in \mathbb{R}^{C \times T}$, with C being the number of channels and T being the number of time samples, they are defined as follows.

A *standard convolutional layer* with weights W , s.t. $W \in \mathbb{R}^{C \times J \times \tau}$, (or $\{W_j\}_{j=1}^J$, s.t. $W_j \in \mathbb{R}^{C \times \tau}$) performs convolution as

$$Y_j = \sum_{c=1}^C X_c * W_{cj} \quad (6.9)$$

where c refers to the c^{th} channel of X and W_j . Y_j is the j^{th} channel of Y . $Y \in \mathbb{R}^{J \times (T-\tau+1)}$, $j \in \{1, \dots, J\}$ and J is the the number of the output channels. An illustration of the convolution in a *standard convolutional layer* is depicted in Figure 6.3.

In a *separable convolutional layer*, the convolution is performed along temporal and spatial dimensions independently. Thus, given the temporal weights $\{\mathbf{u}_j\}_{j=1}^J$, s.t.

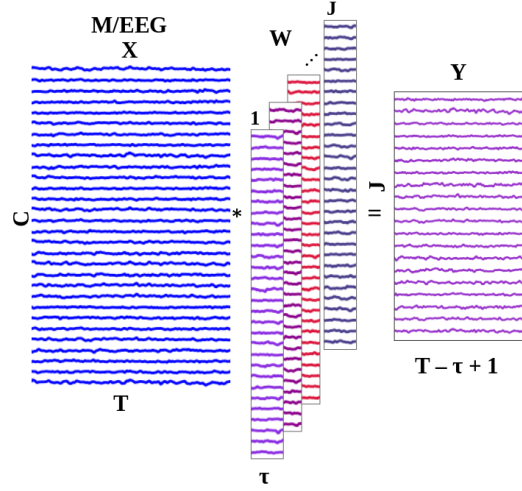


Figure 6.3: Illustration of convolution in a *standard convolutional layer*.

$\mathbf{u}_j \in \mathbb{R}^T$, the temporal convolution is defined as

$$Z_{cj} = X_c * \mathbf{u}_j \quad (6.10)$$

where $Z \in \mathbb{R}^{C \times J_t \times (T - \tau + 1)}$ and $Z_{cj} \in \mathbb{R}^{T - \tau + 1}$. J_t is the number of temporal filters. This is followed by a spatial convolution (correlation more precisely) with $\{\mathbf{v}_j\}_{j=1}^{J_s}$, $\mathbf{v}_j \in \mathbb{R}^{J_t \times C}$ defined as

$$Y_k = \sum_{c=1}^C \sum_{j=1}^{J_t} Z_{cj} \cdot \mathbf{v}_{cjk} \quad (6.11)$$

where $Y \in \mathbb{R}^{J_s \times (T - \tau + 1)}$. An illustration of the convolution in a *separable convolutional layer* is depicted in Figure 6.4.

A *depthwise convolutional layer* is closely related to separable convolutional layer, where after the temporal convolution as given by Eq. 6.10, correlation along spatial dimensions is performed with $\{\mathbf{v}_j\}_{j=1}^{J_t}$, $\mathbf{v}_j \in \mathbb{R}^{C \times D}$, where D is a depth multiplier. Thus the output is obtained as

$$Y_{jd} = \sum_{c=1}^C Z_{cj} \cdot \mathbf{v}_{cjd} \quad (6.12)$$

where $Y \in \mathbb{R}^{J_t \times D \times (T - \tau + 1)}$ and $Y_{jd} \in \mathbb{R}^{(T - \tau + 1)}$. An illustration of the convolution in a *depthwise convolutional layer* is depicted in Figure 6.5 (with $J = J_t$).

The three types of multivariate convolutional layers differ in terms of the number of parameters and the number of multiplication. Assuming that $J = J_t = J_s = C$ and $D = 1$, thus all the layers yield the output of the same size, the number of trainable weights is $C^2 \times \tau$, $C \times \tau + C^3$ and $C \times \tau + C^2$, for the standard, separable and depthwise convolutional layers, respectively. The corresponding number of the multiplications is $C^2 \times \tau \times (T - \tau + 1)$, $C \times \tau \times (T - \tau + 1) + C^2 \times (T - \tau + 1)$ and

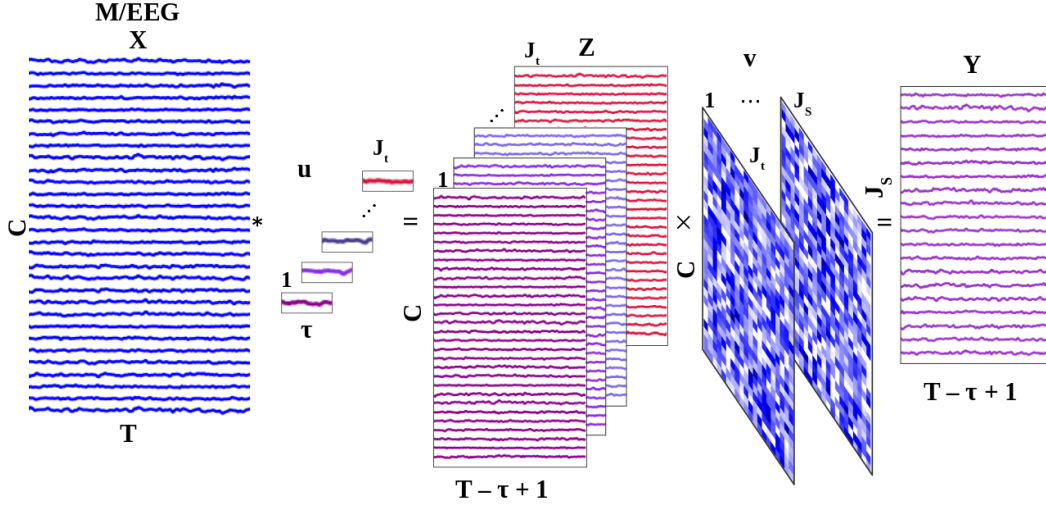


Figure 6.4: Illustration of convolution in a *separable convolutional layer*.

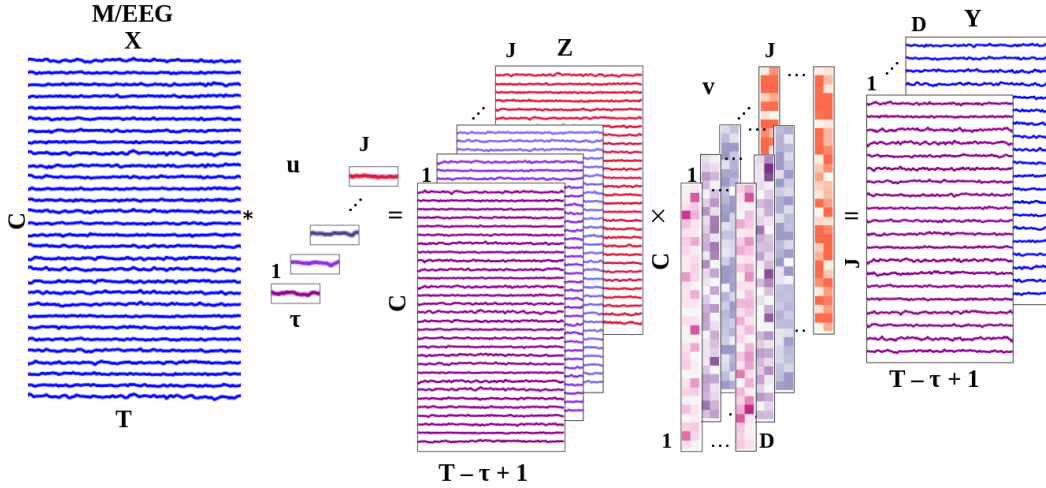


Figure 6.5: Illustration of convolution in a *depthwise convolutional layer*.

$$C \times \tau \times (T - \tau + 1) + C \times (T - \tau + 1).$$

To analyse multi-channel M/EEG data, multiple models with standard, separable and depthwise convolutional layers have been investigated. In [Schirrmester *et al.* 2017], *DeepConvNet* and *ShallowConvNet* have been proposed for classification of motor task and motor-imagery task related EEG signals. In [Lawhern *et al.* 2018], a more compact CNN model termed as *EEGNet* has been proposed for BCI applications.

DeepConvNet [Schirrmester *et al.* 2017] model is composed of four convolutional layers and one fully connected layer. The first layer contains separable convolutions as given in Eqs. 6.10 and 6.11, while the following three contain standard convolu-

tions as in Eq. 6.9. Each convolutional layer is followed by a batch normalization, an Exponential Linear Unit (ELU) non-linearity, a max-pooling and drop-out operations. The features extracted from the last layer are fed into a fully connected network.

ShallowConvNet, a more compact model, was proposed in the same work of [Schirrmester *et al.* 2017]. It contains one separable convolutional layer with longer filters compared to *DeepConvNet* and one fully connected layer. The convolutional layer is followed by a batch normalization, a square non-linearity, average pooling, a logarithmic non-linearity, inspired by filter bank common spatial pattern approach [Ang *et al.* 2008] and a drop-out layer. As in *DeepConvNet*, the extracted features are fed into a fully connected network which performs the final inference.

EEGNet model has been proposed as a compact CNN for EEG BCI applications in [Lawhern *et al.* 2018]. It is composed of two convolutional layers, the former with depthwise convolutions as in Eqs. 6.10 and 6.12 and the latter with separable convolutions as in Eqs. 6.10 and 6.11. In addition to the batch normalization layers applied after each of the convolutional layers, it is also performed after the convolution with the temporal filters in the depthwise convolutional layer. As non-linearity, ELU is used. It is followed by average pooling layer and drop-out layer. The last layer of the model is one fully connected layer.

All the three models, *DeepConvNet*, *ShallowConvNet* and *EEGNet*, apart from the regularization achieved indirectly with batch normalization and drop-out operations, regularize the model weights directly by constraining their maximum norm.

In addition to the three described methods, in the context of passive BCI (classification of cognitive load) a recurrent-CNN has been proposed in [Bashivan *et al.* 2015]. The authors proposed to transform EEG signals into a sequence of topology preserving multi-spectral images, which are used to train the model. The transformation is achieved by projecting spatial component of the signals to 2D images for different power spectrum bands (theta, alpha, beta), where each band is treated as one channel (R, G, B) of a video.

6.4 Conclusion

In this chapter, we have firstly describe modeling of multivariate EEG and MEG signals as sum of rank-1 multivariate signals corresponding to individual brain sources and noise, where temporal courses of the brain activities are modeled as convolution of activation signals and characteristic temporal waveforms. Further, we have provided an overview of the most important inverse problems in EEG and MEG signal analysis. Whereas in the section state of the art, we have provided a more detailed description of the most prominent multivariate convolutional dictionary learning approaches. At the end, an overview of the CNN models used for EEG and MEG signal classification in the BCI is provided.

Rank-1 M/EEG waveform and spatial pattern learning with L_0 constraint

Contents

7.1	Introduction	120
7.2	Method	121
7.2.1	Encoding	122
7.2.2	Decoding	123
7.2.3	Loss and update of the dictionaries	124
7.2.4	Testing	125
7.3	Databases	125
7.4	Implementation details	128
7.5	Results	130
7.6	Conclusion	144

Executive summary

This chapter contains our first contribution in the field of EEG and MEG analysis. We have proposed a model for rank-1 spatial and temporal convolutional dictionary learning with L_0 constraint. Firstly, we introduce constrained least mean square minimization problem we have addressed, followed by a description of multivariate signal encoding and decoding steps, and the process of dictionary update. Since the optimization problem is globally non-convex, we have illustrated the importance of proper initialization of the initial dictionaries. The model is quantitatively compared with rank-1 multivariate convolutional dictionary learning with L_1 constraint on the synthetic data. Qualitative analysis is provided for the real MEG somatosensory data and HCP MEG motor dataset.

7.1 Introduction

Brain activity associated to the cognitive processes, execution of the sensory-motor tasks, and certain neurodegenerative disorders can be often characterized by specific time courses and their location in the cerebral cortex. Thus, the extraction of relevant temporal waveforms and spatial patterns from M/EEG signals is of interest in the active and passive BCI, in the analysis of dynamic brain networks and for a better understanding of the brain disorders. As presented in Chapter 6, assuming that the waveforms are of a transient and recurrent nature [van Ede *et al.* 2018], M/EEG signal $X \in \mathbb{R}^{C \times T}$ measured over C channels and T time instants can be modeled as a sum of rank-1 multivariate signals and additive noise \mathcal{N} [La Tour *et al.* 2018] as:

$$X = \sum_{k=1}^K \mathbf{u}_k \cdot (\mathbf{z}_k * \mathbf{v}_k)^T + \mathcal{N} \quad (7.1)$$

where $\mathbf{v}_k \in \mathbb{R}^\tau$ is a waveform associated to the source k and $\mathbf{z}_k \in \mathbb{R}^{T+\tau-1}$ is a sparse vector with Dirac impulses indicating instants of the activation of the waveform k . $\mathbf{u}_k \in \mathbb{R}^C$ is a topographic map which describes how signals from source k spread over channels. \mathcal{N} is an additive noise which incorporates subject, environment and device related sources of noise. The estimation of $\{\mathbf{v}_k, \mathbf{u}_k, \mathbf{z}_k\}$ from the observed signal X is an ill-posed inverse problem which has been addressed via multivariate convolutional dictionary learning paradigms as described in Chapter 6. In [La Tour *et al.* 2018], the authors proposed a rank-1 spatio-temporal dictionary learning with L_1 sparsity constraint imposed on the activation vectors $\{\mathbf{z}_k\}$. With this regularization term, the estimation of the sparse activation vectors is a convex problem when the atoms in the spatial and temporal dictionaries are fixed. On the other hand, L_0 constraint results in an NP-hard problem with respect to $\{\mathbf{z}_k\}$. In the context of univariate translation noninvariant dictionary learning, sparse vector estimation with L_0 constraint can be solved via Iterative Hard Thresholding (IHT) [Blumensath & Davies 2008] if the dictionary satisfies restricted isometry condition [Candès *et al.* 2006]. The solution can be formulated as follows. Given a univariate signal $\mathbf{x} \in \mathbb{R}^N$ and a dictionary $D \in \mathbb{R}^{N \times K}$, with K being the number of atoms and N being the length of \mathbf{x} , a sparse vector $\mathbf{z}^{i+1} \in \mathbb{R}^K$, in the iteration $i + 1$, is estimated via IHT as

$$\mathbf{z}^{i+1} = H_\lambda(\mathbf{z}^i + D^T(\mathbf{x} - D\mathbf{z}^i)). \quad (7.2)$$

where H_λ is a thresholding operator and $\mathbf{z}^0 = 0$. Although convolution can be written in a form of matrix vector multiplication, by transforming atoms $\{\mathbf{v}_k\}$ into a matrix D , it is clear that one such matrix does not satisfy restricted isometry condition (nearly orthogonal matrix) even only with respect to the thresholding operator. To address this problem, matching pursuit (MP) [Mallat & Zhang 1993, Pati *et al.* 1993] algorithms have been used, which are greedy algorithms, where instead of the thresholding operator, the maximum absolute value (of the inner product) is used to approximate the sparse representations.

In the context of multivariate signal sparse coding, multichannel MP solutions have been proposed in [Gribonval 2003, Durka *et al.* 2005, Barthélemy *et al.* 2012]. Whereas in [Gribonval 2003, Durka *et al.* 2005], multichannel MP solutions have been proposed for the sparse representations given the dictionary of Gabor atoms, in [Barthélemy *et al.* 2012], multichannel MP has been used in dictionary learning. In this chapter, we have investigated a rank-1 convolutional dictionary learning with L_0 constraint. This problem is determined up to waveform shift and rank-1 atom sign. Therefore, as in [La Tour *et al.* 2018], we assume that one source always has activity of the same polarity which is non-negative. As in the standard dictionary learning paradigms, the estimation of the dictionaries and the activation vectors is alternated. To estimate the sparse activations, we have used a greedy approach inspired by the sparse autoencoders, IHT and OMP methods, adjusted to the rank-1 convolutional atoms. It iteratively uses ReLU (thresholding) and maximum operator to estimate the sparse activation vectors. The combination of ReLU and maximum operator for the estimation of the activations has been commonly used in the family of sparse autoencoders [Makhzani & Frey 2013, Makhzani & Frey 2014, Luo *et al.* 2017]. In contrast to the sparse coding in [Gribonval 2003, Durka *et al.* 2005, Barthélemy *et al.* 2012], where the multivariate sparse representations are estimated for univariate temporal atoms (Gabor or learnable) (see Eq. 6.7), our approach estimates univariate sparse representations given the pairs of the spatial and temporal atoms (see Eq. 6.8). Further, whereas in [Gribonval 2003, Durka *et al.* 2005, Barthélemy *et al.* 2012], in each iteration sparse representation for one atom is updated, in our approach in each iteration sparse representations for all pairs of atoms are updated at once. Finally, using ReLU and maximum operator sparse representations are enforced to be non-negative, while in [Gribonval 2003, Durka *et al.* 2005, Barthélemy *et al.* 2012] they can have arbitrary sign. As in [La Tour *et al.* 2018], for fixed activations, the individual update of the spatial and temporal patterns is a convex problem, thus we have used Adam optimizer [Kingma & Ba 2014], which is faster than the traditional stochastic gradient descent.

7.2 Method

We aim to address the multivariate translation-invariant dictionary learning problem from Eq. 6.8, redefined as

$$\begin{aligned} \hat{\mathbf{u}}_k, \hat{\mathbf{v}}_k, \hat{\mathbf{z}}_k &= \underset{\mathbf{u}_k, \mathbf{v}_k, \mathbf{z}_n^k}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^N \left\| X_n - \sum_{k=1}^K \mathbf{z}_n^k * (\mathbf{u}_k \mathbf{v}_k^T) \right\|_2^2 \\ \text{s.t.} \quad & \|\mathbf{z}_n^k\|_0 \leq Q, \mathbf{z}_n^k > \mathbf{0}, \|\mathbf{v}_k\|_2^2 \leq 1 + d, \|\mathbf{u}_k\|_2^2 \leq 1 + d \\ & \text{for } k \in \{1, 2, \dots, K\} \text{ and for } n \in \{1, 2, \dots, N\} \end{aligned}$$

where Q is a parameter that ensures sparsity of the activations $\{\{\mathbf{z}_n^k\}_{k=1}^K\}_{n=1}^N$ and $d \in \mathbb{R}$ is a small constant. Joint estimation of the $\{\mathbf{v}_k, \mathbf{u}_k\}_{k=1}^K$ and $\{\{\mathbf{z}_n^k\}_{k=1}^K\}_{n=1}^N$ is

a non-convex problem, which is in addition NP-hard due to the L_0 norm imposed on the sparse vectors. On the other hand, minimization with respect to $\{\mathbf{v}_k\}_{k=1}^K$ or $\{\mathbf{u}_k\}_{k=1}^K$, while keeping the other two sets of variables fixed is a convex problem. The process of the sparse activation vector encoding and decoding is illustrated in Figure 7.1. Both encoding and decoding steps use the same dictionary atoms $\{\mathbf{v}_k, \mathbf{u}_k\}_{k=1}^K$. Given a sample $X_n \in \mathbb{R}^{C \times T}$, in the encoding process, the sparse codes $\{\mathbf{z}_n^{k,Q}\}_{k=1}^K$ are non-linearly iteratively estimated over Q iterations, while in the decoding process, they are linearly mapped to the signal \hat{X}_n .

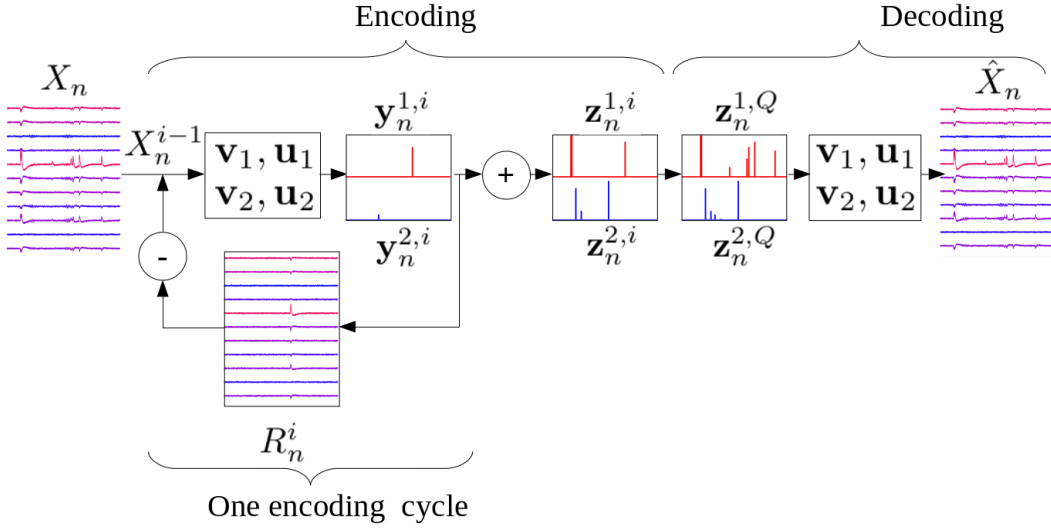


Figure 7.1: Illustration of the encoding and decoding procedures. Estimation of the sparse codes is performed iteratively, where in each encoding cycle at most one activation per source is estimated. After Q encoding cycles, the activations are linearly mapped to a reconstructed signal.

7.2.1 Encoding

In the encoding process, we use correlations along spatial and temporal multivariate signal dimensions to identify the atom activations. Given a multivariate data sample $X_n \in \mathbb{R}^{C \times T}$, correlations with a spatial dictionary of atoms $\{\mathbf{u}_k\}_{k=1}^K$ is given by

$$\mathbf{s}_n^k = X_n^T \mathbf{u}_k \quad \text{for } k \in \{1, \dots, K\} \quad (7.3)$$

where $\mathbf{s}_n^k \in \mathbb{R}^T$. Correlation of $\{\mathbf{s}_n^k\}_{k=1}^K$ with the temporal dictionary of atoms $\{\mathbf{v}_k\}_{k=1}^K$ is given by

$$\mathbf{c}_n^k = \mathbf{s}_n^k * J\mathbf{v}_k \quad \text{for } k \in \{1, \dots, K\} \quad (7.4)$$

where \mathbf{c}_n^k is zero-padded so that $\mathbf{c}_n^k \in \mathbb{R}^T$. $J\mathbf{v}_k$ is reversed version of the atom \mathbf{v}_k .

Iterative estimation of activations. For a sample X_n , the activation vectors $\{\mathbf{z}_n^{k,i} \in \mathbb{R}^{T+\tau-1}\}_{k=1}^K$ in iteration i are estimated as

$$X_n^i = X_n^{i-1} - \sum_{k=1}^K \mathbf{u}_k(\mathbf{y}_n^{k,i} * \mathbf{v}_k)^T = X_n^0 - \sum_{k=1}^K \mathbf{u}_k(\mathbf{z}_n^{k,i} * \mathbf{v}_k)^T \quad (7.5)$$

where $X_n^0 = X_n$ and $\mathbf{z}_n^{k,0} = \mathbf{0}$, and $\mathbf{z}_n^{k,i} = \mathbf{z}_n^{k,i-1} + \mathbf{y}_n^{k,i}$. $\mathbf{y}_n^{k,i}$ is a sparse vector containing at most one activation estimated as follows. Given X_n^{i-1} , we estimate $\mathbf{c}_n^{k,i-1}$ using Equations 7.3 and 7.4. The position of the activation of the k^{th} atom in the i^{th} iteration is performed as $j_n^{k,i} = \text{argmax}(\text{ReLU}(\mathbf{c}_n^{k,i-1}))$ since the activations are constrained to be non-negative. The amplitude of the activation in $\mathbf{y}_n^{k,i}$ is determined as

$$\mathbf{y}_n^{k,i}[j] = \begin{cases} \mathbf{c}_n^{k,i-1}[j] & \text{if } j = j_n^{k,i} \text{ and } \|\mathbf{X}_n^{i-1}[:, \dots, j, \dots] - \mathbf{c}_n^{k,i-1}[j] \mathbf{u}_k \mathbf{v}_k^T\|_2^2 < \|\mathbf{X}_n^{i-1}[:, \dots, j+1, \dots]\|_2^2 \\ 0 & \text{otherwise} \end{cases} \quad (7.6)$$

The vectors $\{\mathbf{y}_n^{k,i}\}_{k=1}^K$ are zero padded so that $\mathbf{y}_n^{k,i} \in \mathbb{R}^{T+\tau-1}$, thus $\mathbf{y}_n^{k,i} * \mathbf{v}_k \in \mathbb{R}^T$. If some multivariate signal $X = \mathbf{u}(\mathbf{z} * \mathbf{v})^T$, where \mathbf{z} contains only one Dirac impulse, peak of its spatio-temporal correlation $\mathbf{c} = \mathbf{u}^T X * J \mathbf{v}$ corresponds to the peak of \mathbf{z} , only if $\|\mathbf{u}\|_2 \|\mathbf{v}\|_2 = 1$. Therefore, since the constraints $\|\mathbf{u}_k\|_2 = 1$ or $\|\mathbf{v}_k\|_2 = 1$ are non-convex, we have constrained the atoms to have norm lower than $1 + d$, where d is a small constant. The step defined in Eq. 7.5 is repeated Q times, ensuring that $\|\mathbf{z}_n^{k,Q}\|_0 \leq Q$. An illustration of one encoding cycle is provided in Figure 7.2.

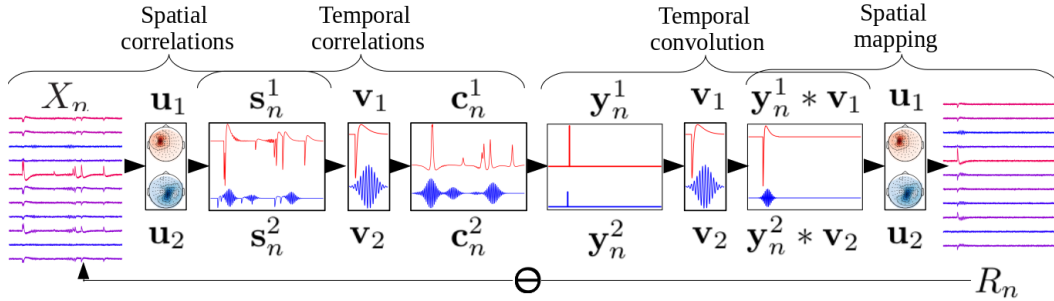


Figure 7.2: Illustration of one encoding cycle with a model containing $K = 2$ pairs of spatial and temporal patterns. For simplicity, superscripts indicating iteration are removed.

7.2.2 Decoding

Once the activations are estimated, they are linearly mapped to the reconstructed signals as

$$\hat{X}_n = \sum_{k=1}^K \mathbf{u}_k(\mathbf{z}_n^k * \mathbf{v}_k)^T. \quad (7.7)$$

Decoding process is illustrated in Figure 7.3.

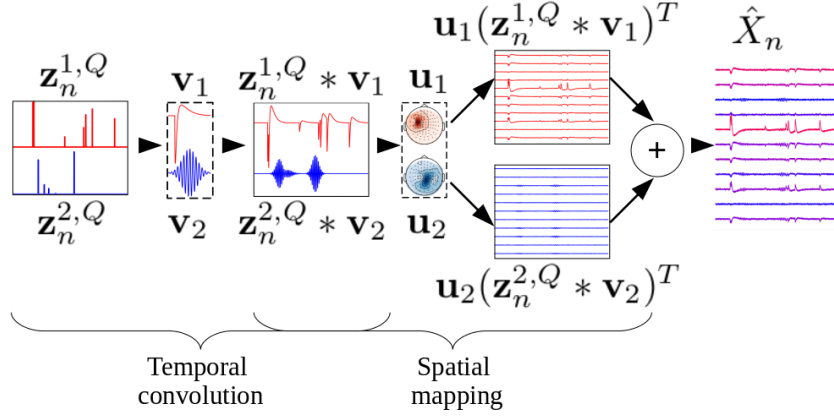


Figure 7.3: Illustration of decoding with $K = 2$ pairs of spatial and temporal patterns.

7.2.3 Loss and update of the dictionaries

If we denote encoding and decoding processes with E and D , respectively, the loss function is defined as the MSE as

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^N \left\| X_n - D(E(X_n | \{\mathbf{u}_k, \mathbf{v}_k\}_{k=1}^K) | \{\mathbf{u}_k, \mathbf{v}_k\}_{k=1}^K) \right\|_2^2 \quad (7.8)$$

or

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^N \left\| X_n - \sum_{k=1}^K \mathbf{u}_k (E(X_n | \{\mathbf{u}_k, \mathbf{v}_k\}_{k=1}^K)_k * \mathbf{v}_k)^T \right\|_2^2. \quad (7.9)$$

Given the estimated activations $\{\{z_n^k\}_{k=1}^K\}_{n=1}^N$, the loss function can be rewritten as

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^N \left\| X_n - \sum_{k=1}^K \mathbf{u}_k (z_n^k * \mathbf{v}_k)^T \right\|_2^2. \quad (7.10)$$

It is the same minimization problem used to estimate dictionaries in [La Tour *et al.* 2018], although not convex jointly, the problem is convex individually with respect to \mathbf{u}_k and \mathbf{v}_k . Gradient of \mathcal{L} with respect to \mathbf{u}_k is

$$\frac{\partial \mathcal{L}}{\partial \mathbf{u}_k} = -2 \frac{1}{N} \sum_{n=1}^N \left(X_n - \sum_{k=1}^K \mathbf{u}_k (z_n^k * \mathbf{v}_k)^T \right)^T (z_n^k * \mathbf{v}_k) \quad (7.11)$$

and gradient of \mathcal{L} with respect to \mathbf{v}_k is

$$\frac{\partial \mathcal{L}}{\partial \mathbf{v}_k[q]} = -2 \frac{1}{N} \sum_{n=1}^N \sum_{j=0}^T \left(\mathbf{u}_k^T \left(X_n - \sum_{k=1}^K \mathbf{u}_k (z_n^k * \mathbf{v}_k)^T \right) [q+j] z_n^k \left[\frac{\tau}{2} + j \right] \right). \quad (7.12)$$

Atoms are updated using the Adam optimizer [Kingma & Ba 2014] where in each training iteration t a weight w is updated as

$$w_{t+1} = w_t - \eta \frac{\nu_t}{\sqrt{s_t + \varepsilon}} g_t \quad (7.13)$$

where

$$\nu_t = \beta_1 \nu_{t-1} + (1 - \beta_1) g_t \quad (7.14)$$

and

$$s_t = \beta_2 s_{t-1} + (1 - \beta_2) g_t^2 \quad (7.15)$$

where η is the learning rate. g_t is a gradient as defined in Eqs. 7.11 and 7.12. ν_t and s_t are gradient's moving mean and moving variance, where β_1 and β_2 are constant determining the contributions of the past and current gradients. ε is small constant ensuring stability of the division. The training is performed by alternating between the update of spatial and temporal atoms.

7.2.4 Testing

During the testing phase, sparse vectors are estimated over P iterations, which does not need to be equal to Q , as it will be discussed in the following section. After each iteration of the sparse vector estimation according to Eqs. 7.5 - 7.6, amplitude of the activations are refined over R steps, where refinements are allowed to be negative. Given a sparse vector $\mathbf{z}_n^{k,i}$ in iteration i and remaining signal X_n^i , in a refinement step r

$$X_n^{i,r} = X_n^{i,r-1} - \sum_{k=1}^K \mathbf{u}_k (\mathbf{y}_n^{k,r} * \mathbf{v}_k)^T \quad (7.16)$$

where $X_n^{i,0} = X_n^i$, $\mathbf{z}_n^{k,i,0} = \mathbf{z}_n^{k,i}$, and $\mathbf{z}_n^{k,i,r} = \mathbf{z}_n^{k,i,r-1} + \mathbf{y}_n^{k,r}$. $\mathbf{y}_n^{k,r}$ is a sparse refinement vector containing at most one activation estimated as follows. Given $X_n^{i,r-1}$, we estimate $\mathbf{c}_n^{k,i,r-1}$ using Equations 7.3 and 7.4. Position of the activation update within sparse vector $\mathbf{z}_n^{k,i,r-1}$ is selected as $j_n^{k,i,r} = \text{argmax}(|\mathbf{c}_n^{k,i,r-1}|)$, such that $\mathbf{z}_n^{k,i,r-1}[j_n^{k,i,r}] \neq 0$. The update is performed as

$$\mathbf{y}_n^{k,r}[j_n^{k,i,r}] = \begin{cases} 0 & \text{if } \mathbf{c}_n^{k,i,r-1}[j_n^{k,i,r}] + \mathbf{z}_n^{k,i,r-1}[j_n^{k,i,r}] \leq 0 \\ \mathbf{c}_n^{k,i,r-1}[j_n^{k,i,r}] & \text{if } \mathbf{c}_n^{k,i,r-1}[j_n^{k,i,r}] + \mathbf{z}_n^{k,i,r-1}[j_n^{k,i,r}] > 0 \end{cases}. \quad (7.17)$$

Allowing the negative refinements during testing phase, is introduced since the amplitudes of the activation vectors obtained via spatio-temporal correlation might contain contributions of the other activations. Whereas, this is the case during the training as well, refinement steps increase training time and the estimation of the activation as in Eqs. 7.5 - 7.6 is sufficient from the point of view of the dictionary updates.

7.3 Databases

We have compared our model with the multivariate convolutional sparse coding algorithm [La Tour *et al.* 2018] on synthetic data and somatosensory MEG data. Furthermore, the model is evaluated on the HCP motor task dataset.

Synthetic dataset

A synthetic MEG dataset is generated using the *MNE* toolbox [Gramfort *et al.* 2013a]. The forward solution is taken from the "sample_audvis-meg-eeg-oct-6-fwd" dataset, which contains 204 MEG gradiometers and 7498 sources. For temporal waveforms, we have used a spike, a sinusoid weighted by a Gaussian window and a saw-tooth signal. The positions of the selected sources, their topographic maps and their corresponding temporal waveforms are illustrated in Figure 7.4. Sparse activation vectors are generated with density 0.01 and a range of amplitudes drawn from a uniform distribution $[0, 1]$. Their duration without zero padding is 5s. Temporal courses are obtained by convolving the zero-padded sparse activations with the temporal waveforms. Their duration is 7s. The sampling rate is 128Hz. The total number of generated samples for training and testing sets is 100. The experiments are conducted on data without noise and data distorted with noise of variance $\sigma = 0.1$. Illustrations of activations and 20 channels corresponding to one generated sample without and with noise are provided in Figure 7.5

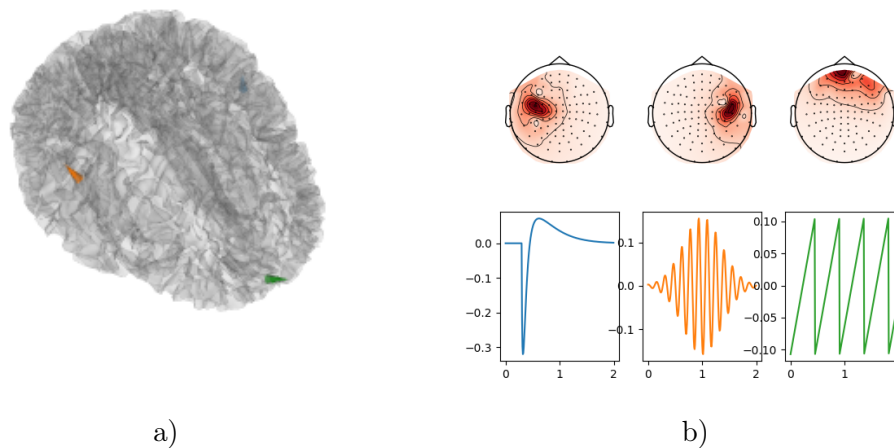


Figure 7.4: Illustration of active sources a) and corresponding waveforms and topographic maps b).

Images generated using: MNE-python [Gramfort *et al.* 2013a]

Motor-task HCP MEG dataset

The motor-task MEG dataset is part of the open HCP dataset [Van Essen *et al.* 2012]. We have selected MEG recordings of five out of 61 subjects acquired over two sessions where participants were guided by visual cues to move either the right hand, left hand, right foot, or left foot. Each session was composed of 42 blocks, where 10 blocks were resting state blocks and 32 blocks were movement blocks (8 block per movement). Each movement block contains 10

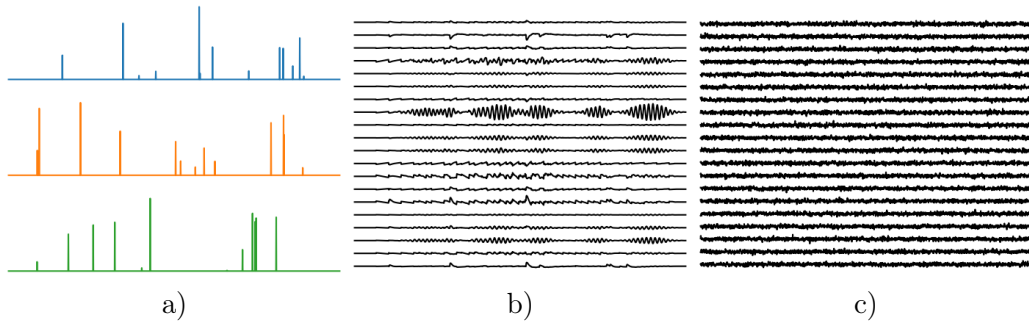


Figure 7.5: Illustrations of activations a) and signals recorded at 20 randomly selected channels without noise b) and with noise $\sigma = 0.1$ c).

movements guided by a visual stimulus. A stimulus lasts $150ms$ and is followed by $1050ms$ of black screen. The number of MEG channels is 248. Sampling frequency is 2034.52 Hz. Signals are segmented into 2.4s long epochs.

To preprocess the raw MEG signals, we have used preprocessing pipeline from the MNE-HCP library [Gramfort *et al.* 2013b]. It included reference correction, filtering with bandpass Butterworth filter of order 4 with cutoff frequencies of 0.5 Hz and 60 Hz, removing of artefacts using ICA and interpolating missing or bad channels. In our experiments, we have subsequently downsampled the signals by a factor 12, given that the signals are low-pass filtered with a cut-off frequency of 60 Hz. Thus, the sampling frequency is ~ 170 Hz. For stability of the model, signals are scaled with the factor $5 \cdot 10^{12}$. The scaling is desirable in order to alleviate the vanishing gradients.

Somatosensory MEG dataset

In the somatosensory MEG dataset, somatosensory EM fields were evoked by electrical stimulation of the median nerve at wrist [Sorrentino *et al.* 2009]. The stimuli were repeated with intervals randomly chosen between $7s$ and $9s$. The MEG signals were acquired with 204 gradiometers and 102 magnetometers with sampling rate of $600Hz$. The dataset was taken from the MNE-python toolbox [Gramfort *et al.* 2013a], and preprocessing was performed as in [La Tour *et al.* 2018], including filtering with two notch filters of $50Hz$ and $100Hz$, downsampling to a sampling frequency of $150Hz$, segmentation into epochs of $6s$ length, epoched signals weighting with a Tukey window, and normalization by their standard deviation. The total number of extracted epochs is 103. In our experiments, we have used only the gradiometer channels as in [La Tour *et al.* 2018].

7.4 Implementation details

Initialization

As the minimization problem from Eq. 7.10 is non-convex, we have investigated how different initializations of spatial and temporal patterns influence convergence of the optimization process. These experiments are conducted on synthetic data without noise and for each initialization type are repeated 50 times. The model includes three pairs of spatial and temporal patterns, whose norm is constrained to $1 + d$, where $d = 0.01$. The maximum number of the activations allowed during training and testing is $Q = P = 40$ and the maximum number of the refinement steps in testing phase is $R = 50$. In the first experiment, we have used random Gaussian $\mathcal{N}(0, n)$ initialization of both the spatial and temporal patterns with different standard deviations $n \in \{1.0, 0.1, 0.01, 0.001\}$. The corresponding learning curves are illustrated in Figure 7.6 a) (left). The MSE between the ground truth and the obtained reconstructions on train and test datasets are illustrated in Figure 7.6 b) (left). In a second and third initialization strategies, temporal waveforms are initialized with a constant normalized to norm 1. In the second strategy, the spatial patterns are initialized with $1 + n\mathcal{U}_c(-1, 1)$, where $\mathcal{U}_c(-1, 1)$ refers to continuous uniform distribution in the range of $[-1, 1]$. In the third strategy, they are initialized with $1 + n\mathcal{U}_d[-1, 1]$, where $\mathcal{U}_d[-1, 1]$ are drawn from discrete uniform distribution $\{-1, 0, 1\}$. In both cases $n \in \{0.1, 0.01, 0.001, 0.0001\}$. After the initialization, as for the temporal patterns, they are normalized to norm 1. Corresponding learning curves for the second and third strategy are illustrated in Figure 7.6 a) (middle, right), while the MSE between ground truth and reconstructions on train and test datasets are illustrated in Figure 7.6 b) (middle, right).

As Figure 7.6 shows, the initialization of the patterns with random values gives very dispersed learning curves with almost no difference between the different standard deviations of the distribution of the initialization values. On the other hand, initialization of the temporal patterns with a constant and spatial patterns with values close to a constant (second and third initialization strategies), yields more coherent learning curves and lower MSEs both on train and test datasets. We can also notice that the losses and MSE decrease with the standard deviation of the uniform distributions.

The impact of Q and P

The maximum number of activations Q determines the number of selected activations with the highest amplitude which contribute to the reconstructed signal during training and thus contribute to the update of the dictionaries. If this number is low, update of the dictionaries will be based on a smaller amount of data segments which correlate the best with the atoms of the dictionaries. Due to the non-convexity of the problem, there is a risk that initial patterns might best correlate to non-representative segments of the signals, leading the minimization process to local minima. On the other hand, if this number is high enough, the

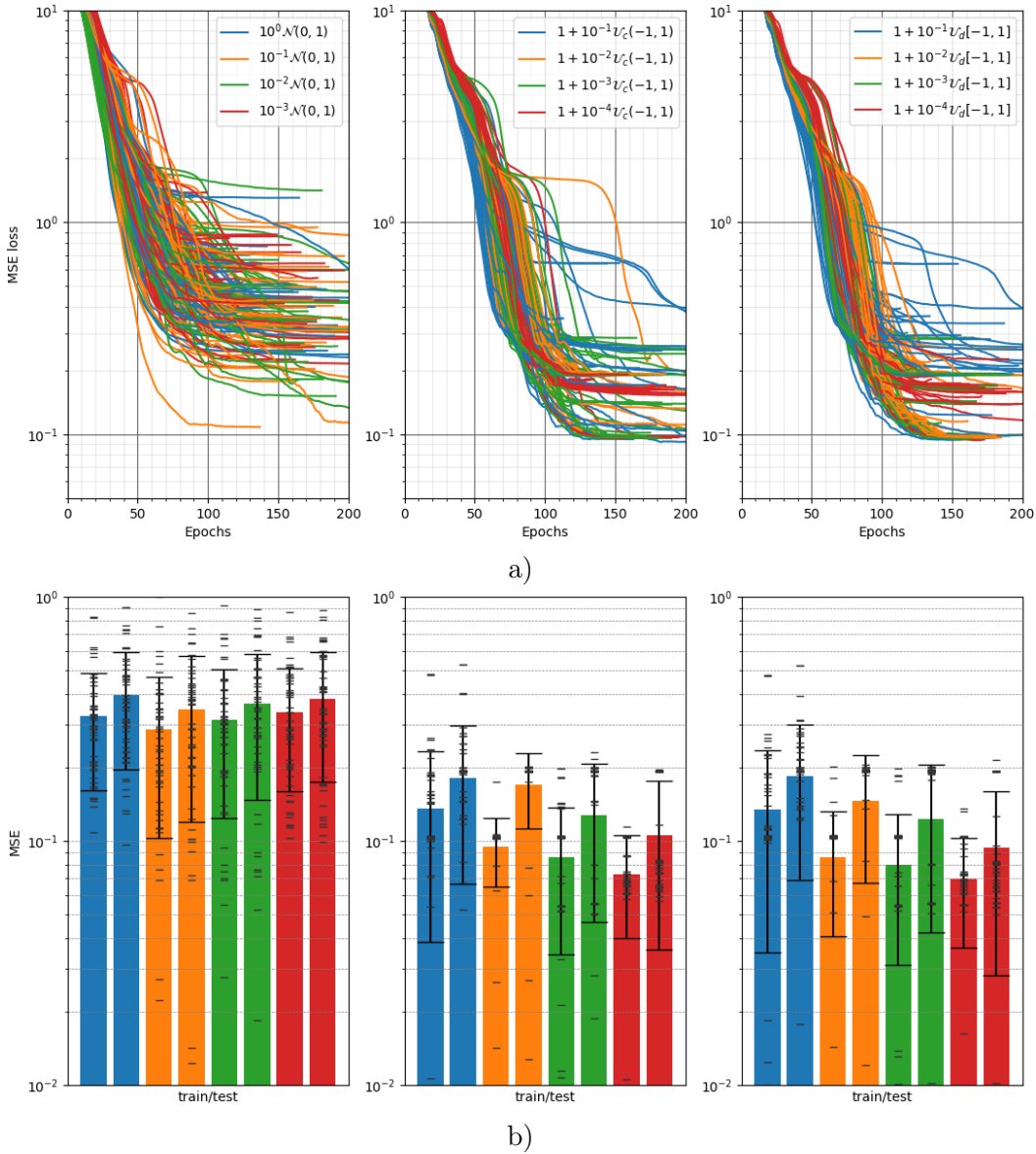


Figure 7.6: Illustration of learning curves a) and MSEs on train and test datasets b) for different initialization strategies for 50 repetitions of the experiments. Random initialization (left), constant initialization of temporal weights and initialization of spatial weights with values drawn from continuous uniform distribution (middle), constant initialization of temporal weights and initialization of spatial weights with values drawn from discrete uniform distribution (right).

update of the atoms is guided with a higher amount of data segments, so among this segments there is a higher chance that some are well-representative and there is more room for a correction of the optimization path. Finally, if the number Q is very high, the algorithm might tend to learn more compact waveforms, especially

when periodic waveforms such as the sawtooth are present in an overall signal. Activation vectors in synthetic data are generated with a density of 0.01, thus a total number of activations per waveform is ~ 12.8 . Firstly, we have investigated how the maximum number of activations during training Q influences the learning process on noiseless data. The learning curves and MSEs estimated on the training and testing data for $Q = 30$ and $Q = 40$ are depicted in Figure 7.7. It shows that the decreasing Q to 30 yields slightly lower MSEs averaged over 50 experiment repetitions, but the MSE standard deviation is higher, when compared to $Q = 40$. When the data is affected by a significant noise, it is of interest to train the dictionaries with the activations of a high amplitude, since those with a lower amplitude might be below or close to the level of noise. The learning curves and MSEs with respect to noiseless ground truth signals, different values of $Q \in \{10, 20\}$ and different values of $P \in \{10, 40\}$ are provided in Figure 7.8. The results indicate that for noisy data, average MSE is lowest for low $Q = P = 10$. In accordance to the results obtained with noiseless data, of $P = 10$, increasing Q from 10 to 20 yields a lower standard deviation of the MSEs.

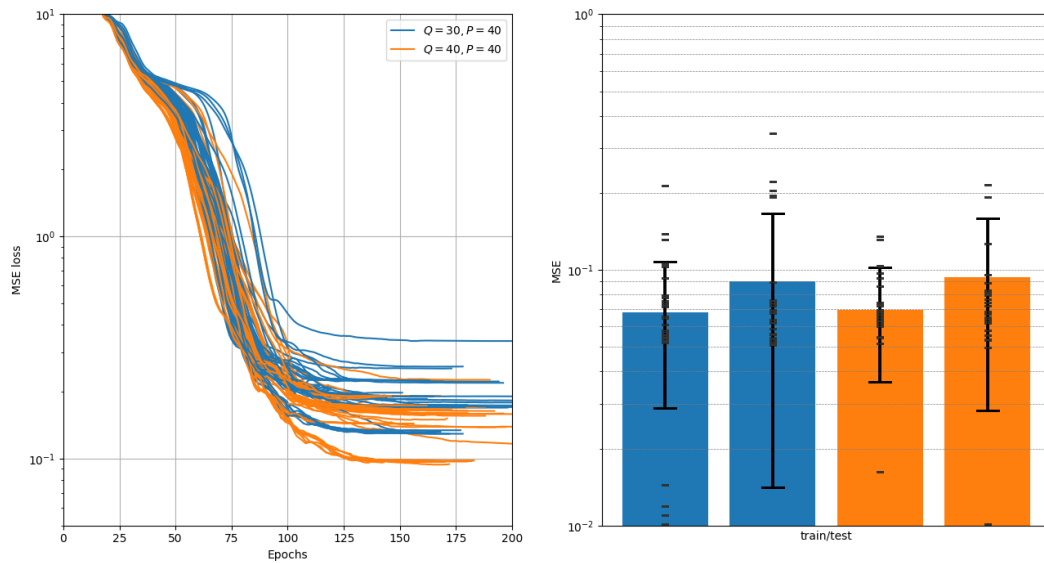


Figure 7.7: Illustration of MSEs on train and test on noiseless datasets for different values of the maximum number of activations during training $Q \in \{30, 40\}$, where test $P = 40$.

