TÉCNICO LISBOA

Identification of EEG Fingerprints of Simultaneous fMRI in Resting State and Motor Imagery

Marta Antunes Evangelista Loures Xavier

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors: Professor Patrícia Margarida Piedade Figueiredo

Doctor Athanasios Vourvopoulos

Examination Committee

Chairperson: Prof. Rita Homem de Gouveia Costanzo Nunes

Supervisor: Prof. Patrícia Margarida Piedade Figueiredo

Member of the Committee: Doctor Carlos Jorge Andrade Mariz Santiago

October 2019

Preface

The work presented in this thesis was performed at the Evolutionary Systems and Biomedical Engineering Laboratory (LaSEEB) of Instituto superior Técnico (Lisbon, Portugal), during the period March-October 2019, under the supervision of Prof. Patrícia Figueiredo and Athanasios Vourvopoulos.

Declaration

I declare that this document is an original work of my own authorship and that it fulfils all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

I would like to acknowledge my supervisors Prof. Patrícia Figueiredo and Athanasios Vourvopoulos for all their support throughout the development of this Thesis and for the valuable knowledge they provided along the way. I am also grateful to all of the teachers that guided me throughout my academic life, and to my family and friends for all the motivation, support and inspiration.

Resumo

Nos últimos anos, as aquisições simultâneas de EEG-fMRI têm sido extensivamente utilizadas para compreender a relação entre o EEG e o fMRI em diversos contextos experimentais. O uso deste tipo de abordagens multimodais tem vindo a ser reconhecido como uma promissora ferramenta, devendo-se o seu sucesso às caracteristicas complementares das duas modalidades em questão. Porém, estas aquisições multimodais têm geralmente associado um alto custo, são desconfortáveis para os pacientes, e utilizam equipamento não-portátil. Tais desvantagens devem-se sobretudo ao uso de fMRI, pelo que têm motivado a procura de soluções capazes de utilizar somente o EEG como substituto para o EEG-fMRI. Neste contexto, o presente trabalho propõe investigar uma estratégia para a extração de *features* relevantes do EEG e aprendizagem dos respetivos pesos, de forma a prever a resposta BOLD simultânea. Os padrões eletrofisiológicos resultantes podem ser referidos como EEG *Fingerprints* (EFPs), um conceito introduzido por Meir-Hasson e colegas (Meir-Hasson et al., 2014). A metodologia aqui proposta e implementada recorreu a métodos de apresendizagem automática, integrando algoritmos de regressão linear, de regularização e de *cross validation*. Foram levadas a cabo duas análises independentes, uma com dados de *resting state* e outra com dados adquiridos durante um ensaio de *motor imagery*.

Palavras-chave: EEG *Fingerprint*, EEG-fMRI, *Resting State*, *Default Mode Network*, *Motor Imagery*, *Features* Espetrais, *Features* de Conetividade, Aprendizagem, Otimização.

Abstract

Over the past few years, simultaneous EEG-fMRI recordings have been largely used to understand the link between EEG and fMRI in multiple conditions. The use of multimodal approaches that combine the two modalities has received recognition a promising new tool, owing to their highly complementary characteristics. However, such multimodal acquisitions are typically costly, non-portable and overall uncomfortable for patients, mostly due to the use of fMRI. This motivated the search for solutions capable of using only the widely available EEG as a surrogate of the simultaneous BOLD signal. Within this scope, this work investigated an integration strategy whereby relevant EEG-features were extracted and their coefficient estimates learnt so as to predict the simultaneous average BOLD signal measured at a specific distributed network. These network-specific EEG patterns were here referred to as EEG Fingerprints (EFPs), a concept introduced by Meir-Hasson and their colleagues (Meir-Hasson et al., 2014). The methodology employed relied on a machine learning approach that included linear regression algorithms and cross validation procedures. Independent analyses were performed for data recorded under two experimental conditions: during resting state; and during a motor imagery task.

Keywords: EEG Fingerprint, Simultaneous EEG-fMRI, Resting State, Default Mode Network, Motor Imagery, Spectral Features, Connectivity Features, Learning, Optimization.

Contents

List of Tables

List of Figures

Acronyms

ACC Anterior Cingulate Cortex.

ACF Auto-Correlation Function.

ADHD Attention Deficit Hyperactivity Disorder.

AIC Akaike Information Criterion.

ANOVA Analysis of Variance.

AP Action Potential.

ATP Adenosine Triphosphate.

BCI Brain-Computer Interface.

BG Basal Ganglia.

BIC Bayesian Information Criterion.