7.5 Results

We have compared our method with a rank-1 multivariate dictionary learning method with L_1 constraint [La Tour *et al.* 2018] on the synthetic data and the so-

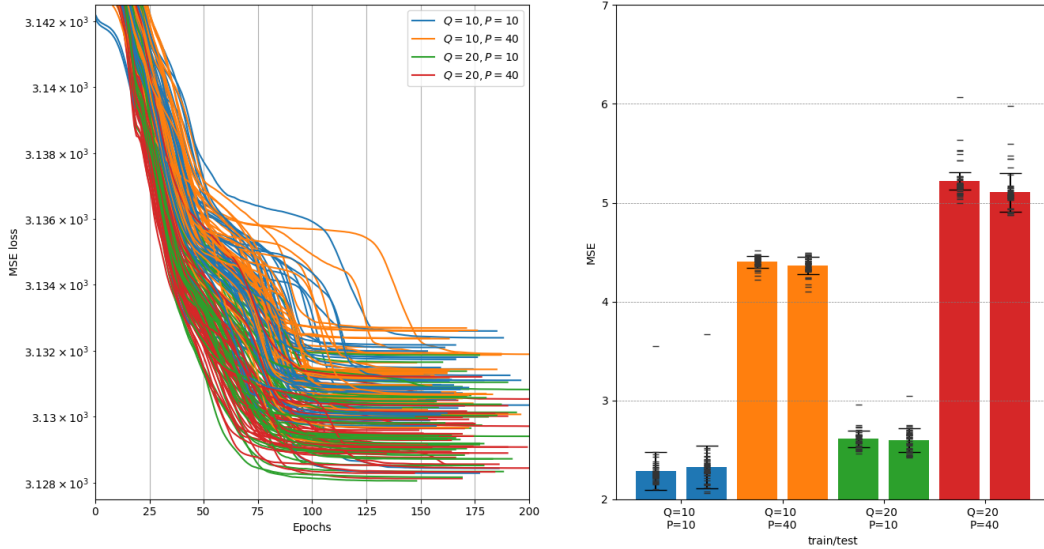


Figure 7.8: Illustration of MSEs on train and test on noisy datasets for different values of the maximum number of activations during training $Q \in \{10, 20\}$, and testing $P \in \{10, 40\}$.

matosensory MEG dataset. Further, we have analyzed the results obtained with our method applied on the motor-task MEG HCP data.

Comparison with the state of the art

Firstly, we have compared the MSE between the ground truth and the reconstructed data on noiseless synthetic data and the MSE between the ground truth and the estimated activation vectors. Since the learned temporal patterns can be shifted compared to the ground truth, the MSE between the activations corresponds to the minimum MSE between the ground truth and corresponding shifted estimated vectors. In this experiment, the maximum number of activations in train and test phase $Q = P = 40$ and the number of refinement steps $R = 50$. The selection of the hyperparameters for MCSC method is given in Appendix C. As illustrated in Figure 7.9, our model yields lower reconstruction errors and has a lower standard deviation. The MSE between activations is lower for MCSC for the waveforms with a narrower support such as spikes and Gaussian weighted sinusoidal waves, but a significantly higher error for the sawtooth waves which have a wide support.

Estimated waveforms are compared in terms of the maximum absolute correlation with ground truth waveforms. In Figure 7.10, we can see that our model estimates Gaussian weighted sines and sawtooths that correlate on average better with the ground truth. This is especially prominent for the sawtooth waveform. MCSC gives better estimates of the spikes on average. In addition, we can see that the standard deviation of the maximum correlation over 50 experiment repetitions is lower with

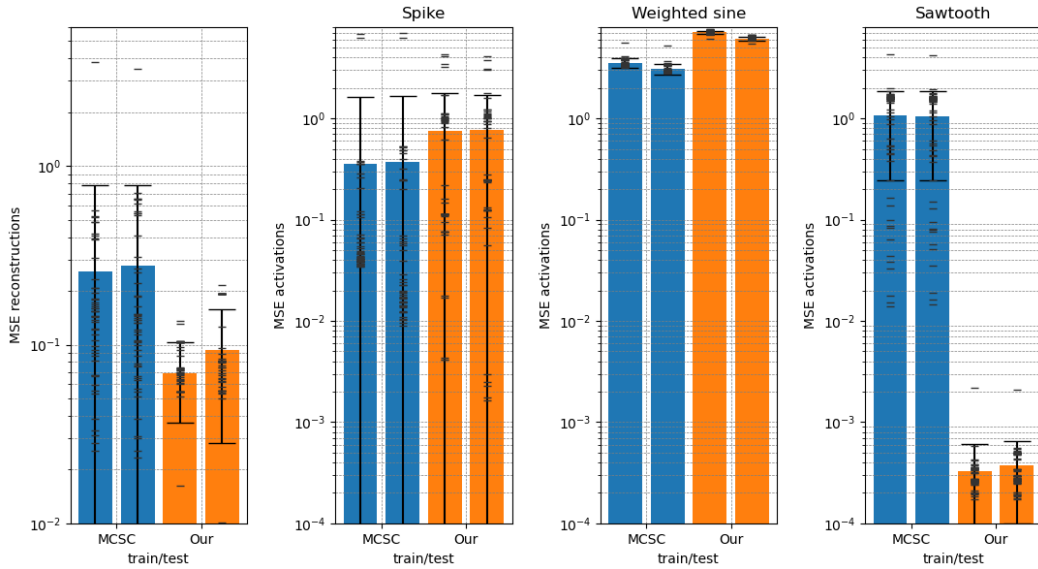


Figure 7.9: Comparison of MSEs between the ground truth and the reconstructed signals and MSEs between the ground truth and the estimated activation vectors on the noiseless data.

our model.

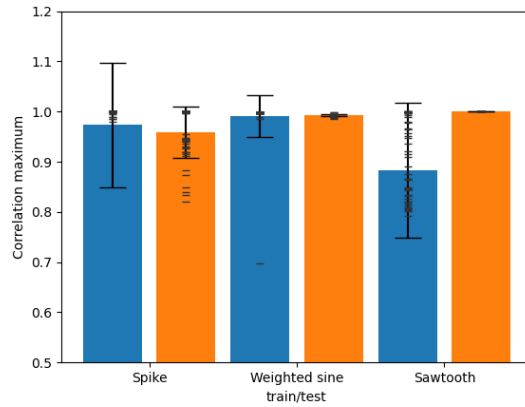


Figure 7.10: Average and standard deviation of the maximum absolute correlation between the ground truth and the estimated waveforms with MCSC (blue) and our method (orange) on noiseless data.

Further, we have visually compared the estimated patterns and the activation vectors for the experiments where the average MSE between the ground truth and the estimated activations is the lowest (Figure 7.11 a)) and the highest (Figure 7.12 a)) and where the reconstruction error is the lowest (Figure 7.11 b)) and the highest (Figure 7.12 b)). As we can notice in Figure 7.11, both methods are able to estimate spatial and temporal patterns which highly resemble to ground truth up to the sign

and shift. The estimated activations for spikes and sawtooth signals also exhibit a high resemblance with ground truth, while the activations for Gaussian weighted sinusoidal waveforms considerably differ (which is in accordance with the results illustrated in Figure 7.9). For the Gaussian weighted sinusoidal waveforms, in the segments with close activations, our model tends to estimate more dense activations with lower amplitudes.

As we can notice in Figure 7.12 a), where the worst results, in terms of activations, are illustrated, our model has difficulty in the estimation of spike pattern and MCSC with the estimation of sawtooth. To compensate these errors, both methods yield denser activation vectors for the corresponding patterns. These errors in the temporal pattern estimation are the ones which appear most commonly over repetitions of the experiments. The comparison of the worst results, in terms of the reconstruction error 7.12 b), show that MCSC failed to separate Gaussian weighted sinusoidal and sawtooth patterns. Also, the results obtained with our method, indicate that a high reconstruction error comes due to the difficulty in the estimation of the activation vectors for the Gaussian weighted sinusoidal.

Models are also compared on synthetic data distorted with Gaussian noise of a standard deviation of 0.1. The selection of the hyperparameters on such data is quite challenging as it requires some prior knowledge. As provided in Appendix C, selecting parameters which minimize MSE between the input noisy signals and the reconstructions may lead to very noisy estimated patterns. Although it is not a real world scenario, to investigate the potential of the models, in this experiment, the hyperparameters are chosen based on the MSE between the noiseless ground truth and the reconstructions estimated on noisy data. The selection of the hyperparameters for MCSC are given in Appendix C. Our model is selected based on the results illustrated in Figure 7.8, thus the maximum number of activations during training and testing phases $Q = P = 10$, while the number of refinement steps $R = 50$. As in the previous experiments, we have firstly compared models in terms of MSEs between the noiseless ground truth and the obtained reconstructions and MSE between the ground truth and the estimated activations. The average MSE and standard deviations are illustrated in Figure 7.13. As it can be seen, the average reconstruction error obtained with our model is slightly lower. On the other hand, the average MSE between the activations for all temporal patterns is significantly lower with MCSC. Contrary to that, the maximum correlations with ground truth patterns are on average higher with our model with a considerable lower standard deviation over the experiment repetitions as depicted in Figure 7.14. The visual comparison of the best and the worse results, according to the mean MSE between the activations and between the reconstruction error, are provided in Figures 7.15 and 7.16. We can notice in both scenarios that the spike patterns are better centered with MCSC, while our model gives smoother temporal patterns which resemble more to the ground truth. Despite the fact that the average MSEs between the activations are much higher with our model, we can notice in the Figure 7.15 and Figure 7.16 that they quite resemble to the ground truth activations, while MCSC yields more spurious low amplitude activations.

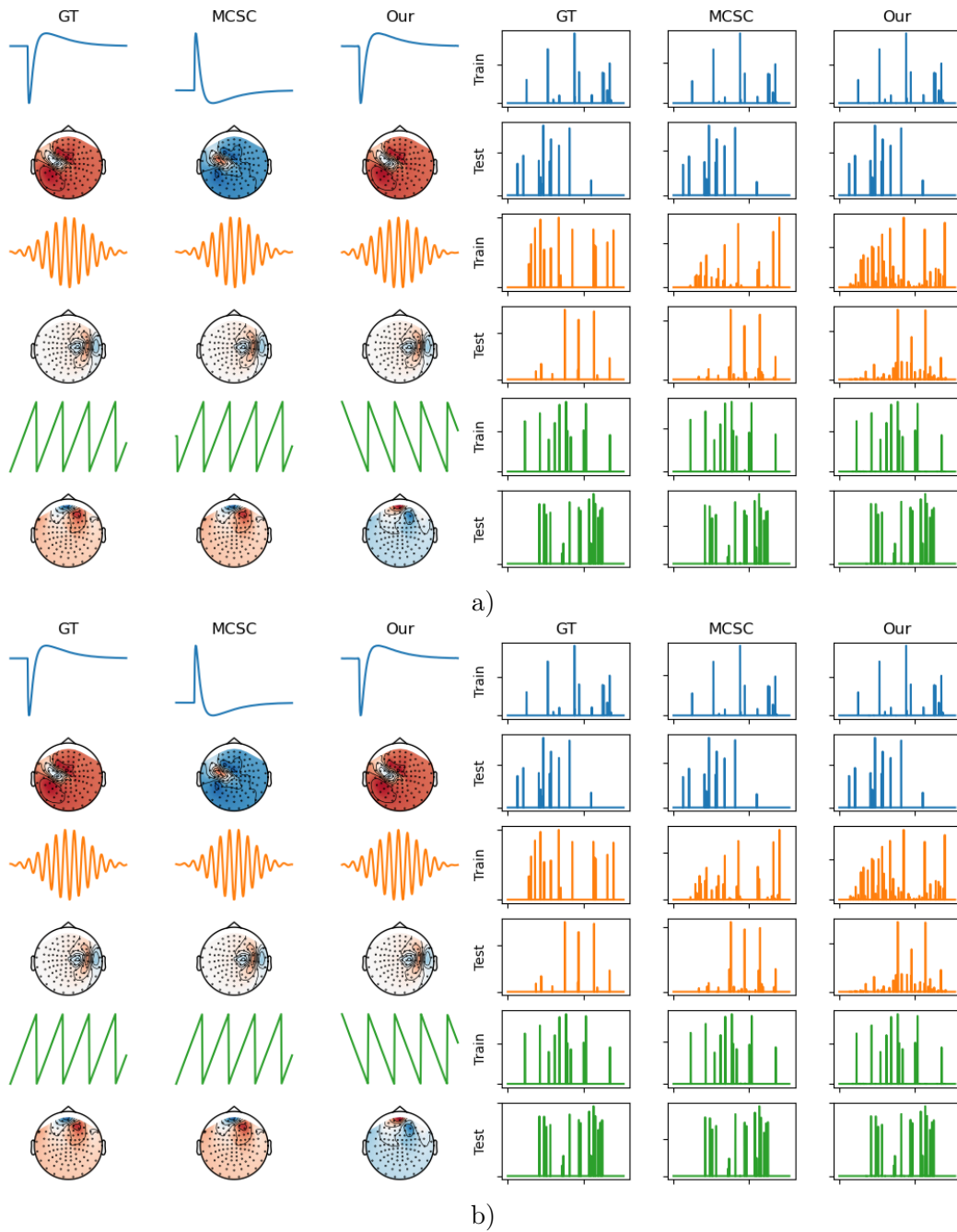


Figure 7.11: Visual comparison of the estimated and the ground truth patterns and the training and testing activation vectors on the experiments where the mean MSE between the ground truth and the estimated activations is the *lowest* a) and where the reconstruction error is the *lowest* b)

Finally, the methods are compared on the somatosensory MEG dataset. As in the experiment presented in [La Tour *et al.* 2018], we have trained a model with 25

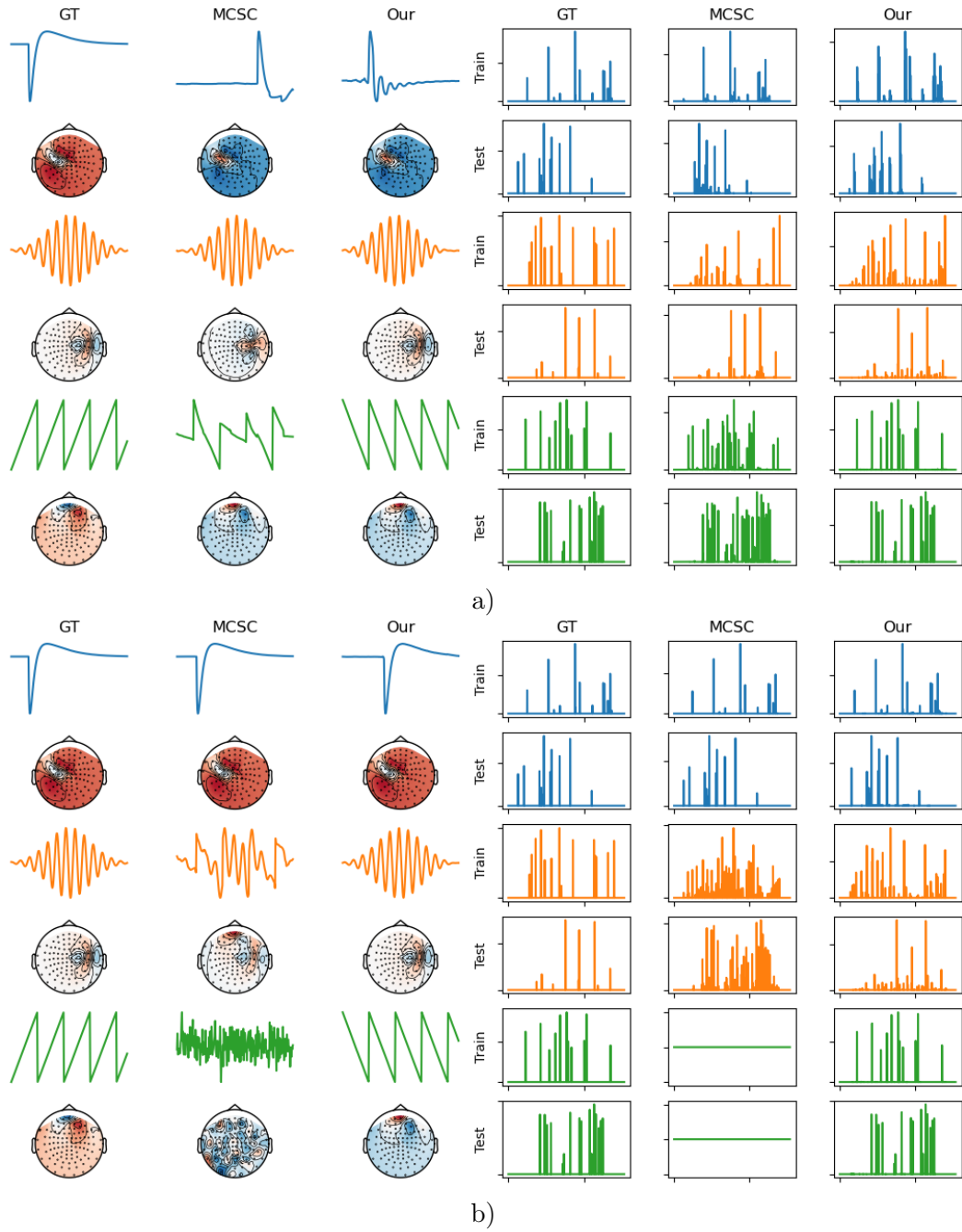


Figure 7.12: Visual comparison of the estimated and the ground truth patterns and the training and testing activation vectors on the experiments where the mean MSE between the ground truth and the estimated activations is the *highest* a) and where the reconstruction error is the *highest* b).

pairs of temporal and spatial patterns. Due to a very large number of atoms, the maximum number of activations per atom pair during train and test $Q = P = 1$

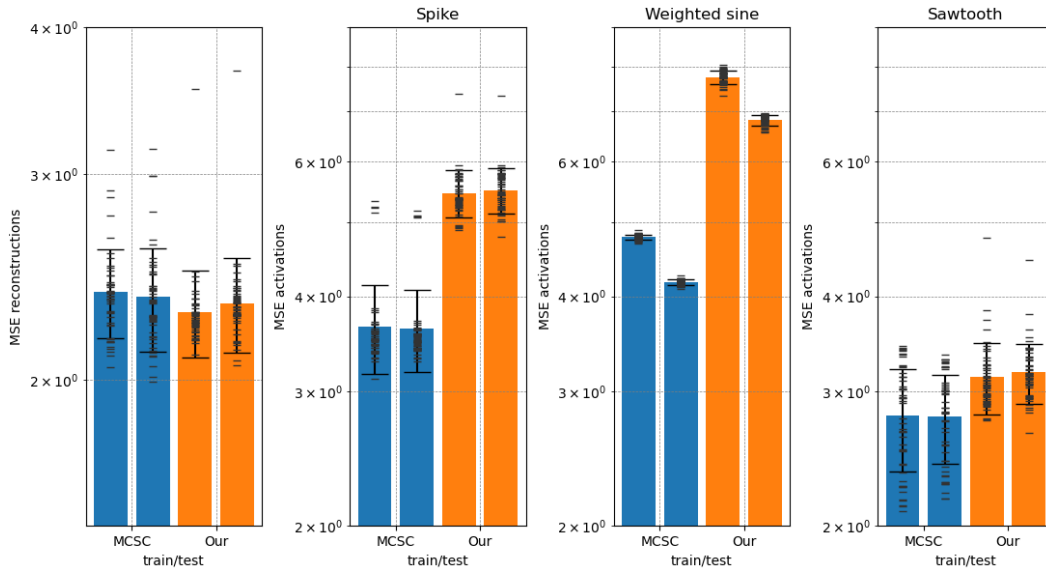


Figure 7.13: Comparison of MSE between ground truth and reconstructed signals and MSE between ground truth and estimated activation vectors on data distorted by Gaussian noise of standard deviation 0.1.

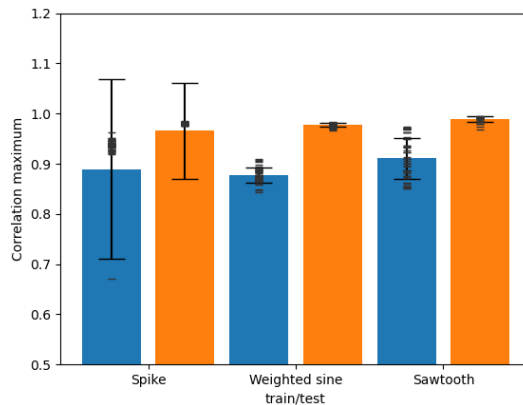


Figure 7.14: Average and standard deviation of maximum absolute correlation between ground truth and estimated waveforms with MCSC (blue) and our (orange) methods on data distorted by Gaussian noise of standard deviation 0.1.

and the maximum number of refinement steps $R = 50$. The length of temporal waveforms in both models is 1s. The average explained variance over epochs is 15.65% and 18.15% for MCSC and our method, respectively. Illustrations of the estimated atoms and activations are given in Figure 7.17. They show that the extracted temporal and spatial patterns between the methods in a great extent visually resemble. A great number of temporal atoms correspond to a special type of α waves, so called μ waves which occur in the sensorimotor cortex and are an

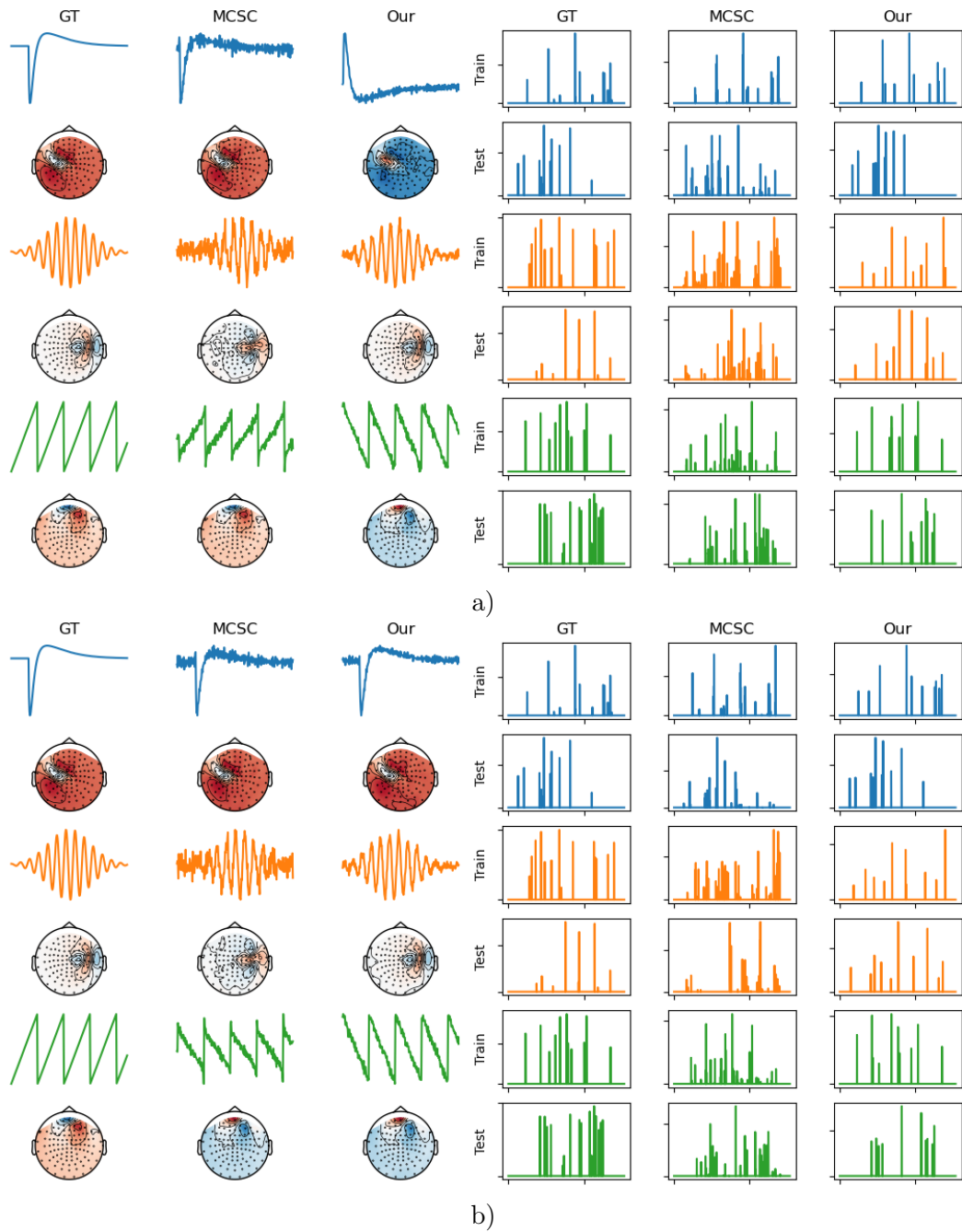


Figure 7.15: Visual comparison of the estimated and the ground truth patterns and the training and testing activation vectors on the experiments where the mean **MSE** between the ground truth and the estimated activations is the *lowest* a) and where the reconstruction error is the *lowest* b).

indicator that the motor system is idling. As expected, the peak of their power spectral density is around $10 - 12Hz$. The highest intensity of the associated spatial

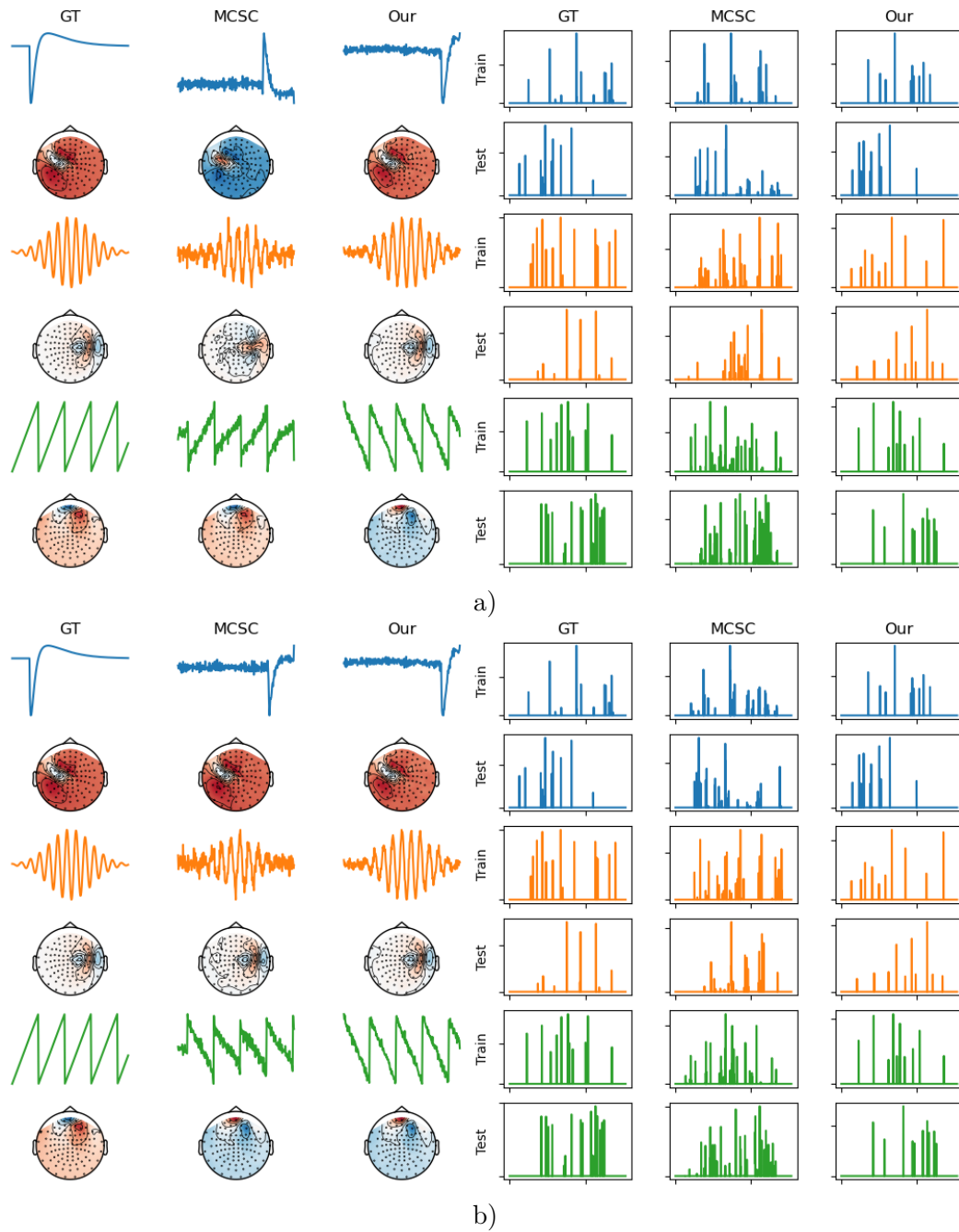


Figure 7.16: Visual comparison of the estimated and the ground truth patterns and the training and testing activation vectors on the experiments where the mean MSE between the ground truth and the estimated activations is the *highest* a) and where the reconstruction error is the *highest* b).

patterns corresponds, to a certain extent, to the location of the sensorimotor cortex. We can also notice few patterns resembling to spikes extracted with our method,

whose power spectral density has peaks in a range below $10Hz$. Their associated spatial patterns have peaks in the prefrontal cortex. In Figure 7.18, we illustrate the distributions of correlations between the estimated rank-1 atoms, where we can see that our model provides less correlated atoms.

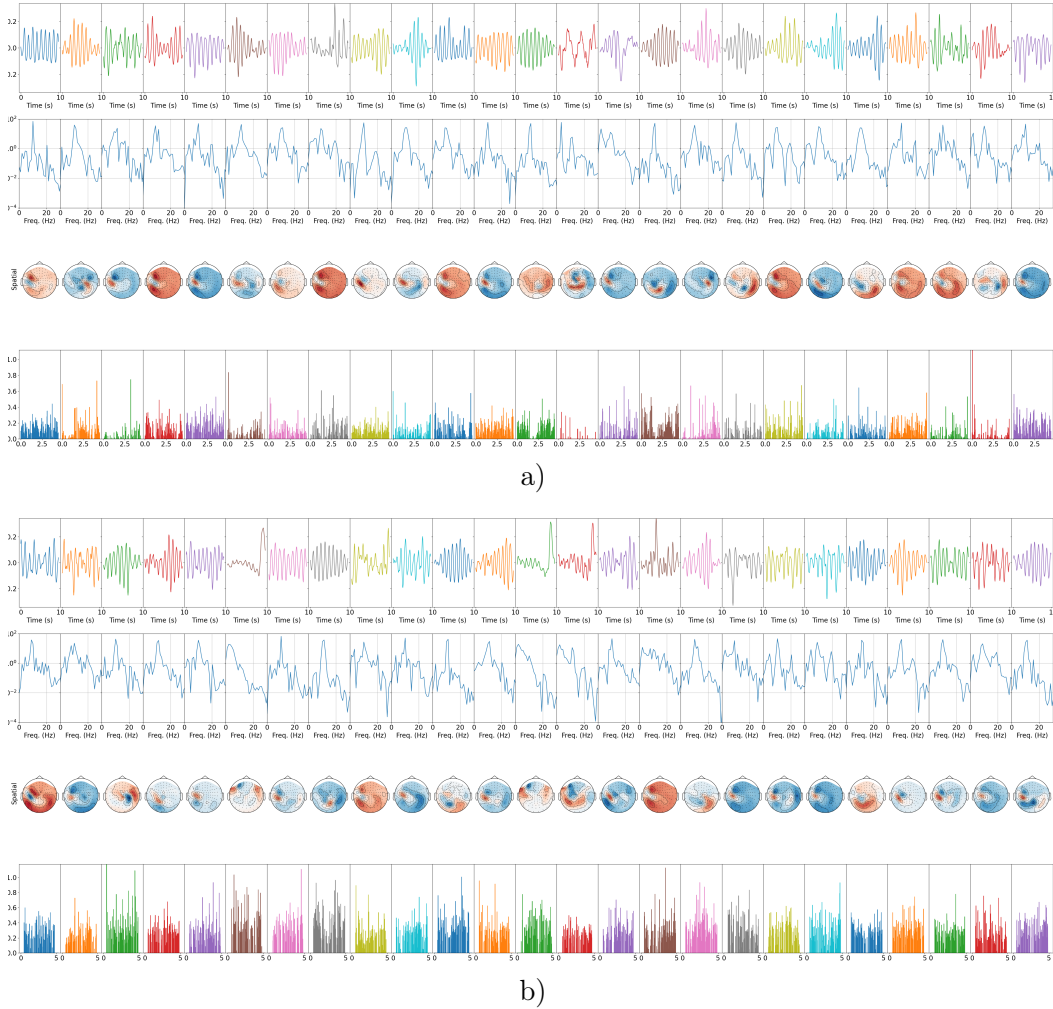


Figure 7.17: Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row) and corresponding activations averaged over epochs (fourth row) obtained with MCSC a) and with our method b).

HCP results

We have trained models with one pair of spatial and temporal atoms, where the maximum number of activations during train Q is 5 and the maximum number of activations during test P is 2. The models are trained on one session and tested on both training and testing sessions. For each subject and each event (left hand,

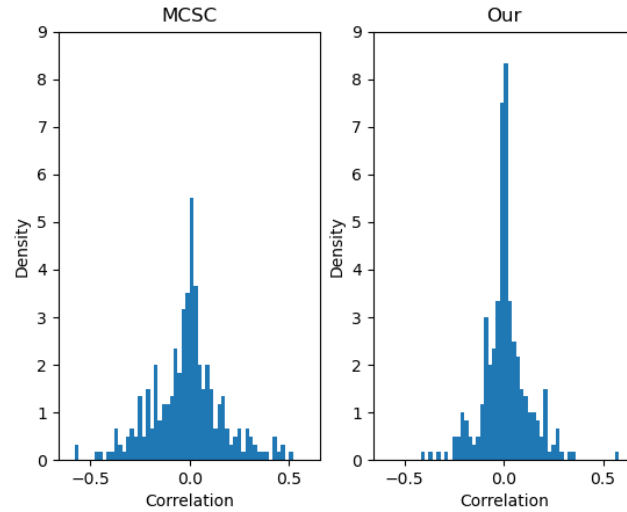


Figure 7.18: Distribution of correlations between different rank-1 atoms obtained with MCSC and our method.

left foot, right hand, right foot movements and fixations), one model is trained. The obtained spatial and temporal patterns, and training and testing activations averaged over epochs are illustrated for five subjects in Figures 7.19, 7.20, 7.21, 7.22 and 7.23. Firstly, we can notice that the spectral composition of the estimated waveforms differ significantly between subjects, while it is similar across different events. Also, average activations on training and testing sessions are consistent. We can notice that spectral components in the range $8 - 12Hz$ are emphasised in the case of the subjects 104012, 108323 and 109123 for all events. For subjects 104012 and 109123, the spectral components below $4Hz$ are of higher amplitudes for events which contain movements compared to fixation/resting state epochs. Apart for the subject 105923, by analysing the average activations, we can notice that for the epochs with movements, show two well separated clusters which correspond to two movements per each epoch. On the other hand, the average activations for fixation epochs are mostly uniformly distributed over time. High peaks at the beginning and end of the average activations are due to proximity to the signal border (taking into account that duration of the signal is $2.4s$ and duration of the temporal patterns is $1s$). If the models are trained with $Q = 3$ and $P = 2$, waveforms tend not to be well centered. On the other hand if $Q = 10$ and $P = 2$ separation of the activations are less specific (illustrations provided in Appendix C).

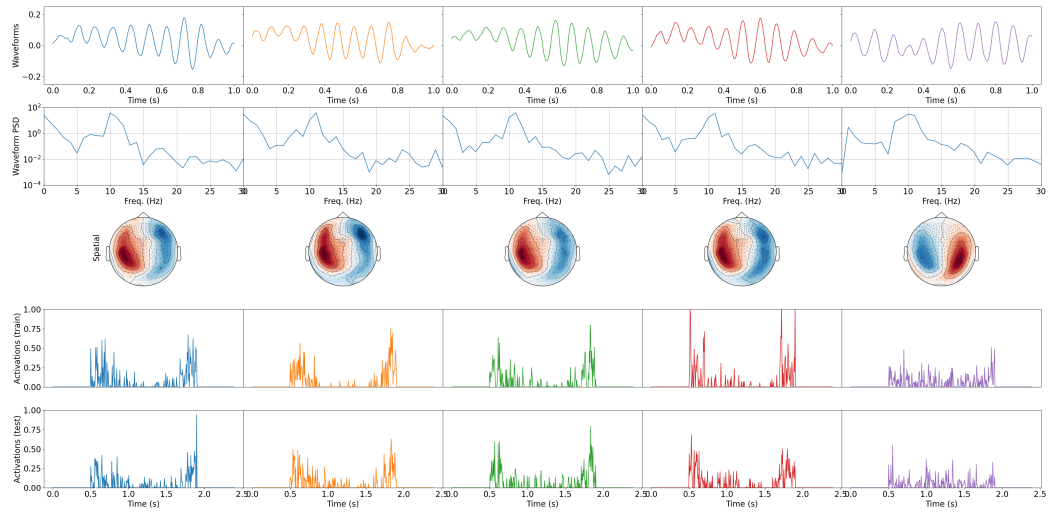


Figure 7.19: **Subject 104012** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. **Left hand** (first column), **left foot** (second column), **right hand** (third column), **right foot** (fourth column) movements, **fixation/resting** (fifth column).

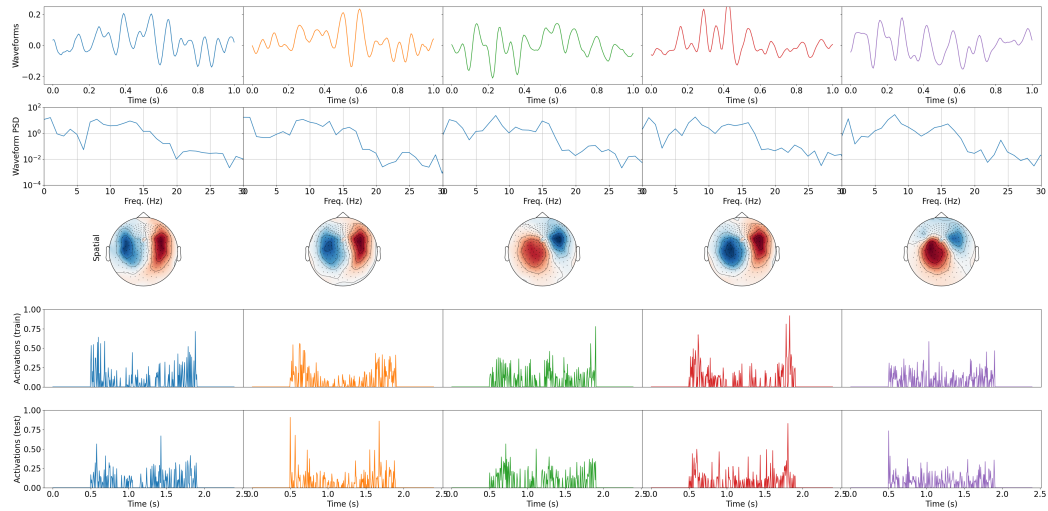


Figure 7.20: **Subject 105923** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. **Left hand** (first column), **left foot** (second column), **right hand** (third column), **right foot** (fourth column) movements, **fixation/resting** (fifth column).

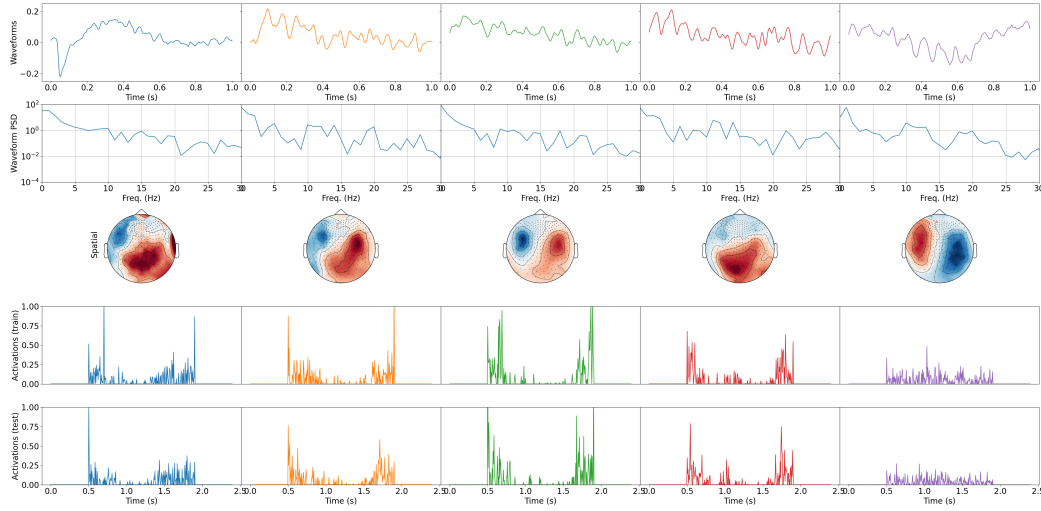


Figure 7.21: **Subject 106521** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. **Left hand** (first column), **left foot** (second column), **right hand** (third column), **right foot** (fourth column) movements, **fixation/resting** (fifth column).

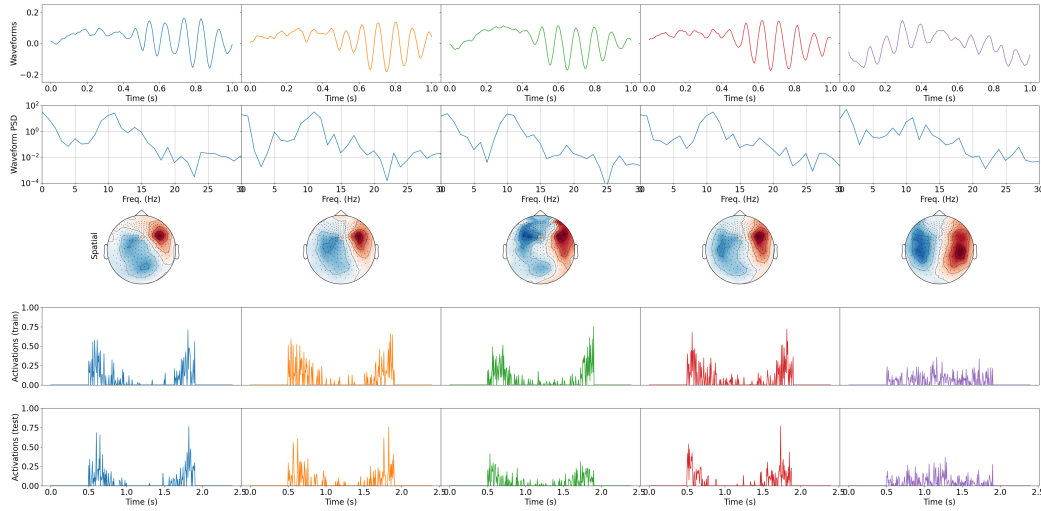


Figure 7.22: **Subject 108323** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. **Left hand** (first column), **left foot** (second column), **right hand** (third column), **right foot** (fourth column) movements, **fixation/resting** (fifth column).

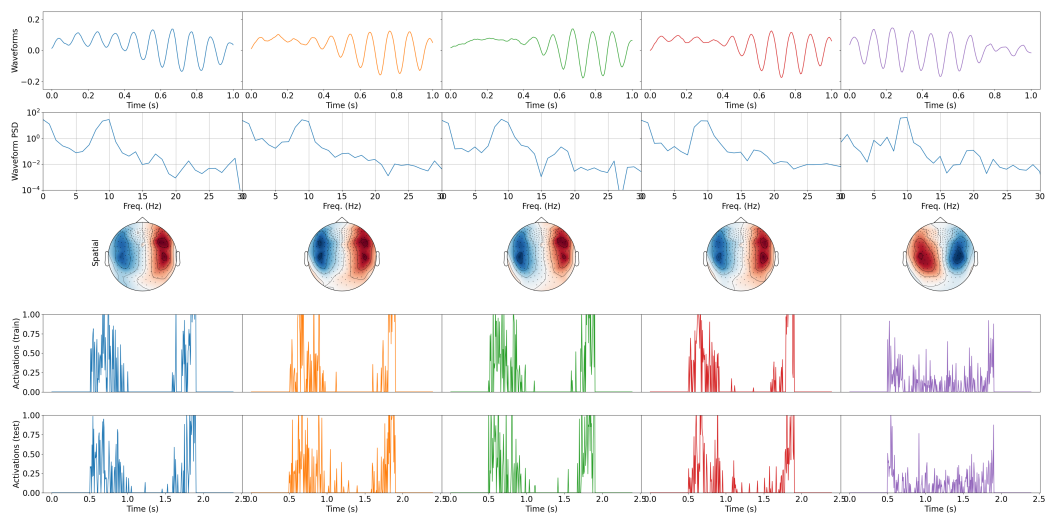


Figure 7.23: **Subject 109123** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. **Left hand** (first column), **left foot** (second column), **right hand** (third column), **right foot** (fourth column) movements, **fixation/resting** (fifth column).

7.6 Conclusion

In this chapter, we have investigated an approach for M/EEG convolutional dictionary learning with L_0 constraint. The model assumes that the multivariate M/EEG signals associated to the individual brain sources are of rank-1 and that they appear always with the same polarity. During the dictionary learning, the sparse activation vectors and the dictionaries are estimated alternatively. The sparse activation vectors are estimated iteratively via an approach inspired by the sparse autoencoders [Makhzani & Frey 2013, Makhzani & Frey 2014, Luo *et al.* 2017], IHT [Blumensath & Davies 2008] and MP [Mallat & Zhang 1993] approaches, adjusted to the convolutional rank-1 spatio-temporal dictionaries. Updates of the spatial and temporal dictionaries are performed independently using adaptive moment estimation (Adam) optimizer [Kingma & Ba 2014]. Since the minimization problem is globally non-convex, we have proposed initialization strategies which decrease chances that the optimization process ends in a local minima. The approach is compared with the state-of-the art MCSC [La Tour *et al.* 2018], an approach with L_1 regularization on the synthetic and somatosensory MEG dataset. The results demonstrated that our method is capable to learn dictionaries which on average better correlate with ground truth, both on noiseless and noisy datasets. This is especially prominent for the waveforms of wide support such as sawtooth waveforms. On the other hand, on average, MCSC yields better estimates of the activation vectors, which is more prominent for noisy data. Qualitative comparison on the somatosensory MEG dataset, showed that our approach is able to learn MEG dictionaries which highly resemble to the ones obtained with MCSC. The analysis of the dictionaries and activations obtained on HCP MEG motor task data with only a single pair of atoms, indicate that the proposed approach is capable to extract motor-task related patterns, which generalize well over an unseen session.

Shallow CNN for M/EEG classification

Contents

8.1	Theory	146
8.2	Method	146
8.2.1	Feature extraction	147
8.2.2	Feature selection and normalization	148
8.2.3	Feature classification	149
8.2.4	Training	149
8.2.5	Validation and test	151
8.3	Experiments	151
8.3.1	Databases	152
8.3.2	Implementation details	153
8.4	Results	157
8.5	Conclusion	162

Executive summary

In this chapter, we present a shallow CNN model for EEG and MEG multivariate signal classification. In this model, in addition to the rank-1 assumption and modeling of time courses as convolution of sparse activation signals and characteristic waveforms, in order to reduce impact of inter-subject and inter-session variabilities, we have assumed that the subject's head can be modeled as a sphere. As traditional BCI pipelines, the model is composed of a feature extraction, selection and classification modules which are presented in Section 8.2. This section also contains details related to the update of trainable parameters. The model is compared with three state-of-the-art CNN models for a passive and active BCI problems on EEG mental workload and MEG motor task signal classification.

8.1 Theory

Apart from being distorted by a significant noise, the main challenge of the analysis of the M/EEG signals comes from inter-subject and inter-session variability. The former one arises from different head geometries between subjects, but also due to different functional properties of the cortex [Saha & Baumert 2020]. Inter-session variability is a consequence of difference in sensor positions between sessions, but an additional variability might also come from the alertness of the subject. This problem has been most effectively addressed using transfer learning paradigms [Lotte *et al.* 2018]. In this work, we propose a regularization of the spatial and temporal feature space in order to reduce inter- and intra-subject variabilities. To achieve this, we have assumed that a head can be modeled with a sphere, thus the spatial topographic maps $\{\mathbf{u}_k\}_{k=1}^K$ can be expressed as

$$\mathbf{u}_k = \sum_{l=0}^B \sum_{m=-l}^l \mathbf{Y}^{lm} \hat{\mathbf{u}}_k^{lm} \quad (8.1)$$

where $\mathbf{Y}^{lm} \in \mathbb{R}^N$ is a discrete real SH basis element of degree l and order m and $\hat{\mathbf{u}}_k^{lm}$ its associated spectral coefficient. B is the signal's bandwidth. $N_B = (B + 1)^2$ is the number of the SH basis elements. Similarly, the temporal waveforms $\{\mathbf{v}_k\}_{k=1}^K$ can be expressed in terms of a discrete cosine basis as

$$\mathbf{v}_k[t] = \sum_{f=0}^F \frac{a_f}{\sqrt{\tau}} \cos\left(\pi f \frac{t+1}{\tau}\right) \hat{\mathbf{v}}_k^f \quad (8.2)$$

where $t = [0, 1, \dots, \tau - 1]^T$, $a_0 = 1$ and $a_f = \sqrt{2}$ if $f \neq 0$. F is the signal's bandwidth that must satisfy $F \leq \tau - 1$. In a matrix-vector notation equations 8.1 and 8.2 can be written as

$$\mathbf{u}_k = Y \hat{\mathbf{u}}_k \quad (8.3) \quad \mathbf{v}_k = C \hat{\mathbf{v}}_k \quad (8.4)$$

where $Y \in \mathbb{R}^{N \times N_B}$ contains the SH basis elements and $C \in \mathbb{R}^{\tau \times (F+1)}$ the discrete cosine basis elements in columns, $\hat{\mathbf{u}}_k \in \mathbb{R}^{N_B}$ and $\hat{\mathbf{v}}_k \in \mathbb{R}^{F+1}$ are the corresponding spectral coefficients. Finally, a multivariate signal X from Eq. 6.3 can be modeled as

$$X = \sum_{k=1}^K [Y \hat{\mathbf{u}}_k] \cdot (\mathbf{z}_k * [C \hat{\mathbf{v}}_k])^T + \mathcal{N}. \quad (8.5)$$

8.2 Method

In this work, we propose a shallow CNN with rank-1 spatial and temporal filters represented in the Fourier domain. The architecture of the model is illustrated in Figure 8.1. As in a majority of the BCI classification pipelines, we can identify a feature extraction step, a feature selection and a feature classification step. Although termed as convolutional, in reality, a CNN use cross-correlation with trainable filters.