BOLD Blood Oxygen Level Dependent.

CBF Cerebral Blood Flow.

CBV Cerebral Blood Volume.

CV Cross-Validation.

dHb Deoxyhemoglobin.

DMN Default Mode Network.

DOF Degrees of Freedom.

ECG Electrocardiogram.

ECoG Electrocorticography.

EEG Electroencephalography.

EFP EEG Fingerprint.

ENR Elastic Net Regularization.

EPI Echo-Planar Imaging.

EPSP Excitatory Postsynaptic Potential.

ERD Event-Related Desynchronization.

ERS Event-Related Synchronization.

EV Explanatory Variable.

FID Free-Induction Decay.

fMRI Functional Magnetic Resonance Imaging.

FSL FMRIB's Software Library.

FWHM Full Width at Half Maximum.

GABA Gamma Aminobutyric Acid.

GFS Global Field Synchronization.

GLM General Linear Model.

GRF Gaussian Random Field.

Hb Oxyhemoglobin.

HRF Hemodynamic Response Function.

ICA Independent Component Analysis.

IPC Imaginary Part of Coherency.

IPSP Inhibitory Postsynaptic Potential.

LC Linear Combination of Band-specific EEG Power.

LFP Local Field Potential.

LOOCV Leave-One-Out CV.

LR Lasso Regression.

LS Least Squares.

LSR Least Squares Regression.

LTI Linear Time-Invariant.

M1 Primary Motor Cortex.

MI Motor Imagery.

mPFC medial Prefrontal Cortex.

MRI Magnetic Resonance Imaging.

MSE Mean Squared Error.

MUA Multi-Unit Activity.

NF Neurofeedback.

NMR Nuclear Magnetic Resonance.

NMSE Normalized Mean Squared Error.

OEF Oxygen Extraction Factor.

PC Principal Component.

PCA Principal Component Analysis.

PCC Posterior Cingulate Cortex.

PET Positron Emission Tomography.

PLV Phase Locking Value.

PMC Premotor Cortex.

PSI Phase Synchronization Index.

PSP Post Synaptic Potential.

rCBF regional Cerebral Blood Flow.

RF Radiofrequency.

RMSF Root Mean Squared Frequency.

ROI Region of Interest.

RR Ridge Regression.

RS Resting State.

rsfMRI Resting State Functional Magnetic Resonance Imaging.

RSN Resting State Network.

RSS Residual Sum of Squares.

S2 Secondary Somatosensory Cortex.

SMA Supplementary Motor Area.

SPM Statistical Parametric Map.

SVD Singular Value Decomposition.

TE Echo Time.

TP Total Power.

TR Repetition Time.

WND Weighted Node Degree.

Chapter 1

Introduction

1.1 Motivation

Electroencephalography (EEG) measures electrical activity of the brain through a set of electrodes placed on the scalp, which capture transient electrical dipoles generated in the cortex that reflect underlying neuronal processes in real time. By means of these fast dynamics, EEG holds a high temporal resolution (order of the milliseconds), desirable for a wide range of practical applications (Wadman and da Silva, 2017). EEG also benefits from being relatively inexpensive, portable, noninvasive and easy to apply. However, it is very sensitive to noise and lacks specificity because of its low spatial resolution (order of the centimeter) and the fact that source localization from EEG suffers from an ill-posed inverse problem.

On the other hand, Functional Magnetic Resonance Imaging (fMRI) reflects the increases in blood flow that accompany neuronal activation. The success of fMRI stems largely from its high spatial resolution (order of the millimeters) and noninvasive nature. However, its temporal resolution (order of the seconds) is limited by the time required to acquire one brain volume and the duration of the hemodynamic response, which acts as a low-pass filter that blurs neural activity (Buxton, 2009). Moreover, fMRI recordings are considerably expensive and time consuming, posing further constraints on the use of this modality for several applications.

Over the past few years, simultaneous EEG-fMRI recordings have been largely used to understand the link between EEG and fMRI in multiple conditions. The use of multimodal approaches that combine these two modalities has received recognition as a promising novel method for several applications. The major advantage lies on the complementary temporal and spatial resolutions inherent to each technique, which together may provide richer information on the neuronal processes under analysis . Additionally, it brings the potential to control for confounding factors that arise from the limitations of each modality. Yet integrating EEG and fMRI data is not a trivial challenge, in fact multiple multimodal data integration methods have been described and compared in recent years (Perronnet et al., 2018; Abreu et al., 2018).

Notably, multimodal EEG-fMRI acquisitions are of special interest for Neurofeedback (NF) based Brain-Computer Interface (BCI) applications, which traditionally rely on either EEG or fMRI recordings.

NF based BCI systems aim to translate features extracted from recorded brain activity into signals able to provide real-time feedback that can be used for neuromodulation. These systems have been exploited as noninvasive techniques to improve neurorehabilitation outcomes in a large spectrum of neurological conditions (Vourvopoulos et al., 2019), for example in the context of rehabilitation or several psychiatric disorders. While combining EEG and fMRI for NF training has been demonstrated to provide a more efficient feedback and better regulation results (Perronnet et al., 2017), such multimodal acquisitions enclose several pitfalls: they are costly, non-portable and overall uncomfortable for patients, mostly due to the use of fMRI. Further, integration of EEG and fMRI signals so as to be fed to the feedback loop should be applicable in real-time (Perronnet et al., 2018).

Within this framework, significant efforts have been employed to reach solutions that combine the complementary advantages of EEG-fMRI multimodal acquisition, whilst using only the widely available EEG for real-time training. Such solutions rely on the ability to use a set of EEG-derived features to simulate the simultaneous BOLD signal.

This work aims to investigate an integration strategy whereby relevant EEG-features are extracted and their coefficient estimates learnt so as to predict the simultaneous average BOLD signal measured at a specific distributed network. These network-specific EEG patterns can be referred to as EEG Fingerprints (EFPs), a concept introduced by Meir-Hasson and their colleagues (Meir-Hasson et al., 2014) in a related study. The hope is that in the future, the estimated EFPs may be applied during real-time neurofeedback training, in order to attain the quality of the results obtained with a NF EEG-fMRI session, without the need to use fMRI.

In this work, independent EFPs are estimated for data recorded under two experimental conditions: during resting state (RS) and during a motor imagery (MI) task. In the former, the BOLD signal considered is extracted from a resting state network, the default mode network (DMN), whereas in the latter the signal is extracted from regions within the motor cortex.

The two conditions studied are considered to be relevant in the field of NF based BCI applications. In fact, EEG and fMRI NF protocols targeting motor imagery patterns are one of the most researched, mostly because of their demonstrated potential for rehabilitation following stroke and injury (Marzbani et al., 2016). On the other hand, the DMN and other resting state networks are known to be altered in multiple clinical conditions, specifically in attention deficit hyperactivity disorder (ADHD) (Rubia et al., 2019) and a number of chronic pain disorders including migraine and neuropathic pain (Edes et al., 2017; Sağ et al., 2018). Although the current amount of research on the role of resting state NF for the treatment of such conditions is very limited, some studies report improvement of RS network patterns, accompanied with decrease of clinical symptoms, following NF therapy in ADHD patients (Rubia et al., 2019). Moreover, several chronic pain conditions have also been proven to respond positively to some types of NF treatment (Marzbani et al., 2016). Altogether, these results indicate that RS patterns may also be promising new targets in the field of NF based therapy.

1.2 Thesis Outline

The document is organized as follows:

- Chapter 2 (Background) presents an overview of the main theoretical concepts within the framework of this work, as well as a review of the current knowledge on the several subjects that it entails. Additionally, the main published studies to date regarding methods that closely relate to the one here analysed are described, and their findings highlighted.
- Chapter 3 (Methods) starts by characterizing the data that was to learn and evaluate the models estimated. Then, the pipeline of the method employed is described, organized in the following steps: data acquisition; standard pre-processing of raw data; post-processing procedures to build the feature space from the EEG data and the output response from the fMRI data; model estimation (i.e., EFP estimation) and evaluation.
- In chapter 4 (Results) the results of model fitting are summarized and discussed. The chapter is organized in three main sections: first, the fitting results of all methods and models assessed are described; then, the estimated EFPs of all models are displayed; finally, the predictive performance of all models is compared.
- Chapter 5 (Conclusions) discusses the main conclusions drawn from the results obtained, and outlines future developments that can be done to further optimize the performance of the methodology studied in this work.

Chapter 2

Background

2.1 EEG and fMRI

2.1.1 Electroencephalography (EEG)

The existence of the electrical activity of the brain was first discovered in monkeys and rabbits by Richard Caton in 1875, long before German psychiatrist Hans Berger demonstrated that the electroencephalogram could be recorded from the human scalp (Berger, 1929). Shortly after, EEG became accepted as a reliable method for the analysis of brain function (Sutter et al., 2017).

Electroencephalography (EEG) is one of the most widely used techniques to mark neural processes, as it non-invasively records electrical activity of the human brain. Electrophysiology methods aim to measure electrical potentials within the brain that are sensitive to the underlying neural activity, thus providing information about brain function with high temporal resolution (Wadman and da Silva, 2017).