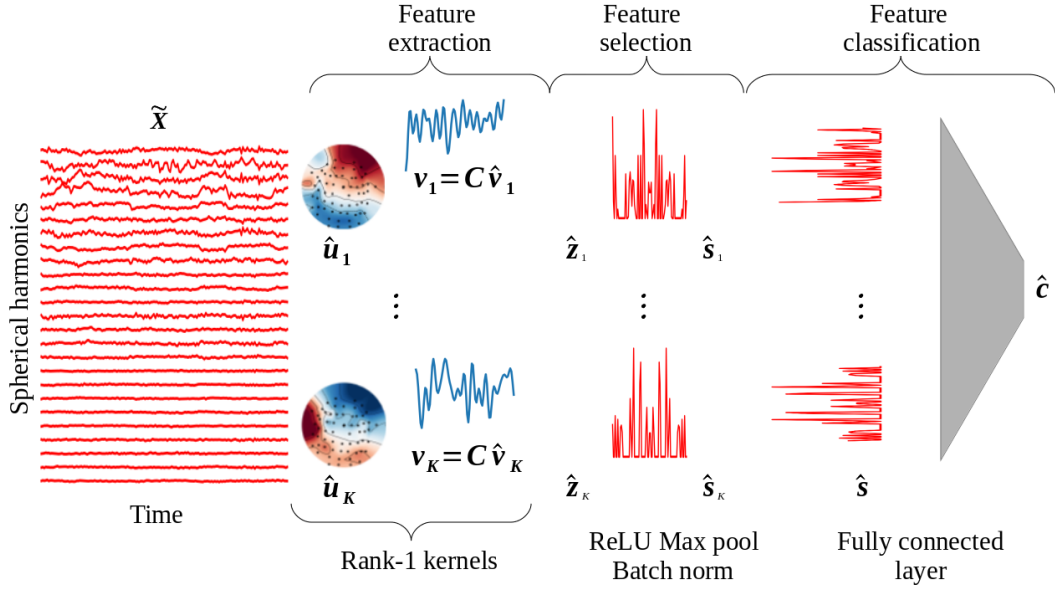


Figure 8.1: Illustration of the shallow rank-1 CNN architecture

8.2.1 Feature extraction

If a M/EEG signal X can be modeled as in Eq. 8.5, its cross-correlations with the spatial and the temporal patterns $\{Y \hat{\mathbf{u}}_k\}_{k=1}^K$ and $\{C \hat{\mathbf{v}}_k\}_{k=1}^K$ represent a measure of their presence in X . Cross-correlation of X with one spatial pattern $Y \hat{\mathbf{u}}_k$ can be written as

$$\mathbf{y}_k = [Y \hat{\mathbf{u}}_k]^T X \quad (8.6)$$

where $\mathbf{y}_k \in \mathbb{R}^T$. Given M/EEG signals from multiple subjects and/or sessions $\{X_i\}$, due to differences in sensor positions, for each session one matrix Y_i containing the SH basis elements needs to be defined. To reduce the computational time and memory requirements during training, we map all the signal samples $\{X_i\}$ to a common Fourier space as

$$\hat{X}_i = Y_i^\dagger X_i \quad (8.7)$$

where $Y_i^\dagger \in \mathbb{R}^{N_L \times N}$ is the pseudo-inverse of matrix Y_i . To solve this problem we have used the least mean square solution penalised with a Laplace-Beltrami term as

$$Y_i^\dagger = (Y_i^T Y_i + \lambda R_{LB})^{-1} Y_i^T \quad (8.8)$$

where R_{LB} is the Laplace-Beltrami regularization term and λ is a parameter which controls the amount of regularization. This solution penalizes more high frequency components, which is desirable as they are more affected by the inter-session and inter-subject variability of the sensor positions. Using equations 8.7 and 8.8, for a sample X_i we re-define cross-correlation from Eq. 8.6 with a spatial pattern \mathbf{u}_k as

$$\hat{\mathbf{y}}_{i,k} = \hat{\mathbf{u}}_k^T Y_i^\dagger X_i. \quad (8.9)$$

Cross-correlation along the temporal axis with a temporal pattern $C\hat{\mathbf{v}}_k$ is defined as

$$\hat{\mathbf{z}}_{i,k} = [JC\hat{\mathbf{v}}_k] * \hat{\mathbf{y}}_{i,k} \quad (8.10)$$

where $J \in \mathbb{R}^{\tau \times \tau}$ is a reversal matrix (ones along antidiagonal) and $\hat{\mathbf{z}}_{i,k} \in \mathbb{R}^{T-\tau+1}$. For each sample X_i , cross-correlations as defined in Eqs. 8.9 and 8.10 are performed with K pairs of spatial and temporal patterns represented in the Fourier domain as $\{\hat{\mathbf{u}}_k, \hat{\mathbf{v}}_k\}_{k=1}^K$, yielding $\{\hat{\mathbf{z}}_{i,k}\}_{k=1}^K$.

8.2.2 Feature selection and normalization

Given the vectors $\{\hat{\mathbf{z}}_{i,k}\}_{k=1}^K$, non-linear feature selection is performed using Rectified Linear Unit (ReLU) and max-pooling operator. ReLU is a simple element-wise thresholding operator which acts as

$$\mathbf{a}_{i,k}[t] = \text{ReLU}(\hat{\mathbf{z}}_{i,k}[t] + b_k) = \begin{cases} \hat{\mathbf{z}}_{i,k}[t] + b_k & \text{if } \hat{\mathbf{z}}_{i,k}[t] + b_k \geq 0 \\ 0 & \text{if } \hat{\mathbf{z}}_{i,k}[t] + b_k < 0 \end{cases} \quad (8.11)$$

where $t \in \{0, 1, \dots, T - \tau + 1\}$ and b_k is a trainable bias term. If we assume that polarity of a brain activity is always the same, discarding negative cross-correlation coefficients with ReLU is justified.

In general, the task of a pooling operator is to summarize input signal over small patches and to provide a feature map of a reduced resolution to the following layer. This is usually achieved by summarizing each patch with its average or maximum value. In our work, we have used the max-pooling operator as it goes along with the assumption that relevant brain activities occur sparsely over time. Given an input vector $\mathbf{a}_{i,k}$ and max-pooling size M , output is obtained as

$$\mathbf{s}_{i,k}[t] = \max\{\mathbf{a}_{k,i}[t'] : t \cdot M \leq t' < \min((t+1) \cdot M, T - \tau + 1)\} \quad (8.12)$$

where $t \in \{0, \dots, \lfloor \frac{T-\tau+1}{M} \rfloor\}$.

Since the spatial and temporal patterns $\{\hat{\mathbf{u}}_k, \hat{\mathbf{v}}_k\}_{k=1}^K$ may poorly correlate with the input signal, corresponding feature maps $\{\mathbf{s}_{i,k}\}_{k=1}^K$ might be very skewed. If for two input samples X_i and X_j belonging to different classes, feature vectors $\mathbf{s}_{i,k}$ and $\mathbf{s}_{j,k}$ are very similar, it means that the pair of spatial and temporal filters $\hat{\mathbf{u}}_k, \hat{\mathbf{v}}_k$ does not have a high discrimination power. Thus, during training these weights will not be significantly updated. To avoid this, we have used batch normalization layer [Ioffe & Szegedy 2015]. Batch normalization layer shifts and scales input feature maps as follows

$$\hat{\mathbf{s}}_{i,k} = \frac{\mathbf{s}_{i,k} - \mu_k}{\sqrt{\sigma_k^2 + \varepsilon}} \quad (8.13)$$

where mean μ_k and standard deviation σ_k differ in the training and the testing phase. During the training phase, features are normalized by their own mean and standard deviation. In the testing phase, features are normalized by the mean and standard deviation estimated during the training phase using moving averages over training data.

8.2.3 Feature classification

Once the feature vectors $\{\hat{\mathbf{s}}_{i,k}\}_{k=1}^K$ are extracted, they are concatenated into feature vector $\hat{\mathbf{s}}_i = [\hat{\mathbf{s}}_{1,i}^T, \dots, \hat{\mathbf{s}}_{K,i}^T]^T$. Classification is performed with a single fully connected layer followed by *softmax* as

$$\hat{\mathbf{c}}_i = \frac{e^{D\hat{\mathbf{s}}_i + \mathbf{b}}}{\|e^{D\hat{\mathbf{s}}_i + \mathbf{b}}\|_1} \quad (8.14)$$

where $D \in \mathbb{R}^{Q \times (K \lfloor \frac{T-\tau+1}{M} \rfloor)}$ and $\mathbf{b} \in \mathbb{R}^Q$, with Q being number of classes.

8.2.4 Training

During the training phase, trainable spatial and temporal patterns $\{\hat{\mathbf{u}}_k, \hat{\mathbf{v}}_k\}_{k=1}^K$ for the feature extraction, biases $\{b_k\}_{k=1}^K$ used in feature selection and classification parameters D and \mathbf{b} are updated via backpropagation by minimizing categorical cross-entropy loss defined as

$$\mathcal{L}(\{X_i, \hat{\mathbf{c}}_i\}_{i=1}^N) = -\frac{1}{N} \sum_{i=1}^N \mathbf{c}_i^T \log_2(\hat{\mathbf{c}}_i) \quad (8.15)$$

where $\mathbf{c}_i \in \mathbb{R}^Q$ is the ground truth vector represented in one-hot format and N is the batch size. During the training phase, moving mean and variance in batch normalization layer for testing phase are updated as follows

$$\mu_k^{it+1} = m\mu_k^{it} + (1-m)\mu_k^{batch} \quad (8.16) \quad \sigma_k^{2it+1} = m\sigma_k^{2it} + (1-m)\sigma_k^{2batch} \quad (8.17)$$

where it refers to the iteration and m is the momentum. In order to reduce the over-fitting, during the training phase a drop-out layer is used before the fully connected layer. Given the feature maps $\{\hat{\mathbf{s}}_i\}$, in each training iteration, the drop-out layer randomly sets a fraction of their entries to zero.

Classifier gradients

In a backpropagation step, the gradients of the loss L with respect to the matrix D and biases \mathbf{b} are given by

$$\frac{\partial \mathcal{L}}{\partial D} = \frac{1}{N} \sum_{i=1}^N \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{c}}_i} \frac{\partial \hat{\mathbf{c}}_i}{\partial D} = -\frac{1}{N} \sum_{i=1}^N \hat{\mathbf{s}}_i^0 (\mathbf{c}_i - \hat{\mathbf{c}}_i)^T \quad (8.18)$$

and

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}} = \frac{1}{N} \sum_{i=1}^N \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{c}}_i} \frac{\partial \hat{\mathbf{c}}_i}{\partial \mathbf{b}} = -\frac{1}{N} \sum_{i=1}^N (\mathbf{c}_i - \hat{\mathbf{c}}_i)^T \quad (8.19)$$

where $\hat{\mathbf{s}}_i^0$ corresponds to the vector $\hat{\mathbf{s}}_i$ after drop-out layer is applied.

Feature extractor gradients

Gradients of the loss L with respect to the bias b_k used in the feature selection step

are obtained as

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial b_k} &= \frac{1}{N} \sum_{i=1}^N \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{c}}_i} \frac{\partial \hat{\mathbf{c}}_i}{\partial \hat{\mathbf{s}}_i^0} \frac{\partial \hat{\mathbf{s}}_i^0}{\partial \hat{\mathbf{s}}_i} \frac{\partial \hat{\mathbf{s}}_i}{\partial \hat{\mathbf{s}}_{i,k}} \sum_{j=1}^N \frac{\partial \hat{\mathbf{s}}_{i,k}}{\partial \mathbf{s}_{j,k}} \frac{\partial \mathbf{s}_{j,k}}{\partial \mathbf{a}_{j,k}} \frac{\partial \mathbf{a}_{j,k}}{\partial b_k} \\ &= \frac{1}{N} \sum_{i=1}^N (\mathbf{c}_i - \hat{\mathbf{c}}_i)^T D \mathcal{T}_i^{dp} \mathcal{T}_k^c \sum_{j=1}^N \frac{\partial \hat{\mathbf{s}}_{i,k}}{\partial \mathbf{s}_{j,k}} \mathcal{T}_{j,k}^p \mathcal{H}(\mathbf{a}_{j,k}) \end{aligned} \quad (8.20)$$

where \mathcal{T}_k^c is an operator (mask) which performs the concatenation of the vectors $\{\hat{\mathbf{s}}_{i,k}\}_{k=1}^K$ to $\hat{\mathbf{s}}_i$ and \mathcal{T}_i^{dp} is an operator (mask) which performs the drop-out operation on the vector $\hat{\mathbf{s}}_i$ producing the vector $\hat{\mathbf{s}}_i^0$. The derivative of the batch normalization function $\frac{\partial \hat{\mathbf{s}}_{i,k}}{\partial \mathbf{s}_{j,k}} \in \mathbb{R}^{\lfloor \frac{T-\tau+1}{M} \rfloor \times \lfloor \frac{T-\tau+1}{M} \rfloor}$ is defined as

$$\frac{\partial \hat{\mathbf{s}}_{i,k}}{\partial \mathbf{s}_{j,k}}[p, q] = \begin{cases} \frac{(N \lfloor \frac{T-\tau+1}{M} \rfloor - 1)(\sigma_k^2 + \varepsilon) - (\mathbf{s}_{i,k}[p] - \mu_k)(\mathbf{s}_{j,k}[q] - \mu_k)}{(N \lfloor \frac{T-\tau+1}{M} \rfloor - 1)\sqrt{\sigma_k^2 + \varepsilon}^3} & \text{if } i = j \text{ and } p = q \\ -\frac{(\sigma_k^2 + \varepsilon) - (\mathbf{s}_{i,k}[p] - \mu_k)(\mathbf{s}_{j,k}[q] - \mu_k)}{(N \lfloor \frac{T-\tau+1}{M} \rfloor - 1)\sqrt{\sigma_k^2 + \varepsilon}^3} & \text{otherwise} \end{cases} \quad (8.21)$$

$\mathcal{T}_{j,k}^p$ is the operator (mask) which performs the max-pooling from the vector $\mathbf{a}_{i,k}$ to the vector $\mathbf{s}_{i,k}$. \mathcal{H} denotes the Heaviside function, which is the gradient of the ReLU function.

The gradients of \mathcal{L} with respect to the temporal filters $\{\hat{\mathbf{v}}_k\}_{k=1}^K$ used in the feature selection step are obtained as

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{v}}_k} &= \frac{1}{N} \sum_{i=1}^N \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{c}}_i} \frac{\partial \hat{\mathbf{c}}_i}{\partial \hat{\mathbf{s}}_i^0} \frac{\partial \hat{\mathbf{s}}_i^0}{\partial \hat{\mathbf{s}}_i} \frac{\partial \hat{\mathbf{s}}_i}{\partial \hat{\mathbf{s}}_{i,k}} \sum_{j=1}^N \frac{\partial \hat{\mathbf{s}}_{i,k}}{\partial \mathbf{s}_{j,k}} \frac{\partial \mathbf{s}_{j,k}}{\partial \mathbf{a}_{j,k}} \frac{\partial \mathbf{a}_{j,k}}{\partial \hat{\mathbf{v}}_k} \\ &= \frac{1}{N} \sum_{i=1}^N (\mathbf{c}_i - \hat{\mathbf{c}}_i)^T D \mathcal{T}_i^{dp} \mathcal{T}_k^c \sum_{j=1}^N \frac{\partial \hat{\mathbf{s}}_{i,k}}{\partial \mathbf{s}_{j,k}} \mathcal{T}_{j,k}^p [\mathcal{H}(\mathbf{a}_{j,k}) \odot (JC * \mathbf{y}_{j,k})] \end{aligned} \quad (8.22)$$

where $JC * \mathbf{y}_{j,k}$ denotes the column-wise correlation between the discrete cosine basis elements which are organized in columns of the matrix C and $\mathbf{y}_{j,k}$. \odot denotes column-wise and element-wise multiplication.

The gradients of \mathcal{L} with respect to the spatial filters $\{\hat{\mathbf{u}}_k\}_{k=1}^K$ used in the feature selection step are obtained as

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{u}}_k} &= \frac{1}{N} \sum_{i=1}^N \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{c}}_i} \frac{\partial \hat{\mathbf{c}}_i}{\partial \hat{\mathbf{s}}_i^0} \frac{\partial \hat{\mathbf{s}}_i^0}{\partial \hat{\mathbf{s}}_i} \frac{\partial \hat{\mathbf{s}}_i}{\partial \hat{\mathbf{s}}_{i,k}} \sum_{j=1}^N \frac{\partial \hat{\mathbf{s}}_{i,k}}{\partial \mathbf{s}_{j,k}} \frac{\partial \mathbf{s}_{j,k}}{\partial \mathbf{a}_{j,k}} \frac{\partial \mathbf{a}_{j,k}}{\partial \hat{\mathbf{u}}_k} \\ &= \frac{1}{N} \sum_{i=1}^N (\mathbf{c}_i - \hat{\mathbf{c}}_i)^T D \mathcal{T}_i^{dp} \mathcal{T}_k^c \sum_{j=1}^N \frac{\partial \hat{\mathbf{s}}_{i,k}}{\partial \mathbf{s}_{j,k}} \mathcal{T}_{j,k}^p [\mathcal{H}(\mathbf{a}_{j,k}) \odot (JC \hat{\mathbf{v}}_k * (\mathbf{Y}_j^\dagger \mathbf{X}_j)^T)] \end{aligned} \quad (8.23)$$

where $J\mathbf{C}\hat{\mathbf{v}}_k * (Y_i^\dagger X_i)^T$ denotes the column-wise correlation between the temporal filters $\mathbf{v}_k = C\hat{\mathbf{v}}_k$ and the input data, whose spatial dimension is transformed into the Fourier domain $(Y_i^\dagger X_i)^T$.

8.2.5 Validation and test

During the validation and the testing phases, the batch normalization is performed using the mean and variance estimated during the training phase as in equations 8.16 and 8.17. Also, during these phases, the drop-out layer is deactivated. The validation accuracy is computed as

$$a_v = \frac{1}{N_v} \sum_{i=1}^{N_v} \mathbf{c}_i^T \operatorname{argmax}_1 \{\hat{\mathbf{c}}_i\} \quad (8.24)$$

where argmax_1 denotes a function which assigns 1 to input's maximum and 0 to other entries and N_v is the number of validation samples. Table 8.1 provides the number of multiplications for the different operations used in the classification process of one sample.

Table 8.1: Number of multiplications per different steps of the entire classification process for one input sample.

Operation	Number of multiplications
Spatial Fourier transform Eq. 8.7	$N_L \times N \times T$
Spatial correlation Eq. 8.9	$K \times N_L \times T$
Temporal correlation Eq. 8.10	$K \times (\tau \times F + \tau \times (T - \tau + 1))$
Batch normalization Eq. 8.13	$K \times \lfloor \frac{T-\tau+1}{M} \rfloor$
Feature classification Eq. 8.14	$Q \times K \times \lfloor \frac{T-\tau+1}{M} \rfloor + Q \times (1 + 3(N_{Ty} - 2))$

* N_{Ty} corresponds to the Taylor series degree used to compute exponentials.

8.3 Experiments

We have compared our method with three state of the art methods, namely *DeepConvNet* and *ShallowConvNet* proposed by [Schirrneister *et al.* 2017] and *EEGNet* proposed in [Lawhern *et al.* 2018]. Methods are compared on two datasets - on the problem of mental workload classification from EEG signals for a passive BCI and on the classification of motor-task MEG data. For each dataset, two labeled sessions per subject are available. Since in the BCI applications it is common that the algorithm is tuned to the recordings of the user, methods are compared for two experimental setups:

- *Subject blind experiment*: subjects used in training and validation do not exist in the testing set.
- *Subject aware experiment*: sessions used in training and validation do not exist in the testing data.

8.3.1 Databases

Mental workload EEG dataset for passive BCI

We used the open mental workload EEG dataset provided in the "Passive BCI Hackathon" organized during the Neuroergonomics 2021 conference [Hinss *et al.* 2021]. The dataset contains EEG recordings of 15 subjects acquired over three sessions where participants were asked to perform a Multi-Attribute Task Battery-II (MATB-II) task developed by NASA. Since the labels of the third session are not publicly available, we have used only two sessions in our experiments. In each session, participants were asked to perform four sub-tasks (system monitoring, tracking, resource management and communications) to create three mental workload difficulties, which are recorded during five minutes long sessions. They are labeled with 'easy', 'medium' and 'difficult' labels. The number of available EEG channels is 61 and sampling frequency 500 Hz. Each session is segmented into 447 2s long epochs. Signals are band-pass filtered with FIR filters with cut-off frequencies 1 Hz and 40 Hz. Artefacts coming from other organs are removed and the signals are downsampled to the sampling rate of 250 Hz.

In our experiments, we have subsequently downsampled the signals by a factor 3, given that the signals have been low-pass filtered with a cut-off frequency of 40 Hz. Thus, the sampling frequency is approximately 83 Hz. Signals are scaled with the factor $5 \cdot 10^4$ to avoid dead neurons. For the *subject blind* setup, we have used 9 subjects for training, 3 for validation and 3 for testing. Correspondingly, for the *subject aware* experiment, we have used one session from each of the 3 subjects for validation and from each of the 3 subjects for testing. The split into train, valid and test is randomly repeated three times.

Motor-task MEG dataset

The motor-task MEG dataset is part of the open HCP [Van Essen *et al.* 2012] dataset. The dataset contains MEG recordings of 61 subjects acquired over two sessions where participants were guided by visual cues to move either the right hand, left hand, right foot, or left foot. Each session was composed of 42 blocks, where 10 blocks were resting state blocks and 32 blocks were movement blocks (8 blocks per movement). Each movement block contains 10 movements guided by a visual stimulus. A stimulus lasts 150ms and is followed by 1050ms of black screen. The number of MEG channels is 248 and the initial sampling frequency is 2034.52 Hz. Signals are segmented into 2.4s long epochs.

To preprocess the raw MEG signals, we have used the MNE-HCP library [Gramfort *et al.* 2013b]. Preprocessing included reference correction, filtering with bandpass Butterworth filter of order 4 with cutoff frequencies of 0.5 Hz and 60 Hz, removing of artefacts using ICA and interpolation of missing or bad channels. In our experiments, we have subsequently downsampled the signals by a factor 12, given that the signals are low-pass filtered with cut-off frequency of 60 Hz. Thus, the sampling frequency is ~ 170 Hz. Again, for stability of the model, signals are

scaled with the factor $5 \cdot 10^{12}$. In the *subject blind* setup, we have used 20 subjects for training, 10 for validation and 31 for testing. In the *subject aware* experiment setup, one session from each of the 10 subjects were used for validation and one session from each of the 31 subjects for testing.

8.3.2 Implementation details

All models are implemented with the *tensorflow* library [Abadi *et al.* 2016]. Loss function of all models is categorical cross entropy and they are trained using Adam optimizer [Kingma & Ba 2014].

In the experiments with motor task MEG data, the models are trained over 200 epochs with batch size 64 and initial learning rate of 0.001. If the difference between validation classification accuracy averaged over two sequential blocks of three epochs is greater than 10^{-4} , the learning rate is reduced by a factor of 0.9. Since the number of trials belonging to *fixation/resting state* is higher compared to the other four classes, at each epoch 1280 samples are randomly selected from each of the five classes over the entire training subset. In each epoch there is 100 iterations. The spatial component of the signals is transformed to the Fourier domain using the pseudo-inverse of the SH basis as in Eq. 8.7 obtained with a Laplace-Beltrami regularization as in Eq. 8.8 and a regularization weight $\lambda = 0.001$. The spatial component bandwidth B is varied between 6 and 12. This transformation reduces the spatial dimensionality from 248 channels to $N_B \in \{49, 81, 121, 144\}$ SH coefficients, for bandwidths $B \in \{6, 8, 10, 12\}$, respectively. The length of the temporal filters \mathbf{v}_k is 85 samples which corresponds to approximately 0.5s. They are represented in terms of DCT coefficients as in Eq. 8.4. Maximum bandwidth of the temporal filter representation is varied between $F \in \{10, 20, 30, 40\}$ Hz. Pooling step used to select features as in Eq. 8.12 is $M = 10$.

In the experiments with mental workload EEG data, the models are trained over 100 epochs with batch size 64 and the initial learning rate of 0.0005. If the difference between validation classification accuracy averaged over two sequential blocks of three epochs is greater than 10^{-4} , the learning rate is reduced by a factor of 0.9. As the classes in this dataset are balanced, the models are trained on the entire training dataset. As in the experiment with MEG data, the SH coefficients are estimated using a Laplace-Beltrami regularization with $\lambda = 0.001$. Due to a lower number of sensors and a lower signal to noise ratio, in the case of the EEG signals, the spatial component bandwidth is varied between 2 and 4. This transformation reduces the spatial dimensionality from 61 channels to $N_B \in \{9, 16, 25\}$ SH coefficients, for bandwidths $B \in \{2, 3, 4\}$, respectively. In this experiment, the length of the temporal filters \mathbf{v}_k is 42 samples which also corresponds to approximately 0.5s and their bandwidth is varied between $F \in \{5, 10, 15\}$ Hz. Pooling step used to select features as in Eq. 8.12 is $M = 20$.

To select hyper-parameters of the models, namely the bandwidths of the spatial and the temporal patterns B and F , and the number of rank-1 kernels K , we have firstly analysed validation curves. Figure 8.2 illustrates validation curves for *subject blind*

and *subject aware* motor task MEG experiments, for a fixed number $K = 50$ of kernels and varying bandwidths B and F . We can notice that in both experimental setups, and for all spatial bandwidths B , limiting temporal filters to bandwidth of 10 Hz results in a lower validation accuracy. This can be explained by the fact that μ waves, which are present in the motor cortex and are suppressed when a motor task is performed, have a frequency range of 8 – 12 Hz. For $F \geq 10$ Hz, we can observe that the validation curves corresponding to $B = 6$ are on average lower than the curves corresponding to $B \geq 6$. This is more prominent in *subject aware* experimental set-up. This is a consequence of a higher inter-subject variability of the spatial components compared to the intra-subject one. The best model with the lowest number of parameters, in the *subject blind* experiment is the model with $B = 8$ and $F = 30$. In the *subject aware* this is clearly the model with $B = 10$ and $F = 30$.

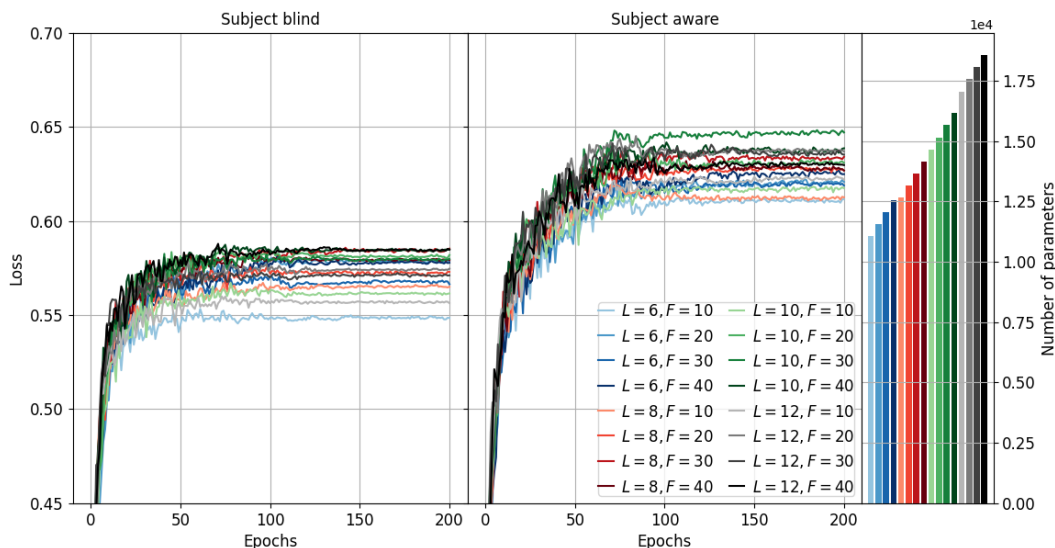


Figure 8.2: Validation curves for the motor task MEG classification problem for fixed number of rank-1 kernels $K = 50$ and different spatial and temporal kernel bandwidths B and F , and corresponding number of trainable parameters.

For the selected hyper-parameters B and F , we have further analyzed validation curves, when the number of rank-1 kernels K increases. In Figure 8.3, validation curves are depicted for different values of $K \in \{50, 100, 200, 300, 400, 500\}$. In the *subject blind* setup, we can notice that increasing the number of kernels does not necessarily and significantly improve the validation accuracy. On the other hand, consistent improvements can be observed in the *subject aware* experiment. This indicates that in addition to patterns common to all subject, the more room (kernels) a model is given, the more subject-specific patterns it is able to learn.

The mental workload EEG dataset is smaller, the signal-to-noise ratio of EEG is lower and the number of sensors is smaller, thus a training of a neural network model

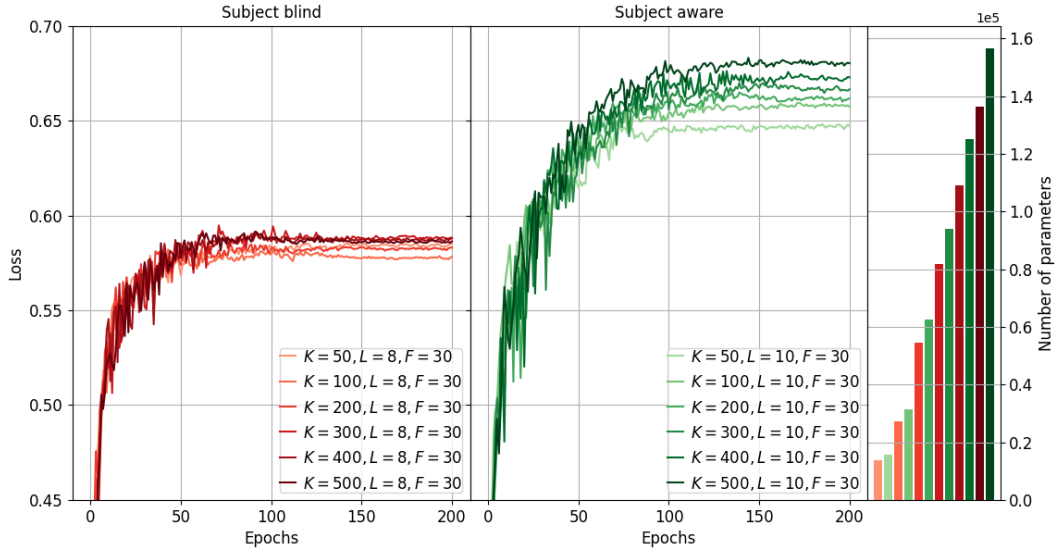


Figure 8.3: Validation curves for motor task MEG classification problem for different number of kernels K and their fixed spatial and temporal bandwidths L and F , and corresponding number of trainable parameters.

on such data is quite challenging. To select hyper-parameters, the experiments are repeated three times for three random splits of dataset into training, validation and testing subsets. In Figures 8.4 and 8.5, different lines styles (full, '-' and '-.-') correspond to different random splits. Plots in Figure 8.4 illustrate validation curves for the *subject blind* and the *subject aware* mental workload EEG experiments, split-wise and averaged, for a fixed number of kernels $K = 50$ and varying bandwidths B and F . Firstly, we can observe that increasing spatial bandwidths B results in more dispersed validation curves over different random splits of dataset and can lead to an overfitting. This is especially visible for $B = 4$ in the *subject aware* validation curves. In the *subject blind* experiment we can notice that on average, validation curves over all spatial and temporal bandwidths are rather close, where the models with $F = 5$ result in slightly higher validation accuracy. On the other hand, in the *subject aware* experiments we can notice that models with $F = 5$ give the lowest validation accuracy and the ones with $F = 10$ the highest. This goes along with the fact that *theta* waves (4-8 Hz) are associated to mental fatigue and workload [Gevins *et al.* 1997]. To select the best model we have used the averages over random splits of the validation accuracies in the last epoch. In the *subject blind* experiment, the model with $B = 2$ and $F = 5$ is selected as the best one, while in the *subject aware* experiment, the best one is the model with $B = 2$ and $F = 10$. For the selected hyper-parameters B and F , we have further analyzed the validation curves for an increasing number of kernels K . In Figure 8.5, validation curves are depicted for different values of $K \in \{50, 100, 200, 300, 400, 500, 1000\}$. In both, *subject blind* and *subject aware* setups, we can notice that increase of the number

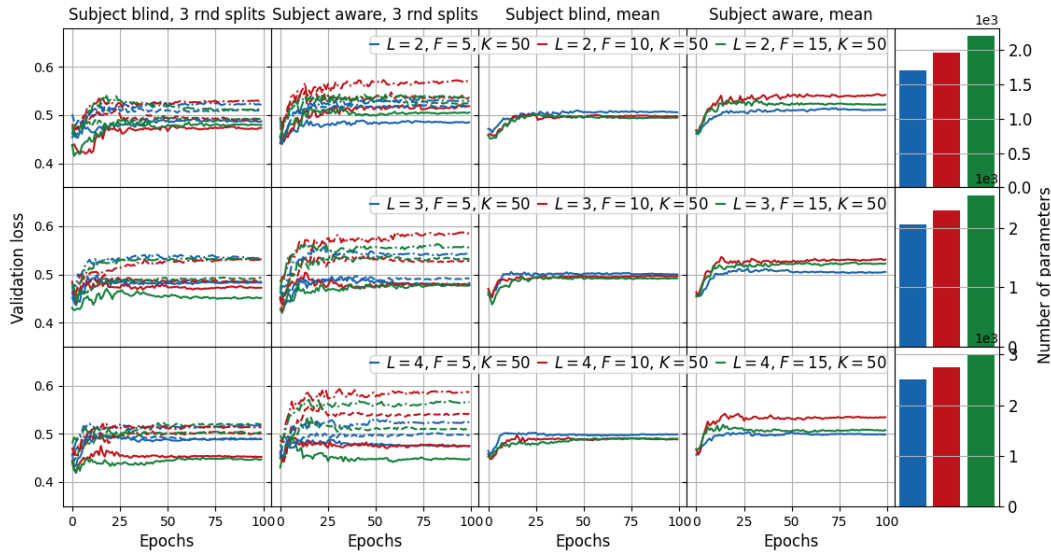


Figure 8.4: Validation curves for mental workload EEG classification problem for fixed number of kernels $K = 50$, and different spatial and temporal kernel bandwidths L and F , and corresponding number of trainable parameters.

of kernels, on average, improves validation accuracy. Contrary to the MEG motor task experiment, where these improvements are more significant in the *subject aware* setup, here that is not the case. This might indicate that the inter-session variability in the case of mental workload EEG signals is more significant and that the improvement in validation accuracy between *subject blind* and *subject aware* model training is rather a consequence of the increase of training data than in learning of subject specific patterns.

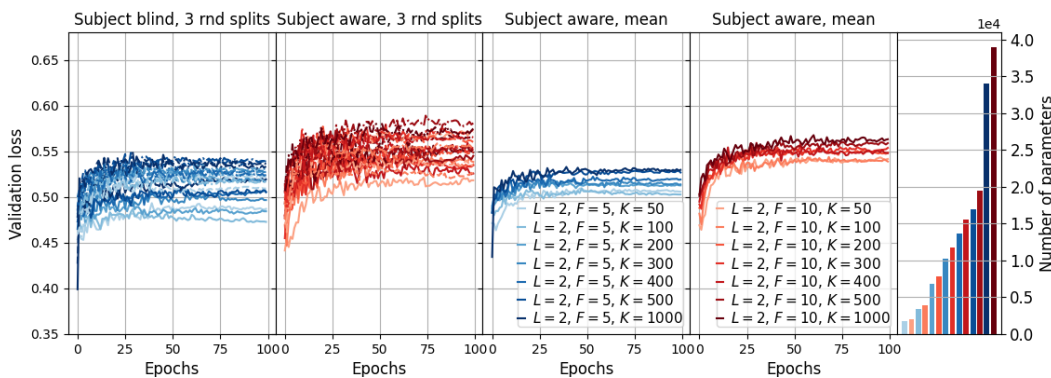


Figure 8.5: Validation curves for mental workload EEG classification problem for different number of kernels K and their fixed spatial and temporal bandwidths L and F , and corresponding number of trainable parameters.

Selection of the hyper-parameters used in compared methods, namely *DeepConvNet* and *ShallowConvNet* [Schirrmester et al. 2017] and *EEGNet* [Lawhern et al. 2018] is provided in Appendix D.

8.4 Results

The results are compared quantitatively in terms of confusion matrices and classification accuracy. Given the importance of model’s speed and memory requirements for the real time applications with portable processors in BCI, the models are also compared in terms of the number of trainable parameters and the number of multiplications.

In Figures 8.6 and 8.7 confusion matrices are given for the *subject blind* and *subject aware* MEG motor task experiments averaged over five repetitions of the experiments. We can observe, that apart from the *fixation* class, all models have high sensitivity (true positive rate) with respect to the *right hand* movement class. On the other side, classification of the *right foot* movements appears to be the most challenging one and they are mostly misclassified into the *left foot* and the *right hand* classes. Compared with the *subject blind* training, *subject aware* training most significantly impacts classification of the *right foot* movements by reducing misclassifications into the *right hand* and the *fixation* classes, while the missclassification into the *left foot* class still remains. The *subject aware* training also significantly improves the classification of the *left hand* movements by reducing the misclassifications into the *left foot*, the *right hand* and the *fixation* classes. Comparing the confusion matrices in both experiments, we can notice that our model with a higher number of parameters exhibits higher sensitivity with respect to the *left hand* class. In the *subject blind* experiments, sensitivity is higher also with respect to the *right hand* movement, but lower for the *left foot* class. In the *subject aware* experiments, our model has higher sensitivity for the *right foot* class, while for the *left foot* class sensitivity of *EEGNet* is significantly higher than with other models.

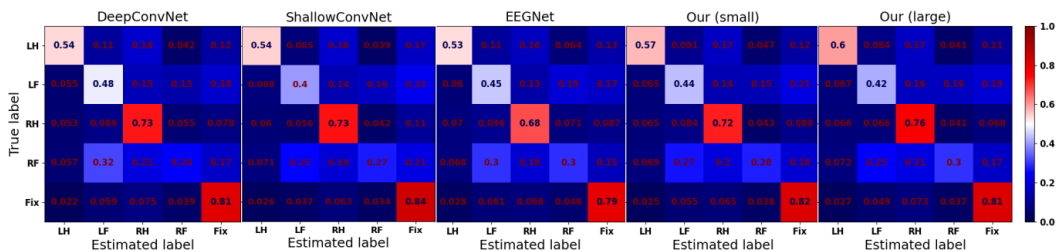


Figure 8.6: Confusion matrices for *DeepConvNet*, *ShallowConvNet*, *EEGNet*, *Our (small)* and *Our (large)* models obtained in MEG motor task the *subject blind* experiments averaged over five experiment repetitions.

In Tables 8.2 and 8.3, classification accuracy is compared for the *subject blind* and

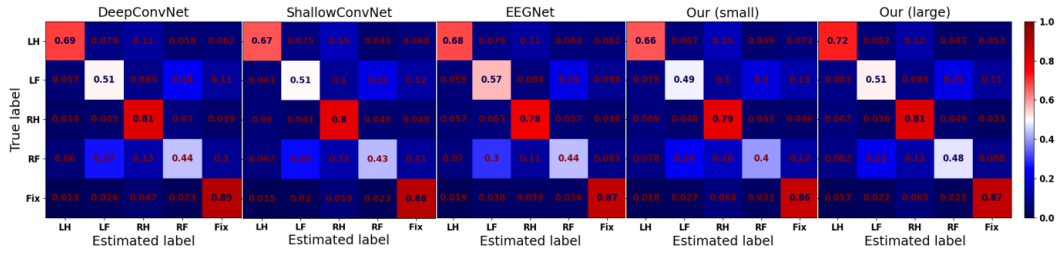


Figure 8.7: Confusion matrices for *DeepConvNet*, *ShallowConvNet*, *EEGNet*, *Our (small)* and *Our (large)* models obtained in MEG motor task *subject aware* experiments averaged over five experiment repetitions.

the *subject aware* MEG motor task experiments for five repetitions of the experiments. In the *subject blind* experiments, we can observe that our model with a small number of trainable parameters can achieve the same performance as significantly larger models *DeepConvNet*, *ShallowConvNet* and *EEGNet*. The larger model, while still having a significantly lower number of parameters compared to *DeepConvNet* and *ShallowConvNet*, leads to an average improvement of at least 1.5%. In the *subject aware* experiments, we can notice that our small model does not have enough capacity to capture subject specific patterns, while the larger model results in a slight improvement of the classification accuracy compared to other models.

Table 8.2: Classification accuracy for *DeepConvNet*, *ShallowConvNet*, *EEGNet*, *Our (small)* and *Our (large)* models obtained in the MEG motor task *subject blind* experiments for five experiment repetitions.

Experiment	<i>subject blind</i>				
Model	1 st run	2 nd run	3 rd run	4 th run	5 th run
<i>DeepConvNet</i>	0.576	0.576	0.573	0.575	0.573
<i>ShallowConvNet</i>	0.576	0.578	0.575	0.575	0.576
<i>EEGNet</i>	0.560	0.567	0.561	0.566	0.569
<i>Our (small)</i>	0.585	0.574	0.578	0.579	0.580
<i>Our (large)</i>	0.595	0.593	0.590	0.588	0.596

Figure 8.8 shows a comparison of the classification accuracies on the testing and validation data versus the number of parameters and the number of multiplications required for the classification of one data sample. The number of multiplications only counts multiplications in convolutional and batch normalization layers (not multiplications required in non-linear layers). Since for the models that are selected as the best ones, based on validation accuracy of one run of the experiments, model training is repeated four more times, for these models we have provided average accuracy (depicted with full circles) and accuracy for each experiment run (depicted with vertical dash lines). Firstly, we can observe that in the *subject blind* experi-

Table 8.3: Classification accuracy for *DeepConvNet*, *ShallowConvNet*, *EEGNet*, *Our(small)* and *Our(large)* models obtained in MEG motor task *subject aware* experiments for five experiment repetitions.

Experiment	<i>subject aware</i>				
Model	1 st run	2 nd run	3 rd run	4 th run	5 th run
<i>DeepConvNet</i>	0.684	0.686	0.682	0.680	0.684
<i>ShallowConvNet</i>	0.678	0.674	0.669	0.672	0.671
<i>EEGNet</i>	0.677	0.678	0.678	0.678	0.683
<i>Our (small)</i>	0.651	0.652	0.656	0.656	0.658
<i>Our (large)</i>	0.693	0.690	0.689	0.691	0.692

ments, our model achieves a high classification accuracy with a significantly lower number of trainable parameters than *DeepConvNet* and *ShallowConvNet*, and with a comparable number of parameters for *EEGNet*. In the *subject aware* training, differences in classification accuracy between our models and *EEGNet* models are less significant for comparable number of parameters. When comparing the number of multiplications, we can notice that all comparing models require at least 10 times more multiplications to achieve accuracy comparable to the one obtained with our models. The reason for such a high number of multiplications in *DeepConvNet*, *ShallowConvNet* and *EEGNet* lies in the way the first convolutional layer with separable and depthwise correlations is defined. Assuming K temporal filters and N channels of the input MEG signal, these models perform correlation of each channel with each temporal filter. This means that for filter length τ and MEG signal length T , there is $N \times K \times (T - \tau + 1) \times \tau$ multiplications. Further, in *DeepConvNet* and *ShallowConvNet*, for each of the K temporal filters, there is K spatial filters of length N , so the number of multiplication is $N \times K \times (T - \tau + 1) \times K$. On the other side, for *EEGNet*, for each one of the K temporal filters there is D spatial filters, thus the number of multiplications is $N \times K \times (T - \tau + 1) \times D$. On the other hand, in our model, assuming a spatial bandwidth L , to transform spatial component of the input MEG signal to Fourier domain the number of multiplications is $(L + 1)^2 \times T \times N$. Contrary to the other models, we firstly perform spatial correlation with K spatial filters which requires $(L + 1)^2 \times T \times K$ multiplications. To transform temporal filters of bandwidth F to signal domain $K \times F \times \tau$ multiplications is required. For each one of the K spatial filters, there is one temporal filter, thus the number of multiplications required for correlations is $K \times (T - \tau + 1) \times \tau$.

Furthermore, we have quantitatively compared results on the problem of EEG mental workload classification. In Figures 8.9 and 8.10, confusion matrices are provided for the *subject blind* and the *subject aware* experiments averaged over three random splits of entire dataset and five repetitions for each of the split. In both experiments, we can observe that models exhibit high sensitivity with respect to the *Easy* class. In the *subject blind* experiment, we can see that our model missclassifies *Easy*

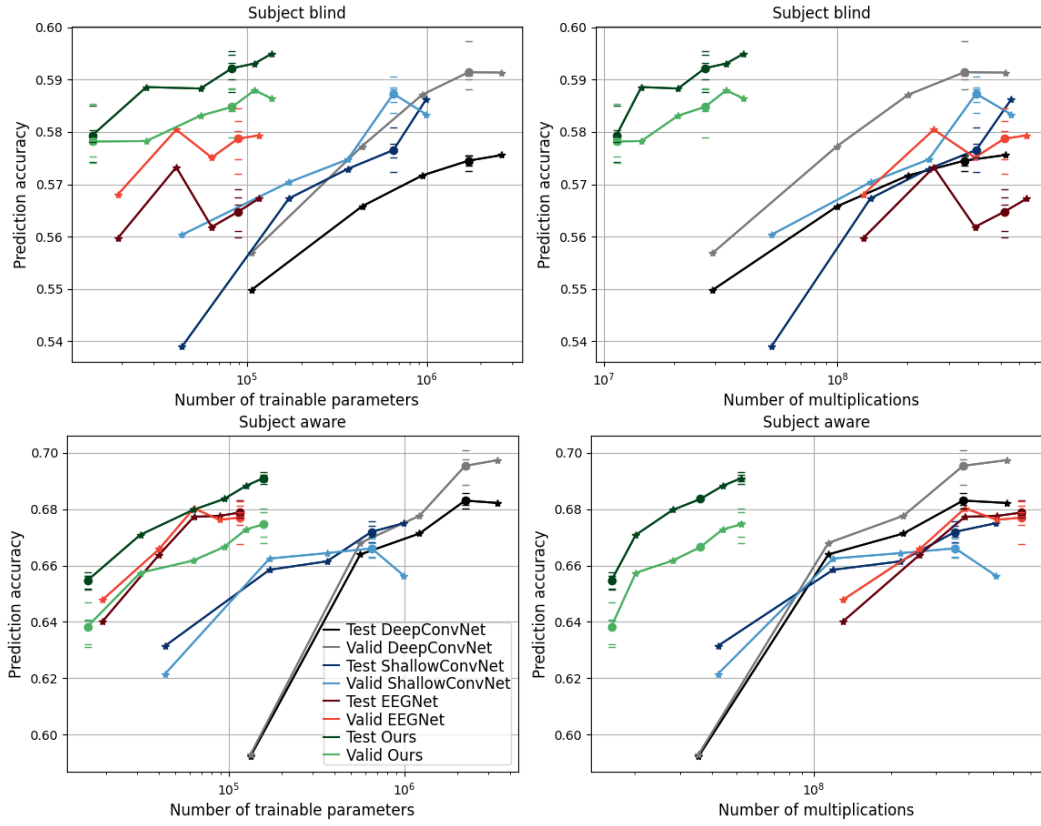


Figure 8.8: Comparison of classification accuracy on test and validation data with respect to the number of trainable parameters and the number of multiplications for the MEG motor task *subject blind* and *subject aware* experiments.

samples mostly in *Medium* class, while the other models tend to misclassify them into *Difficult* class. It has the highest sensitivity with respect to the *Medium* class, but the lowest with respect to the *Difficult* class, with a difference that majority of misclassified samples are classified in *Medium* class in contrast to *DeepConvNet* and *EEGNet*. In the *subject aware* experiments, our model has the highest sensitivity with respect to the *Easy* class, while the sensitivity is noticeably lower for *Medium* class compared to *ShallowConvNet*. In Tables 8.4 and 8.5, classification accuracy is compared for the *subject blind* and *subject aware* EEG mental workload experiments for five repetitions of the experiments averaged over three dataset splits. In *subject blind* experiment, we can observe that classification accuracies of *ShallowConvNet* and *Our* model are comparable and slightly better than ones obtained with *DeepConvNet* and *EEGNet*. On the other hand, the differences between *ShallowConvNet* and *Our* on one side and *DeepConvNet* and *EEGNet* on the other side are more significant in the *subject aware* experiment setup. Although these classification accuracies seem very low, they are comparable to the results obtained in a challenge *Passive BCI Hackathon* [Roy et al.], where the winning model [Pang et al. 2021] has achieved accuracy 54.26% with a difference that amount of training data was

higher compared to the data used in these analysis (in the challenge, two sessions for 15 subjects had labels and the labels of the third session have been hidden).

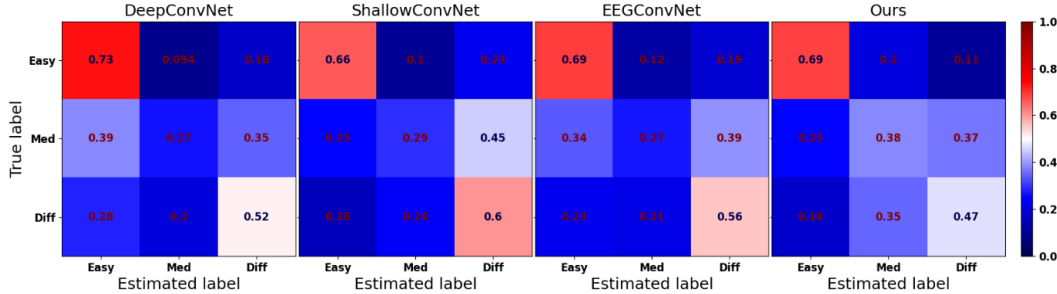


Figure 8.9: Confusion matrices for *DeepConvNet*, *ShallowConvNet*, *EEGNet* and *Our* models obtained in EEG mental workload task *subject blind* experiments averaged over five experiment repetitions and over three random splits of dataset.

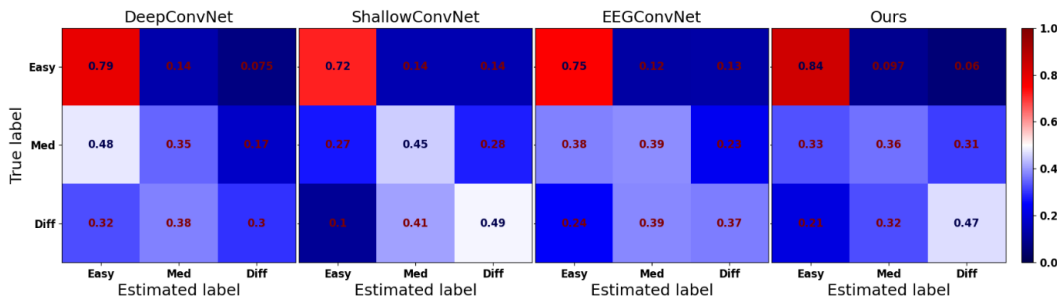


Figure 8.10: Confusion matrices for *DeepConvNet*, *ShallowConvNet*, *EEGNet* and *Our* models obtained in EEG mental workload task *subject aware* experiments averaged over five experiment repetitions and over three random splits of dataset.

Table 8.4: Classification accuracy for *DeepConvNet*, *ShallowConvNet*, *EEGNet* and *Our* models obtained in EEG mental workload task *subject blind* experiments for five experiment repetitions averaged over three random splits of data.

Experiment	<i>subject blind</i>				
	1 st run	2 nd run	3 rd run	4 th run	5 th run
<i>DeepConvNet</i>	0.510	0.496	0.512	0.504	0.502
<i>ShallowConvNet</i>	0.520	0.520	0.522	0.510	0.531
<i>EEGNet</i>	0.494	0.504	0.508	0.508	0.516
<i>Our</i>	0.518	0.508	0.514	0.513	0.517

Table 8.5: Classification accuracy for *DeepConvNet*, *ShallowConvNet*, *EEGNet* and *Our* models obtained in EEG mental workload task *subject aware* experiments for five experiment repetitions averaged over three random splits of data.

Experiment	<i>subject aware</i>				
Model	1 st run	2 nd run	3 rd run	4 th run	5 th run
<i>DeepConvNet</i>	0.486	0.475	0.472	0.485	0.476
<i>ShallowConvNet</i>	0.556	0.544	0.550	0.534	0.574
<i>EEGNet</i>	0.495	0.493	0.517	0.504	0.514
<i>Our</i>	0.556	0.554	0.553	0.559	0.563

8.5 Conclusion

In this chapter, a shallow CNN model for M/EEG data classification is presented. The model takes into account properties of the M/EEG signals. Firstly, by assuming that a head can be modeled with a sphere, spatial components of M/EEG signals are represented as a linear combination of spherical harmonics. Such representation allows dimensionality reduction along spatial dimension making the model more robust with respect to inter- and intra-subject variabilities. Furthermore, by exploiting the fact that a brain activity can be represented by a rank-1 spatio-temporal multivariate signal, we have used in our model rank-1 trainable weights. Since temporal courses of certain brain activities spread over certain frequency bandwidths, temporal kernels are regularized by being represented in terms of discrete cosine basis with a predefined bandwidth. Finally, we have demonstrated that the proposed method can achieve high accuracy with very small number of parameters and at the same time being very time efficient, what makes it suitable for portable devices with limited RAM and processors. In the future work, we will investigate potential of our model in the subject specific BCI scenarios.

Conclusions and perspectives

In this thesis, we have investigated convolutional machine learning models tailored to the properties of the examined structural and functional neuroimaging data, namely of the **dMRI**, **EEG**, and **MEG** signals. Aiming to exploit the learning capacity of the machine learning models, such as **CNNs**, while taking into account the high data variability and/or low data quantity, we have studied the models which are adapted to the domain and properties of the acquired data. This is achieved by equipping the models with a prior knowledge about the data.

For the local analysis of **dMRI** signals acquired with a q-space sampling schemes, we have investigated the spherical rotation equivariant **CNN** models which take into account real and spherical nature of the **dMRI** signals, their rotation equivariance with respect to the underlying tissue microstructures, antipodal symmetry and uniform random distribution of the sampling points. Firstly, we have studied a spherical U-net model for the **fODF** estimation from **dMRI**. The model is designed with convolutional layers realized in Fourier domain and traditional ReLU non-linearities realized in signal domain. Non-linear activations are crucial in **CNN** models as they allow the identification of the most salient features and in such a way guide the training process of the model. In order to avoid the aliasing introduced by non-linearities applied in the signal domain, in our second contribution, we have proposed a rotation equivariant **CNN** with quadratic non-linearities realized in Fourier domain for the **dMRI** local inferences, such as microstructure parameter estimation, tissue classification, and axon bundle count. In addition to evaluations performed on synthetic data, in both projects, the models are evaluated on the real **HCP dMRI** signals resampled to reduced sampling schemes which is more clinically desirable.

In the context of M/EEG signal analysis, the models are based on the assumptions that the acquired multivariate signals can be represented as a sum of rank-1 multivariate signals corresponding to the individual sources and additive noise. In addition, assuming a recurrence and transience of the brain waveforms, the temporal courses can be represented as a convolution of the sparse activation vectors and the characteristic brain waveforms. Finally, by approximating a head with a sphere, M/EEG topographic maps on the measuring sensors can be seen as spherical signals. In this part of the thesis, we have firstly studied a rank-1 spatio-temporal convolutional dictionary learning with sparse codes penalized by an L_0 norm. In addition to the quantitative analysis on the synthetic data, the model was evaluated in the subject wise extraction of the temporal and spatial patterns on the somatosensory

and HCP motor-task MEG datasets. In our second contribution, we have proposed a shallow rank-1 CNN model for the classification of multivariate M/EEG signals. In addition to rank-1 constraint, the model performs regularization by representing spatial and temporal components of the trainable parameters in Fourier domain, in SH and discrete cosine bases, respectively. The model has been evaluated on the problem of mental workload classification from EEG signals containing a small number of subjects and on the motor-task classification from the HCP MEG dataset.

Perspectives

The work presented in this thesis raised multiple questions and ideas which will be investigated in our future work.

Concretely, in the context of rotation equivariant CNNs, although the non-linearities of a quadratic nature realized in Fourier domain showed potential, their main drawback is that they are not bounded. However, being able to defined x^2 functions in the Fourier domain, leads us to an idea to approximate via the Taylor series more traditional DL non-linearities such as sigmoid $\frac{1}{1+e^x}$ and hyperbolic tangent $\frac{e^x - e^{-x}}{e^x + e^{-x}}$. Furthermore, as the real ground truth of biophysically inspired micro-structure parameters does not exist, it would be of interest to design models of high generalization power which can be trained on synthetic data.

Our future work related to the spatio-temporal dictionary learning, might focus in conducting of more quantitative studies on real data such as the analysis of epileptic seizures. In addition, the extraction of the spatio-temporal patterns associated to individual brain sources, opens the door to the analysis of dynamic brain networks. Finally, we will investigate potential of our shallow rank-1 CNN model in the subject specific BCI scenarios.

Publications

Conference and workshop papers:

- Sara Sedlar, Théodore Papadopoulo, Rachid Deriche, Samuel Deslauriers-Gauthier. Diffusion MRI fiber orientation distribution function estimation using voxel-wise spherical U-net. International MICCAI Workshop 2020 - Computational Diffusion MRI, Oct 2020, Lima, Peru
- Sara Sedlar, Abib Alimi, Théodore Papadopoulo, Rachid Deriche, Samuel Deslauriers-Gauthier. A spherical convolutional neural network for white matter structure imaging via dMRI. MICCAI 2021 - 24th International Conference on Medical Image Computing and Computer Assisted Intervention, Sep 2021, Strasbourg / Virtual

Poster communications:

- *Sara Sedlar, Johann Benerradi, Côme Le Breton, Rachid Deriche, Théodore Papadopoulo et al.* Rank-1 CNN for mental workload classification from EEG. Neuroergonomics conference, Sep 2021, Munich (virtual event), Germany
- *Sara Sedlar, Samuel Deslauriers-Gauthier, Rachid Deriche, Théodore Papadopoulo.* Shallow convolutional neural network with rank-1 Fourier domain weights for brain signal classification. Proceedings of SophIA 2022, Nov 2022, Sophia Antipolis, France

In preparation for journal:

- *Sara Sedlar, Théodore Papadopoulo, Rachid Deriche, Samuel Deslauriers-Gauthier.* A spherical convolutional neural network for white matter structure imaging via dMRI.
- *Sara Sedlar, Rachid Deriche, Samuel Deslauriers-Gauthier, Théodore Papadopoulo.* Shallow convolutional neural network with rank-1 Fourier domain weights for brain signal classification.

Code contributions

- Spherical U-net for dMRI fiber orientation distribution function estimation (Chapter 4)
https://gitlab.inria.fr/ssedlar/spherical_unet
- Fourier domain spherical CNN for dMRI local analysis (Chapter 5)
https://gitlab.inria.fr/ssedlar/fourier_s2cnn
- Rank-1 M/EEG waveform and spatial pattern learning with L_0 constraint (Chapter 7)
https://gitlab.inria.fr/ssedlar/st_cdl_l0
- Shallow CNN for M/EEG classification (Chapter 8)
https://gitlab.inria.fr/ssedlar/shallow_cnn_meeg

Appendices

S^2 and $SO(3)$ signal related derivations appendix

Spherical harmonics

Definition of the complex spherical harmonics

The complex SH basis element $Y_l^m : S^2 \rightarrow \mathbb{C}$ is defined as

$$Y_l^m(\mathbf{r}) = Y_l^m(\theta, \phi) = \sqrt{\frac{(2l+1)(l-m)!}{2\pi(l+m)!}} P_l^m(\cos\theta) e^{jm\phi} \quad (\text{A.1})$$

where $P_l^m : [-1, 1] \rightarrow \mathbb{R}$ is associated Legendre polynomial of degree l and order m , defined in closed form as

$$P_l^m = (-1)^m 2^l (1 - (\cos\theta)^2)^{\frac{m}{2}} \sum_{k=m}^l \frac{k!}{(k-m)!} (\cos\theta)^{k-m} \binom{l}{k} \binom{\frac{l+k-1}{2}}{l}. \quad (\text{A.2})$$

Definition of the real spherical harmonics

The real SH [Homeier & Steinborn 1996] basis elements can be defined as

$$Y_{lm} = \begin{cases} \sqrt{2}(-1)^m \text{Im}[Y_l^{|m|}] & \text{if } m < 0 \\ Y_l^0 & \text{if } m = 0. \\ \sqrt{2}(-1)^m \text{Re}[Y_l^m] & \text{if } m > 0 \end{cases} \quad (\text{A.3})$$

If the complex SH basis elements of degree l are placed into columns of a matrix $Y_l^{\mathbb{C}}$ in the order $\{-l, -(l-1), \dots, -1, 0, 1, \dots, (l-1), l\}$, then the real SH basis elements of degree l can be obtained as

$$[Y_l^{\mathbb{R}}]^T = U_l [Y_l^{\mathbb{C}}]^T \quad (\text{A.4})$$

where $U_l \in \mathbb{C}^{(2l+1) \times (2l+1)}$ is unitary matrix defined as in [Homeier & Steinborn 1996]

$$U_l = \frac{1}{\sqrt{2}} \begin{bmatrix} j & 0 & \dots & 0 & \dots & 0 & (-1)^{-l+1}j \\ 0 & j & \dots & 0 & \dots & (-1)^{-l}j & 0 \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots & \vdots \\ \vdots & \vdots & \dots & \sqrt{2} & \dots & \vdots & \vdots \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots & \vdots \\ 0 & 1 & \dots & 0 & \dots & (-1)^{l-1} & 0 \\ 1 & 0 & \dots & 0 & \dots & 0 & (-1)^l \end{bmatrix}. \tag{A.5}$$

Rotation of S^2 functions

The complex Wigner-D matrices

The complex Wigner-D matrix is defined as

$$D_l^{mn}(R(\phi, \theta, \psi)) = d_l^{mn}(\theta)e^{-jm\phi}e^{-jn\psi} \quad (\text{A.6})$$

where d_l^{mn} is small Wigner-d matrix defined as

$$d_l^{mn}(\theta) = [(l+m)!(l-m)!(l+n)!(l-n)!]^{\frac{1}{2}} \sum_{s=s_{min}}^{s_{max}} \left[\frac{(-1)^{m-n+s} \left(\cos \frac{\theta}{2}\right)^{2l+n-m-2s} \left(\sin \frac{\theta}{2}\right)^{m-n+2s}}{(l+n-s)!s!(m-n+s)!(l-m-s)!} \right] \quad (\text{A.7})$$

where $s_{min} = \max(0, n-m)$ and $s_{max} = \min(l+n, l-m)$. We refer to l as Wigner-D matrix or RH degree and to m and n as to their orders. $R(\phi, \theta, \psi) \in SO(3)$ is rotation matrix with $\phi, \psi \in [0, 2\pi)$ and $\theta \in [0, \pi]$.

Rotation of the complex S^2 functions

Rotation of an L^2 signal $s : S^2 \rightarrow \mathbb{C}$ of bandwidth B by angle $R = R(\phi, \theta, \psi) \in SO(3)$, such that $g(\mathbf{r}) = Rs(\mathbf{r})$ can be written as (*cite)

$$\begin{aligned} s(R^{-1}\mathbf{r}) &= \sum_{l=0}^B \sum_{m=-l}^{m=l} \hat{s}_l^m Y_l^m(R^{-1}\mathbf{r}) = \sum_{l=0}^B \sum_{m=-l}^{m=l} \hat{s}_l^m \sum_{k=-l}^{k=l} D_l^{km}(R) Y_l^k(\mathbf{r}) \\ &= \sum_{l=0}^B \sum_{k=-l}^{k=l} \left(\sum_{m=-l}^{m=l} D_l^{km}(R) \hat{s}_l^m \right) Y_l^k(\mathbf{r}) = \sum_{l=0}^B \sum_{k=-l}^{k=l} [D_l(R) \hat{\mathbf{s}}_l]^k Y_l^k(\mathbf{r}) \quad (\text{A.8}) \\ &= \sum_{l=0}^B \sum_{k=-l}^{k=l} g_l^k Y_l^k(\mathbf{r}) = g(\mathbf{r}) \end{aligned}$$

where Y_l^m is the complex SH basis element of degree l and order m . D_l^{mn} is the complex Wigner-D matrix (RH basis element) of degree l and orders m and n .

The real Wigner-D matrices

The real RH basis elements (Wigner-D matrices) can be expressed as

$$D_{000}^{\mathbb{R}} = D_{00}^0 \quad (\text{A.9})$$

$$D_{lm0}^{\mathbb{R}} = \begin{cases} \sqrt{2}(-1)^m \text{Im}[D_l^{|m|0}], & m < 0 \\ \sqrt{2}(-1)^m \text{Re}[D_l^{m0}], & m > 0 \end{cases} \quad \text{and} \quad D_{l0n}^{\mathbb{R}} = \begin{cases} -\sqrt{2}(-1)^n \text{Im}[D_l^{0|n|}], & n < 0 \\ \sqrt{2}(-1)^n \text{Re}[D_l^{0n}], & n > 0 \end{cases} \quad (\text{A.10})$$

$$D_{lmn}^{\mathbb{R}} = \begin{cases} (-1)^{m+n} \operatorname{Re}[D_l^{mn}] + (-1)^m \operatorname{Re}[D_l^{m-n}], & m > 0, n > 0 \\ (-1)^{m+n} \operatorname{Im}[D_l^{m|n|}] + (-1)^m \operatorname{Im}[D_l^{mn}], & m > 0, n < 0 \\ (-1)^{m+n} \operatorname{Im}[D_l^{|m|n}] + (-1)^m \operatorname{Im}[D_l^{|m|-n}], & m < 0, n > 0 \\ (-1)^{m+n} \operatorname{Re}[D_l^{|m||n|}] - (-1)^m \operatorname{Re}[D_l^{|m|n}], & m < 0, n < 0 \end{cases}. \quad (\text{A.11})$$

As noted in [Homeier & Steinborn 1996], a consequence of unitarity of the matrix U_l from Eq. A.4 is identity $Y_l^{\mathbb{R}T}(\theta_1, \phi_1) Y_l^{\mathbb{R}}(\theta_2, \phi_2) = Y_l^T(\theta_1, \phi_1) Y_l^*(\theta_2, \phi_2)$. By defining

$$Y_l^{\mathbb{R}}(\theta_1, \phi_1) = Y_l^{\mathbb{R}}(\theta, \phi) \quad \text{and} \quad Y_l(\theta_1, \phi_1) = Y_l(\theta, \phi) \quad (\text{A.12})$$

and

$$Y_l^{\mathbb{R}}(\theta_2, \phi_2) = D_l^{\mathbb{R}}(R) Y_l^{\mathbb{R}}(\theta, \phi) \quad \text{and} \quad Y_l(\theta_2, \phi_2) = D_l(R) Y_l(\theta, \phi) \quad (\text{A.13})$$

we obtain real Wigner-D matrix $D_l^{\mathbb{R}}(R)$ as follows

$$\begin{aligned} Y_l^{\mathbb{R}T}(\theta, \phi) D_l^{\mathbb{R}}(R) Y_l^{\mathbb{R}}(\theta, \phi) \\ &= (U_l Y_l(\theta, \phi))^T D_l^{\mathbb{R}}(R) U_l Y_l(\theta, \phi) = Y_l^T(\theta, \phi) U_l^T D_l^{\mathbb{R}}(R) U_l Y_l(\theta, \phi) \\ &= Y_l^T(\theta, \phi) U_l^T D_l^{\mathbb{R}}(R) U_l^* Y_l^*(\theta, \phi) = Y_l^T(\theta, \phi) D_l^*(R) Y_l^*(\theta, \phi) \end{aligned} \quad (\text{A.14})$$

and

$$U_l^T D_l^{\mathbb{R}}(R) U_l^* = D_l^*(R) \quad \text{and} \quad D_l^{\mathbb{R}}(R) = U_l^* D_l^*(R) U_l^T = U_l D_l^*(R) U_l^H \quad (\text{A.15})$$

where we used the property that $Y_l^{\mathbb{R}}(\theta, \phi) = Y_l^{\mathbb{R}*}(\theta, \phi)$ and $D_l^{\mathbb{R}}(R) = D_l^{\mathbb{R}*}(R)$ in equations A.14 and A.15.

Rotation of the real S^2 functions

In analogy to the rotation of the complex S^2 functions from Eq. A.8 and using the real Wigner-D matrices defined in Eq. A.15, we define rotation of the real S^2 functions. Rotation of an \mathbb{L}^2 signal $s : S^2 \rightarrow \mathbb{R}$ of bandwidth B by angle $R = R(\phi, \theta, \psi) \in SO(3)$, such that $g(\mathbf{r}) = Rs(\mathbf{r})$ can be written as

$$\begin{aligned} s(R^{-1}\mathbf{r}) &= \sum_{l=0}^B \sum_{m=-l}^{m=l} \hat{s}_{lm} Y_{lm}(R^{-1}\mathbf{r}) = \sum_{l=0}^B \sum_{m=-l}^{m=l} \hat{s}_{lm} \sum_{k=-l}^{k=l} D_{lkm}(R) Y_{lk}(\mathbf{r}) \\ &= \sum_{l=0}^B \sum_{k=-l}^{k=l} \left(\sum_{m=-l}^{m=l} D_{lkm}(R) \hat{s}_{lm} \right) Y_{lk}(\mathbf{r}) = \sum_{l=0}^B \sum_{k=-l}^{k=l} [D_l(R) \hat{\mathbf{s}}_l]_k Y_{lk}(\mathbf{r}) \\ &= \sum_{l=0}^B \sum_{k=-l}^{k=l} g_{lk} Y_{lk}(\mathbf{r}) = g(\mathbf{r}) \end{aligned} \quad (\text{A.16})$$

where Y_{lm} is the real SH basis element of degree l and order m . D_{lmn} is the real Wigner-D matrix (RH basis element) of degree l and orders m and n .

Convolutions of S^2 , zonal and $SO(3)$ functions

As we are dealing with real signals and we have defined real SH and RH basis, we provide derivations of convolutions between real functions only.

Convolution of an S^2 and a zonal function

Convolution between a spherical and a zonal function results in a function whose domain is S^2 . Given a signal $f : S^2 \rightarrow \mathbb{R}$ and a zonal signal $g : S^2 \rightarrow \mathbb{R}$ s.t. $g(\theta, \phi) = g(\theta, 0) \forall \phi \in [0, 2\pi)$ and $\forall \theta \in [0, \pi)$, of bandwidths B , convolution is given by [Kostelec & Rockmore 2008]

$$\begin{aligned}
[f * g](\mathbf{r}) &= [f * g](\theta, \phi) = \int_{S^2} f(\mathbf{r}') g(R^{-1}(\phi, \theta, 0)\mathbf{r}') d\mathbf{r}' \\
&= \int_{S^2} \sum_{l'=0}^B \sum_{m=-l}^l \hat{f}_{lm} Y_{lm}(\mathbf{r}') \sum_{l'=0}^B \hat{g}_{l'} Y_{l'0}(R^{-1}(\phi, \theta, 0)\mathbf{r}') d\mathbf{r}' \\
&= \int_{S^2} \sum_{l=0}^B \sum_{m=-l}^l \hat{f}_{lm} Y_{lm}(\mathbf{r}') \sum_{l'=0}^B \hat{g}_{l'} \sum_{k=-l'}^{l'} D_{l'k0}(\phi, \theta, 0) Y_{l'k}(\mathbf{r}') d\mathbf{r}' \\
&= \sum_{l=0}^B \sum_{m=-l}^l \hat{f}_{lm} \sum_{l'=0}^B \hat{g}_{l'} \sum_{k=-l'}^{l'} D_{l'k0}(R(0, \theta, \phi)) \int_{S^2} Y_{lm}(\mathbf{r}') Y_{l'k}(\mathbf{r}') d\mathbf{r}' \quad (\text{A.17}) \\
&= \sum_{l=0}^B \sum_{m=-l}^l \hat{f}_{lm}^m \sum_{l'=0}^B \hat{g}_{l'} \sum_{k=-l'}^{l'} D_{l'k0}(R(0, \theta, \phi)) \delta_{ll'} \delta_{mk} \\
&= \sum_{l=0}^B \sum_{m=-l}^l D_{lm0}(R(0, \theta, \phi)) \hat{f}_{lm} \hat{g}_l \\
&= \sum_{l=0}^B \sqrt{\frac{4\pi}{2l+1}} \hat{g}_l \sum_{m=-l}^l Y_{lm}(\theta, \phi) \hat{f}_{lm}
\end{aligned}$$

where \hat{f}_{lm} is the real SH coefficient of degree l and order m of the function f and \hat{g}_l is ZH coefficient of degree l of the function g .

Convolution of S^2 functions

Given two \mathbb{L}^2 signals $f, g : S^2 \rightarrow \mathbb{R}$ of bandwidth B , convolution between them is defined as [Driscoll & Healy 1994]

$$\begin{aligned}
[f * g](R) &= \int_{S^2} f(\mathbf{r})g(R^{-1}\mathbf{r})d\mathbf{r} \\
&= \int_{S^2} \sum_{l=0}^B \sum_{m=-l}^l \hat{f}_{lm} Y_{lm}(\mathbf{r}) \sum_{l'=0}^B \sum_{n=-l'}^{l'} \hat{g}_{l'n} Y_{l'n}(R^{-1}\mathbf{r})d\mathbf{r} \\
&= \int_{S^2} \sum_{l=0}^B \sum_{m=-l}^l \hat{f}_{lm} Y_{lm}(\mathbf{r}) \sum_{l'=0}^B \sum_{n=-l'}^{l'} \hat{g}_{l'n} \sum_{k=-l'}^{l'} D_{l'kn}(R) Y_{l'k}(\mathbf{r})d\mathbf{r} \\
&= \sum_{l=0}^B \sum_{m=-l}^l \hat{f}_{lm} \sum_{l'=0}^B \sum_{n=-l'}^{l'} \hat{g}_{l'n} \sum_{k=-l'}^{l'} D_{l'kn}(R) \int_{S^2} Y_{lm}(\mathbf{r}) Y_{l'k}(\mathbf{r})d\mathbf{r} \\
&= \sum_{l=0}^B \sum_{m=-l}^l \hat{f}_{lm} \sum_{l'=0}^B \sum_{n=-l'}^{l'} \hat{g}_{l'n} \sum_{k=-l'}^{l'} D_{l'kn}(R) \delta_{l'l} \delta_{mn} \\
&= \sum_{l=0}^B \sum_{m=-l}^l \sum_{n=-l}^l D_{lmn}(R) \hat{f}_{lm} \hat{g}_{ln}
\end{aligned} \tag{A.18}$$

where $R = R(\phi, \theta, \psi) \in SO(3)$. $\hat{f}_{lm}, \hat{g}_{ln}$ are the real SH coefficients of degree l and orders m and n of the functions f and g and $D_{lmn} : SO(3) \rightarrow \mathbb{R}$ is an element of the real RH basis (Wigner-D matrix) of degree l and orders m and n .

Convolution between $SO(3)$ signals

Convolution between two $SO(3)$ signals results in a signal whose domain is also $SO(3)$. Given two \mathbb{L}^2 functions function $f, g : SO(3) \rightarrow \mathbb{R}$ of bandwidth B convo-

lution between them is defined as [Kostelec & Rockmore 2008]

$$\begin{aligned}
[f * g](Q) &= \int_{SO(3)} f(R)g(Q^{-1}R)dR = \\
&\int_{SO(3)} \sum_{l=0}^B \sum_{m=-l}^l \sum_{n=-l}^l \hat{F}_{lmn} D_{lmn}(R) \sum_{l'=0}^B \sum_{m'=-l'}^{l'} \sum_{n'=-l'}^{l'} \hat{G}_{l'm'n'} D_{l'm'n'}(Q^{-1}R)dR = \\
&\int_{SO(3)} \sum_{l=0}^B \sum_{m=-l}^l \sum_{n=-l}^l \hat{F}_{lmn} D_{lmn}(R) \sum_{l'=0}^B \sum_{m'=-l'}^{l'} \sum_{n'=-l'}^{l'} \hat{G}_{l'm'n'} \sum_{k=-l'}^{l'} D_{l'km'}(Q) D_{l'kn'}(R)dR = \\
&\sum_{l=0}^B \sum_{m=-l}^l \sum_{n=-l}^l \hat{F}_{lmn} \sum_{l'=0}^B \sum_{m'=-l'}^{l'} \sum_{n'=-l'}^{l'} \hat{G}_{l'm'n'} \sum_{k=-l'}^{l'} D_{l'km'}(Q) \int_{SO(3)} D_{lmn}(R) D_{l'kn'}(R)dR = \\
&\sum_{l=0}^B \sum_{m=-l}^l \sum_{n=-l}^l \hat{F}_{lmn} \sum_{l'=0}^B \sum_{m'=-l'}^{l'} \sum_{n'=-l'}^{l'} \hat{G}_{l'm'n'} \sum_{k=-l'}^{l'} D_{l'km'}(Q) \frac{8\pi^2}{2l+1} \delta_{ll'} \delta_{mk} \delta_{nn'} = \\
&\sum_{l=0}^B \sum_{m=-l}^l \sum_{n=-l}^l \hat{F}_{lmn} \sum_{l'=0}^B \sum_{m'=-l'}^{l'} \sum_{n'=-l'}^{l'} \hat{G}_{l'm'n'} D_{l'mm'}(Q) \frac{8\pi^2}{2l+1} \delta_{ll'} \delta_{nn'} = \\
&\sum_{l=0}^B \sum_{m=-l}^l \sum_{n=-l}^l \hat{F}_{lmn} \sum_{m'=-l}^l \hat{G}_{lm'n} D_{lmm'}(Q) \frac{8\pi^2}{2l+1} = \\
&\sum_{l=0}^B \frac{8\pi^2}{2l+1} \sum_{m=-l}^l \sum_{m'=-l}^l D_{lmm'}(Q) \sum_{n=-l}^l \hat{F}_{lmn} \hat{G}_{lm'n} = \\
&\sum_{l=0}^B \frac{8\pi^2}{2l+1} \sum_{m=-l}^l \sum_{n=-l}^l D_{lmn}(Q) \sum_{k=-l}^l \hat{F}_{lmk} \hat{G}_{lnk}
\end{aligned}
\tag{A.19}$$

Quadratic functions

Product of S^2 signals

Multiplication of two spherical signals in S^2 domain results in a signal whose domain is also S^2 . Given two \mathbb{L}^2 functions function $f, g : S^2 \rightarrow \mathbb{C}$ of bandwidths B_f and B_g , their product is defined as [Kondor *et al.* 2018]

$$\begin{aligned}
h &= [f \times g] = \sum_{l'=0}^{B_f} \sum_{m'=-l'}^{l'} \hat{f}_{l'}^{m'} Y_{l'}^{m'} \sum_{l''=0}^{B_g} \sum_{m''=-l''}^{l''} \hat{g}_{l''}^{m''} Y_{l''}^{m''} = \\
&= \sum_{l'=0}^{B_f} \sum_{l''=0}^{B_g} \sum_{l=|l'-l''|}^{l'+l''} \sum_{m'=-l'}^{l'} \sum_{m''=-l''}^{l''} \hat{f}_{l'}^{m'} \hat{g}_{l''}^{m''} \sqrt{\frac{(2l'+1)(2l''+1)}{4\pi(2l+1)}} \mathcal{C}_{l',m',l'',m''}^{l,m'+m''} \mathcal{C}_{l',0,l'',0}^{l,0} Y_l^{m'+m''} = \\
&= \sum_{l=0}^{B_f+B_g} \sum_{m=-l}^l \sum_{l'=0}^{B_f} \sum_{l''=0}^{B_g} \sum_{m'=-l'}^{l'} \sum_{m''=-l''}^{l''} \hat{f}_{l'}^{m'} \hat{g}_{l''}^{m''} \sqrt{\frac{(2l'+1)(2l''+1)}{4\pi(2l+1)}} \mathcal{C}_{l',m',l'',m''}^{l,m} \mathcal{C}_{l',0,l'',0}^{l,0} Y_l^m = \\
&= \sum_{l=0}^{B_f+B_g} \sum_{m=-l}^l \hat{h}_l^m Y_l^m
\end{aligned} \tag{A.20}$$

where $\mathcal{C}_{l',q',l'',q''}^{l,q} \in \mathbb{R}$ is Clebsch-Gordan coefficient associated to complex SH basis elements, such that $\mathcal{C}_{l',q',l'',q''}^{l,q} \neq 0$ only when $q' + q'' = q$. If the Clebsch-Gordan coefficients are stored in a sparse matrix $\mathcal{C}_{l',l''}^{l,q} \in \mathbb{R}^{(2l'+1)(2l''+1) \times (2l+1)}$, Eq. A.20 can be written in a more elegant way as

$$\hat{h}_l = \sum_{l',l''} \sqrt{\frac{(2l'+1)(2l''+1)}{4\pi(2l+1)}} \mathcal{C}_{l',0,l'',0}^{l,0} \mathcal{C}_{l',l''}^{l,0T} [\hat{f}_{l'} \otimes \hat{g}_{l''}] \quad \text{s.t.} \quad |l' - l''| \leq l \leq l' + l'' \tag{A.21}$$

where $\hat{h}_k, \hat{f}_k, \hat{g}_k \in \mathbb{C}^{2k+1}$ are the vector with complex SH coefficients of degree k . In analogy, given two \mathbb{L}^2 functions function $f, g : S^2 \rightarrow \mathbb{R}$ of bandwidths B_f and B_g , their product is defined as

$$\begin{aligned}
h &= [f \times g] = \sum_{l'=0}^{B_f} \sum_{m'=-l'}^{l'} \hat{f}_{m',l'} Y_{m',l'} \sum_{l''=0}^{B_g} \sum_{m''=-l''}^{l''} \hat{g}_{m'',l''} Y_{m'',l''} = \\
&= \sum_{l'=0}^{B_f} \sum_{l''=0}^{B_g} \sum_{l=|l'-l''|}^{l'+l''} \sum_{m'=-l'}^{l'} \sum_{m''=-l''}^{l''} \hat{f}_{m',l'} \hat{g}_{m'',l''} \sqrt{\frac{(2l'+1)(2l''+1)}{4\pi(2l+1)}} \mathcal{C}_{l',m',l'',m''}^{l,m'+m''} \mathcal{C}_{l',0,l'',0}^{l,0} Y_{m'+m'',l} = \\
&= \sum_{l=0}^{B_f+B_g} \sum_{m=-l}^l \sum_{l'=0}^{B_f} \sum_{l''=0}^{B_g} \sum_{m'=-l'}^{l'} \sum_{m''=-l''}^{l''} \hat{f}_{m',l'} \hat{g}_{m'',l''} \sqrt{\frac{(2l'+1)(2l''+1)}{4\pi(2l+1)}} \mathcal{C}_{l',m',l'',m''}^{l,m} \mathcal{C}_{l',0,l'',0}^{l,0} Y_{m,l} = \\
&= \sum_{l=0}^{B_f+B_g} \sum_{m=-l}^l \hat{h}_{m,l} Y_{m,l}
\end{aligned} \tag{A.22}$$

and $C_{l',q',l'',q''}^{l,q} \in \mathbb{R}$ is Clebsch-Gordan coefficient associated to real SH basis elements. If the Clebsch-Gordan coefficients are stored in a sparse matrix $C_{l',l''}^l \in \mathbb{R}^{(2l'+1)(2l''+1) \times (2l+1)}$, Eq. A.22 can be written in matrix-vector notation as

$$\hat{\mathbf{h}}_l = \sum_{l',l''} \sqrt{\frac{(2l'+1)(2l''+1)}{4\pi(2l+1)}} C_{l',0,l'',0}^{l,0} C_{l',l''}^{l,q}{}^T [\hat{\mathbf{f}}_{l'} \otimes \hat{\mathbf{g}}_{l'']} \quad \text{s.t.} \quad |l' - l''| \leq l \leq l' + l'' \quad (\text{A.23})$$

where $\hat{\mathbf{h}}_k, \hat{\mathbf{f}}_k, \hat{\mathbf{g}}_k \in \mathbb{R}^{2k+1}$ are the vector with real SH coefficients of degree k .

Denoting with $C_{l',q',l'',q''}^{l,q} \in \mathbb{R}$ and with $C_{l',q',l'',q''}^{l,q} \in \mathbb{R}$ Clebsch-Gordan coefficient associated to *complex* and *real* SH basis elements, respectively, the real Clebsch-Gordan coefficients can be derived as

$$Y_{m',l'} Y_{m'',l''} = 2(-1)^{m'+m} \begin{cases} \text{Im}[Y_{l'}^{m'}] \text{Im}[Y_{l''}^{m''}] & \text{if } m' < 0, \quad m'' < 0 \\ \text{Im}[Y_{l'}^{m'}] \text{Re}[Y_{l''}^{m''}] & \text{if } m' < 0, \quad m'' > 0 \\ \text{Re}[Y_{l'}^{m'}] \text{Im}[Y_{l''}^{m''}] & \text{if } m' > 0, \quad m'' < 0 \\ \text{Re}[Y_{l'}^{m'}] \text{Re}[Y_{l''}^{m''}] & \text{if } m' > 0, \quad m'' > 0 \\ \frac{1}{\sqrt{2}} \text{Im}[Y_{l'}^{m'}] Y_{l''}^{m''} & \text{if } m' < 0, \quad m'' = 0 \\ \frac{1}{\sqrt{2}} \text{Re}[Y_{l'}^{m'}] Y_{l''}^{m''} & \text{if } m' > 0, \quad m'' = 0 \\ \frac{1}{\sqrt{2}} Y_{l'}^{m'} \text{Im}[Y_{l''}^{m''}] & \text{if } m' = 0, \quad m'' < 0 \\ \frac{1}{\sqrt{2}} Y_{l'}^{m'} \text{Re}[Y_{l''}^{m''}] & \text{if } m' = 0, \quad m'' > 0 \\ \frac{1}{2} Y_{l'}^{m'} Y_{l''}^{m''} & \text{if } m' = 0, \quad m'' = 0 \end{cases} \quad (\text{A.24})$$

using that $\text{Im}[Y_l^m] = \frac{Y_l^m - Y_l^{m*}}{2}$, $\text{Re}[Y_l^m] = \frac{Y_l^m(\mathbf{r}) + Y_l^{m*}}{2}$ and $Y_l^{m*} = (-1)^m Y_l^{-m}$, it

can be obtained that

$$C_{l',m',l'',m''}^{l,m} = c \begin{cases} C_{l',m',l'',m''}^{l,m} - (-1)^{m'} C_{l',-m',l'',m''}^{l,m} - (-1)^{m''} C_{l',m',l'',-m''}^{l,m} + (-1)^{m'+m''} C_{l',-m',l'',-m''}^{l,m} & \text{if } m' < 0, m'' < 0 \\ C_{l',m',l'',m''}^{l,m} - (-1)^{m'} C_{l',-m',l'',m''}^{l,m} + (-1)^{m''} C_{l',m',l'',-m''}^{l,m} - (-1)^{m'+m''} C_{l',-m',l'',-m''}^{l,m} & \text{if } m' < 0, m'' > 0 \\ C_{l',m',l'',m''}^{l,m} + (-1)^{m'} C_{l',-m',l'',m''}^{l,m} - (-1)^{m''} C_{l',m',l'',-m''}^{l,m} - (-1)^{m'+m''} C_{l',-m',l'',-m''}^{l,m} & \text{if } m' < 0, m'' > 0 \\ C_{l',m',l'',m''}^{l,m} + (-1)^{m'} C_{l',-m',l'',m''}^{l,m} + (-1)^{m''} C_{l',m',l'',-m''}^{l,m} + (-1)^{m'+m''} C_{l',-m',l'',-m''}^{l,m} & \text{if } m' > 0, m'' > 0 \\ \sqrt{2}(C_{l',m',l'',m''}^{l,m} - (-1)^{m'} C_{l',-m',l'',m''}^{l,m}) & \text{if } m' < 0, m'' = 0 \\ \sqrt{2}(C_{l',m',l'',m''}^{l,m} + (-1)^{m'} C_{l',-m',l'',m''}^{l,m}) & \text{if } m' > 0, m'' = 0 \\ \sqrt{2}(C_{l',m',l'',m''}^{l,m} - (-1)^{m'} C_{l',m',l'',-m''}^{l,m}) & \text{if } m' = 0, m'' < 0 \\ \sqrt{2}(C_{l',m',l'',m''}^{l,m} + (-1)^{m'} C_{l',m',l'',-m''}^{l,m}) & \text{if } m' = 0, m'' > 0 \\ 2C_{l',m',l'',m''}^{l,m} & \text{if } m' = 0, m'' = 0 \end{cases} \quad (\text{A.25})$$

where $c = \frac{1}{2}(-1)^{m'+m''}$.

Conversion between the sparse matrices $C_{l',l''}^l, C_{l',l''}^l \in \mathbb{R}^{(2l'+1)(2l''+1) \times (2l+1)}$ used in equations A.21 and A.23 can be derived from

$$\begin{aligned} U_l^H \hat{\mathbf{h}}_l &= \sum_{l',l''} \sqrt{\frac{(2l'+1)(2l''+1)}{4\pi(2l+1)}} C_{l',0,l'',0}^{l,0} C_{l',l''}^l{}^T [U_{l'}^H \hat{\mathbf{f}}_{l'} \otimes U_{l''}^H \hat{\mathbf{g}}_{l''}] \quad \text{s.t. } |l' - l''| \leq l \leq l' + l'' \\ &= \sqrt{\frac{(2l'+1)(2l''+1)}{4\pi(2l+1)}} C_{l',0,l'',0}^{l,0} C_{l',l''}^l{}^T [U_{l'}^H \otimes U_{l''}^H] [\hat{\mathbf{f}}_{l'} \otimes \hat{\mathbf{g}}_{l''}] \quad \text{s.t. } |l' - l''| \leq l \leq l' + l'' \end{aligned} \quad (\text{A.26})$$

thus

$$C_{l',l''}^l{}^T = \begin{cases} \text{Re} \left[U_l C_{l',l''}^l{}^T [U_{l'}^H \otimes U_{l''}^H] \right] & l_1 + l_2 + l \text{ is even} \\ \text{Im} \left[U_l C_{l',l''}^l{}^T [U_{l'}^H \otimes U_{l''}^H] \right] & l_1 + l_2 + l \text{ is odd} \end{cases} \quad (\text{A.27})$$

Product of $SO(3)$ signals

Multiplication of two $SO(3)$ signals in $SO(3)$ domain results in a signal whose domain is also $SO(3)$. Given two \mathbb{L}^2 functions function $f, g : SO(3) \rightarrow \mathbb{R}$ of bandwidths B_f and B_g , their product is defined as (*cite)

$$\begin{aligned}
[f \times g](R) &= \sum_{l'=0}^{B_f} \sum_{m'=-l'}^{l'} \sum_{n'=-l'}^{l'} \hat{F}_l^{m'n'} D_l^{m'n'}(R) \sum_{l''=0}^{B_g} \sum_{m''=-l''}^{l''} \sum_{n''=-l''}^{l''} \hat{G}_{l''}^{m''n''} D_{l''}^{m''n''}(R) = \\
&= \sum_{l'=0}^{B_f} \sum_{m'=-l'}^{l'} \sum_{n'=-l'}^{l'} \hat{F}_l^{m'n'} \sum_{l''=0}^{B_g} \sum_{m''=-l''}^{l''} \sum_{n''=-l''}^{l''} \hat{G}_{l''}^{m''n''} D_l^{m'n'}(R) D_{l''}^{m''n''}(R) = \\
&= \sum_{l'=0}^{B_f} \sum_{m'=-l'}^{l'} \sum_{n'=-l'}^{l'} \hat{F}_l^{m'n'} \sum_{l''=0}^{B_g} \sum_{m''=-l''}^{l''} \sum_{n''=-l''}^{l''} \hat{G}_{l''}^{m''n''} \sum_{l=|l'-l''|}^{l'+l''} C_{l',m',l'',m''}^{l,m'+m''} C_{l',n',l'',n''}^{l,n'+n''} D_l^{(m'+m'')(n'+n'')}(R) = \\
&= \sum_{l'=0}^{B_f} \sum_{l''=0}^{B_g} \sum_{l=|l'-l''|}^{l'+l''} \sum_{m'=-l'}^{l'} \sum_{n'=-l'}^{l'} \sum_{m''=-l''}^{l''} \sum_{n''=-l''}^{l''} \hat{F}_l^{m'n'} \hat{G}_{l''}^{m''n''} C_{l',m',l'',m''}^{l,m'+m''} C_{l',n',l'',n''}^{l,n'+n''} D_l^{(m'+m'')(n'+n'')}(R) = \\
&= \sum_{l=0}^{B_f+B_g} \sum_{m=-l}^l \sum_{n=-l}^l \sum_{l'=0}^{B_f} \sum_{l''=0}^{B_g} \sum_{m'=-l'}^{l'} \sum_{n'=-l'}^{l'} \sum_{m''=-l''}^{l''} \sum_{n''=-l''}^{l''} \hat{F}_l^{m'n'} \hat{G}_{l''}^{m''n''} C_{l',m',l'',m''}^{l,m'+m''} C_{l',n',l'',n''}^{l,n'+n''} D_l^{mn}(R) = \\
&= \sum_{l=0}^{B_f+B_g} \sum_{m=-l}^l \sum_{n=-l}^l \hat{H}_l^{mn} D_l^{mn}(R)
\end{aligned} \tag{A.28}$$

Microstructure estimation experiments appendix

MLP hyperparameter selection for microstructure parameter estimation

In this section we provide details related to the hyperparameter selection for the MLP model introduced by [Golkov *et al.* 2016]. We have evaluated models of two sizes and depths, namely *MLP1* composed of four layers of sizes $60 \times 256, 256 \times 256, 256 \times 256, 256 \times n_{out}$ and *MLP2* composed of seven layers of sizes $60 \times 256, 256 \times 192, 192 \times 128, 128 \times 64, 64 \times 32, 32 \times 16, 16 \times n_{out}$, where $n_{out} = 3$ for NODDI and $n_{out} = 2$ for SMT. Also, we have trained models with two different initial learning rates, 0.001 and 0.0001. The original method uses drop out rate of 0.1, thus we have evaluated the model *MLP1* with different drop out rates of 0.1, 0.05 and 0.0 for NODDI parameter estimation and found that the models without drop out (0.0) have much better performance regardless of the number of training subjects. Also, instead of stochastic gradient descent used in the original work, we have found that the Adam optimizer gives better performance. Illustrations of the validation losses for NODDI and SMT parameter estimation and corresponding number of trainable parameters, for the experiments with the number of training subjects 1, 3, 5 are provided in Figure B.1 and for 10, 15, 30 training subjects in Figure B.2.

For a comparison with another methods on the problem of NODDI parameter estimation, for the number of training subjects 1, 3, 5 and 10, we have selected *MLP1* with $lr = 0.001$ and for 15 and 30 subjects the same model with $lr = 0.0001$. For the SMT parameter estimation, for the number of training subjects 1 we have selected *MLP1* with $lr = 0.001$, for 3, 5, 10 subjects *MLP2* with $lr = 0.001$, while for 15 and 30 the same model with $lr = 0.0001$.

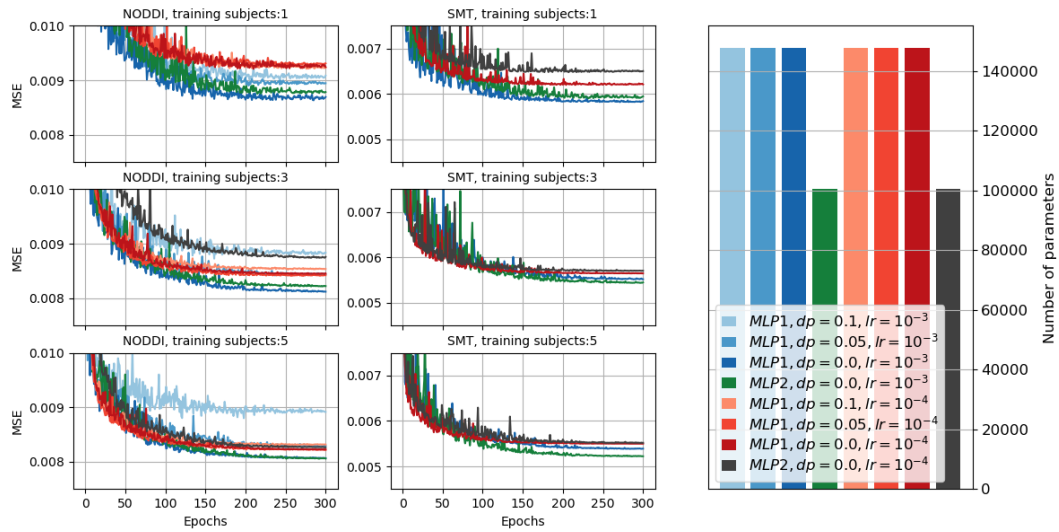


Figure B.1: Validation losses for NODDI (left) and SMT (middle) parameter estimation and corresponding number of trainable parameters (right) for the number of training subjects 1, 3, 5.

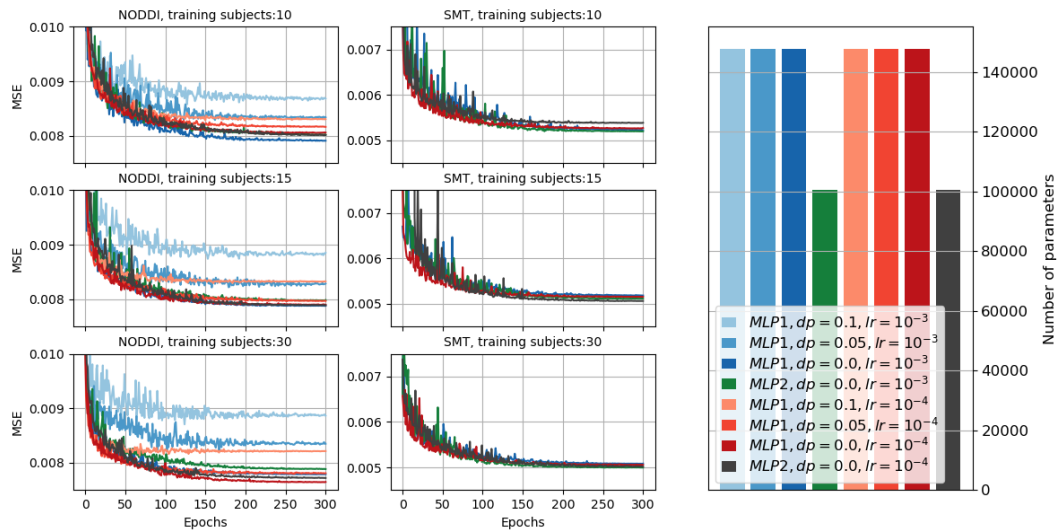


Figure B.2: Validation losses for NODDI (left) and SMT (middle) parameter estimation and corresponding number of trainable parameters (right) for the number of training subjects 10, 15, 30.