The electroencephalogram is generated by the summed electrical activities of neuron populations, with a small contribution from glial cells. The neurons are excitable cells that, upon stimulation, produce changing membrane potentials that propagate and ultimately generate extracellular electrical potentials, generally known as field potentials. These fields may be recorded locally by the means of microelectrodes at short distance from the source, or at longer distances by the means of electrodes placed directly on the cortical surface (Electrocorticography (ECoG)) or on the scalp (scalp EEG or just EEG). The amplitudes of scalp EEG lie between 10 and 100 μV and its spectrum extends from 0.1 Hz to 100.0 Hz (Krishnan et al., 2017).

Neural Basis of the Electroencephalogram

The electrical currents generated by neurons upon stimulation are essentially ionic currents at the level of cellular membranes, i.e., transmembrane currents. Two main forms of neural activation may be distinguished (Nunez and Cutillo, 1995): the slow changes in membrane potential due to synaptic activation, mediated by neurotransmitters, and the fast depolarization of neuronal membranes, mediated by sodium and potassium voltage-dependent ionic channels, yielding a voltage dependent membrane conductance.

The latter results in action potentials (AP), which are rapid changes in membrane potentials from their negative resting potential (typically around -60 mV) to a positive potential, followed by a quick repolarization. APs start at the axon hillock, where the neuron's body meets the axon, and generate impulses that propagate along the axons. Regarding the first, slower, post synaptic potentials (PSPs), these can be excitatory (EPSPs) or inhibitory (IPSPs). In general terms, EPSPs are generated by excitatory neurotransmitters (e.g., glutamate), released by excitatory neurons (e.g., glutamatergic) into the synaptic gap, whereas IPSPs are generated by inhibitory neurotransmitters (e.g., GABA), released by inhibitory neurons (e.g., GABAergic). These neurotransmitters ultimately trigger the flow of positive (in the case of glutamate) or negative (in the case of GABA) ions inwards the cell, generating the mentioned transmembrane currents. Postsynaptic potentials start at the level of the dendrite and undergo spatial and temporal summation. At the axon hillock, excitatory and inhibitory inputs are integrated, and an action potential is generated, i.e. the neuron fires, if the integration is above a given threshold.

As a result of the transmembrane currents triggered by EPSPs, active sinks are generated in the extracellular medium at the level of excitatory synapses, since the flow of positive charges into the interior of the cell leaves behind uncompensated negative charges. Importantly, because no accumulation of charge occurs, the currents that flow at the active synaptic sites are compensated by opposite currents elsewhere along the soma-dendritic membrane, which form distributed passive sources. Likewise, IPSPs are associated with active sources at the level of the synapse and distributed passive sinks along the soma-dendritic membrane. In both cases, these current patterns create dipolar sink-source configurations in the extracellular medium around active neurons (Wadman and da Silva, 2017). While postsynaptic potentials can be accurately described by a single current dipole oriented along the dendrite, action potentials at the axon lead to a pair of opposite current dipoles: one at the depolarization side and another at the repolarization side. Hence, postsynaptic potentials are more likely than action potentials to be sources of field potentials that can be measured at distance. This is not only because, since dipoles extend over a larger portion of the membrane, postsynaptic potentials attenuate less with distance, but also because they last significantly longer than action potentials, allowing for temporal summation.

At a macroscopic level, the field potential of a population of neurons is roughly given by the sum of the field potentials of the individual neurons, thus also behaving like that of a dipole layer. The EEG signal recorded at the scalp arises mainly from cortical pyramidal cell postsynaptic potentials, referred to as local field potentials (LFPs). Although other sources (e.g., other neurons and glial cells) and forms of electrical potentials (e.g., action potentials, associated with multi-unit activity; MUA) within the cortex exist, their contribution to extracranial electrical fields appears to be very limited. Pyramidal neurons contain long dendrites, called apical dendrites, that branch as they extend towards the cortical surface, forming open fields, as classically described by Lorente de Nó (de Nó, 1995). Additionally, these populations are arranged in palisades, with the main axes of the dendritic trees parallel to each other and perpendicular to the cortical surface. Hence, when pyramidal neurons are activated with a certain degree of synchrony, the longitudinal intra and extracellular currents add, yielding a current along the main axes of neurons. Current flows passively in the medium by volume conduction, with a distribution

depending on the conductivity profile, producing an electric field that can be detected in the scalp for large synchronous populations (Wadman and da Silva, 2017).