MEDN hyperparameter selection for microstructure parameter estimation

In this section we provide details related to the hyperparameter selection for the MEDN model introduced by [Ye 2017]. This model is strictly designed for NODDI parameter estimation. We have evaluated the models for different number of iterations 6, 8, 10 used in the approximation of iterative hard thresholding, as described in Chapter 3, and for two different initial learning rates, 0.001 and 0.0001. Illustrations of the validation losses for NODDI parameter estimation and corresponding number of trainable parameters, for the experiments with the number of training subjects 1, 3, 5 are provided in Figure B.3 and for 10, 15, 30 training subjects in Figure B.4. According to the validation curves, we have observed that the model sometimes experiences instabilities with higher learning rates, thus the update of trainable weights stops.

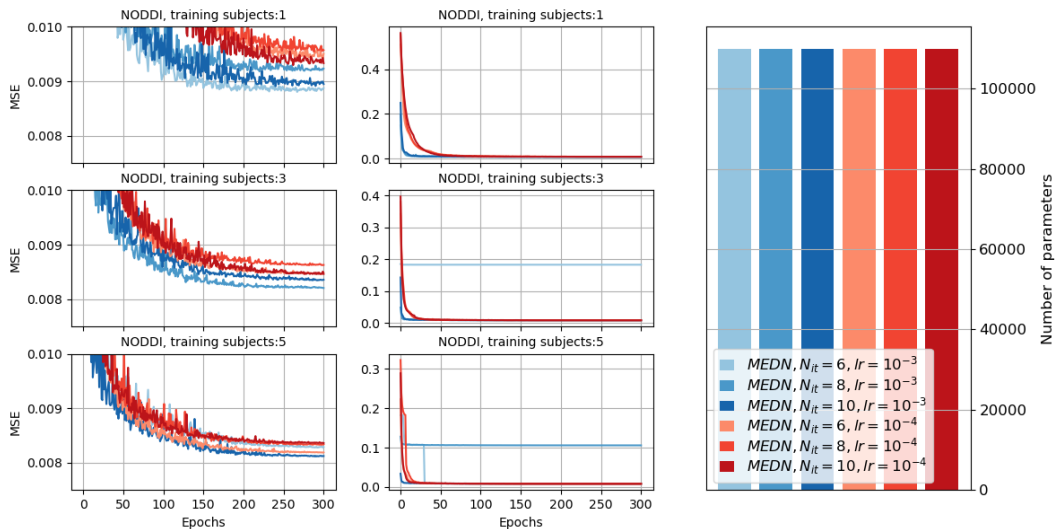


Figure B.3: Validation losses for NODDI parameter estimation, illustrated within a range $[0.0075, 0.01]$ (left), without range limit to illustrate instabilities (middle) and corresponding number of trainable parameters (right) for the number of training subjects 1, 3, 5.

For a comparison with another methods, for the number of training subjects 1, 3 and 5, we have selected models with $N_{it} = 6$, $N_{it} = 8$ and $N_{it} = 10$, respectively with $lr = 0.001$. For 10, 15 and 30 training subjects, we have selected model with $N_{it} = 10$, for 10 subjects with $lr = 0.001$ and for 15 and 30 with $lr = 0.0001$.

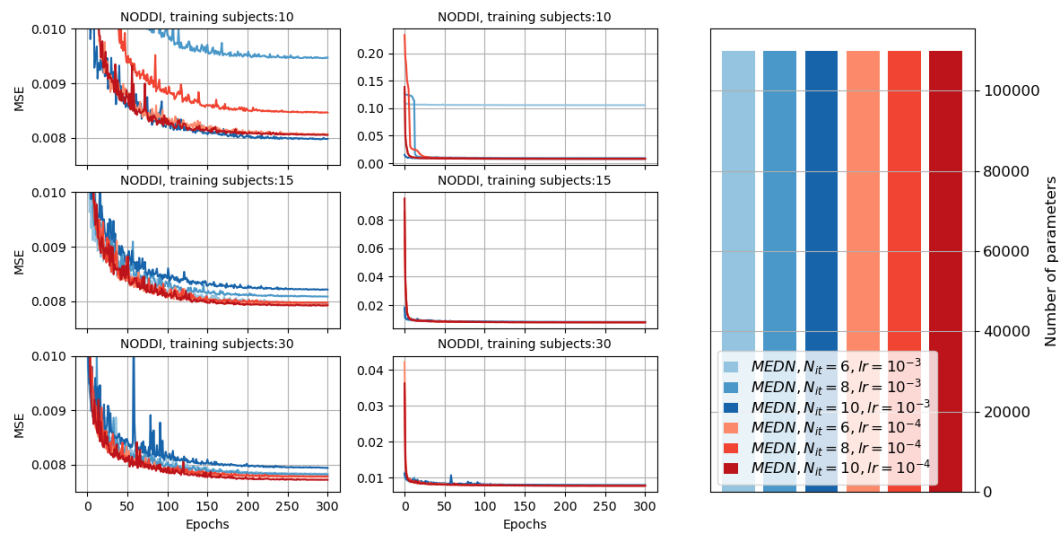


Figure B.4: Validation losses for NODDI parameter estimation, illustrated within a range $[0.0075, 0.01]$ (left), without rage limit to illustrate instabilities (middle) and corresponding number of trainable parameters (right) for the number of training subjects 10, 15, 30.

Hyperparameter selection for microstructure parameter estimation for our models

We have evaluated our models for different input bandwidths and for different depths. All models have the same denoising layer composed of two trainable matrices of size 60×60 and four fully connected layers with the number of output neurons 128, 128, 128, n_{out} at the end which take as input rotation invariant features and based on them perform parameter estimation. Model $Fourier_S^2_SO(3)_1$ contains three convolutional layers of input and output bandwidths (6, 4), (4, 2), (2, 0) with the input and output number of channels (2, 8), (8, 16), (16, 32). Model $Fourier_S^2_SO(3)_2$ contains three convolutional layers of input and output bandwidths (8, 4), (4, 2), (2, 0) with the input and output number of channels (2, 8), (8, 16), (16, 32). Model $Fourier_S^2_SO(3)_3$ contains four convolutional layers of input and output bandwidths (8, 6), (6, 4), (4, 2), (2, 0) with the input and output number of channels (2, 4), (4, 8), (8, 16), (16, 32). Model $Fourier_S^2_zonal_1$ contains three convolutional layers of input and output bandwidths (6, 4), (4, 2), (2, 0) with the input and output number of channels (2, 20), (20, 40), (40, 80). Model $Fourier_S^2_zonal_2$ contains four convolutional layers of input and output bandwidths (8, 6), (6, 4), (4, 2), (2, 0) with the input and output number of channels (2, 12), (12, 24), (24, 48), (48, 96). Illustrations of the validation losses for NODDI and SMT parameter estimation and corresponding number of trainable parameters, for the experiments with the number of training subjects 1, 3, 5 are provided in Figure B.5 and for 10, 15, 30 training subjects in Figure B.6.

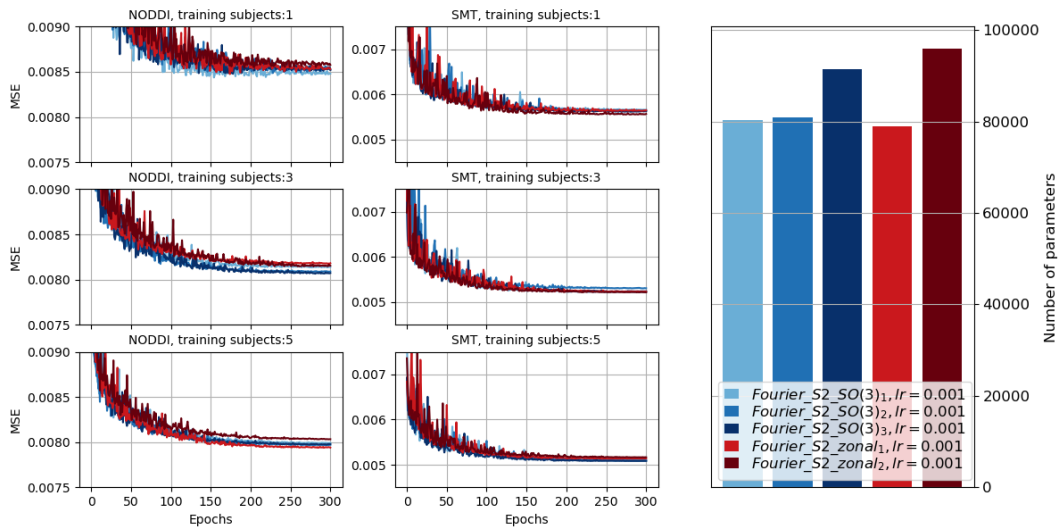


Figure B.5: Validation losses for NODDI (left) and SMT (middle) parameter estimation and corresponding number of trainable parameters (right) for the number of training subjects 1, 3, 5.

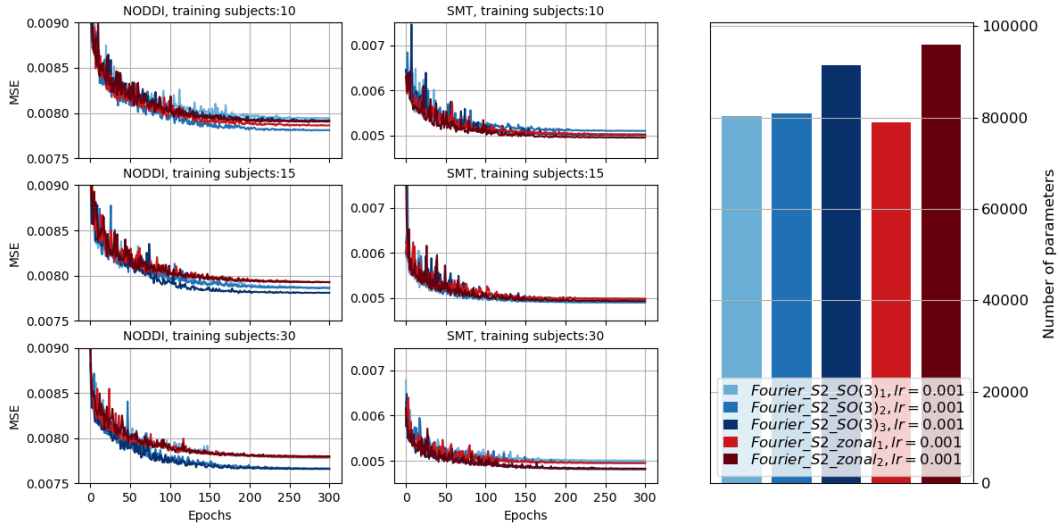


Figure B.6: Validation losses for NODDI (left) and SMT (middle) parameter estimation and corresponding number of trainable parameters (right) for the number of training subjects 10, 15, 30.

Since the differences between validation losses for different $Fourier_S^2_SO(3)$ and $Fourier_S^2_zonal$ are smaller, for all subjects and for both NODDI and SMT parameter estimation we have selected, $Fourier_S^2_SO(3)_3$ and $Fourier_S^2_zonal_1$.

MLP+ hyperparameter selection for microstructure parameter estimation

We have extended the model MLP [Golkov *et al.* 2016] to the version termed as MLP+ which as input take dMRI signals from a neighbourhood of size $3 \times 3 \times 3$. We have evaluated models of two sizes and depths, namely *MLP1+* composed of four layers of sizes $60 \times 27 \times 256, 256 \times 256, 256 \times 256, 256 \times n_{out}$ and *MLP2* composed of seven layers of sizes $60 \times 27 \times 256, 256 \times 192, 192 \times 128, 128 \times 64, 64 \times 32, 32 \times 16, 16 \times n_{out}$, where $n_{out} = 3$ for NODDI and $n_{out} = 2$ for SMT. The models are trained with three different initial learning rates 0.001, 0.0005 and 0.0001. Illustrations of the validation losses and corresponding number of trainable parameters, for the experiments with the number of training subjects 1, 3, 5 are provided in Figure B.7.

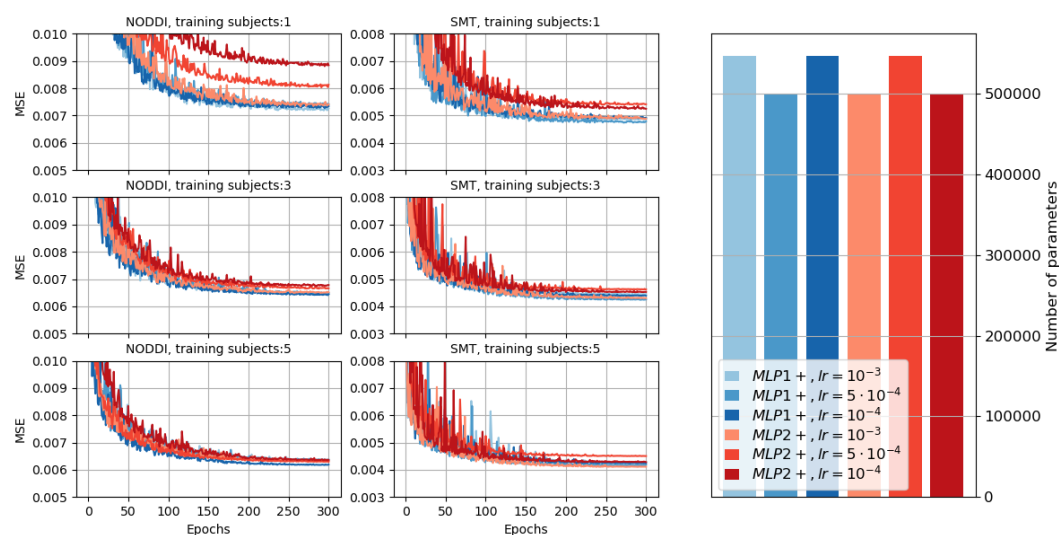


Figure B.7: Validation losses for NODDI (left) and SMT (middle) parameter estimation and corresponding number of trainable parameters (right) for the number of training subjects 1, 3, 5.

For a comparison with other approaches, we have selected *MLP1+* with $lr = 0.0001$ for NODDI parameter estimation and *MLP2+* with $lr = 0.001$ for SMT parameter estimation.

MEDN+ hyperparameter selection for microstructure parameter estimation

In the work presented in [Ye 2017], in analogy to MEDN, a model termed as MEDN+ is introduced. It has the same architecture as MEDN with a difference that it takes as input dMRI signals from neighbourhood $3 \times 3 \times 3$. The model MEDN+ is evaluated for three different initial learning rates 0.001, 0.0005 and 0.0001. Illustrations of the validation losses for NODDI parameter estimation and corresponding number of trainable parameters, for the experiments with the number of training subjects 1, 3, 5 are provided in Figure B.8. As for MEDN, according to the validation curves, we have observed that the model sometimes experiences instabilities with higher learning rates, thus the update of trainable weights stops.

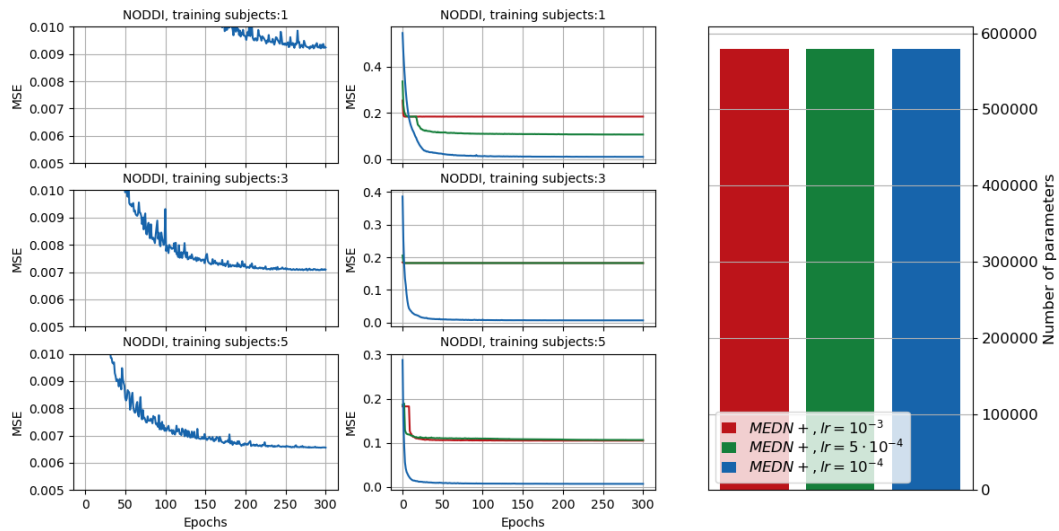


Figure B.8: Validation losses for NODDI parameter estimation, illustrated within a range $[0.005, 0.01]$ (left), without range limit to illustrate instabilities (middle) and corresponding number of trainable parameters (right) for the number of training subjects 1, 3, 5.

Clearly, for a comparison with other approaches, we have selected *MEDN+* with $lr = 0.0001$.

MescNet hyperparameter selection for microstructure parameter estimation

The model MescNet introduced in [Ye *et al.* 2019] is designed for the estimation of arbitrary microstructure parameters, thus it is evaluated on both problems of NODDI and SMT parameter estimation. It is evaluated for three different initial learning rates 0.001, 0.0005 and 0.0001. Illustrations of the validation losses for NODDI and SMT parameter estimation and corresponding number of trainable parameters, for the experiments with the number of training subjects 1, 3, 5 are provided in Figure B.9. As MEDN and MEDN+, the model exhibit instabilities for higher learning rates, thus those curves are not visible in the illustrated ranges.

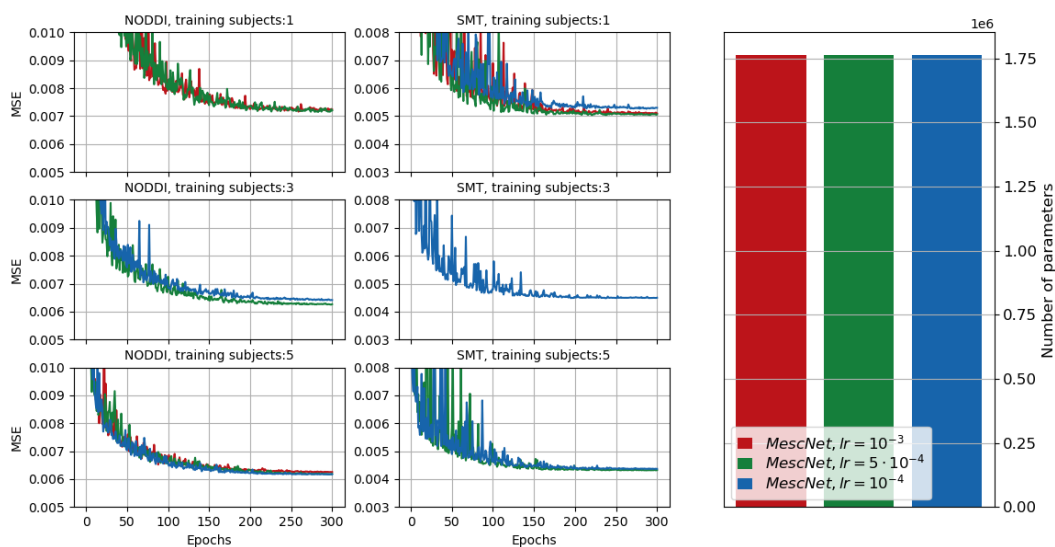


Figure B.9: Validation losses for NODDI (left) and SMT (middle) parameter estimation and corresponding number of trainable parameters (right) for the number of training subjects 1, 3, 5.

For a comparison with other approaches, we have selected *MescNet* with $lr = 0.0005$, except for SMT parameter estimation trained on 3 subjects where the selected model is trained with $lr = 0.0001$.

MescNetSepDict hyperparameter selection for microstructure parameter estimation

As MescNet, MescNetSepDict introduced in [Ye *et al.* 2020] is designed for the estimation of arbitrary microstructure parameters. It represents optimization of the model MescNet in terms of the number of parameters, however this comes with a highly increased computational time. It is evaluated for three different initial learning rates 0.001, 0.0005 and 0.0001. Illustrations of the validation losses for NODDI and SMT parameter estimation and corresponding number of trainable parameters, for the experiments with the number of training subjects 1, 3, 5 are provided in Figure B.10. As MEDN, MEDN+ and MescNet, the model exhibit sometimes instabilities, thus those curves are not visible in the illustrated ranges.

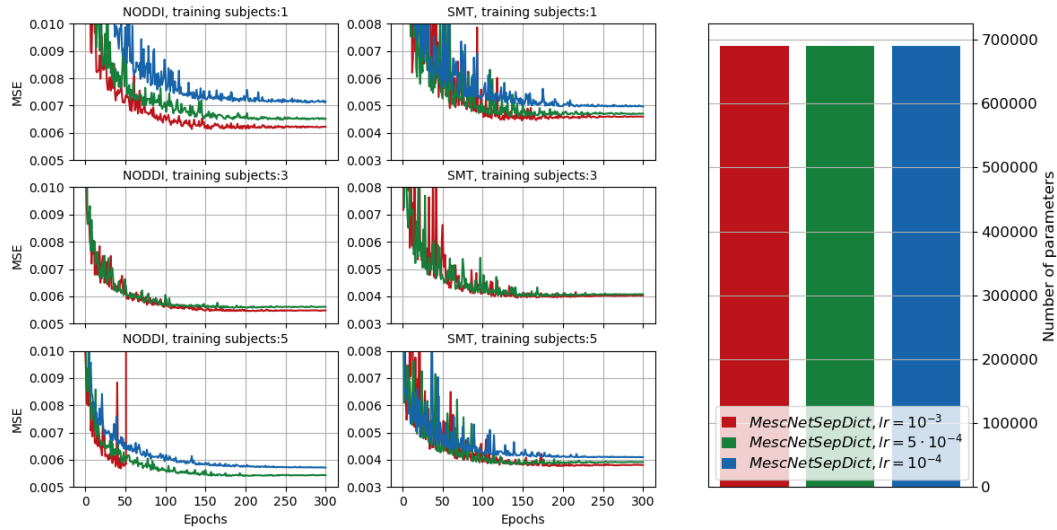


Figure B.10: Validation losses for NODDI (left) and SMT (middle) parameter estimation and corresponding number of trainable parameters (right) for the number of training subjects 1, 3, 5.

For a comparison with other approaches, we have selected *MescNetSepDict* with $lr = 0.001$, except for NODDI parameter estimation trained on 5 subjects where the selected model is trained with $lr = 0.0005$.

Hyperparameter selection for microstructure parameter estimation for our models

In analogy to MLP+ and MEDN+, we have designed *Fourier_S²_SO(3)+* and *Fourier_S²_zonal+* models which take as input dMRI signals from neighbourhood of size $3 \times 3 \times 3$. As single voxel models, they have the same denoising layer composed of two trainable matrices of size 60×60 and four fully connected layers with the number of output neurons 128, 128, 128, n_{out} at the end. Model *Fourier_S²_SO(3)+* contains four convolutional layers of input and output bandwidths (8, 6), (6, 4), (4, 2), (2, 0) with the input and output number of channels $(2 \times 27, 8)$, (8, 16), (16, 32), (32, 64). Model *Fourier_S²_zonal+* contains four convolutional layers of input and output bandwidths (8, 6), (6, 4), (4, 2), (2, 0) with the input and output number of channels $(2 \times 27, 16)$, (16, 32), (32, 64), (64, 128). In the model *Fourier_S²_SO(3)+*, since the number of rotation invariant features extracted from the first SH coefficients (after denoising) is $2 \times 27 \times 5 = 270$ is much larger than the number of rotation invariant features extracted from the following layers after *SO(3)* non-linearities $8 \times 4, 16 \times 3, 32 \times 2, 64 \times 1$, the input rotation invariant features are projected to a vector of length 64 with a trainable matrix of size 270×60 prior to concatenation to the features from other layers. In *Fourier_S²_zonal+*, the number of rotation invariant features extracted after *S²* non-linearities is $16 \times 4, 32 \times 3, 64 \times 2, 128$, thus the rotation invariant features extracted from the first SH coefficients (after denoising) is concatenated directly to them. Illustrations of the validation losses for NODDI and SMT parameter estimation and corresponding number of trainable parameters, for the experiments with the number of training subjects 1, 3, 5 are provided in Figure B.11.

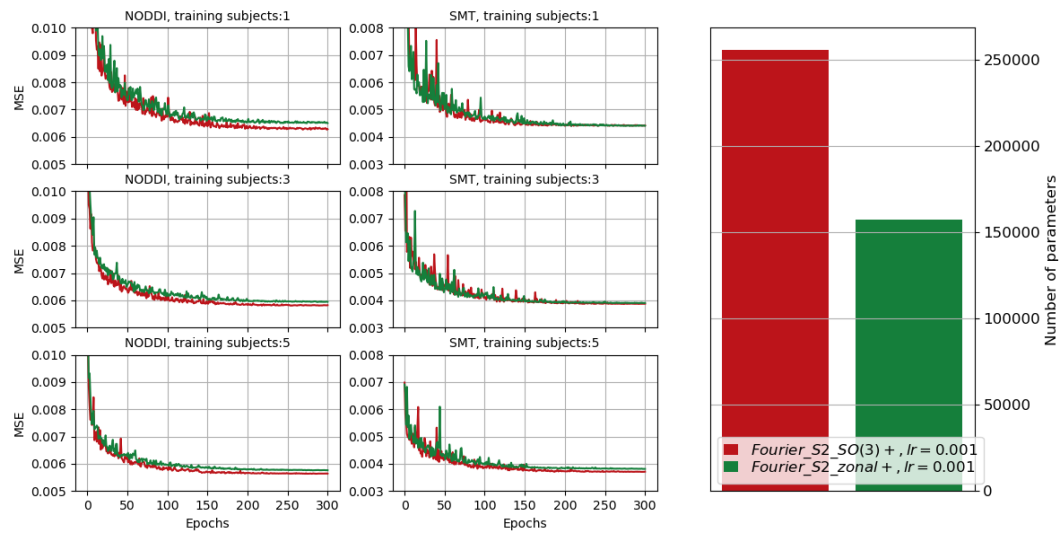


Figure B.11: Validation losses for NODDI (left) and SMT (middle) parameter estimation and corresponding number of trainable parameters (right) for the number of training subjects 1, 3, 5.

Dictionary learning experiments

appendix

Multivariate Convolutional Sparse Coding (MCSC) hyperparameter selection for noiseless data

To select the hyperparameters for MCSC, we have performed a grid search on four parameters. λ which controls sparsity of the activations, ε a stopping criterion (if the cost descent after an update of dictionary and activations is smaller than ε). ε_z tolerance of the solver for the estimation of the activations (locally greedy coordinate descent (LGCD) solver was used). ε_D of the solver for the update of dictionary (alternate adaptive solver was used). Experiments are repeated 10 times to select the hyperparameters. To perform comparison with our approach, the experiments are repeated again 40 times for the best configuration of the parameters. Maximum number of iterations for all parameter configurations is 400. The MSE and standard deviations for different parameters are given in Tables C.1, C.2, C.3, C.4 and C.5.

Table C.1: $\varepsilon_z = 10^{-4}$, $\varepsilon = 10^{-8}$, $\varepsilon_D = 10^{-8}$

$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$	$\lambda = 0.9$
0.1367 \pm 0.0679	0.1581 \pm 0.0998	0.1576 \pm 0.0632	0.2538 \pm 0.1438	0.2357 \pm 0.1552	0.2468 \pm 0.0890
0.1511 \pm 0.0822	0.1871 \pm 0.1375	0.1580 \pm 0.0773	0.2914 \pm 0.1922	0.2630 \pm 0.2051	0.2336 \pm 0.0860

Table C.2: $\varepsilon_z = 10^{-5}$, $\varepsilon = 10^{-8}$, $\varepsilon_D = 10^{-8}$

$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$	$\lambda = 0.9$
0.2082 \pm 0.1408	0.1309 \pm 0.0279	0.2016 \pm 0.1308	0.1811 \pm 0.0905	0.1326 \pm 0.0999	0.1435 \pm 0.0483
0.2435 \pm 0.2024	0.1176 \pm 0.0239	0.2190 \pm 0.1797	0.1732 \pm 0.0931	0.1416 \pm 0.1357	0.1381 \pm 0.0527

Table C.3: $\varepsilon = 10^{-8}$, $\varepsilon_D = 10^{-8}$

$\lambda = 0.4, \varepsilon_z = 10^{-3}$	$\lambda = 0.3, \varepsilon_z = 10^{-3}$	$\lambda = 0.8, \varepsilon_z = 10^{-6}$	$\lambda = 0.9, \varepsilon_z = 10^{-6}$
0.2241 \pm 0.1114	0.2634 \pm 0.2098	0.1395 \pm 0.0621	0.1402 \pm 0.0918
0.2525 \pm 0.1355	0.3202 \pm 0.2629	0.1349 \pm 0.0571	0.1366 \pm 0.0879

MCSC hyperparameter selection for noisy data

As for noiseless data, to select the hyperparameters for MCSC applied on noisy data, we have performed a grid search on four parameters. Parameters are selected based

Table C.4: $\varepsilon_z = 10^{-5}$, $\varepsilon_D = 10^{-8}$

$\lambda = 0.5, \varepsilon = 10^{-9}$	$\lambda = 0.5, \varepsilon = 10^{-7}$	$\lambda = 0.8, \varepsilon = 10^{-9}$	$\lambda = 0.8, \varepsilon = 10^{-7}$
0.2174 \pm 0.1421	0.1244 \pm0.0478	0.1840 \pm 0.1046	0.2080 \pm 0.1107
0.2641 \pm 0.2006	0.1200 \pm0.0602	0.1891 \pm 0.1413	0.2127 \pm 0.1472

Table C.5: $\varepsilon_z = 10^{-5}$, $\varepsilon = 10^{-7}$

$\lambda = 0.5, \varepsilon_D = 10^{-9}$	$\lambda = 0.5, \varepsilon_D = 10^{-7}$
0.1288 \pm 0.0889	0.1659 \pm 0.0799
0.1425 \pm 0.1177	0.1740 \pm 0.1194

on reconstruction MSE computed with respect to noiseless ground truth signals which are given in Tables C.6, C.7, C.8, C.9 and C.10.

Table C.6: $\varepsilon_z = 10^{-5}$, $\varepsilon = 10^{-8}$, $\varepsilon_D = 10^{-8}$

$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$
2.9842 \pm 0.2601	2.2990 \pm 0.1859	2.4923 \pm 0.1773	4.0116 \pm 1.5745	6.7545 \pm 2.0822	8.6524 \pm 1.7048
2.8564 \pm 0.2831	2.2999 \pm 0.23028	2.5246 \pm 0.1980	3.9704 \pm 1.3986	6.5052 \pm 1.9051	8.3485 \pm 1.4671

Table C.7: $\varepsilon_z = 10^{-4}$, $\varepsilon = 10^{-8}$, $\varepsilon_D = 10^{-8}$

$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$
2.7965 \pm 0.1466	2.3316 \pm 0.1645	2.6254 \pm 0.1639	3.5966 \pm 1.4074
2.6385 \pm 0.1646	2.3051 \pm 0.1573	2.6665 \pm 0.1945	3.6517 \pm 1.3504

Table C.8: $\varepsilon_z = 10^{-6}$, $\varepsilon = 10^{-8}$, $\varepsilon_D = 10^{-8}$

$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$
2.7811 \pm 0.1815	2.3578 \pm 0.1285	2.6275 \pm 0.19479	3.2454 \pm 1.1326
2.6547 \pm 0.1993	2.3557 \pm 0.1502	2.6735 \pm 0.2204	3.2946 \pm 1.1355

Table C.9: $\varepsilon_z = 10^{-5}$, $\varepsilon_D = 10^{-8}$

$\lambda = 0.4, \varepsilon = 10^{-9}$	$\lambda = 0.4, \varepsilon = 10^{-7}$
2.3304 \pm 0.1064	2.3005 \pm 0.1639
2.2951 \pm 0.1376	2.2513 \pm 0.2192

Table C.10: $\varepsilon_z = 10^{-5}$, $\varepsilon = 10^{-8}$

$\lambda = 0.4, \varepsilon_D = 10^{-9}$	$\lambda = 0.4, \varepsilon_D = 10^{-7}$
2.2935 \pm0.1049	2.3428 \pm 0.1347
2.2562 \pm0.1330	2.3314 \pm 0.1515

Since ground truth noiseless signals are not available in the real scenario, we have investigated whether selection of parameters can be based on reconstruction MSE computed with respect to noisy available signals and concluded that some prior knowledge for parameter selection is required. The MSE and standard deviations for different parameters are given in Tables C.11, C.12, C.13, C.14, C.15 and C.16.

Table C.11: $\varepsilon_z = 10^{-5}$, $\varepsilon = 10^{-8}$, $\varepsilon_D = 10^{-8}$

$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$
3109.898 \pm 0.788	3121.307 \pm 1.192	3128.235 \pm 1.039
3110.205 \pm 0.941	3122.229 \pm 1.169	3128.432 \pm 0.511

Table C.12: $\varepsilon_z = 10^{-5}$, $\varepsilon = 10^{-8}$, $\varepsilon_D = 10^{-8}$

$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$	$\lambda = 0.7$
3129.933 \pm 0.940	3131.443 \pm 0.702	3133.805 \pm 1.750	3136.926 \pm 2.533
3130.829 \pm 0.969	3132.013 \pm 0.603	3134.207 \pm 2.030	3137.245 \pm 2.037

Table C.13: $\varepsilon_z = 10^{-4}$, $\varepsilon = 10^{-8}$, $\varepsilon_D = 10^{-8}$

$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$
3109.1191 \pm0.8951	3120.9702 \pm 0.9151	3127.5977 \pm 1.0059
3109.9946 \pm1.2312	3121.6355 \pm 1.2983	3128.268 \pm 1.0959

Table C.14: $\varepsilon_z = 10^{-3}$, $\varepsilon = 10^{-8}$, $\varepsilon_D = 10^{-8}$

$\lambda = 0.1$
3109.587 \pm 1.2439
3111.091 \pm 1.4872

Table C.15: $\varepsilon_z = 10^{-4}$, $\varepsilon_D = 10^{-8}$

$\lambda = 0.1, \varepsilon = 10^{-7}$	$\lambda = 0.1, \varepsilon = 10^{-9}$
3109.274 \pm 0.786	3109.6804 \pm 1.252
3110.586 \pm 1.222	3110.848 \pm 0.9401

Table C.16: $\varepsilon_z = 10^{-4}$, $\varepsilon = 10^{-8}$

$\lambda = 0.1, \varepsilon_D = 10^{-7}$	$\lambda = 0.1, \varepsilon_D = 10^{-9}$
3110.058 1.484	3109.657 \pm 1.141
3110.297 1.506	3110.540 \pm 0.909

HCP $Q=3$ and $P=2$

Illustrations of the learned spatial and temporal patterns obtained with our approach and the corresponding activation vectors. Models contain 1 pair of spatial and temporal atoms. The maximum number of activations during train $Q = 3$ and during test $P = 2$. The models are trained one one session corresponding to one event (left hand, left foot, right hand, right foot movements and fixation).

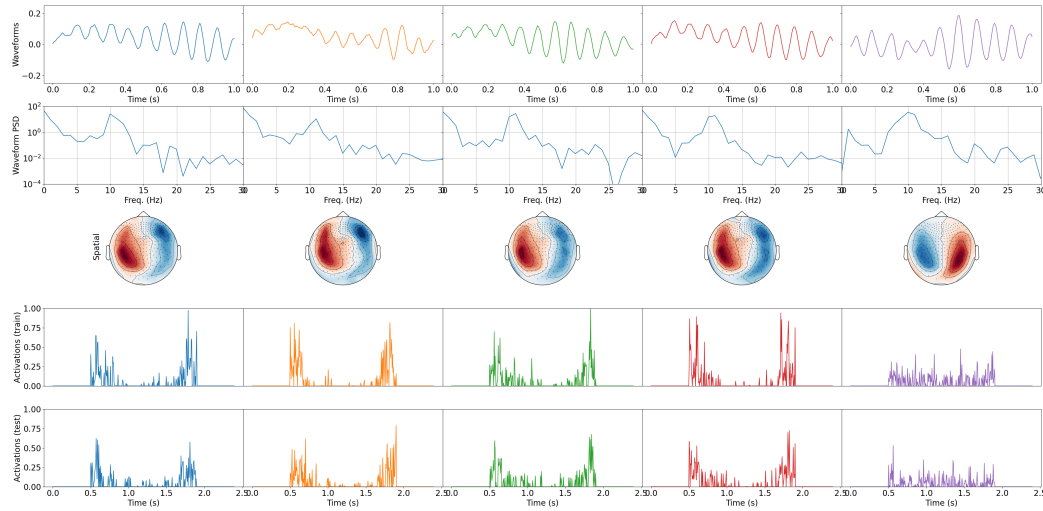


Figure C.1: **Subject 104012** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. **Left hand** (first column), **left foot** (second column), **right hand** (third column), **right foot** (fourth column) movements, **fixation/resting** (fifth column).

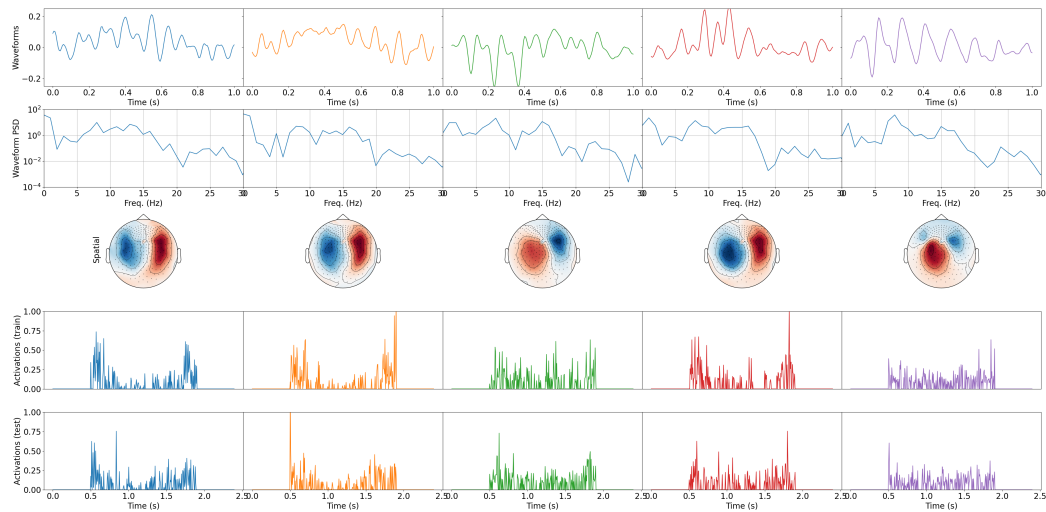


Figure C.2: **Subject 105923** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. **Left hand** (first column), **left foot** (second column), **right hand** (third column), **right foot** (fourth column) movements, **fixation/resting** (fifth column).

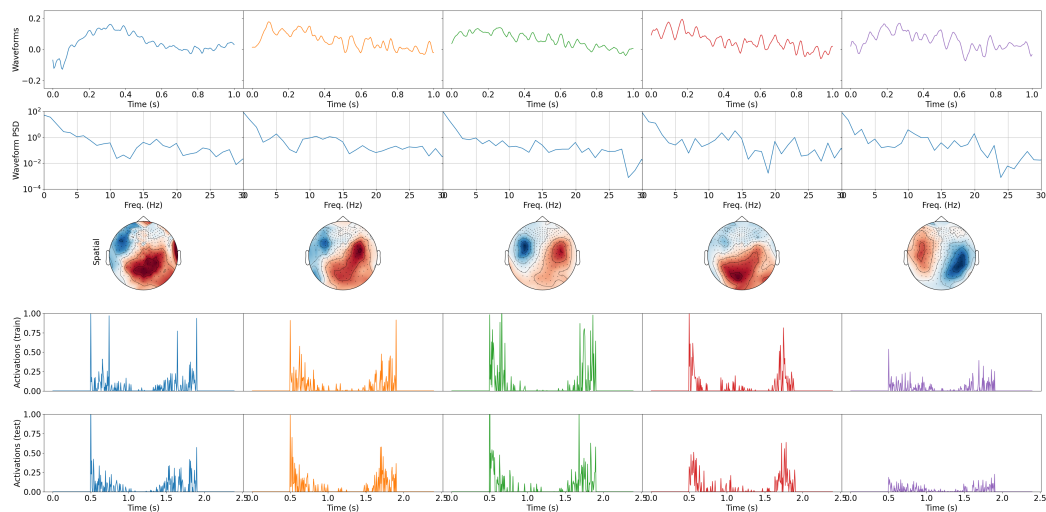


Figure C.3: **Subject 106521** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. **Left hand** (first column), **left foot** (second column), **right hand** (third column), **right foot** (fourth column) movements, **fixation/resting** (fifth column).

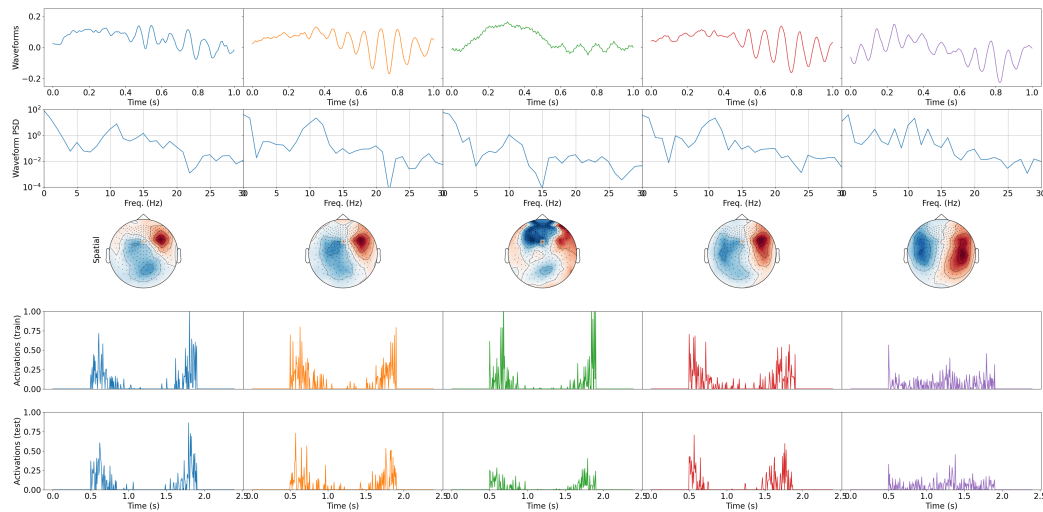


Figure C.4: **Subject 108323** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. **Left hand** (first column), **left foot** (second column), **right hand** (third column), **right foot** (fourth column) movements, **fixation/resting** (fifth column).

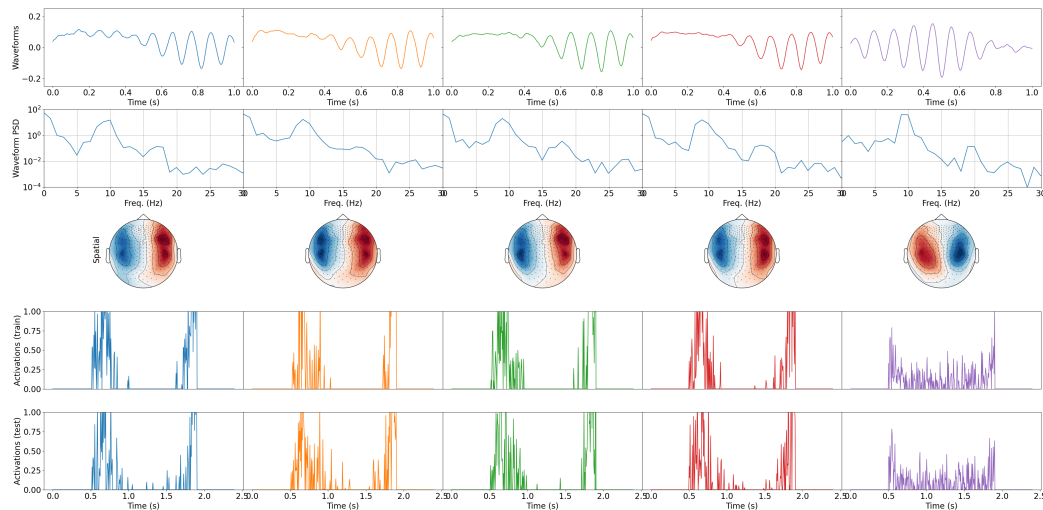


Figure C.5: **Subject 109123** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. **Left hand** (first column), **left foot** (second column), **right hand** (third column), **right foot** (fourth column) movements, **fixation/resting** (fifth column).

HCP $Q=10$ and $P=2$

Illustrations of the learned spatial and temporal patterns obtained with our approach and the corresponding activation vectors. Models contain 1 pair of spatial and temporal atoms. The maximum number of activations during train $Q = 10$ and during test $P = 2$. The models are trained one one session corresponding to one event (left hand, left foot, right hand, right foot movements and fixation).

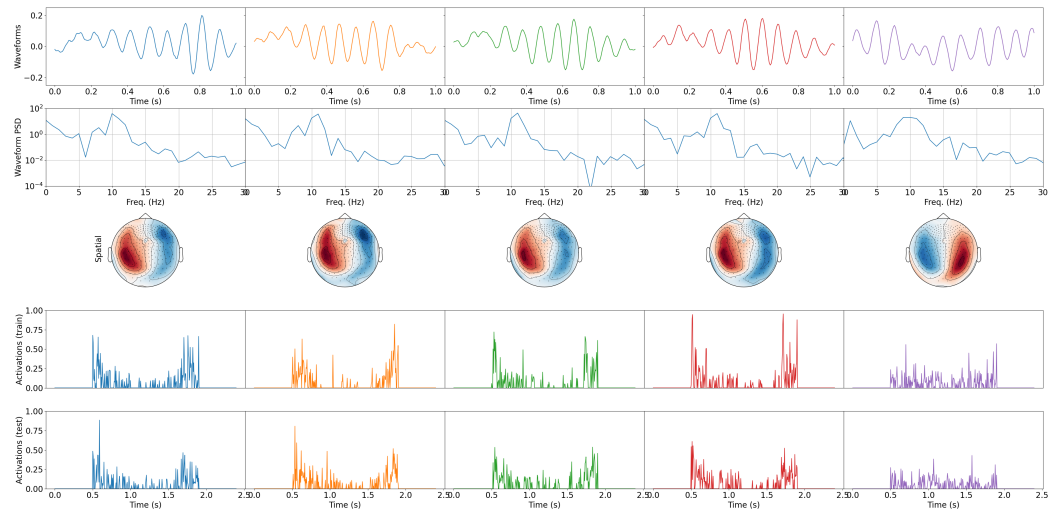


Figure C.6: **Subject 104012** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. **Left hand** (first column), **left foot** (second column), **right hand** (third column), **right foot** (fourth column) movements, **fixation/resting** (fifth column).

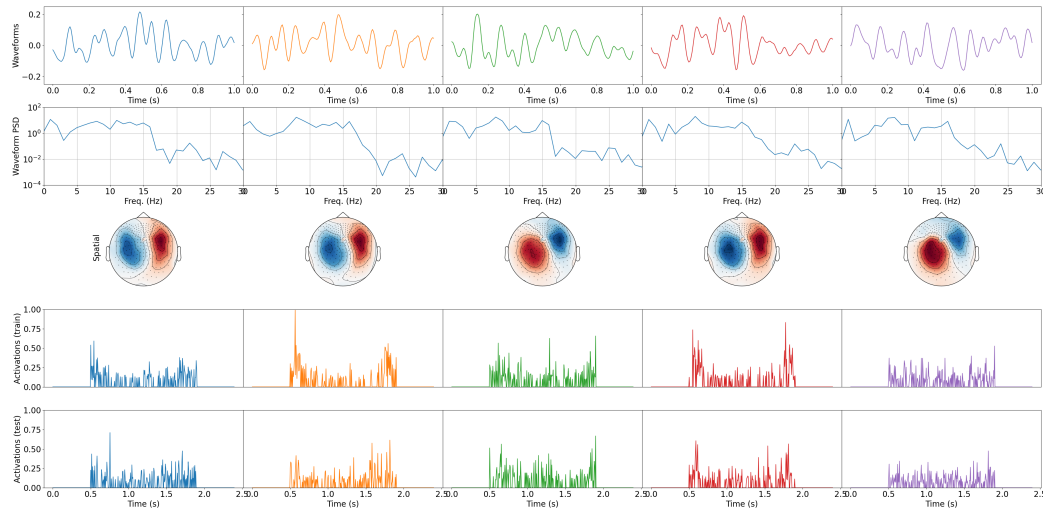


Figure C.7: **Subject 105923** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. **Left hand** (first column), **left foot** (second column), **right hand** (third column), **right foot** (fourth column) movements, **fixation/resting** (fifth column).

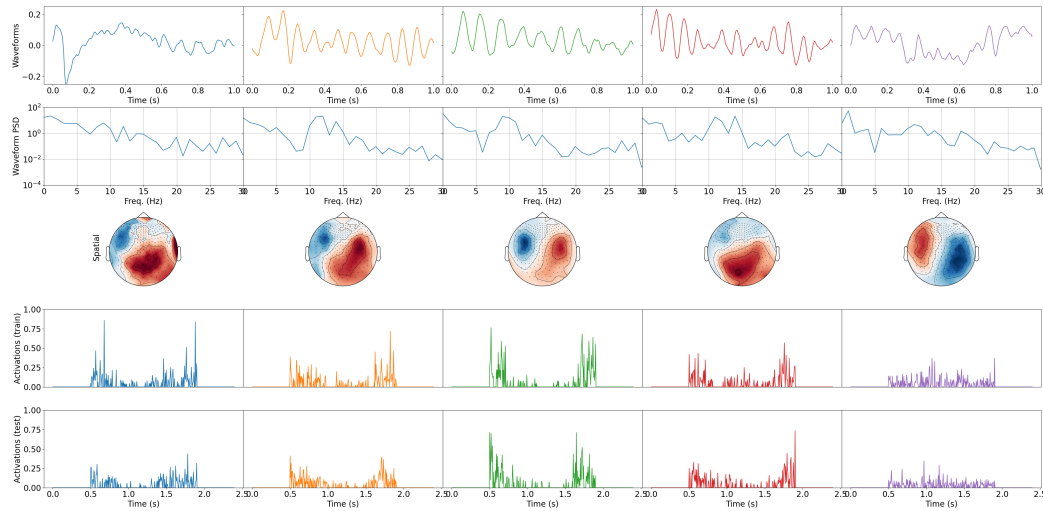


Figure C.8: **Subject 106521** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. **Left hand** (first column), **left foot** (second column), **right hand** (third column), **right foot** (fourth column) movements, **fixation/resting** (fifth column).

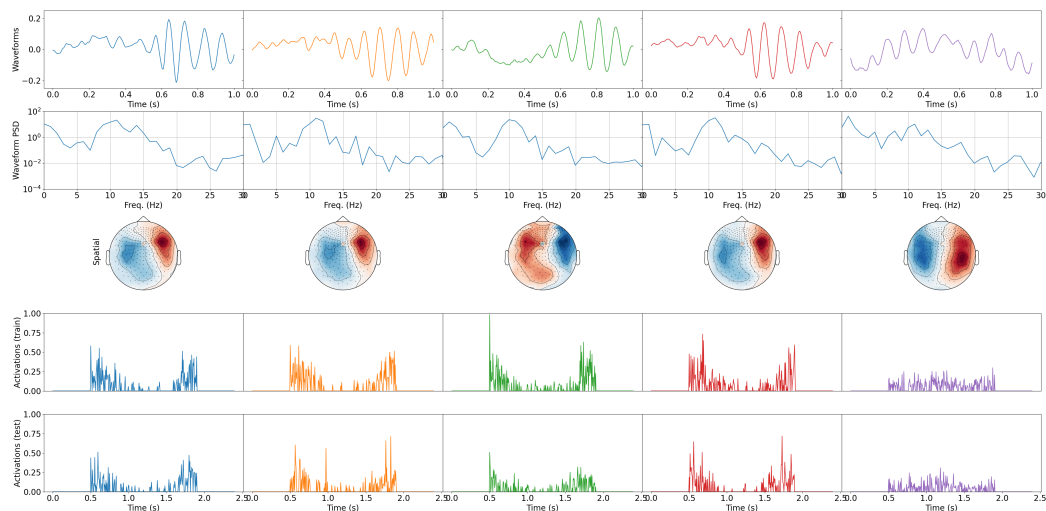


Figure C.9: **Subject 108323** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. **Left hand** (first column), **left foot** (second column), **right hand** (third column), **right foot** (fourth column) movements, **fixation/resting** (fifth column).

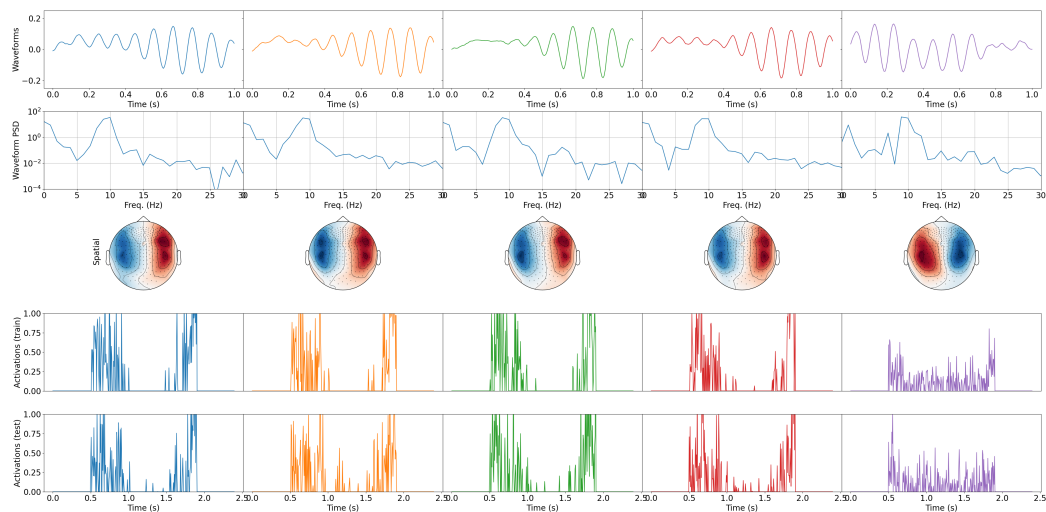


Figure C.10: **Subject 109123** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. **Left hand** (first column), **left foot** (second column), **right hand** (third column), **right foot** (fourth column) movements, **fixation/resting** (fifth column).

HCP, 10 atoms, $Q=5$ and $P=2$

Illustrations of the learned spatial and temporal patterns obtained with our approach and the corresponding activation vectors. Models contain 10 pairs of spatial and temporal atoms. The maximum number of activations during train $Q = 5$ and during test $P = 2$. The models are trained one one session corresponding to one event (left hand, left foot, right hand, right foot movements and fixation).

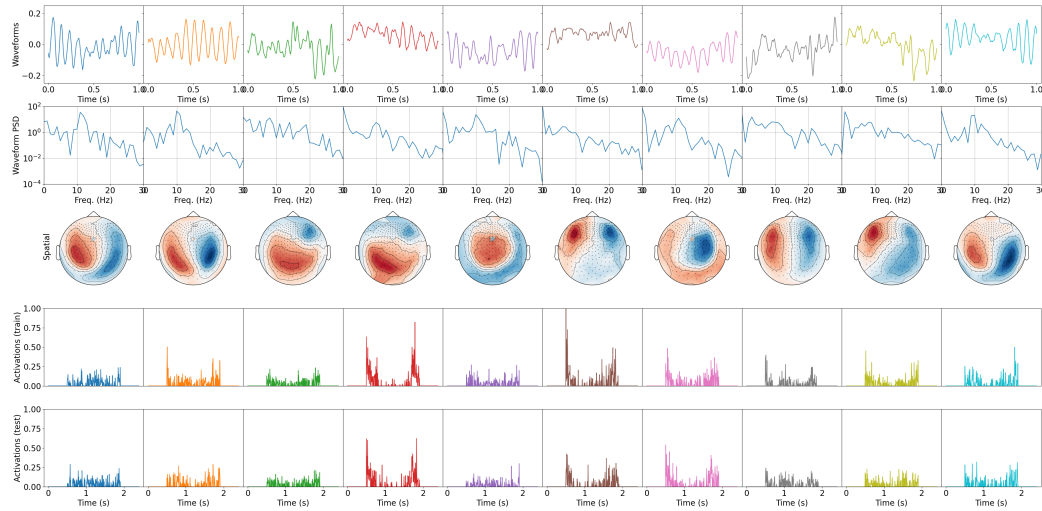


Figure C.11: **Subject 104012, Left hand** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

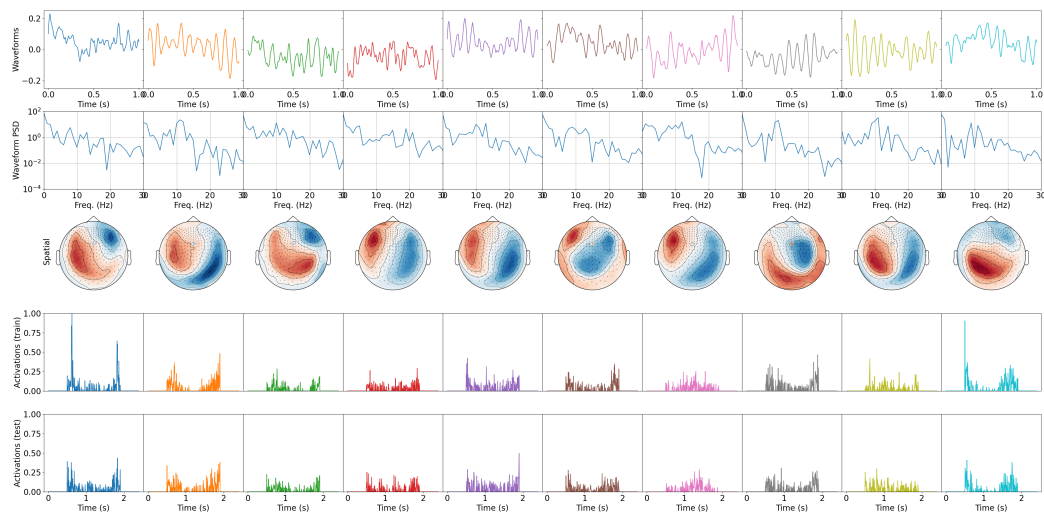


Figure C.12: **Subject 104012, Left foot** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

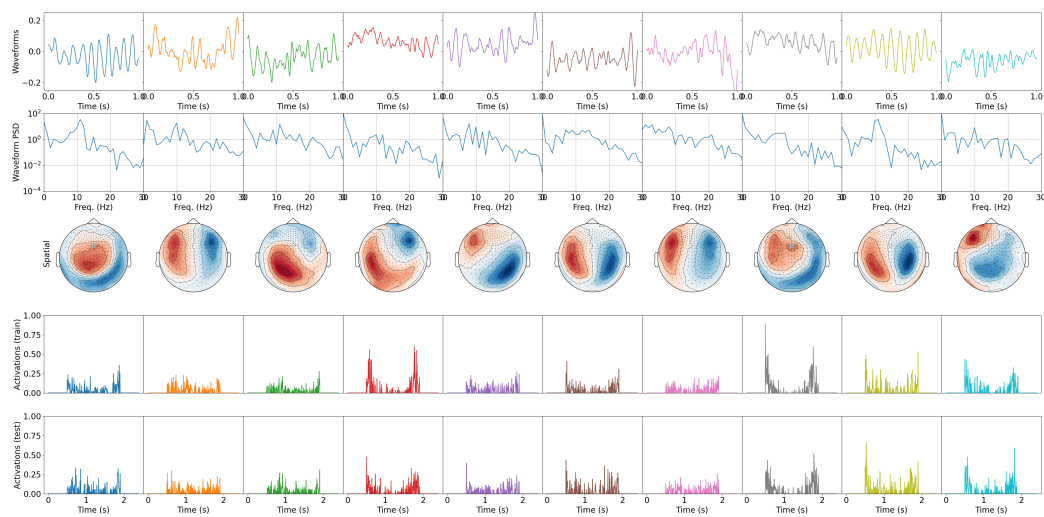


Figure C.13: **Subject 104012, Right hand** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

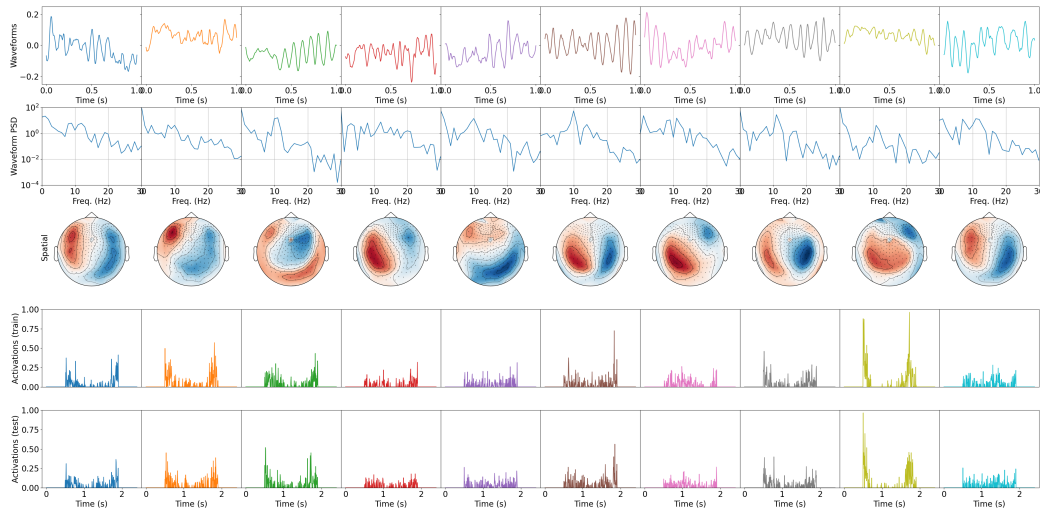


Figure C.14: **Subject 104012, Right foot** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

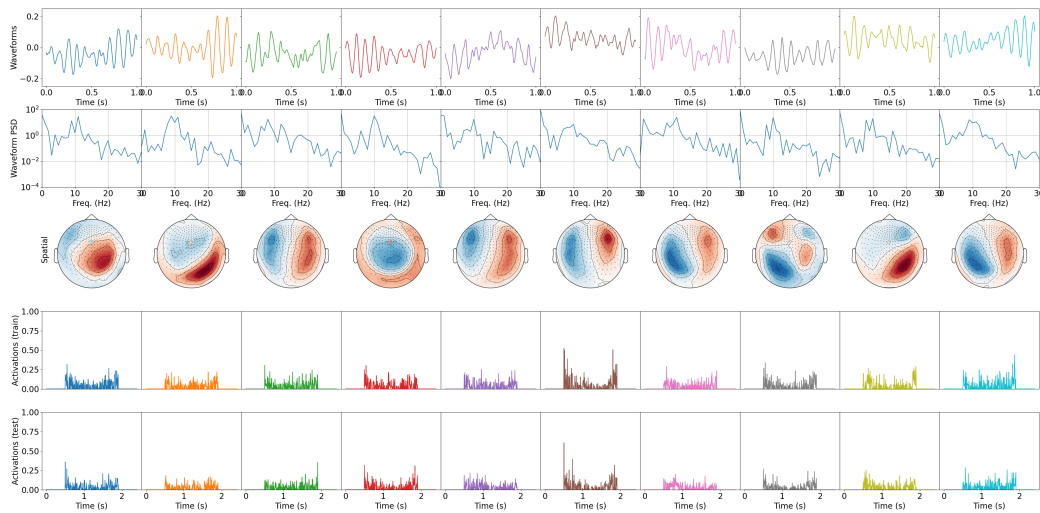


Figure C.15: **Subject 104012, Fixation** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

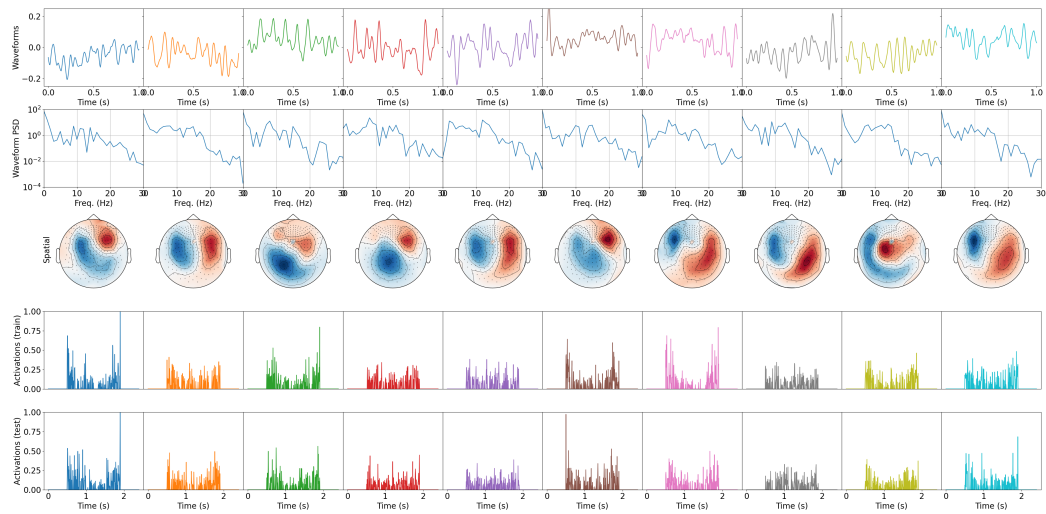


Figure C.16: **Subject 105923, Left hand** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

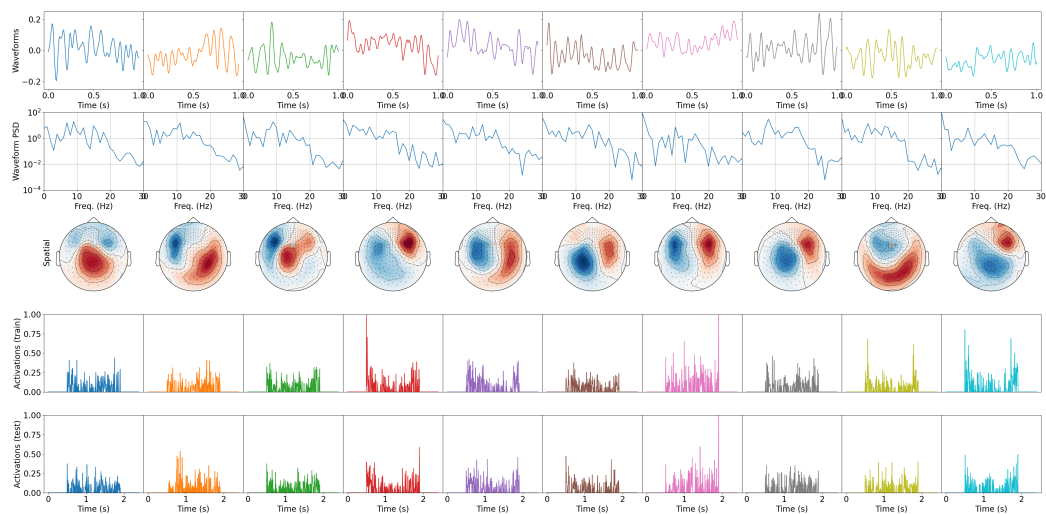


Figure C.17: **Subject 105923, Left foot** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

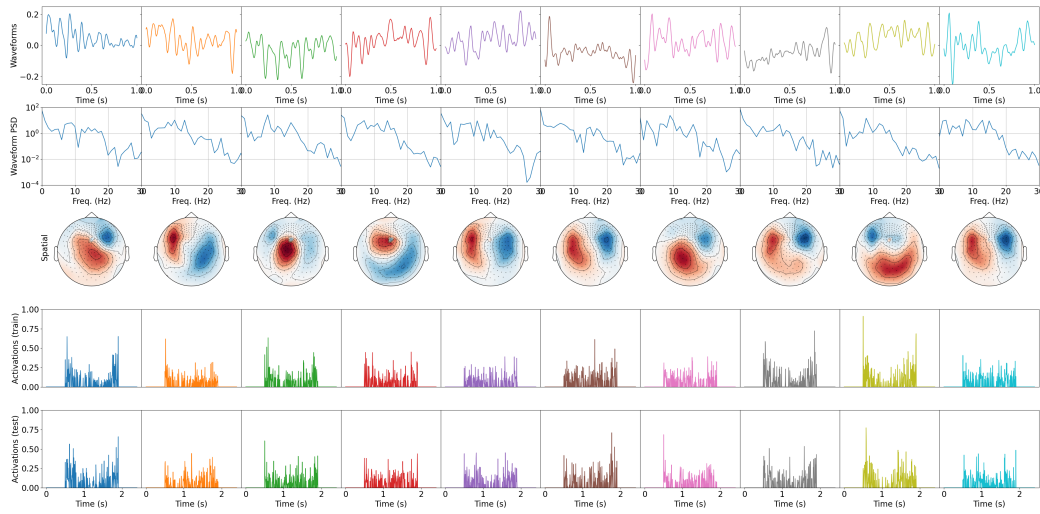


Figure C.18: **Subject 105923, Right hand** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

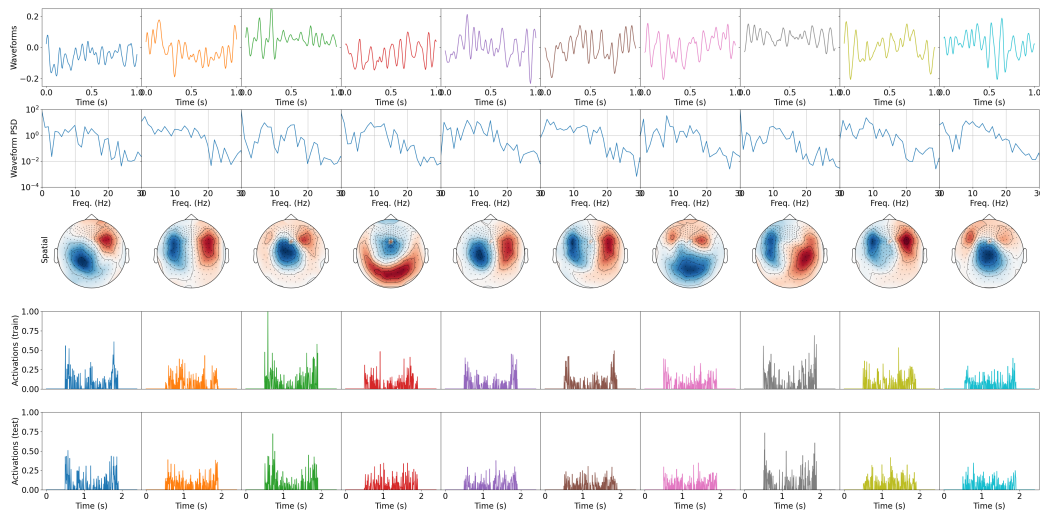


Figure C.19: **Subject 105923, Right foot** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

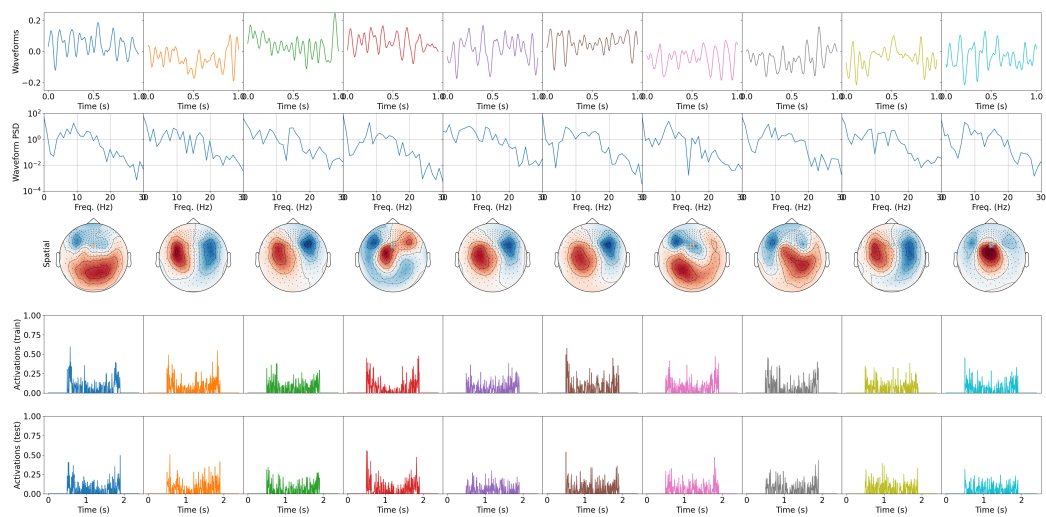


Figure C.20: **Subject 105923, Fixation** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

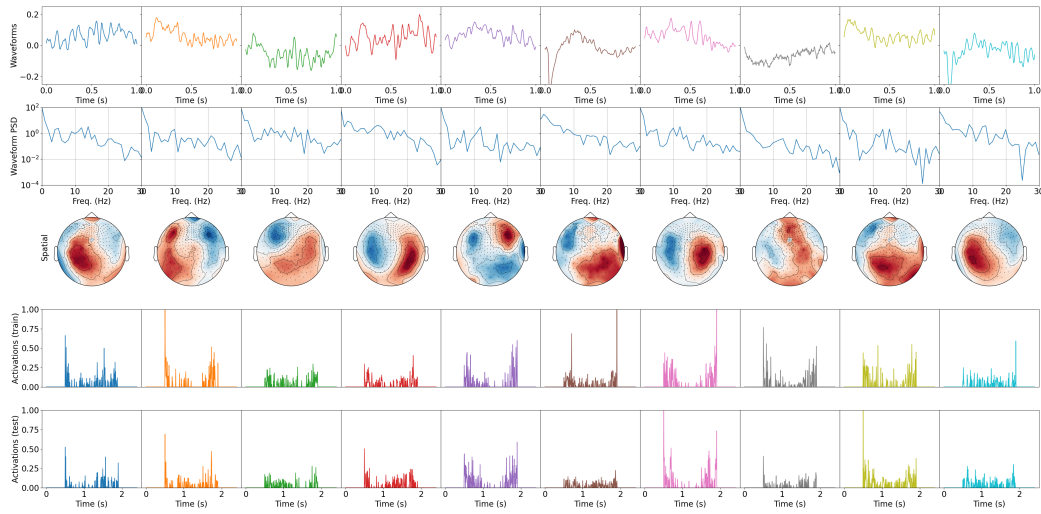


Figure C.21: **Subject 106521, Left hand** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

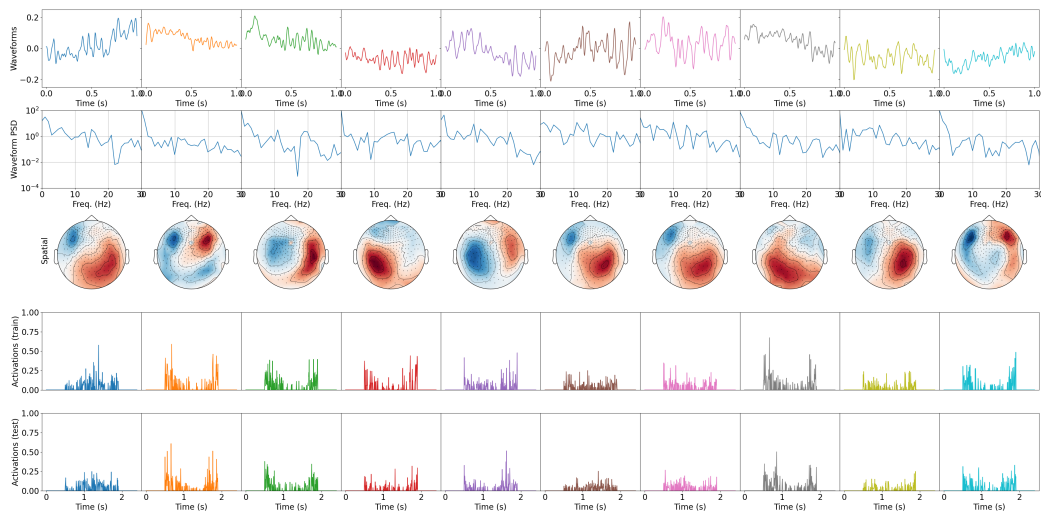


Figure C.22: **Subject 106521, Left foot** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

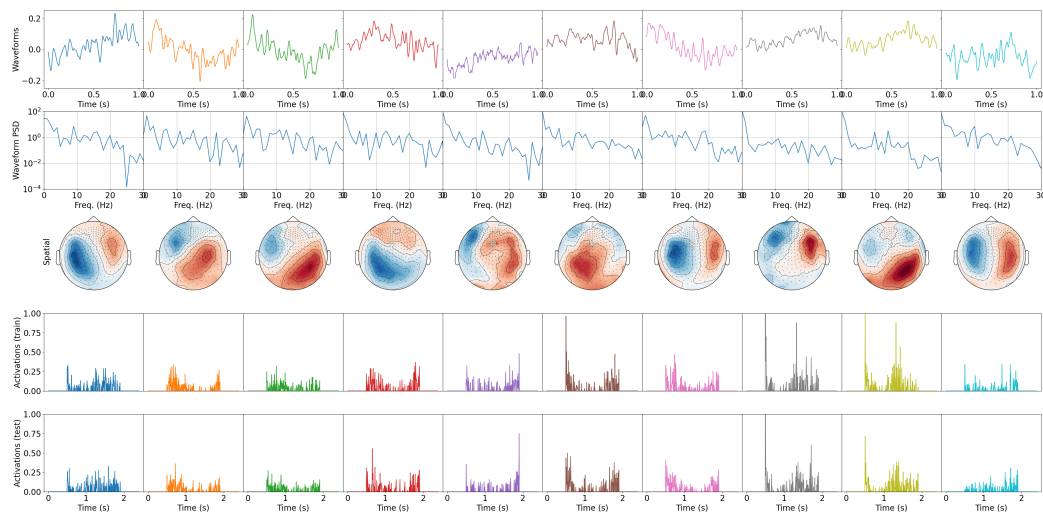


Figure C.23: **Subject 106521, Right hand** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

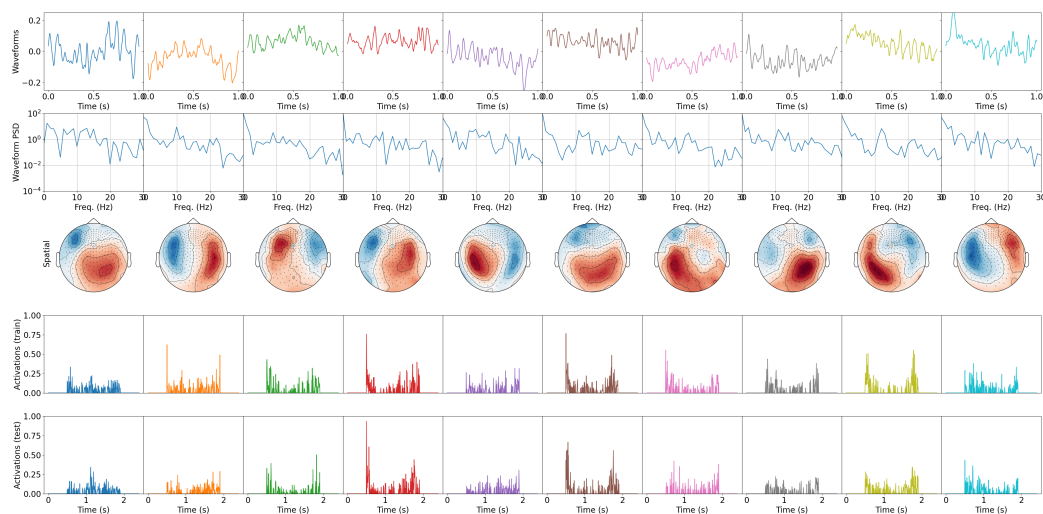


Figure C.24: **Subject 106521, Right foot** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

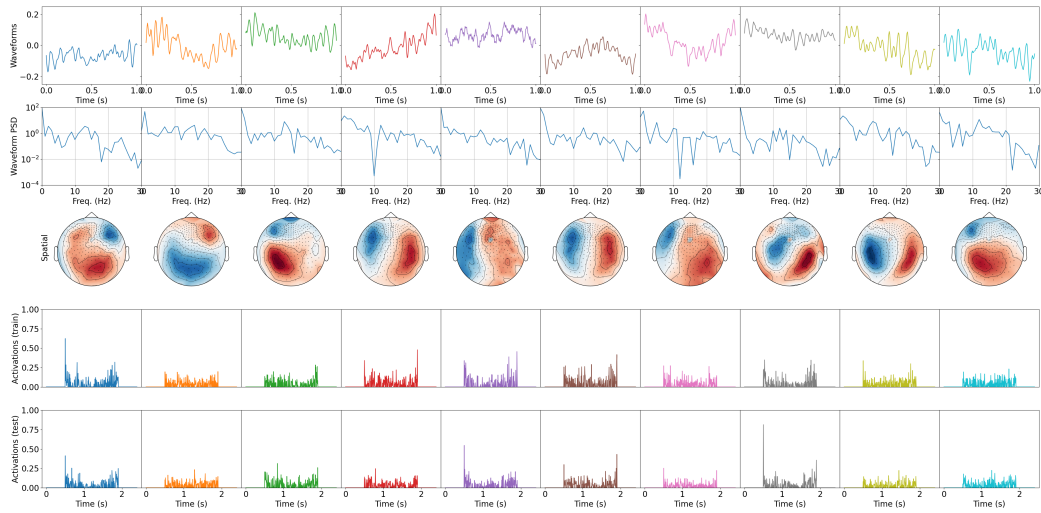


Figure C.25: **Subject 106521, Fixation** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

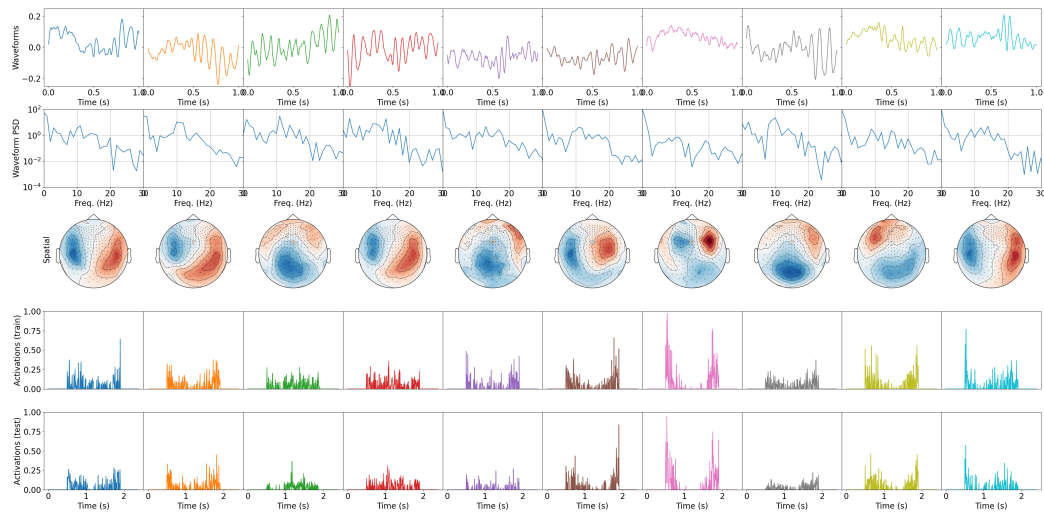


Figure C.26: **Subject 108323, Left hand** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

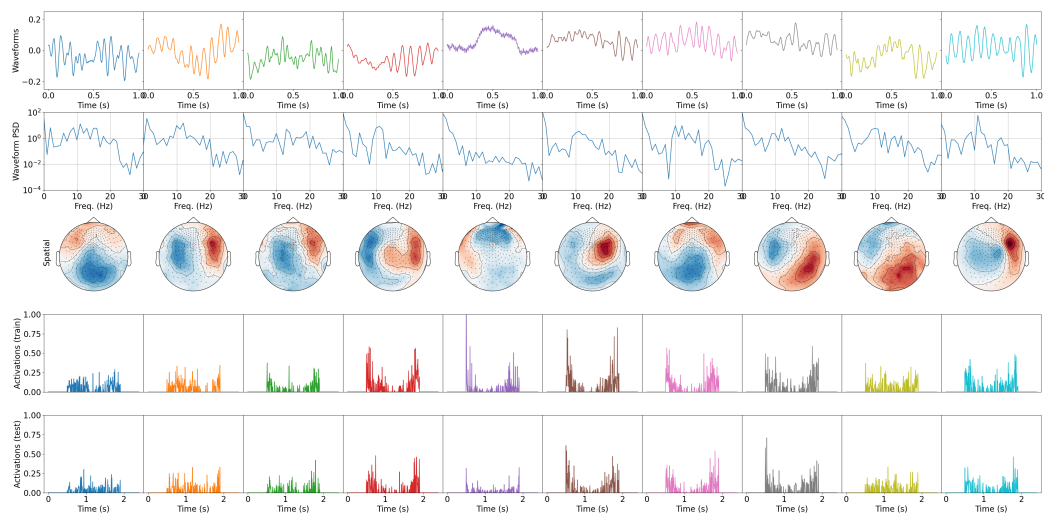


Figure C.27: **Subject 108323, Left foot** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

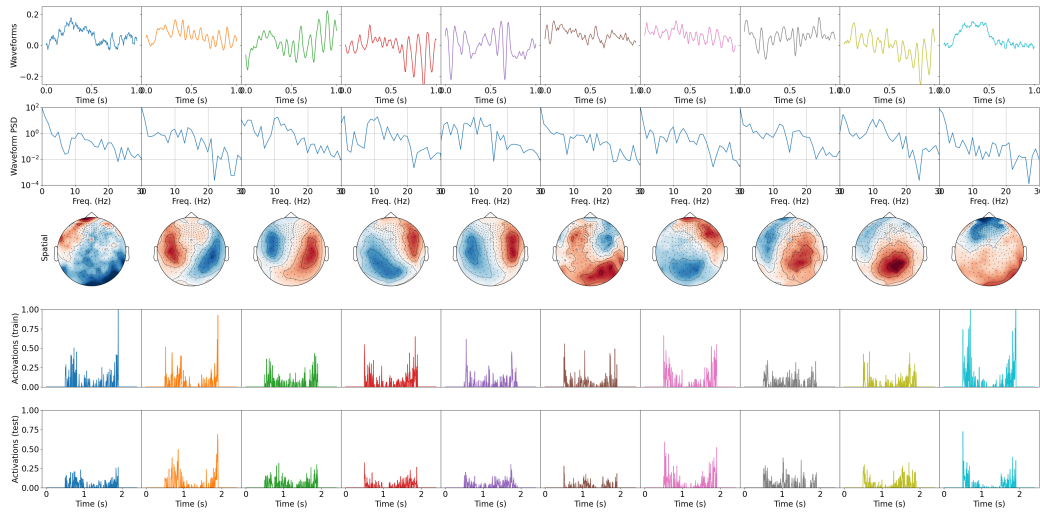


Figure C.28: **Subject 108323, Right hand** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

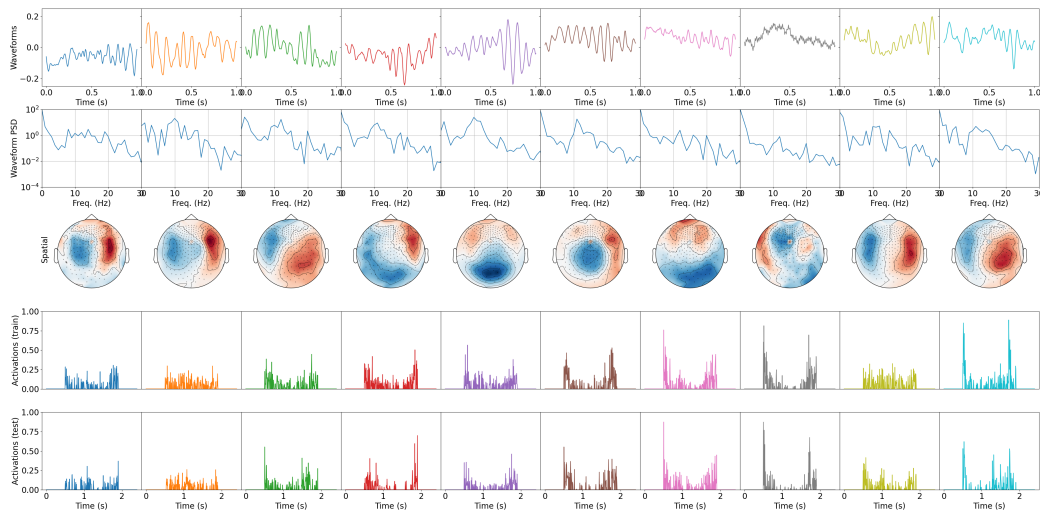


Figure C.29: **Subject 108323, Right foot** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

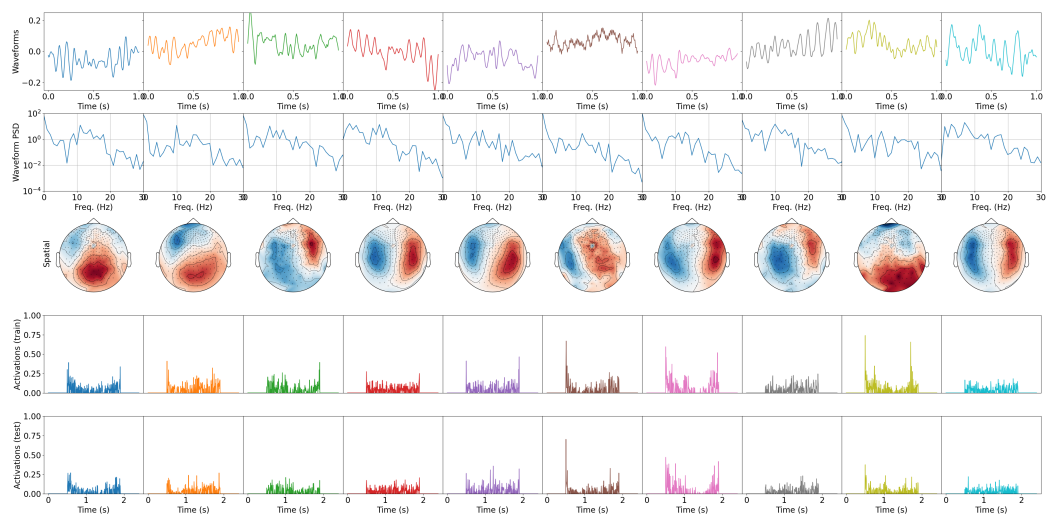


Figure C.30: **Subject 108323, Fixation** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

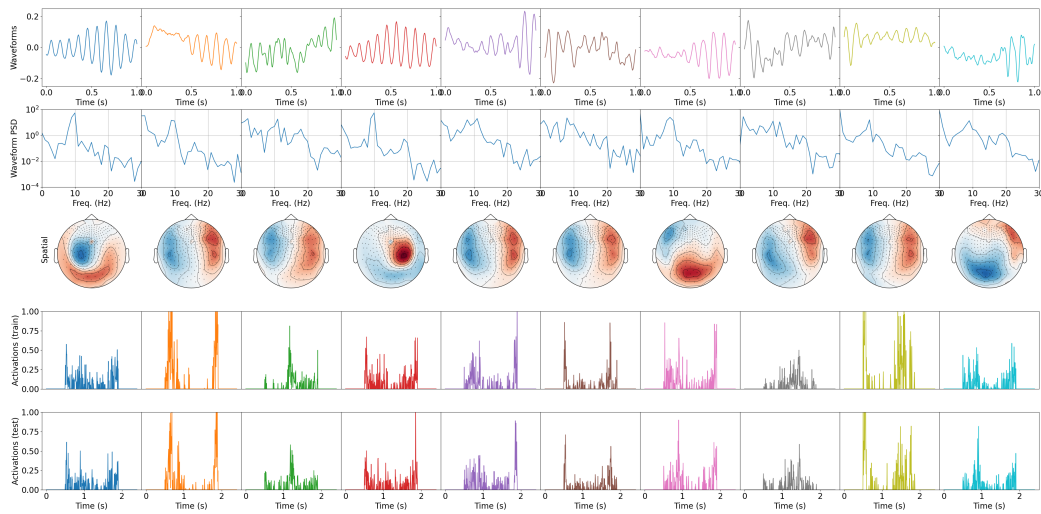


Figure C.31: **Subject 109123, Left hand** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

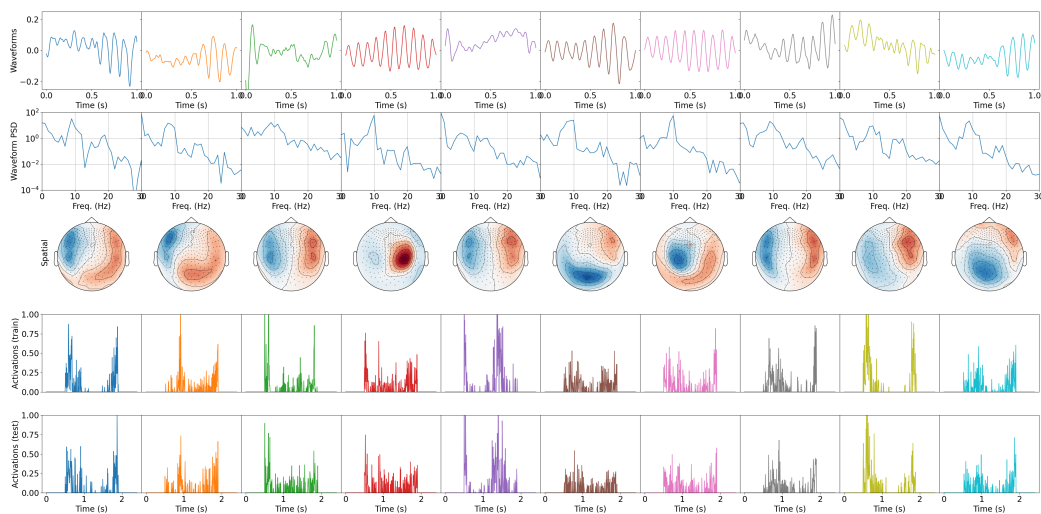


Figure C.32: **Subject 109123, Left foot** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

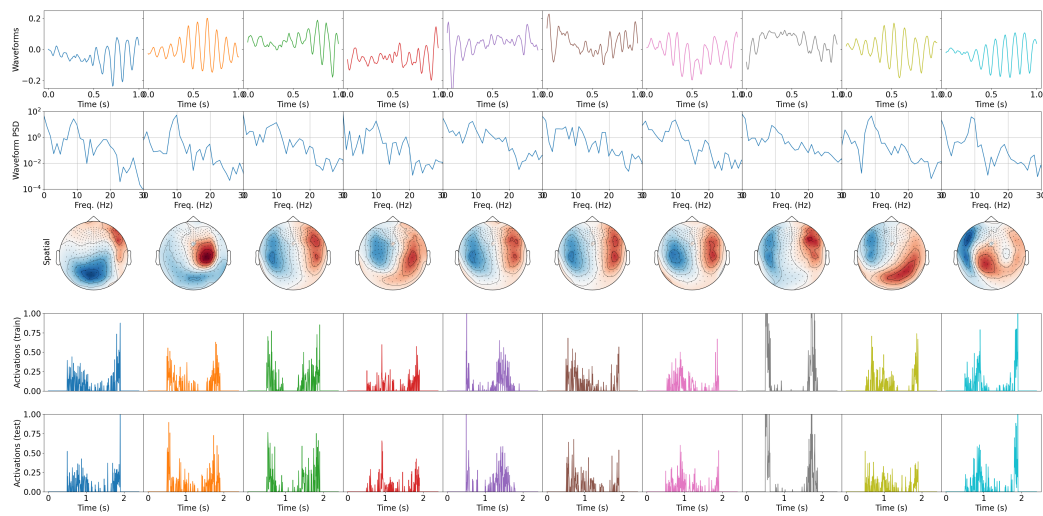


Figure C.33: **Subject 109123, Right hand** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

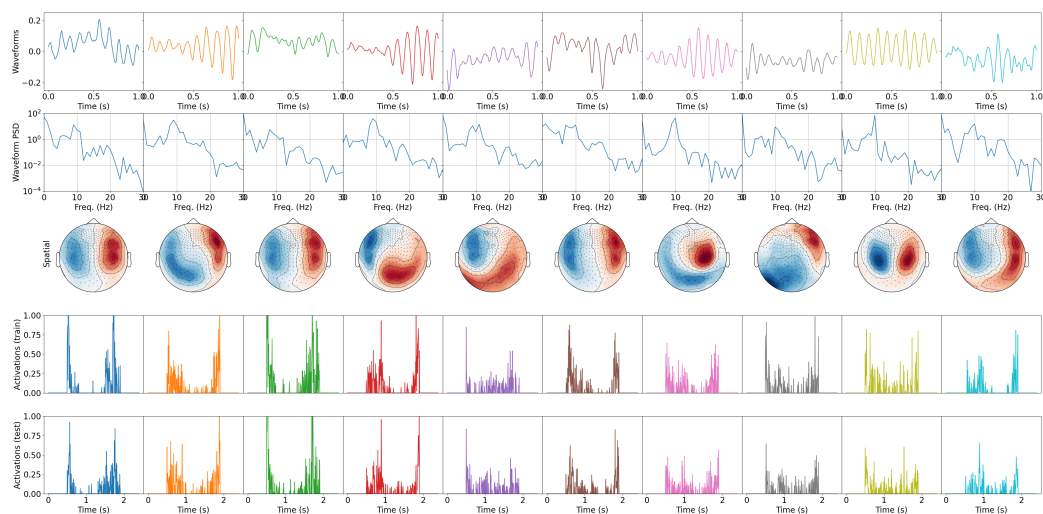


Figure C.34: **Subject 109123, Right foot** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

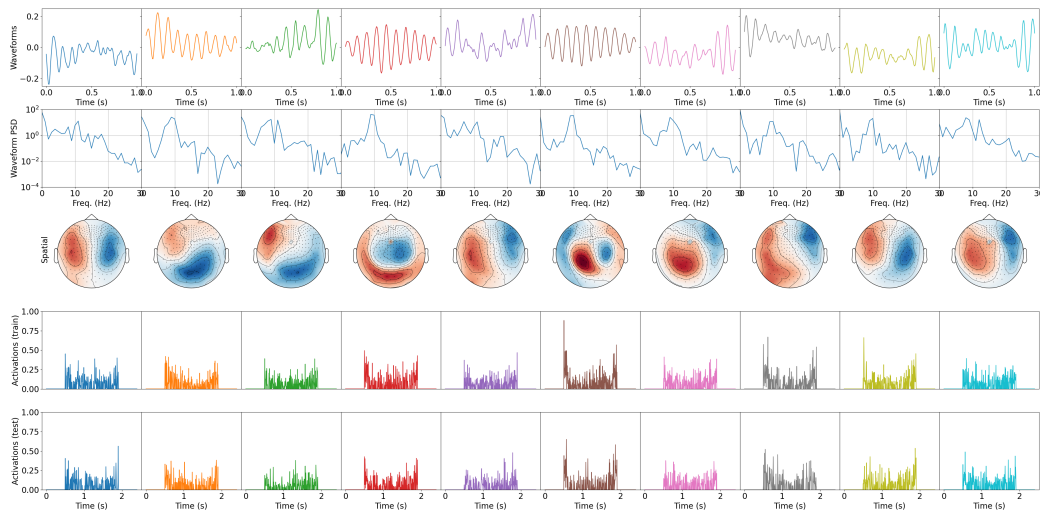


Figure C.35: **Subject 109123, Fixation** Illustration of estimated temporal patterns (first row), their power spectral density (second row), spatial patterns (third row), activations on training session (fourth row) and activations on testing session (fifth row) obtained with our method. Each column corresponds to different atom.

M/EEG classification experiments

appendix

DeepConvNet

In this subsection we provide details on *DeepConvNet* [Schirrmester *et al.* 2017] hyperparameter search and the number of multiplications. The number of multiplications per layer are given in Table D.1.

Table D.1: *DeepConvNet* number of multiplications per different steps of entire classification process for one input sample. T is the input signal length. N is the number of channels of the input signal. K is the number of temporal kernels. k_L is the kernel length. $T_1 = \lfloor \frac{T-k_L+1}{p} \rfloor$. $T_2 = \lfloor \frac{T_1-k_L+1}{p} \rfloor$. $T_3 = \lfloor \frac{T_2-k_L+1}{p} \rfloor$. p is pooling size. Q is the number of output classes.

Operation	Number of multiplications
Temporal correlation	$N \times (T - k_L + 1) \times k_L \times K$
Spatial correlation	$N \times (T - k_L + 1) \times K \times K$
Batch normalization	$2 \times (T - k_L + 1) \times K$
Exponential Linear Unit	$(T - k_L + 1)/2 \times K \times (1 + 3(N_{Ty} - 2))$
Temporal correlation	$(T_1 - k_L + 1) \times K \times 2K$
Batch normalization	$2 \times (T_1 - k_L + 1) \times 2K$
Exponential Linear Unit	$(T_1 - k_L + 1) \times K \times (1 + 3(N_{Ty} - 2))$
Temporal correlation	$(T_2 - k_L + 1) \times 2K \times 4K$
Batch normalization	$2 \times (T_2 - k_L + 1) \times 4K$
Exponential Linear Unit	$(T_2 - k_L + 1) \times 2K \times (1 + 3(N_{Ty} - 2))$
Temporal correlation	$(T_3 - k_L + 1) \times 4K \times 8K$
Batch normalization	$2 \times (T_3 - k_L + 1) \times 8K$
Exponential Linear Unit	$(T_3 - k_L + 1) \times 4K \times (1 + 3(N_{Ty} - 2))$
Feature classification	$Q \times (T_3 - k_L + 1) \times 8K + Q \times (1 + 3(N_{Ty} - 2))$

* N_{Ty} corresponds to Taylor series degree used to compute exponential

Illustration of validation curves *DeepConvNet* models for different hyperparameters are provided in Figures D.1 and D.2 for MEG experiment. K refers to the number of convolutional kernels in the first layer, where in the each following this number is increased by a factor of two. k_L corresponds to the convolutional filter length and

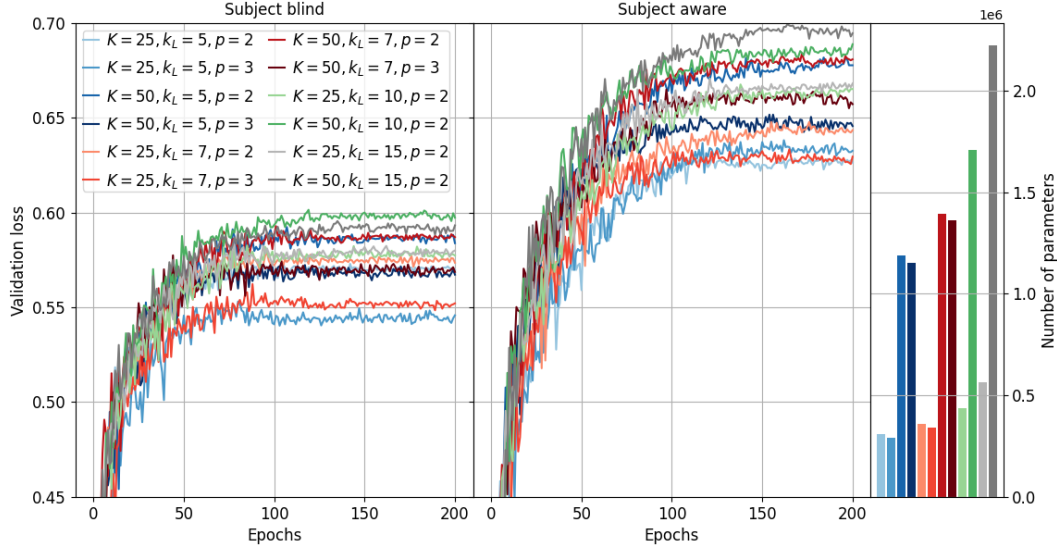


Figure D.1: *DeepConvNet* validation curves for *subject blind* training (left) and *subject aware* training (right) for motor task MEG classification problem for fixed $K = 50$, different lengths of temporal kernels k_L and different pooling sizes p .

p to max pooling size after each convolution layer. We can notice that validation accuracy is lower for pooling step 3. Increase of K leads to significant accuracy improvement, while increase of k_L from 5, 7 to 10, 15 leads to finer improvements. In the *subject blind* experiment set-up, the model with $K = 50$, $k_L = 10$ and $p = 2$ is selected as the best one. The number of trainable parameters is $\sim 1.71 \times 10^6$. In *subject aware* experiment set-up, the model with $K = 50$, $k_L = 15$ and $p = 2$ is selected as the best one. The number of trainable parameters is $\sim 2.22 \times 10^6$. Illustration of the validation curves *DeepConvNet* models for different hyperparameters are provided in Figures D.3 and D.4 for EEG experiment. In the *subject blind* experiment set-up, the model with $K = 100$, $k_L = 3$ and $p = 2$ is selected as the best one. The number of trainable parameters is $\sim 1.89 \times 10^6$. In *subject aware* experiment set-up, the model with $K = 100$, $k_L = 3$ and $p = 3$ is selected as the best one. The number of trainable parameters is $\sim 1.89 \times 10^6$. Learning rate is 0.0005.

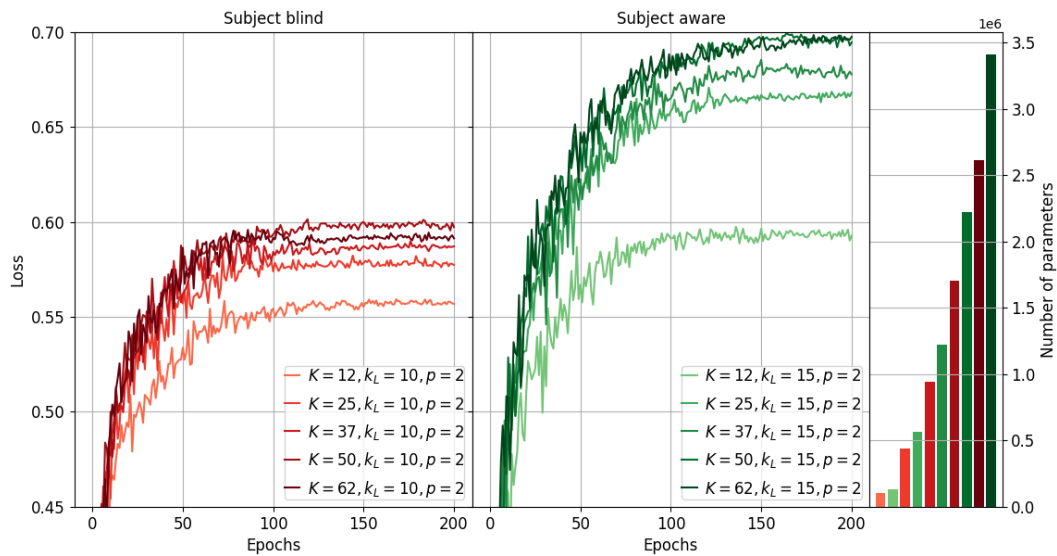


Figure D.2: *DeepConvNet* validation curves for *subject blind* training (left) and *subject aware* training (right) for motor task MEG classification problem for different K , and fixed lengths of temporal kernels k_L and pooling sizes p .

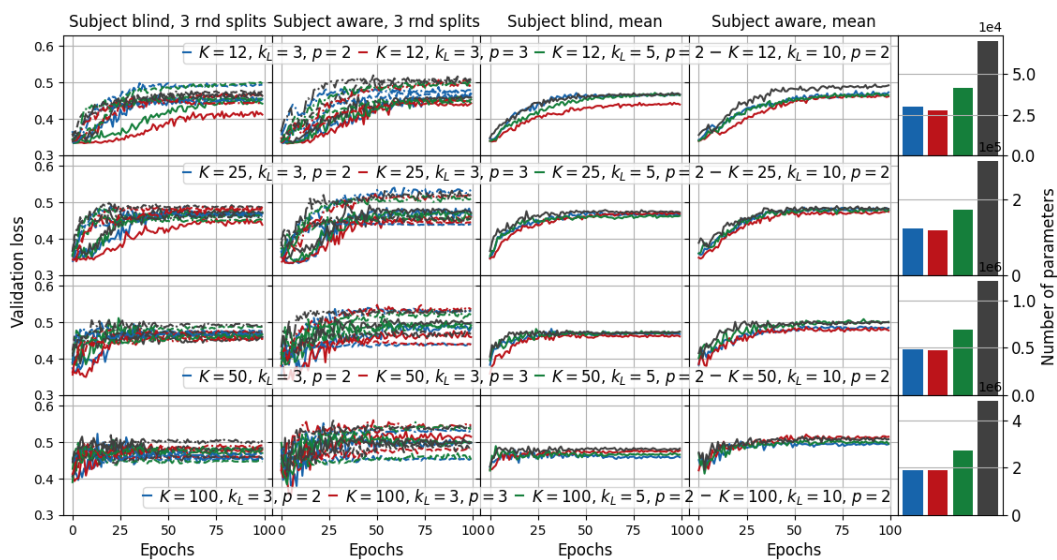


Figure D.3: *DeepConvNet* validation curves for *subject blind* training and *subject aware* training for mental workload EEG classification problem. Learning rate 0.0005

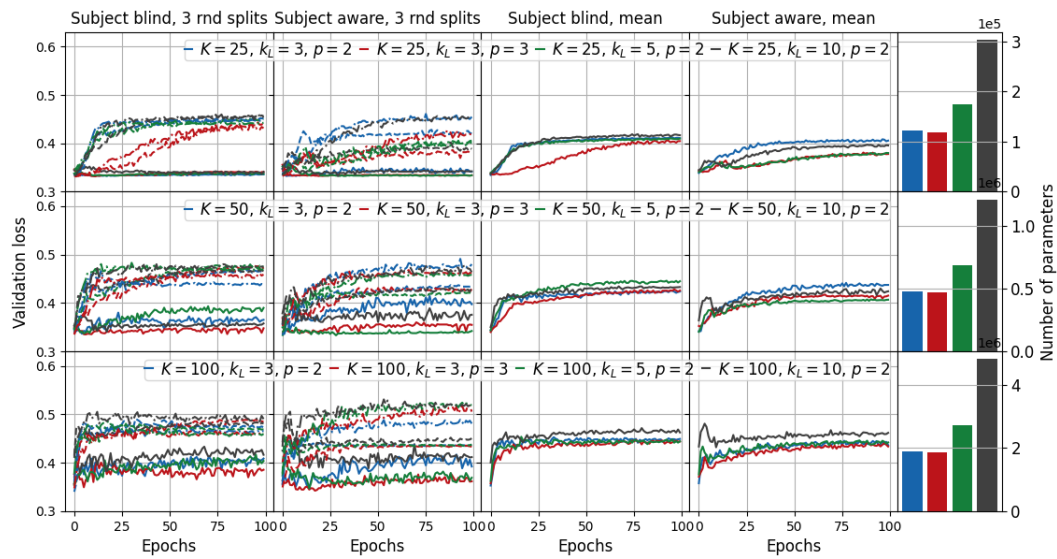


Figure D.4: *DeepConvNet* validation curves for *subject blind* training and *subject aware* training for mental workload EEG classification problem. Learning rate 0.0001

ShallowConvNet

In this subsection we provide details on *ShallowConvNet* [Schirrmeister *et al.* 2017] hyperparameter search and the number of multiplications. The number of multiplications per layer are given in Table D.2.

Table D.2: *ShallowConvNet* number of multiplications per different steps of entire classification process for one input sample. T is the input signal length. N is the number of channels of the input signal. K is the number of temporal kernels. k_L is the kernel length. p is pooling size. Q is the number of output classes.

Operation	Number of multiplications
Temporal correlation	$N \times (T - k_L + 1) \times k_L \times K$
Spatial correlation	$N \times (T - k_L + 1) \times K \times K$
Batch normalization	$2 \times (T - k_L + 1) \times K$
Square activation	$(T - k_L + 1) \times K$
Average pooling	$\frac{5^{(T-k_L+1)}}{p} \times K$
Logarithmic activation	$\frac{5^{(T-k_L+1)}}{p} \times K \times 3(N_{Ty} - 2)$
Feature classification	$Q \times \frac{5^{(T-k_L+1)}}{p} \times K + Q \times (1 + 3(N_{Ty} - 2))$

* N_{Ty} corresponds to Taylor series degree used to compute exponential and logarithm

Illustration of the validation curves for *ShallowConvNet* models for different hyperparameters are provided in Figure D.5 for MEG experiment. k_L refers to the length of convolutional kernels and p to average pooling size. The number of convolutional kernels is $K = 50$. Contrary to the *DeepConvNet* where pooling size corresponds to the pooling stride, in *ShallowConvNet* pooling stride is $p/5$. We can notice that validation accuracy is higher for longer convolutional kernels and smaller pooling size. We can also notice that in *subject blind* training there is overfitting after 50th epoch in majority of the models. To decrease overfitting, models are trained with convolutional kernels constrained to norm lower than 1, whereas the default norm bound is 2. The models are trained for $p = 15$ and the corresponding validation curves are depicted in Figures D.6 and D.7. Decrease of norm bound yields a slight improvement in *subject aware* training as well. In the *subject blind* experiment set-up, the model with $k_L = 35$ and $p = 15$ is selected as the best one, with $\sim 0.652 \times 10^6$ parameters. In the *subject aware* experiment set-up, the model with $k_L = 25$ and $p = 15$ is selected as the best one, with $\sim 0.652 \times 10^6$ parameters. We have also observed that although decreasing learning rate can lead to smoother validation loss, the curve flattens at lower accuracy.

Illustration of the validation curves for *ShallowConvNet* models for different hyperparameters are provided in Figures D.8, D.9 and D.10 for EEG experiment. In the *subject blind* experiment set-up, the model with $K = 50$, $k_L = 15$ and $p = 15$ is selected as the best one, with $\sim 0.16 \times 10^6$ parameters. In the *subject aware* exper-

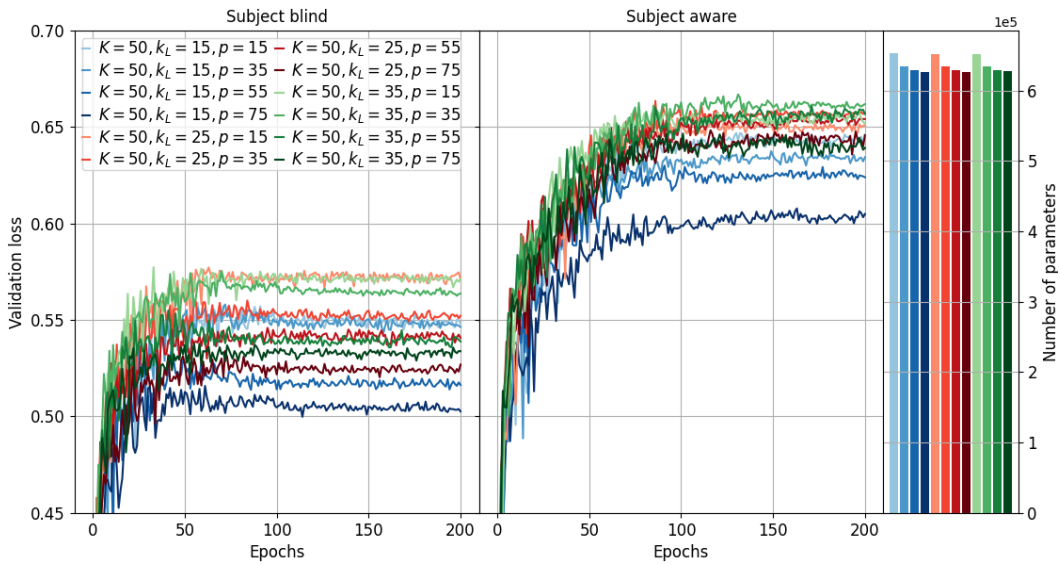


Figure D.5: *ShallowConvNet* validation curves for *subject blind* training (left) and *subject aware* training (right) for motor task MEG classification problem.

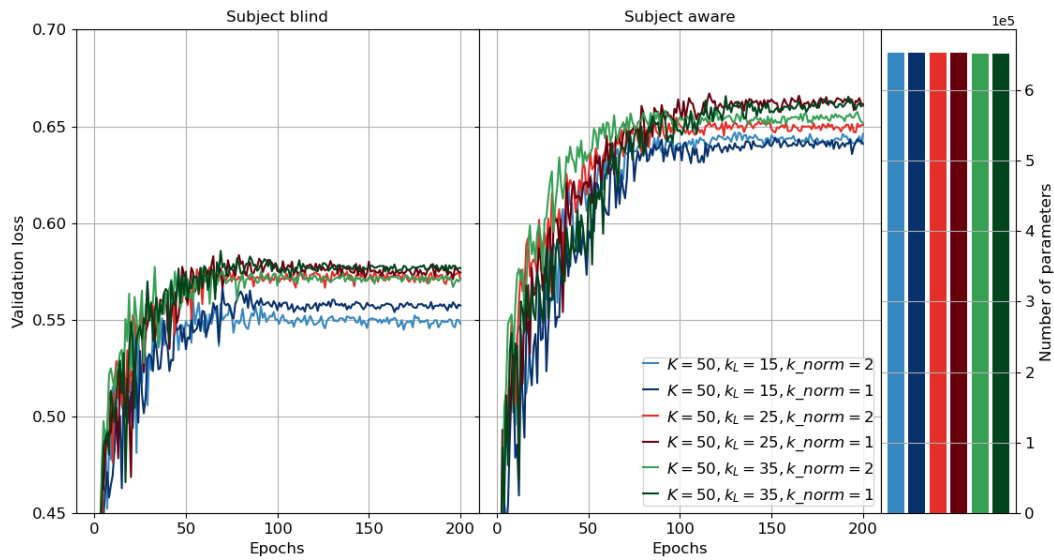


Figure D.6: *ShallowConvNet* validation curves for *subject blind* training (left) and *subject aware* training (right) for motor task MEG classification problem.

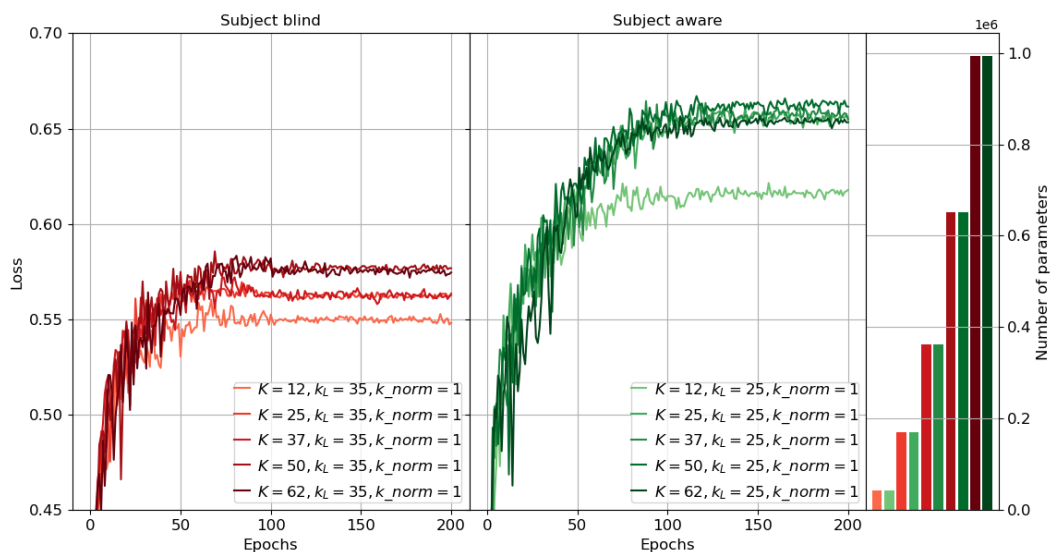


Figure D.7: *ShallowConvNet* validation curves for *subject blind* training (left) and *subject aware* training (right) for motor task MEG classification problem.

iment set-up, the model with $K = 25$, $k_L = 15$ and $p = 10$ is selected as the best one, with $\sim 0.04 \times 10^6$ parameters. Learning rate is 0.0001.

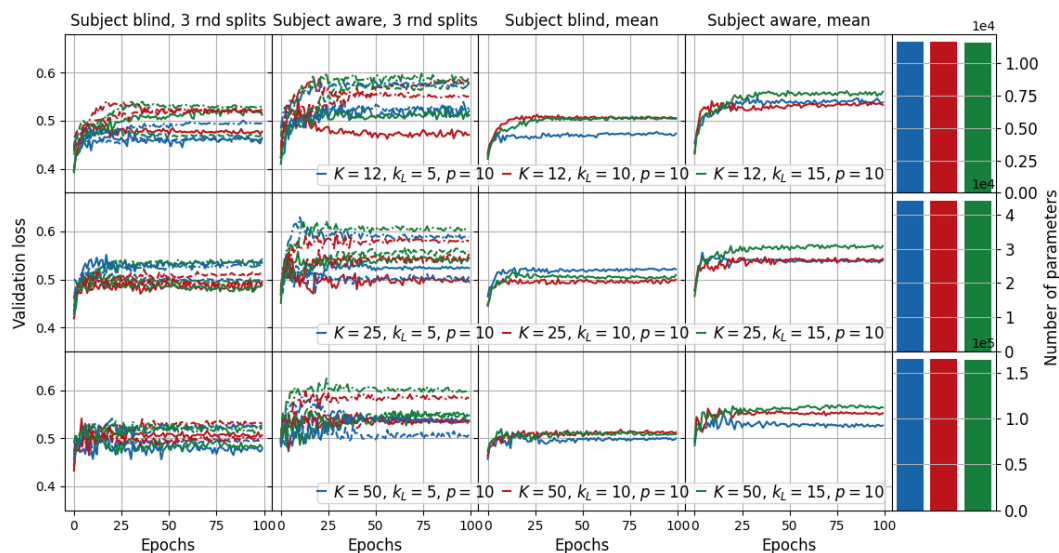


Figure D.8: *ShallowConvNet* validation curves for *subject blind* training and *subject aware* training for mental workload EEG classification problem. Learning rate 0.0001

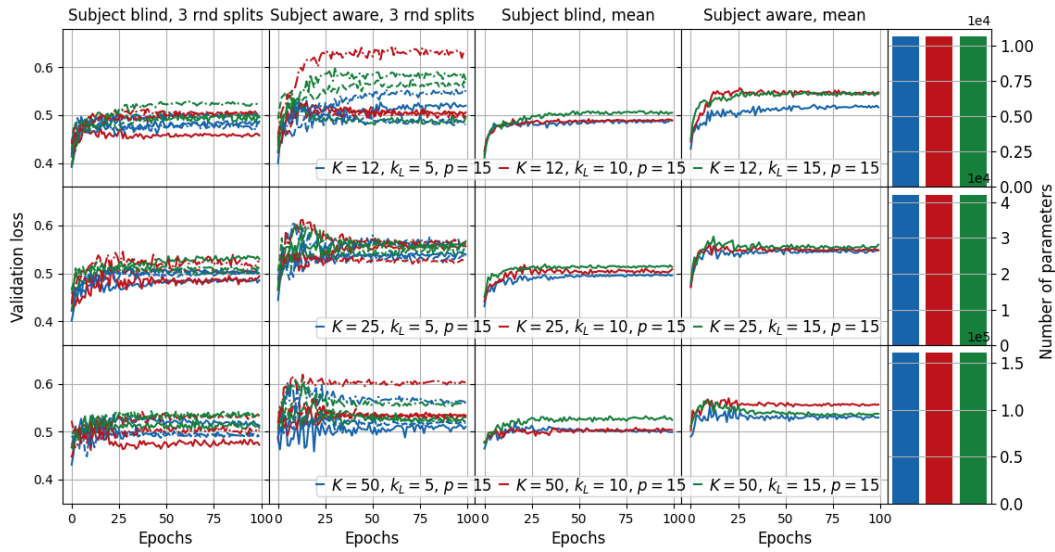


Figure D.9: *ShallowConvNet* validation curves for mental workload EEG classification problem. Learning rate 0.0001

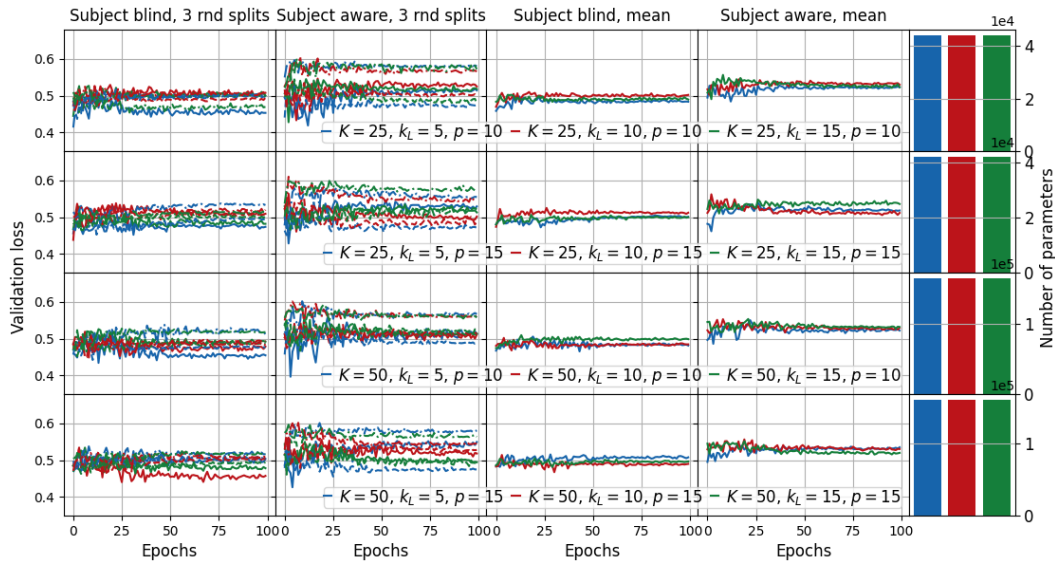


Figure D.10: *ShallowConvNet* validation curves for *subject blind* training and *subject aware* training for mental workload EEG classification problem. Learning rate 0.0005

EEGNet

In this subsection we provide details on *EEGNet* [Lawhern *et al.* 2018] hyperparameter search and the number of multiplications. The number of multiplications per layer are given in Table D.3.

Table D.3: *EEGNet* number of multiplications per different steps of entire classification process for one input sample. T is the input signal length. N is the number of channels of the input signal. K is the number of temporal kernels. k_L is the kernel length. p_1 is the pooling size after the temporal convolution. p_2 is the pooling size after the spatial convolution. Q is the number of output classes.

Operation	Number of multiplications
Temporal correlation	$N \times T \times k_L \times K$
Batch normalization	$2 \times N \times T \times K$
Spatial correlation	$N \times T \times K \times 2K$
Batch normalization	$2 \times T \times 2K$
Exponential Linear Unit	$T \times 2K \times (1 + 3(N_{Ty} - 2))$
Average pool	$\lfloor \frac{T}{p_1} \rfloor \times 2K$
Separable correlation	$\lfloor \frac{T}{p_1} \rfloor \times 16 \times 2K + \lfloor \frac{T}{p_1} \rfloor \times 2K \times 2K$
Batch normalization	$2 \times \lfloor \frac{T}{p_1} \rfloor \times 2K$
Exponential Linear Unit	$\lfloor \frac{T}{p_1} \rfloor \times 2K \times (1 + 3(N_{Ty} - 2))$
Average pool	$\lfloor \frac{\lfloor \frac{T}{p_1} \rfloor}{p_2} \rfloor \times 2K$
Feature classification	$Q \times \lfloor \frac{\lfloor \frac{T}{p_1} \rfloor}{p_2} \rfloor \times 2K + Q \times (1 + 3(N_{Ty} - 2))$

* N_{Ty} corresponds to Taylor series degree used to compute exponential

Illustration of validation curves *EEGNet* models for different hyperparameters are provided in Figures D.11 and D.12 for MEG experiment. In *subject blind* experiment $k_L = 85, p_1 = 2, p_2 = 4, K = 64$. The norm constraint on fully connected layer is 0.5. The number of parameters is 0.088 In *subject aware* experiment $k_L = 85, p_1 = 2, p_2 = 4, K = 80, norm_{rate} = 0.5$. The number of parameters is 0.115. The norm constraint on fully connected layer is 0.5. $dp1$ refers to the standard drop-out operation and $dp2$ to the spatial dropout. Illustration of validation curves *EEGNet* models for different hyperparameters are provided in Figures D.13 and D.14 for EEG experiment. In *subject blind* experiment $K = 32, k_L = 42, p_1 = 2, p_2 = 4, K = 32$. In *subject aware* experiment $K = 32, k_L = 83, p_1 = 2, p_2 = 4$. The norm constraint on fully connected layer is 0.25.

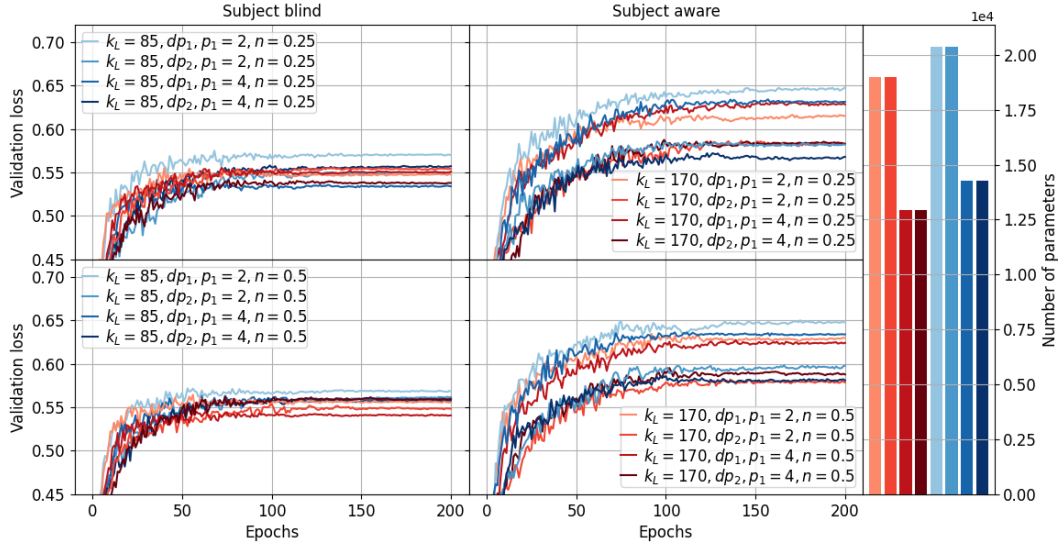


Figure D.11: *EEGNet* validation curves for *subject blind* training (left) and *subject aware* training (right) for motor task MEG classification problem. The curves are illustrated for the norm constraint on fully connected layer 0.25 (default) and 0.5, for different lengths of convolutional filters k_L and different pooling sizes p_1 and $p_2 = 2p_1$ and different dropout approaches (dp_1, dp_2).

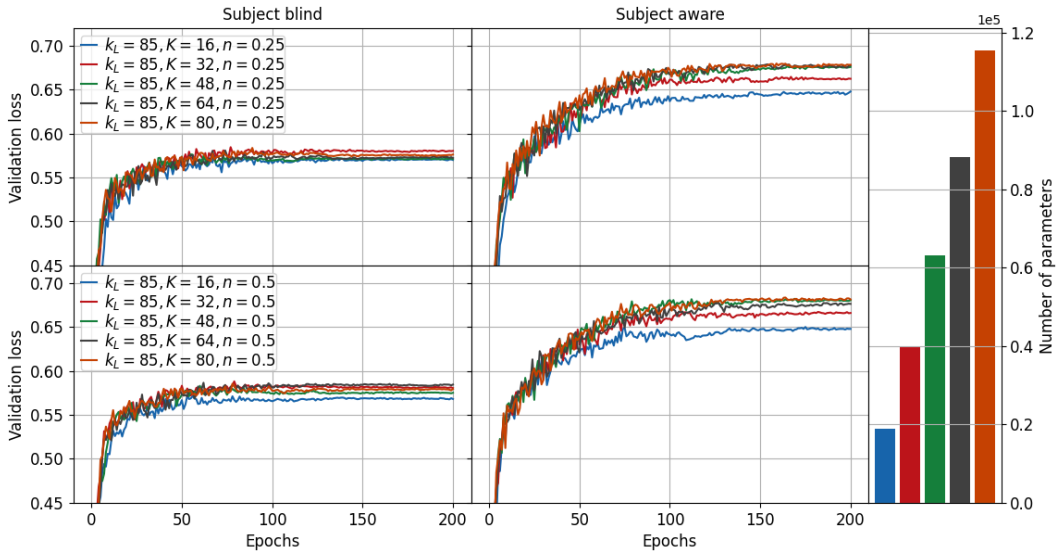


Figure D.12: *EEGNet* validation curves for *subject blind* training (left) and *subject aware* training (right) for motor task MEG classification problem. The curves are illustrated for fixed lengths of convolutional kernels $k_L = 85$, fixed $p_1 = 2, p_2 = 4$ and dropout type dp_1 , and varying number of kernels K .

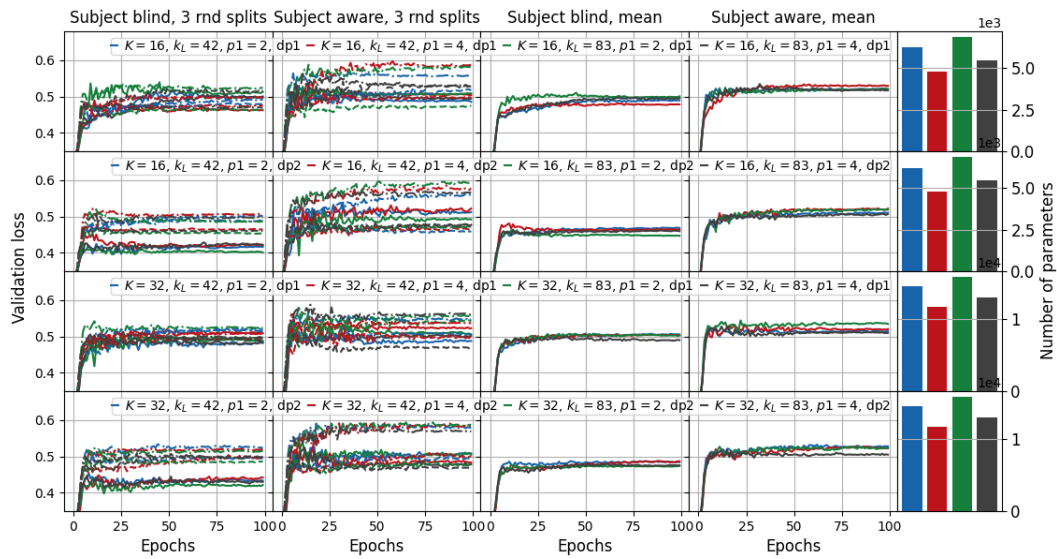


Figure D.13: *EEGNet* validation curves for *subject blind* training and *subject aware* training for mental workload EEG classification problem. The norm constraint on the fully connected layer is 0.25 (default). The curves are illustrated for different lengths of convolutional filters k_L and different pooling sizes p_1 and $p_2 = 2p_1$ and different dropout (dp_1, dp_2) approaches. Learning rate is 0.0005

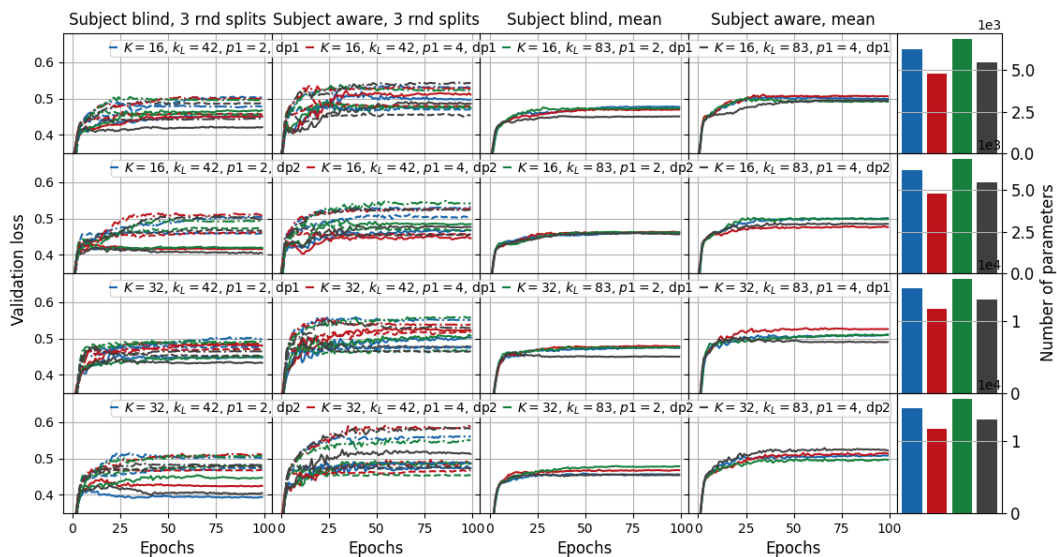


Figure D.14: *EEGNet* validation curves for *subject blind* training and *subject aware* training for mental workload EEG classification problem. The norm constraint on the fully connected layer is 0.25 (default). The curves are illustrated for different lengths of convolutional filters k_L and different pooling sizes p_1 and $p_2 = 2p_1$ and different dropout (dp_1, dp_2) approaches. Learning rate is 0.0001

Bibliography

- [Abadi *et al.* 2015] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu and Xiaoqiang Zheng. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, 2015. Software available from tensorflow.org.
- [Abadi *et al.* 2016] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard *et al.* *Tensorflow: a system for large-scale machine learning*. In OSDI, volume 16, pages 265–283, 2016.
- [Abreu *et al.* 2019] Rodolfo Abreu, Alberto Leal and Patrícia Figueiredo. *Identification of epileptic brain states by dynamic functional connectivity analysis of simultaneous EEG-fMRI: a dictionary learning approach*. Scientific reports, vol. 9, no. 1, pages 1–18, 2019.
- [Acharya *et al.* 2018] U Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan and Hojjat Adeli. *Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals*. Computers in biology and medicine, vol. 100, pages 270–278, 2018.
- [Aggarwal *et al.* 2019] Hemant K Aggarwal, Merry P Mani and Mathews Jacob. *MoDL-MUSSELS: model-based deep learning for multishot sensitivity-encoded diffusion MRI*. IEEE transactions on medical imaging, vol. 39, no. 4, pages 1268–1277, 2019.
- [Aharon *et al.* 2006] Michal Aharon, Michael Elad and Alfred Bruckstein. *K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation*. IEEE Transactions on signal processing, vol. 54, no. 11, pages 4311–4322, 2006.
- [Ahlfors *et al.* 2010] Seppo P Ahlfors, Jooman Han, Fa-Hsuan Lin, Thomas Witzel, John W Belliveau, Matti S Hämäläinen and Eric Halgren. *Cancellation of EEG and MEG signals generated by extended and distributed sources*. Human brain mapping, vol. 31, no. 1, pages 140–149, 2010.

- [Alexander *et al.* 2002] DC Alexander, GJ Barker and SR Arridge. *Detection and modeling of non-Gaussian apparent diffusion coefficient profiles in human brain data*. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, vol. 48, no. 2, pages 331–340, 2002.
- [Allison *et al.* 2007] Brendan Z Allison, Elizabeth Winter Wolpaw and Jonathan R Wolpaw. *Brain-computer interface systems: progress and prospects*. Expert review of medical devices, vol. 4, no. 4, pages 463–474, 2007.
- [Ang *et al.* 2008] Kai Keng Ang, Zheng Yang Chin, Haihong Zhang and Cuntai Guan. *Filter bank common spatial pattern (FBCSP) in brain-computer interface*. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), pages 2390–2397. IEEE, 2008.
- [Antelis *et al.* 2013] Javier M Antelis, Luis Montesano, Ander Ramos-Murguialday, Niels Birbaumer and Javier Minguez. *On the usage of linear regression models to reconstruct limb kinematics from low frequency EEG signals*. PloS one, vol. 8, no. 4, page e61976, 2013.
- [Aricò *et al.* 2018] Pietro Aricò, Gianluca Borghini, Gianluca Di Flumeri, Nicolina Sciaraffa and Fabio Babiloni. *Passive BCI beyond the lab: current trends and future directions*. Physiological measurement, vol. 39, no. 8, page 08TR02, 2018.
- [Asadzadeh *et al.* 2020] Shiva Asadzadeh, Tohid Yousefi Rezaii, Soosan Beheshti, Azra Delpak and Saeed Meshgini. *A systematic review of EEG source localization techniques and their applications on diagnosis of brain abnormalities*. Journal of neuroscience methods, vol. 339, page 108740, 2020.
- [Assaf & Basser 2005] Yaniv Assaf and Peter J Basser. *Composite hindered and restricted model of diffusion (CHARMED) MR imaging of the human brain*. Neuroimage, vol. 27, no. 1, pages 48–58, 2005.
- [Assaf *et al.* 2004] Yaniv Assaf, Raisa Z Freidlin, Gustavo K Rohde and Peter J Basser. *New modeling and experimental framework to characterize hindered and restricted water diffusion in brain white matter*. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, vol. 52, no. 5, pages 965–978, 2004.
- [Aurlien *et al.* 2004] H Aurlien, IO Gjerde, JH Aarseth, G Eldøen, B Karlsen, H Skeidsvoll and NE Gilhus. *EEG background activity described by a large computerized database*. Clinical Neurophysiology, vol. 115, no. 3, pages 665–673, 2004.
- [Azevedo *et al.* 2009] Frederico AC Azevedo, Ludmila RB Carvalho, Lea T Grinberg, José Marcelo Farfel, Renata EL Ferretti, Renata EP Leite, Wilson Jacob Filho, Roberto Lent and Suzana Herculano-Houzel. *Equal numbers*

- of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain.* *Journal of Comparative Neurology*, vol. 513, no. 5, pages 532–541, 2009.
- [Banerjee *et al.* 2019] Monami Banerjee, Rudrasis Chakraborty, Derek Archer, David Vaillancourt and Baba C Vemuri. *DMR-CNN: A CNN Tailored For DMR Scans With Applications To PD Classification.* In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pages 388–391. IEEE, 2019.
- [Barachant *et al.* 2010] Alexandre Barachant, Stéphane Bonnet, Marco Congedo and Christian Jutten. *Riemannian geometry applied to BCI classification.* In International conference on latent variable analysis and signal separation, pages 629–636. Springer, 2010.
- [Barthélemy *et al.* 2012] Quentin Barthélemy, Anthony Larue, Aurélien Mayoue, David Mercier and Jérôme I Mars. *Shift & 2D rotation invariant sparse coding for multivariate signals.* *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pages 1597–1611, 2012.
- [Barthélemy *et al.* 2013] Quentin Barthélemy, Cedric Gouy-Pailler, Yoann Isaac, Antoine Souloumiac, Anthony Larue and Jérôme I Mars. *Multivariate temporal dictionary learning for EEG.* *Journal of neuroscience methods*, vol. 215, no. 1, pages 19–28, 2013.
- [Bashivan *et al.* 2015] Pouya Bashivan, Irina Rish, Mohammed Yeasin and Noel Codella. *Learning representations from EEG with deep recurrent-convolutional neural networks.* arXiv preprint arXiv:1511.06448, 2015.
- [Basser *et al.* 1994] Peter J Basser, James Mattiello and Denis LeBihan. *MR diffusion tensor spectroscopy and imaging.* *Biophysical journal*, vol. 66, no. 1, pages 259–267, 1994.
- [Basser *et al.* 2000] Peter J Basser, Sinisa Pajevic, Carlo Pierpaoli, Jeffrey Duda and Akram Aldroubi. *In vivo fiber tractography using DT-MRI data.* *Magnetic resonance in medicine*, vol. 44, no. 4, pages 625–632, 2000.
- [Bastiani *et al.* 2019] Matteo Bastiani, Jesper LR Andersson, Lucilio Cordero-Grande, Maria Murgasova, Jana Hutter, Anthony N Price, Antonios Makropoulos, Sean P Fitzgibbon, Emer Hughes, Daniel Rueckert *et al.* *Automated processing pipeline for neonatal diffusion MRI in the developing Human Connectome Project.* *NeuroImage*, vol. 185, pages 750–763, 2019.
- [Behrens *et al.* 2003] Timothy EJ Behrens, Mark W Woolrich, Mark Jenkinson, Heidi Johansen-Berg, Rita G Nunes, Stuart Clare, Paul M Matthews, J Michael Brady and Stephen M Smith. *Characterization and propagation of uncertainty in diffusion-weighted MR imaging.* *Magnetic Resonance in*

- Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, vol. 50, no. 5, pages 1077–1088, 2003.
- [Behrens *et al.* 2007] Timothy EJ Behrens, H Johansen Berg, Saad Jbabdi, Matthew FS Rushworth and Mark W Woolrich. *Probabilistic diffusion tractography with multiple fibre orientations: What can we gain?* neuroimage, vol. 34, no. 1, pages 144–155, 2007.
- [Belaoucha *et al.* 2015] Brahim Belaoucha, Jean-Marc Lina, Maureen Clerc and Théodore Papadopoulo. *MEM-diffusion MRI framework to solve MEEG inverse problem*. In 2015 23rd European Signal Processing Conference (EU-SIPCO), pages 1875–1879. IEEE, 2015.
- [Bhattacharyya *et al.* 2010] Saugat Bhattacharyya, Anwesha Khasnobish, Somsirsa Chatterjee, Amit Konar and DN Tibarewala. *Performance analysis of LDA, QDA and KNN algorithms in left-right limb movement classification from EEG data*. In 2010 International conference on systems in medicine and biology, pages 126–131. IEEE, 2010.
- [Binder 2015] Jeffrey R Binder. *The Wernicke area: Modern evidence and a reinterpretation*. Neurology, vol. 85, no. 24, pages 2170–2175, 2015.
- [Bloch 1946] Felix Bloch. *Nuclear induction*. Physical review, vol. 70, no. 7-8, page 460, 1946.
- [Blumensath & Davies 2008] Thomas Blumensath and Mike E Davies. *Iterative thresholding for sparse approximations*. Journal of Fourier analysis and Applications, vol. 14, no. 5-6, pages 629–654, 2008.
- [Blumensath & Davies 2009] Thomas Blumensath and Mike E Davies. *Iterative hard thresholding for compressed sensing*. Applied and computational harmonic analysis, vol. 27, no. 3, pages 265–274, 2009.
- [Bowyer *et al.* 2020] Susan M Bowyer, Andrew Zillgitt, Margaret Greenwald and Renee Lajiness-O’Neill. *Language mapping with magnetoencephalography: an update on the current state of clinical research and practice with considerations for clinical practice guidelines*. Journal of Clinical Neurophysiology, vol. 37, no. 6, pages 554–563, 2020.
- [Broad *et al.* 2018] Rebecca J Broad, Matt C Gabel, Nicholas G Dowell, David J Schwartzman, Anil K Seth, Hui Zhang, Daniel C Alexander, Mara Cercignani and P Nigel Leigh. *Neurite orientation and dispersion density imaging (NODDI) detects cortical and corticospinal tract degeneration in ALS*. J Neurol Neurosurg Psychiatry, pages jnnp–2018, 2018.
- [Brown 1828] Robert Brown. A brief account of microscopical observations made... on the particles contained in the pollen of plants, and on the general existence of active molecules in organic and inorganic bodies. 1828.

- [Bubb *et al.* 2018] Emma J Bubb, Claudia Metzler-Baddeley and John P Aggleton. *The cingulum bundle: anatomy, function, and dysfunction*. Neuroscience & Biobehavioral Reviews, vol. 92, pages 104–127, 2018.
- [Callaghan *et al.* 1988] Paul T Callaghan, CD Eccles and Y Xia. *NMR microscopy of dynamic displacements: k-space and q-space imaging*. Journal of Physics E: Scientific Instruments, vol. 21, no. 8, page 820, 1988.
- [Callaghan 1993] Paul T Callaghan. Principles of nuclear magnetic resonance microscopy. Oxford University Press on Demand, 1993.
- [Candès *et al.* 2006] Emmanuel J Candès, Justin Romberg and Terence Tao. *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*. IEEE Transactions on information theory, vol. 52, no. 2, pages 489–509, 2006.
- [Caruyer *et al.* 2013] Emmanuel Caruyer, Christophe Lenglet, Guillermo Sapiro and Rachid Deriche. *Design of multishell sampling schemes with uniform coverage in diffusion MRI*. Magnetic resonance in medicine, vol. 69, no. 6, pages 1534–1540, 2013.
- [Cheplygina *et al.* 2019] Veronika Cheplygina, Marleen de Bruijne and Josien PW Pluim. *Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis*. Medical image analysis, vol. 54, pages 280–296, 2019.
- [Cohen *et al.* 2018] Taco S Cohen, Mario Geiger, Jonas Köhler and Max Welling. *Spherical CNNs*. arXiv preprint arXiv:1801.10130, 2018.
- [Collobert *et al.* 2002] Ronan Collobert, Samy Bengio and Johnny Mariéthoz. *Torch: a modular machine learning software library*. Technical report, Idiap, 2002.
- [Conner *et al.* 2018] Andrew K Conner, Robert G Briggs, Goksel Sali, Meherzad Rahimi, Cordell M Baker, Joshua D Burks, Chad A Glenn, James D Battiste and Michael E Sughrue. *A connectomic atlas of the human cerebrum—chapter 13: tractographic description of the inferior fronto-occipital fasciculus*. Operative Neurosurgery, vol. 15, no. suppl_1, pages S436–S443, 2018.
- [da Silva 2013] Fernando Lopes da Silva. *EEG and MEG: relevance to neuroscience*. Neuron, vol. 80, no. 5, pages 1112–1128, 2013.
- [de Beeck & Nakatani 2019] Hans Op de Beeck and Chie Nakatani. Introduction to human neuroimaging. Cambridge University Press, 2019.
- [De Santis *et al.* 2017] Silvia De Santis, Tobias Granberg, Russell Ouellette, Constantina A Treaba, Qiuyun Fan, Elena Herranz, Caterina Mainero and Nicola

- Toschi. *Early axonal damage in normal appearing white matter in multiple sclerosis: Novel insights from multi-shell diffusion MRI*. In Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE, pages 3024–3027. IEEE, 2017.
- [de Vico Fallani *et al.* 2014] Fabrizio de Vico Fallani, Jonas Richiardi, Mario Chavez and Sophie Achard. *Graph analysis of functional brain networks: practical issues in translational neuroscience*. Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 369, no. 1653, page 20130521, 2014.
- [Deriche 2016] Rachid Deriche. *Computational brain connectivity mapping: A core health and scientific challenge*, 2016.
- [Descoteaux *et al.* 2007] Maxime Descoteaux, Elaine Angelino, Shaun Fitzgibbons and Rachid Deriche. *Regularized, fast, and robust analytical Q-ball imaging*. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, vol. 58, no. 3, pages 497–510, 2007.
- [Descoteaux *et al.* 2014] Maxime Descoteaux, Cyril Poupon, D Belvic and K Belvic. *Comprehensive biomedical physics*. Comprehensive Biomedical Physics, pages 81–97, 2014.
- [Driscoll & Healy 1994] James R Driscoll and Dennis M Healy. *Computing Fourier transforms and convolutions on the 2-sphere*. Advances in applied mathematics, vol. 15, no. 2, pages 202–250, 1994.
- [Dunkley *et al.* 2015] BT Dunkley, L Da Costa, A Bethune, R Jetly, EW Pang, MJ Taylor and SM Doesburg. *Low-frequency connectivity is associated with mild traumatic brain injury*. NeuroImage: Clinical, vol. 7, pages 611–621, 2015.
- [Durka *et al.* 2005] Piotr J Durka, Artur Matysiak, Eduardo Martínez Montes, Pedro Valdés Sosa and Katarzyna J Blinowska. *Multichannel matching pursuit and EEG inverse solutions*. Journal of neuroscience methods, vol. 148, no. 1, pages 49–59, 2005.
- [Efron *et al.* 2004] Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani. *Least angle regression*. The Annals of statistics, vol. 32, no. 2, pages 407–499, 2004.
- [Eichenbaum *et al.* 1993] Howard Eichenbaum *et al.* Memory, amnesia, and the hippocampal system. MIT press, 1993.
- [Eichert *et al.* 2019] Nicole Eichert, Lennart Verhagen, Davide Folloni, Saad Jbabdi, Alexandre A Khrapitchev, Nicola R Sibson, Dante Mantini, Jerome Sallet and Rogier B Mars. *What is special about the human arcuate fasciculus?*

- Lateralization, projections, and expansion.* Cortex, vol. 118, pages 107–115, 2019.
- [Elad & Aharon 2006] Michael Elad and Michal Aharon. *Image denoising via sparse and redundant representations over learned dictionaries.* IEEE Transactions on Image processing, vol. 15, no. 12, pages 3736–3745, 2006.
- [Elaldi *et al.* 2021] Axel Elaldi, Neel Dey, Heejong Kim and Guido Gerig. *Equivariant Spherical Deconvolution: Learning Sparse Orientation Distribution Functions from Spherical Data.* arXiv preprint arXiv:2102.09462, 2021.
- [Erickson *et al.* 2017] Bradley J Erickson, Panagiotis Korfiatis, Zeynettin Akkus and Timothy L Kline. *Machine learning for medical imaging.* Radiographics, vol. 37, no. 2, page 505, 2017.
- [Esteves *et al.* 2018] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia and Kostas Daniilidis. *Learning so (3) equivariant representations with spherical cnns.* In Proceedings of the European Conference on Computer Vision (ECCV), pages 52–68, 2018.
- [Fenlon *et al.* 2021] Laura R Fenlon, Rodrigo Suarez, Zorana Lynton and Linda J Richards. *The evolution, formation and connectivity of the anterior commissure.* In Seminars in cell & developmental biology. Elsevier, 2021.
- [Fick *et al.* 2016] Rutger HJ Fick, Demian Wassermann, Emmanuel Caruyer and Rachid Deriche. *MAPL: Tissue microstructure estimation using Laplacian-regularized MAP-MRI and its application to HCP data.* NeuroImage, vol. 134, pages 365–385, 2016.
- [Fick *et al.* 2019] Rutger HJ Fick, Demian Wassermann and Rachid Deriche. *The Dmipy Toolbox: Diffusion MRI Multi-Compartment Modeling and Microstructure Recovery Made Easy.* Frontiers in neuroinformatics, vol. 13, page 64, 2019.
- [Fick 1855] Adolf Fick. *Ueber diffusion.* Annalen der Physik, vol. 170, no. 1, pages 59–86, 1855.
- [Filippi *et al.* 2019] Massimo Filippi, Wolfgang Brück, Declan Chard, Franz Fazekas, Jeroen JG Geurts, Christian Enzinger, Simon Hametner, Tanja Kuhlmann, Paolo Preziosa, Àlex Rovira *et al.* *Association between pathological and MRI findings in multiple sclerosis.* The Lancet Neurology, vol. 18, no. 2, pages 198–210, 2019.
- [Foerster 1936] Otfried Foerster. *The motor cortex in man in the light of Hughlings Jackson's doctrines.* Brain, vol. 59, no. 2, pages 135–159, 1936.
- [Fox & Raichle 2007] Michael D Fox and Marcus E Raichle. *Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging.* Nature reviews neuroscience, vol. 8, no. 9, pages 700–711, 2007.

- [Garcia-Cardona & Wohlberg 2018] Cristina Garcia-Cardona and Brendt Wohlberg. *Convolutional dictionary learning: A comparative review and new algorithms*. IEEE Transactions on Computational Imaging, vol. 4, no. 3, pages 366–381, 2018.
- [Gevins *et al.* 1997] Alan Gevins, Michael E Smith, Linda McEvoy and Daphne Yu. *High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice*. Cerebral cortex (New York, NY: 1991), vol. 7, no. 4, pages 374–385, 1997.
- [Golkov *et al.* 2016] Vladimir Golkov, Alexey Dosovitskiy, Jonathan I Sperl, Marion I Menzel, Michael Czisch, Philipp Sämann, Thomas Brox and Daniel Cremers. *Q-space deep learning: twelve-fold shorter and model-free diffusion MRI scans*. IEEE transactions on medical imaging, vol. 35, no. 5, pages 1344–1351, 2016.
- [Gordillo *et al.* 2013] Nelly Gordillo, Eduard Montseny and Pilar Sobrevilla. *State of the art survey on MRI brain tumor segmentation*. Magnetic resonance imaging, vol. 31, no. 8, pages 1426–1438, 2013.
- [Gorski *et al.* 2005] Krzysztof M Gorski, Eric Hivon, Anthony J Banday, Benjamin D Wandelt, Frode K Hansen, Mstvos Reinecke and Matthias Bartelmann. *HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere*. The Astrophysical Journal, vol. 622, no. 2, page 759, 2005.
- [Gramfort *et al.* 2013a] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen *et al.* *MEG and EEG data analysis with MNE-Python*. Frontiers in neuroscience, page 267, 2013.
- [Gramfort *et al.* 2013b] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen and Matti S. Hämäläinen. *MEG and EEG Data Analysis with MNE-Python*. Frontiers in Neuroscience, vol. 7, no. 267, pages 1–13, 2013.
- [Grave de Peralta Menendez *et al.* 2000] R Grave de Peralta Menendez, SL Gonzalez Andino, S Morand, CM Michel and T Landis. *Imaging the electrical activity of the brain: ELECTRA*. Human brain mapping, vol. 9, no. 1, pages 1–12, 2000.
- [Gribonval 2003] Rémi Gribonval. *Piecewise linear source separation*. In Wavelets: Applications in Signal and Image Processing X, volume 5207, pages 297–310. SPIE, 2003.

- [Grienberger *et al.* 2015] Christine Grienberger, Xiaowei Chen and Arthur Konnerth. *Dendritic function in vivo*. Trends in neurosciences, vol. 38, no. 1, pages 45–54, 2015.
- [Grooms *et al.* 2017] Joshua K Grooms, Garth J Thompson, Wen-Ju Pan, Jacob Billings, Eric H Schumacher, Charles M Epstein and Shella D Keilholz. *Infraslow electroencephalographic and dynamic resting state network activity*. Brain connectivity, vol. 7, no. 5, pages 265–280, 2017.
- [Grunwald *et al.* 1999] Martin Grunwald, Thomas Weiss, Werner Krause, Lothar Beyer, Reinhard Rost, Ingmar Gutberlet and Hermann-Josef Gertz. *Power of theta waves in the EEG of human subjects increases during recall of haptic information*. Neuroscience Letters, vol. 260, no. 3, pages 189–192, 1999.
- [Hämäläinen & Ilmoniemi 1994] Matti S Hämäläinen and Risto J Ilmoniemi. *Interpreting magnetic fields of the brain: minimum norm estimates*. Medical & biological engineering & computing, vol. 32, no. 1, pages 35–42, 1994.
- [Hämäläinen *et al.* 1993] Matti Hämäläinen, Riitta Hari, Risto J Ilmoniemi, Jukka Knuutila and Olli V Lounasmaa. *Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain*. Reviews of modern Physics, vol. 65, no. 2, page 413, 1993.
- [Hamner *et al.* 2011] Benjamin Hamner, Ricardo Chavarriaga and Jose del R Millán. *Learning dictionaries of spatial and temporal EEG primitives for brain-computer interfaces*. In Workshop on Structured Sparsity: Learning and Inference, ICML 2011, number CONF, 2011.
- [Hari & Puce 2017] Riitta Hari and Aina Puce. Meg-eeG primer. Oxford University Press, 2017.
- [Herbet *et al.* 2018] Guillaume Herbet, Ilyess Zemmoura and Hugues Duffau. *Functional anatomy of the inferior longitudinal fasciculus: from historical reports to current hypotheses*. Frontiers in neuroanatomy, vol. 12, page 77, 2018.
- [Herculano-Houzel 2012] Suzana Herculano-Houzel. *The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost*. Proceedings of the National Academy of Sciences, vol. 109, no. Supplement 1, pages 10661–10668, 2012.
- [Herman *et al.* 2008] Pawel Herman, Girijesh Prasad, Thomas Martin McGinnity and Damien Coyle. *Comparative analysis of spectral approaches to feature extraction for EEG-based motor imagery classification*. IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 16, no. 4, pages 317–326, 2008.

- [Hess *et al.* 2006] Christopher P Hess, Pratik Mukherjee, Eric T Han, Duan Xu and Daniel B Vigneron. *Q-ball reconstruction of multimodal fiber orientations using the spherical harmonic basis*. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 56, no. 1, pages 104–117, 2006.
- [Hinss *et al.* 2021] Marcel F Hinss, Bertille Somon, Frédéric Dehais and Raphaëlle N Roy. *Open EEG Datasets for Passive Brain-Computer Interface Applications: Lacks and Perspectives*. In 2021 10th International IEEE/EMBS Conference on Neural Engineering (NER), pages 686–689. IEEE, 2021.
- [Hitziger *et al.* 2017] Sebastian Hitziger, Maureen Clerc, Sandrine Saillet, Christian Bénar and Théodore Papadopoulos. *Adaptive waveform learning: a framework for modeling variability in neurophysiological signals*. *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pages 4324–4338, 2017.
- [Homeier & Steinborn 1996] Herbert HH Homeier and E Otto Steinborn. *Some properties of the coupling coefficients of real spherical harmonics and their relation to Gaunt coefficients*. *Journal of Molecular Structure: THEOCHEM*, vol. 368, pages 31–37, 1996.
- [Hong *et al.* 2019] Yoonmi Hong, Geng Chen, Pew-Thian Yap and Dinggang Shen. *Multifold acceleration of diffusion MRI via deep learning reconstruction from slice-undersampled data*. In *International Conference on Information Processing in Medical Imaging*, pages 530–541. Springer, 2019.
- [Hyvärinen & Oja 2000] Aapo Hyvärinen and Erkki Oja. *Independent component analysis: algorithms and applications*. *Neural networks*, vol. 13, no. 4-5, pages 411–430, 2000.
- [Ioffe & Szegedy 2015] Sergey Ioffe and Christian Szegedy. *Batch normalization: Accelerating deep network training by reducing internal covariate shift*. In *International conference on machine learning*, pages 448–456. PMLR, 2015.
- [Iskan *et al.* 2011] Zafer Iskan, Zümray Dokur and Tamer Demiralp. *Classification of electroencephalogram signals with combined time and frequency features*. *Expert Systems with Applications*, vol. 38, no. 8, pages 10499–10505, 2011.
- [Jbabdi *et al.* 2015] Saad Jbabdi, Stamatios N Sotiropoulos, Suzanne N Haber, David C Van Essen and Timothy E Behrens. *Measuring macroscopic brain connections in vivo*. *Nature neuroscience*, vol. 18, no. 11, pages 1546–1555, 2015.
- [Jelescu *et al.* 2015] Ileana O Jelescu, Jelle Veraart, Vitria Adisetiyo, Sarah S Milla, Dmitry S Novikov and Els Fieremans. *One diffusion acquisition and different white matter models: how does microstructure change in human early development based on WMTI and NODDI?* *Neuroimage*, vol. 107, pages 242–256, 2015.

- [Jeurissen *et al.* 2014] Ben Jeurissen, Jacques-Donald Tournier, Thijs Dhollander, Alan Connelly and Jan Sijbers. *Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion MRI data*. NeuroImage, vol. 103, pages 411–426, 2014.
- [Johns 2014] Paul Johns. Clinical neuroscience e-book. Elsevier Health Sciences, 2014.
- [Jones 2010] Derek K Jones. Diffusion mri. Oxford University Press, 2010.
- [Jost *et al.* 2005] Philippe Jost, Pierre Vandergheynst, Sylvain Lesage and Rémi Gribonval. *Learning redundant dictionaries with translation invariance property: the MoTIF algorithm*. In SPARS’05-Workshop on Signal Processing with Adaptive Sparse Structured Representations, pages 1–3, 2005.
- [Jrad *et al.* 2016] Nisrine Jrad, Amar Kachenoura, Isabelle Merlet, Fabrice Bartolomei, Anca Nica, Arnaud Biraben and Fabrice Wendling. *Automatic detection and classification of high-frequency oscillations in depth-EEG signals*. IEEE Transactions on Biomedical Engineering, vol. 64, no. 9, pages 2230–2240, 2016.
- [Kaden *et al.* 2016] Enrico Kaden, Nathaniel D Kelm, Robert P Carson, Mark D Does and Daniel C Alexander. *Multi-compartment microscopic diffusion imaging*. NeuroImage, vol. 139, pages 346–359, 2016.
- [Kamida *et al.* 2016] Akira Kamida, Kenta Shimabayashi, Masayoshi Oguri, Toshihiro Takamori, Naoyuki Ueda, Yuki Koyanagi, Naoko Sannomiya, Haruki Nagira, Saeko Ikunishi, Yuiko Hattori *et al.* *EEG power spectrum analysis in children with ADHD*. Yonago acta medica, vol. 59, no. 2, page 169, 2016.
- [Kauhanen *et al.* 2006] Laura Kauhanen, Tommi Nykopp, Janne Lehtonen, Pasi Jylanki, Jukka Heikkonen, Pekka Rantanen, Hannu Alaranta and Mikko Sams. *EEG and MEG brain-computer interface for tetraplegic patients*. IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 14, no. 2, pages 190–193, 2006.
- [Keller *et al.* 2009] Simon S Keller, Timothy Crow, Anne Foundas, Katrin Amunts and Neil Roberts. *Broca’s area: nomenclature, anatomy, typology and asymmetry*. Brain and language, vol. 109, no. 1, pages 29–48, 2009.
- [Kingma & Ba 2014] Diederik P Kingma and Jimmy Ba. *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980, 2014.
- [Kojčić *et al.* 2021] Ivana Kojčić, Théodore Papadopoulo, Rachid Deriche and Samuel Deslauriers-Gauthier. *Incorporating transmission delays supported by diffusion MRI in MEG source reconstruction*. In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pages 64–68. IEEE, 2021.

- [Kołodziej *et al.* 2012] Marcin Kołodziej, Andrzej Majkowski and Remigiusz J Rak. *Linear discriminant analysis as EEG features reduction technique for brain-computer interfaces*. *Przeład Elektrotechniczny*, vol. 88, no. 3, pages 28–30, 2012.
- [Kondor *et al.* 2018] Risi Kondor, Zhen Lin and Shubhendu Trivedi. *Clebsch-gordan nets: a fully fourier space spherical convolutional neural network*. arXiv preprint arXiv:1806.09231, 2018.
- [Kong & Wang 2012] Shu Kong and Donghui Wang. *A dictionary learning approach for classification: Separating the particularity and the commonality*. In European conference on computer vision, pages 186–199. Springer, 2012.
- [Kostelec & Rockmore 2008] Peter J Kostelec and Daniel N Rockmore. *FFTs on the rotation group*. *Journal of Fourier analysis and applications*, vol. 14, no. 2, pages 145–179, 2008.
- [Kowsky 1986] WS Kowsky. *A quadrature formula over the sphere with application to high resolution spherical harmonic analysis*. *Bull. Gdod*, vol. 60, pages 1–14, 1986.
- [Kreutz-Delgado *et al.* 2003] Kenneth Kreutz-Delgado, Joseph F Murray, Bhaskar D Rao, Kjersti Engan, Te-Won Lee and Terrence J Sejnowski. *Dictionary learning algorithms for sparse representation*. *Neural computation*, vol. 15, no. 2, pages 349–396, 2003.
- [Krusienski *et al.* 2007] Dean J Krusienski, Gerwin Schalk, Dennis J McFarland and Jonathan R Wolpaw. *A μ -Rhythm Matched Filter for Continuous Control of a Brain-Computer Interface*. *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 2, pages 273–280, 2007.
- [Kumar *et al.* 2015] T Sunil Kumar, Vivek Kanhangad and Ram Bilas Pachori. *Classification of seizure and seizure-free EEG signals using local binary patterns*. *Biomedical Signal Processing and Control*, vol. 15, pages 33–40, 2015.
- [La Tour *et al.* 2018] Tom Dupré La Tour, Thomas Moreau, Mainak Jas and Alexandre Gramfort. *Multivariate convolutional sparse coding for electro-magnetic brain signals*. arXiv preprint arXiv:1805.09654, 2018.
- [Lawhern *et al.* 2018] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung and Brent J Lance. *EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces*. *Journal of neural engineering*, vol. 15, no. 5, page 056013, 2018.
- [Le Bihan *et al.* 1986] Denis Le Bihan, Eric Breton, Denis Lallemand, Philippe Grenier, Emmanuel Cabanis and Maurice Laval-Jeantet. *MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders*. *Radiology*, vol. 161, no. 2, pages 401–407, 1986.

- [Le Bihan *et al.* 2006] Denis Le Bihan, Cyril Poupon, Alexis Amadon and Franck Lethimonnier. *Artifacts and pitfalls in diffusion MRI*. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine, vol. 24, no. 3, pages 478–488, 2006.
- [Lee *et al.* 2006] Po-Lei Lee, Jen-Chuen Hsieh, Chi-Hsun Wu, Kuo-Kai Shyu, Shyan-Shiou Chen, Tzu-Chen Yeh and Yu-Te Wu. *The brain computer interface using flash visual evoked potential and independent component analysis*. Annals of biomedical engineering, vol. 34, no. 10, pages 1641–1654, 2006.
- [Lenroot & Giedd 2006] Rhoshel K Lenroot and Jay N Giedd. *Brain development in children and adolescents: insights from anatomical magnetic resonance imaging*. Neuroscience & biobehavioral reviews, vol. 30, no. 6, pages 718–729, 2006.
- [Lin *et al.* 2019] Zhichao Lin, Ting Gong, Kewen Wang, Zhiwei Li, Hongjian He, Qiqi Tong, Feng Yu and Jianhui Zhong. *Fast learning of fiber orientation distribution function for MR tractography using convolutional neural network*. Medical physics, vol. 46, no. 7, pages 3101–3116, 2019.
- [Lindberg *et al.* 2019] Daniel M Lindberg, Nicholas V Stence, Joseph A Grubenhoff, Terri Lewis, David M Mirsky, Angie L Miller, Brent R O’Neill, Kathleen Grice, Peter M Mourani and Desmond K Runyan. *Feasibility and accuracy of fast MRI versus CT for traumatic brain injury in young children*. Pediatrics, vol. 144, no. 4, 2019.
- [Long *et al.* 2015] Jonathan Long, Evan Shelhamer and Trevor Darrell. *Fully convolutional networks for semantic segmentation*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015.
- [Lotte & Roy 2019] Fabien Lotte and Raphaëlle N Roy. *Brain–computer interface contributions to neuroergonomics*. In Neuroergonomics, pages 43–48. Elsevier, 2019.
- [Lotte *et al.* 2007] Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche and Bruno Arnaldi. *A review of classification algorithms for EEG-based brain–computer interfaces*. Journal of neural engineering, vol. 4, no. 2, page R1, 2007.
- [Lotte *et al.* 2018] Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy and Florian Yger. *A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update*. Journal of neural engineering, vol. 15, no. 3, page 031005, 2018.
- [Luo *et al.* 2017] Wei Luo, Jun Li, Jian Yang, Wei Xu and Jian Zhang. *Convolutional sparse autoencoders for image classification*. IEEE transactions on neural networks and learning systems, vol. 29, no. 7, pages 3289–3294, 2017.

- [Madsen *et al.* 2004] Kaj Madsen, Hans Bruun Nielsen and Ole Tingleff. *Methods for non-linear least squares problems*. 2004.
- [Makhzani & Frey 2013] Alireza Makhzani and Brendan Frey. *K-sparse autoencoders*. arXiv preprint arXiv:1312.5663, 2013.
- [Makhzani & Frey 2014] Alireza Makhzani and Brendan Frey. *A winner-take-all method for training sparse convolutional autoencoders*. In NIPS Deep Learning Workshop. Citeseer, 2014.
- [Mallat & Zhang 1993] Stéphane G Mallat and Zhifeng Zhang. *Matching pursuits with time-frequency dictionaries*. IEEE Transactions on signal processing, vol. 41, no. 12, pages 3397–3415, 1993.
- [McEwen & Wiaux 2011] Jason D McEwen and Yves Wiaux. *A novel sampling theorem on the sphere*. IEEE Transactions on Signal Processing, vol. 59, no. 12, pages 5876–5887, 2011.
- [Mehra & Moshirfar 2021] Divy Mehra and Majid Moshirfar. *Neuroanatomy, Optic Tract*. StatPearls [Internet], 2021.
- [Metsomaa *et al.* 2016] Johanna Metsomaa, Jukka Sarvas and Risto Juhani Ilmoniemi. *Blind source separation of event-related EEG/MEG*. IEEE Transactions on Biomedical Engineering, vol. 64, no. 9, pages 2054–2064, 2016.
- [Miller *et al.* 2002] Earl K Miller, David J Freedman and Jonathan D Wallis. *The prefrontal cortex: categories, concepts and cognition*. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, vol. 357, no. 1424, pages 1123–1136, 2002.
- [Milner & Goodale 2006] David Milner and Mel Goodale. *The visual brain in action*, volume 27. OUP Oxford, 2006.
- [Minaee *et al.* 2018] Shervin Minaee, Yao Wang, Anna Choromanska, Sohae Chung, Xiuyuan Wang, Els Fieremans, Steven Flanagan, Joseph Rath and Yvonne W Lui. *A deep unsupervised learning approach toward MTBI identification using diffusion MRI*. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 1267–1270. IEEE, 2018.
- [Mitchell *et al.* 1945] HH Mitchell, TS Hamilton, FR Steggerda and HW Bean. *The chemical composition of the adult human body and its bearing on the biochemistry of growth*. Journal of Biological Chemistry, vol. 158, no. 3, pages 625–637, 1945.
- [Moreau *et al.* 2018] Thomas Moreau, Laurent Oudre and Nicolas Vayatis. *Dicod: Distributed convolutional coordinate descent for convolutional sparse coding*. In International Conference on Machine Learning, pages 3626–3634. PMLR, 2018.

- [Morell & Quarles 1999] Pierre Morell and Richard H Quarles. *The myelin sheath*. Basic Neurochemistry: Molecular, Cellular and Medical Aspects, vol. 6, 1999.
- [Mosher & Leahy 1998] John C Mosher and Richard M Leahy. *Recursive MUSIC: a framework for EEG and MEG source localization*. IEEE Transactions on Biomedical Engineering, vol. 45, no. 11, pages 1342–1354, 1998.
- [Mosher *et al.* 1992] John C Mosher, Paul S Lewis and Richard M Leahy. *Multiple dipole modeling and localization from spatio-temporal MEG data*. IEEE transactions on biomedical engineering, vol. 39, no. 6, pages 541–557, 1992.
- [Müller *et al.* 2021] Philip Müller, Vladimir Golkov, Valentina Tomassini and Daniel Cremers. *Rotation-Equivariant Deep Learning for Diffusion MRI*. arXiv preprint arXiv:2102.06942, 2021.
- [Nauta *et al.* 2021] Ilse M Nauta, Shanna D Kulik, Lucas C Breedt, Anand JC Eijlers, Eva MM Strijbis, Dirk Bertens, Prejaas Tewarie, Arjan Hillebrand, Cornelis J Stam, Bernard MJ Uitdehaage *et al.* *Functional brain network organization measured with magnetoencephalography predicts cognitive decline in multiple sclerosis*. Multiple Sclerosis Journal, vol. 27, no. 11, pages 1727–1737, 2021.
- [Ning *et al.* 2018] Lipeng Ning, Elisenda Bonet-Carne, Francesco Grussu, Farshid Seppehrband, Enrico Kaden, Jelle Veraart, Stefano B Blumberg, Can Son Khoo, Marco Palombo, Jaume Coll-Font *et al.* *Muti-shell diffusion MRI harmonisation and enhancement challenge (MUSHAC): progress and results*. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 217–224. Springer, 2018.
- [Nocedal & Wright 2006] Jorge Nocedal and Stephen Wright. Numerical optimization. Springer Science & Business Media, 2006.
- [Olshausen & Field 1997] Bruno A Olshausen and David J Field. *Sparse coding with an overcomplete basis set: A strategy employed by V1?* Vision research, vol. 37, no. 23, pages 3311–3325, 1997.
- [Orrison *et al.* 2017] William W Orrison, Jeffrey Lewine, John Sanders and Michael F Hartshorne. Functional brain imaging. Elsevier Health Sciences, 2017.
- [O’Shea & Nash 2015] Keiron O’Shea and Ryan Nash. *An introduction to convolutional neural networks*. arXiv preprint arXiv:1511.08458, 2015.
- [Özarslan *et al.* 2013] Evren Özarslan, Cheng Guan Koay, Timothy M Shepherd, Michal E Komlosh, M Okan İrfanoğlu, Carlo Pierpaoli and Peter J Basser. *Mean apparent propagator (MAP) MRI: a novel diffusion imaging method for mapping tissue microstructure*. NeuroImage, vol. 78, pages 16–32, 2013.

- [Panagiotaki *et al.* 2014] Eletheria Panagiotaki, Simon Walker-Samuel, Bernard Siow, S Peter Johnson, Vineeth Rajkumar, R Barbara Pedley, Mark F Lythgoe and Daniel C Alexander. *Noninvasive quantification of solid tumor microstructure using VERDICT MRI*. *Cancer research*, vol. 74, no. 7, pages 1902–1912, 2014.
- [Pang *et al.* 2021] Liping Pang, Liang Guo, Jie Zhang, Xiaoru Wanyan, Hongquan Qu and Xin Wang. *Subject-specific mental workload classification using EEG and stochastic configuration network (SCN)*. *Biomedical Signal Processing and Control*, vol. 68, page 102711, 2021.
- [Papageorgakis *et al.* 2017] Christos Papageorgakis, Sebastian Hitziger and Théodore Papadopoulo. *Dictionary learning for multidimensional data*. In *Proceedings of GRETSI 2017*, 2017.
- [Pascual-Marqui *et al.* 1994] Roberto D Pascual-Marqui, Christoph M Michel and Dietrich Lehmann. *Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain*. *International Journal of psychophysiology*, vol. 18, no. 1, pages 49–65, 1994.
- [Pati *et al.* 1993] Yagyensh Chandra Pati, Ramin Rezaifar and Perinkulam Sambamurthy Krishnaprasad. *Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition*. In *Proceedings of 27th Asilomar conference on signals, systems and computers*, pages 40–44. IEEE, 1993.
- [Penfield & Rasmussen 1950] Wilder Penfield and Theodore Rasmussen. *The cerebral cortex of man; a clinical study of localization of function*. 1950.
- [Peng *et al.* 2021] Hong Peng, Cancheng Li, Jinlong Chao, Tao Wang, Chengjian Zhao, Xiaoning Huo and Bin Hu. *A novel automatic classification detection for epileptic seizure based on dictionary learning and sparse representation*. *Neurocomputing*, vol. 424, pages 179–192, 2021.
- [Peters *et al.* 1976] A Peters, SL Palay and H Webster. *DeF. The Fine Structure of the Nervous System: The Neurons and Supporting Cells*, pages 162–166, 1976.
- [Peterson *et al.* 2018] Diana C Peterson, Vamsi Reddy and Renee N Hamel. *Neuroanatomy, auditory pathway*. 2018.
- [Pezoulas *et al.* 2020] Vasileios Pezoulas, Themis Exarchos and Dimitrios I Fotiadis. *Medical data sharing, harmonization and analytics*. Academic Press, 2020.
- [Pickles 1998] James Pickles. *An introduction to the physiology of hearing*. Brill, 1998.

- [Pineda *et al.* 2000] Jaime A Pineda, BZ Allison and A Vankov. *The effects of self-movement, observation, and imagination on/spl mu/rhythms and readiness potentials (RP's): toward a brain-computer interface (BCI)*. IEEE Transactions on Rehabilitation Engineering, vol. 8, no. 2, pages 219–222, 2000.
- [Pittau & Vulliemoz 2015] Francesca Pittau and Serge Vulliemoz. *Functional brain networks in epilepsy: recent advances in noninvasive mapping*. Current opinion in neurology, vol. 28, no. 4, pages 338–343, 2015.
- [Ramirez *et al.* 2010] Ignacio Ramirez, Pablo Sprechmann and Guillermo Sapiro. *Classification and clustering via dictionary learning with structured incoherence and shared features*. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 3501–3508. IEEE, 2010.
- [Rea 2015] Paul Rea. *Essential clinical anatomy of the nervous system*. Academic Press, 2015.
- [Rehman & Al Khalili 2019] Amna Rehman and Yasir Al Khalili. *Neuroanatomy, occipital lobe*. 2019.
- [Ronneberger *et al.* 2015] Olaf Ronneberger, Philipp Fischer and Thomas Brox. *U-net: Convolutional networks for biomedical image segmentation*. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.
- [Roy & Shukla 2019] Vandana Roy and Shailja Shukla. *Designing efficient blind source separation methods for EEG motion artifact removal based on statistical evaluation*. Wireless Personal Communications, vol. 108, no. 3, pages 1311–1327, 2019.
- [Roy *et al.*] Raphaëlle N Roy, Marcel F Hinss, Ludovic Darmet, Simon Ladouce, Emilie S Jahanpour, Bertille Somon, Xiaoqi Xu, Nicolas Drougard, Frédéric Dehais and Fabien Lotte. *Retrospective on the First Passive Brain-Computer Interface Competition on Cross-Session Workload Estimation*. Frontiers in Neuroergonomics, page 4.
- [Saha & Baumert 2020] Simanto Saha and Mathias Baumert. *Intra-and inter-subject variability in EEG-based sensorimotor brain computer interface: a review*. Frontiers in computational neuroscience, page 87, 2020.
- [Sarvas 1987] Jukka Sarvas. *Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem*. Physics in Medicine & Biology, vol. 32, no. 1, page 11, 1987.
- [Schacter 1977] Daniel L Schacter. *EEG theta waves and psychological phenomena: A review and analysis*. Biological psychology, vol. 5, no. 1, pages 47–82, 1977.

- [Schirrmeyer *et al.* 2017] Robin Tibor Schirrmeyer, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard and Tonio Ball. *Deep learning with convolutional neural networks for EEG decoding and visualization*. Human brain mapping, vol. 38, no. 11, pages 5391–5420, 2017.
- [Schneider *et al.* 2017] Torben Schneider, Wallace Brownlee, Hui Zhang, Olga Ciccarelli, David H Miller and Claudia Gandini Wheeler-Kingshott. *Sensitivity of multi-shell NODDI to multiple sclerosis white matter changes: a pilot study*. Functional neurology, vol. 32, no. 2, page 97, 2017.
- [Sedlar *et al.* 2020] Sara Sedlar, Théodore Papadopoulo, Rachid Deriche and Samuel Deslauriers-Gauthier. *Diffusion MRI fiber orientation distribution function estimation using voxel-wise spherical U-net*. In Computational Diffusion MRI, MICCAI Workshop, 2020.
- [Sedlar *et al.* 2021] Sara Sedlar, Abib Alimi, Théodore Papadopoulo, Rachid Deriche and Samuel Deslauriers-Gauthier. *A spherical convolutional neural network for white matter structure imaging via dMRI*. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 529–539. Springer, 2021.
- [Smith & Webb 2010] Nadine Barrie Smith and Andrew Webb. Introduction to medical imaging: physics, engineering and clinical applications. Cambridge university press, 2010.
- [Smith *et al.* 2004] Stephen M Smith, Mark Jenkinson, Mark W Woolrich, Christian F Beckmann, Timothy EJ Behrens, Heidi Johansen-Berg, Peter R Bannister, Marilena De Luca, Ivana Drobnjak, David E Flitney *et al.* *Advances in functional and structural MR image analysis and implementation as FSL*. Neuroimage, vol. 23, pages S208–S219, 2004.
- [Snell 2010] Richard S Snell. Clinical neuroanatomy. Lippincott Williams & Wilkins, 2010.
- [Solomon *et al.* 2014] Eldra Solomon, Charles Martin, Diana W Martin and Linda R Berg. Biology. Cengage Learning, 2014.
- [Sorrentino *et al.* 2009] Alberto Sorrentino, Lauri Parkkonen, Annalisa Pascarella, Cristina Campi and Michele Piana. *Dynamical MEG source modeling with multi-target Bayesian filtering*. Human brain mapping, vol. 30, no. 6, pages 1911–1921, 2009.
- [Sors *et al.* 2018] Arnaud Sors, Stéphane Bonnet, Sébastien Mirek, Laurent Vercueil and Jean-François Payen. *A convolutional neural network for sleep stage scoring from raw single-channel EEG*. Biomedical Signal Processing and Control, vol. 42, pages 107–114, 2018.

- [Spielman *et al.* 2020] Rose M Spielman, William Jenkins and Marilyn Lovett. *Psychology 2e*. 2020.
- [Sporns *et al.* 2005] Olaf Sporns, Giulio Tononi and Rolf Kötter. *The human connectome: a structural description of the human brain*. PLoS computational biology, vol. 1, no. 4, page e42, 2005.
- [Sprechmann & Sapiro 2010] Pablo Sprechmann and Guillermo Sapiro. *Dictionary learning and sparse coding for unsupervised clustering*. In 2010 IEEE international conference on acoustics, speech and signal processing, pages 2042–2045. IEEE, 2010.
- [Stam 2010] CJ Stam. *Use of magnetoencephalography (MEG) to study functional brain networks in neurodegenerative disorders*. Journal of the neurological sciences, vol. 289, no. 1-2, pages 128–134, 2010.
- [Standring 2020] Susan Standring. *Gray’s anatomy e-book: the anatomical basis of clinical practice*. Elsevier Health Sciences, 2020.
- [Stejskal & Tanner 1965] Edward O Stejskal and John E Tanner. *Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient*. The journal of chemical physics, vol. 42, no. 1, pages 288–292, 1965.
- [Subasi & Gursoy 2010] Abdulhamit Subasi and M Ismail Gursoy. *EEG signal classification using PCA, ICA, LDA and support vector machines*. Expert systems with applications, vol. 37, no. 12, pages 8659–8666, 2010.
- [Supratak *et al.* 2017] Akara Supratak, Hao Dong, Chao Wu and Yike Guo. *Deep-SleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG*. IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 25, no. 11, pages 1998–2008, 2017.
- [Szafer *et al.* 1995] Aaron Szafer, Jianhui Zhong and John C Gore. *Theoretical model for water diffusion in tissues*. Magnetic resonance in medicine, vol. 33, no. 5, pages 697–712, 1995.
- [Tariq *et al.* 2016] Maira Tariq, Torben Schneider, Daniel C Alexander, Claudia A Gandini Wheeler-Kingshott and Hui Zhang. *Bingham–NODDI: mapping anisotropic orientation dispersion of neurites using diffusion MRI*. Neuroimage, vol. 133, pages 207–223, 2016.
- [Tibshirani 1996] Robert Tibshirani. *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pages 267–288, 1996.
- [Tillmann 2014] Andreas M Tillmann. *On the computational intractability of exact and approximate dictionary learning*. IEEE Signal Processing Letters, vol. 22, no. 1, pages 45–49, 2014.

- [Torrey 1956] Henry C Torrey. *Bloch equations with diffusion terms*. Physical review, vol. 104, no. 3, page 563, 1956.
- [Tošić & Frossard 2011] Ivana Tošić and Pascal Frossard. *Dictionary learning*. IEEE Signal Processing Magazine, vol. 28, no. 2, pages 27–38, 2011.
- [Tournier *et al.* 2004] J-Donald Tournier, Fernando Calamante, David G Gadian and Alan Connelly. *Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution*. Neuroimage, vol. 23, no. 3, pages 1176–1185, 2004.
- [Tournier *et al.* 2007] J-Donald Tournier, Fernando Calamante and Alan Connelly. *Robust determination of the fibre orientation distribution in diffusion MRI: non-negativity constrained super-resolved spherical deconvolution*. Neuroimage, vol. 35, no. 4, pages 1459–1472, 2007.
- [Tournier *et al.* 2011] Jacques-Donald Tournier, Susumu Mori and Alexander Leemans. *Diffusion tensor imaging and beyond*. Magnetic resonance in medicine, vol. 65, no. 6, page 1532, 2011.
- [Tournier *et al.* 2019] J-Donald Tournier, Robert Smith, David Raffelt, Rami Tabbara, Thijs Dhollander, Maximilian Pietsch, Daan Christiaens, Ben Jeurissen, Chun-Hung Yeh and Alan Connelly. *MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation*. NeuroImage, page 116137, 2019.
- [Tsinalis *et al.* 2016] Orestis Tsinalis, Paul M Matthews, Yike Guo and Stefanos Zafeiriou. *Automatic sleep stage scoring with single-channel EEG using convolutional neural networks*. arXiv preprint arXiv:1610.01683, 2016.
- [Tuch 2004] David S Tuch. *Q-ball imaging*. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, vol. 52, no. 6, pages 1358–1372, 2004.
- [Ullah *et al.* 2018] Ihsan Ullah, Muhammad Hussain, Hatim Aboalsamhet *al.* *An automated system for epilepsy detection using EEG brain signals based on deep learning approach*. Expert Systems with Applications, vol. 107, pages 61–71, 2018.
- [van Ede *et al.* 2018] Freek van Ede, Andrew J Quinn, Mark W Woolrich and Anna C Nobre. *Neural oscillations: sustained rhythms or transient burst-events?* Trends in Neurosciences, vol. 41, no. 7, pages 415–417, 2018.
- [Van Essen *et al.* 2012] David C Van Essen, Kamil Ugurbil, Edward Auerbach, Deanna Barch, Timothy EJ Behrens, Richard Bucholz, Acer Chang, Liyong Chen, Maurizio Corbetta, Sandra W Curtisset *al.* *The Human Connectome Project: a data acquisition perspective*. Neuroimage, vol. 62, no. 4, pages 2222–2231, 2012.

- [Van Essen *et al.* 2013] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium *et al.* *The WU-Minn human connectome project: an overview*. *Neuroimage*, vol. 80, pages 62–79, 2013.
- [Vanhatalo *et al.* 2004] Sampsa Vanhatalo, J Matias Palva, MD Holmes, JW Miller, Juha Voipio and Kai Kaila. *Infraslow oscillations modulate excitability and interictal epileptic activity in the human cortex during sleep*. *Proceedings of the National Academy of Sciences*, vol. 101, no. 14, pages 5053–5057, 2004.
- [Vatta *et al.* 2010] Federica Vatta, Fabio Meneghini, Fabrizio Esposito, Stefano Mininell and Francesco Di Salle. *Realistic and spherical head modeling for EEG forward problem solution: a comparative cortex-based analysis*. *Computational intelligence and neuroscience*, vol. 2010, 2010.
- [Vingerhoets 2014] Guy Vingerhoets. *Contribution of the posterior parietal cortex in reaching, grasping, and using objects and tools*. *Frontiers in psychology*, vol. 5, page 151, 2014.
- [Von Der Heide *et al.* 2013] Rebecca J Von Der Heide, Laura M Skipper, Elizabeth Klobusicky and Ingrid R Olson. *Dissecting the uncinate fasciculus: disorders, controversies and a hypothesis*. *Brain*, vol. 136, no. 6, pages 1692–1707, 2013.
- [Wang *et al.* 2010] Huijuan Wang, Linda Douw, J Martin Hernandez, JC Reijneveld, CJ Stam and P Van Mieghem. *Effect of tumor resection on the characteristics of functional brain networks*. *Physical Review E*, vol. 82, no. 2, page 021924, 2010.
- [Wang *et al.* 2016] Xuhui Wang, Sudhir Pathak, Lucia Stefaneanu, Fang-Cheng Yeh, Shiting Li and Juan C Fernandez-Miranda. *Subcomponents and connectivity of the superior longitudinal fasciculus in the human brain*. *Brain Structure and Function*, vol. 221, no. 4, pages 2075–2092, 2016.
- [Watson 2018] Brendon O Watson. *Cognitive and physiologic impacts of the infraslow oscillation*. *Frontiers in systems neuroscience*, vol. 12, page 44, 2018.
- [Wedeen *et al.* 2000] VJ Wedeen, TG Reese, DS Tuch, MR Weigel, JG Dou, RM Weiskoff and D Chessler. *Mapping fiber orientation spectra in cerebral white matter with Fourier-transform diffusion MRI*. In *Proceedings of the 8th Annual Meeting of ISMRM, Denver*, page 82, 2000.
- [Wedeen *et al.* 2005] Van J Wedeen, Patric Hagmann, Wen-Yih Isaac Tseng, Timothy G Reese and Robert M Weisskoff. *Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging*. *Magnetic resonance in medicine*, vol. 54, no. 6, pages 1377–1386, 2005.
- [Wei *et al.* 2007] Qingguo Wei, Yijun Wang, Xiaorong Gao and Shangkai Gao. *Amplitude and phase coupling measures for feature extraction in an EEG-based*

- brain-computer interface*. Journal of neural engineering, vol. 4, no. 2, page 120, 2007.
- [Wilkins *et al.* 2015] Bryce Wilkins, Namgyun Lee, Niharika Gajawelli, Meng Law and Natasha Leporé. *Fiber estimation and tractography in diffusion MRI: development of simulated brain images and comparison of multi-fiber analysis methods at clinical b-values*. Neuroimage, vol. 109, pages 341–356, 2015.
- [Wu *et al.* 2016a] Dongrui Wu, Vernon J Lawhern, Stephen Gordon, Brent J Lance and Chin-Teng Lin. *Driver drowsiness estimation from EEG signals using online weighted adaptation regularization for regression (OwARR)*. IEEE Transactions on Fuzzy Systems, vol. 25, no. 6, pages 1522–1535, 2016.
- [Wu *et al.* 2016b] Yupeng Wu, Dandan Sun, Yong Wang and Yibao Wang. *Subcomponents and connectivity of the inferior fronto-occipital fasciculus revealed by diffusion spectrum imaging fiber tracking*. Frontiers in neuroanatomy, vol. 10, page 88, 2016.
- [Ye *et al.* 2012] Wenxing Ye, Sharon Portnoy, Alireza Entezari, Stephen J Blackband and Baba C Vemuri. *An efficient interlaced multi-shell sampling scheme for reconstruction of diffusion propagators*. IEEE transactions on medical imaging, vol. 31, no. 5, pages 1043–1050, 2012.
- [Ye *et al.* 2019] Chuyang Ye, Xiuli Li and Jingnan Chen. *A deep network for tissue microstructure estimation using modified LSTM units*. Medical image analysis, vol. 55, pages 49–64, 2019.
- [Ye *et al.* 2020] Chuyang Ye, Yuxing Li and Xiangzhu Zeng. *An improved deep network for tissue microstructure estimation with uncertainty quantification*. Medical image analysis, vol. 61, page 101650, 2020.
- [Ye 2017] Chuyang Ye. *Estimation of tissue microstructure using a deep network inspired by a sparse reconstruction framework*. In International Conference on Information Processing in Medical Imaging, pages 466–477. Springer, 2017.
- [Yeatman *et al.* 2014] Jason D Yeatman, Kevin S Weiner, Franco Pestilli, Ariel Rokem, Aviv Mezer and Brian A Wandell. *The vertical occipital fasciculus: a century of controversy resolved by in vivo measurements*. Proceedings of the National Academy of Sciences, vol. 111, no. 48, pages E5214–E5223, 2014.
- [Yu *et al.* 2014] Xinyang Yu, Pharino Chum and Kwee-Bo Sim. *Analysis the effect of PCA for feature reduction in non-stationary EEG based motor imagery of BCI system*. Optik, vol. 125, no. 3, pages 1498–1502, 2014.
- [Zander & Kothe 2011] Thorsten O Zander and Christian Kothe. *Towards passive brain-computer interfaces: applying brain-computer interface technology to*

- human-machine systems in general*. Journal of neural engineering, vol. 8, no. 2, page 025005, 2011.
- [Zhang *et al.* 2001] Yongyue Zhang, Michael Brady and Stephen Smith. *Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm*. IEEE transactions on medical imaging, vol. 20, no. 1, pages 45–57, 2001.
- [Zhang *et al.* 2012] Hui Zhang, Torben Schneider, Claudia A Wheeler-Kingshott and Daniel C Alexander. *NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain*. NeuroImage, vol. 61, no. 4, pages 1000–1016, 2012.
- [Zhang *et al.* 2018] Ziheng Zhang, Yanyu Xu, Jingyi Yu and Shenghua Gao. *Saliency detection in 360 videos*. In Proceedings of the European conference on computer vision (ECCV), pages 488–503, 2018.
- [Zhang *et al.* 2021] Fan Zhang, Anna Breger, Kang Ik Kevin Cho, Lipeng Ning, Carl-Fredrik Westin, Lauren J O’Donnell and Ofer Pasternak. *Deep learning based segmentation of brain tissue from diffusion MRI*. NeuroImage, vol. 233, page 117934, 2021.
- [Zhao *et al.* 2019] Fenqiang Zhao, Shunren Xia, Zhengwang Wu, Dingna Duan, Li Wang, Weili Lin, John H Gilmore, Dinggang Shen and Gang Li. *Spherical U-Net on cortical surfaces: methods and applications*. In International Conference on Information Processing in Medical Imaging, pages 855–866. Springer, 2019.
- [Zhou *et al.* 2012] Wei Zhou, Ya Yang and Zhuliang Yu. *Discriminative dictionary learning for EEG signal classification in Brain-computer interface*. In 2012 12th International Conference on Control Automation Robotics & Vision (ICARCV), pages 1582–1585. IEEE, 2012.
- [Zhukov *et al.* 2000] Leonid Zhukov, David Weinstein and Chris Johnson. *Independent component analysis for EEG source localization*. IEEE Engineering in Medicine and Biology Magazine, vol. 19, no. 3, pages 87–96, 2000.
- [Ziegler *et al.* 2014] Erik Ziegler, Sarah L Chellappa, Giulia Gaggioni, Julien QM Ly, Gilles Vandewalle, Elodie André, Christophe Geuzaine and Christophe Phillips. *A finite-element reciprocity solution for EEG forward modeling with realistic individual head models*. NeuroImage, vol. 103, pages 542–551, 2014.
- [Zou *et al.* 2019] Liang Zou, Xun Chen, Ge Dang, Yi Guo and Z Jane Wang. *Removing muscle artifacts from EEG data via underdetermined joint blind source separation: A simulation study*. IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 67, no. 1, pages 187–191, 2019.

-
- [Zucchelli *et al.* 2020] Mauro Zucchelli, Samuel Deslauriers-Gauthier and Rachid Deriche. *A computational Framework for generating rotation invariant features and its application in diffusion MRI*. Medical image analysis, vol. 60, page 101597, 2020.
- [Zucchelli *et al.* 2021] Mauro Zucchelli, Samuel Deslauriers-Gauthier and Rachid Deriche. *Investigating the effect of DMRI signal representation on fully-connected neural networks brain tissue microstructure estimation*. In IEEE International Symposium on Biomedical Imaging (ISBI)(2021), 2021.