Acquisition of the Electroencephalogram

The electrodes placed on the scalp record cortical activities of the brain regions that are close to them. The placement of scalp electrodes is traditionally done according to the standard system 10-20 (ame, 1961), in which electrodes are placed at distances of 10% or 20% the total distance between the nasion (lower forehead between the eyes) and the inion (midline of the back of the head).

Standard EEG sets include 32, 64, 128 or 256 channels. EEG channels are named using letters, which distinguish between brain regions, and numbers, which distinguish between hemispheres of the brain and distance to the midline. Specifically, letters F, P, T and O and C refer respectively to the frontal, parietal, temporal, occipital and central areas, and Fp refers to the frontal pole. Odd numbers are associated with the left hemisphere, even numbers with the right hemisphere, and the lowercase z refers to electrodes placed on the midline. Moreover, the further the electrodes are placed laterally from the midline, the higher the subscript number. A standard EEG 32-channel montage is illustrated in figure 2.1.

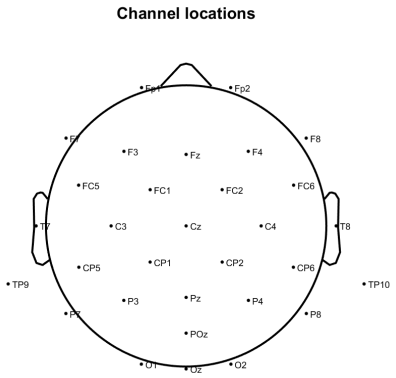


Figure 2.1: Electroencephalogram 32-channel montage, according to a standard 10-20 system.

Because of the electrically noisy environment where recordings commonly take place, differential amplifiers are essential for high-quality EEG recordings, enabling to cancel external noise. By using a common reference montage, in which all channels share the same reference channel, it is possible to re-reference the data to the average channel. The advantage of average reference rests on the fact that the reference is not locally dependent (Schomer et al., 2017).

Interpretation of EEG recordings requires first assigning EEG activity to specific electric positions on the scalp and then relating the scalp localization to the likely source in the cortex. The latter process is commonly known as reconstruction of the EEG signal and can only be achieved by understanding and modeling how the EEG signal is conducted through the brain, a process referred to as volume conduction. To reach the scalp, electrical currents generated in the cortex pass several layers of tissues with

different geometry and electrical properties, which causes spatial spread of the electromagnetic fields generated. This is aggravated when performing scalp measurements (EEG), because the skull further incurs spatial blurring on the electric potential distribution at the scalp (Lopes Da Silva, 2013). The term volume conduction has been colloquially adopted to describe this field spread phenomenon, despite this not being its strict meaning. Hence, due to the effect of volume conduction, the reconstruction of the EEG signal, also termed the inverse problem, is an ill-posed problem. This means that, without any constraints, an unlimited number of solutions may be reached, i.e., multiple combinations of sources can result in the same potential distribution at the scalp.

Rhythmical Activities of the Electroencephalogram

Waves of irregular periodicity, termed oscillations, are one of the essential components of the electroencephalogram. Each of these oscillations can be characterized by its peak frequency, bandwidth, and power. Hence, EEG spectra are traditionally described in terms of five well-established frequency bands, also referred to as brain rhythms: delta (0.4-4Hz), theta (4-8Hz), alpha (8-13Hz), beta (13-30Hz) and gamma (above 30 Hz). Despite the advantages of this traditional description, it also encloses numerous pitfalls, notably the fact that a given spectral component might ambiguously result either from the presence in the EEG of oscillations or from the contribution of particular waveforms, not necessarily rhythmic, to the EEG spectrum (Amzica and da Silva, 2017).

At a cellular level, oscillations arise because they are the most efficient way to establish synchrony between interconnected neural assemblies (Buzsáki, 2009). As Herrmann et al. (Herrmann et al., 2016) points out, over the course of time multiple authors have suggested that specific EEG oscillations contribute to different cognitive functions, depending on the brain region they occur in and on parameters such as amplitude, frequency, phase and coherence. This is supported by the knowledge that specific brain regions perform different functions, along with the evidence that coherent EEG oscillations in two distant brain regions may reflect the functional cooperation of these two regions (Siegel et al., 2012). Importantly, low-frequency oscillations tend to represent the cooperative activity of large-scale neuronal networks, whereas high-frequency oscillations mostly reflect cooperation between nearby cortical regions. This phenomenon is thought to be the basis for the observed power-law distribution of the EEG spectra (Singer, 1993): power tends to fall off with increasing frequency.

To be simplistic, the role of EEG oscillations in different cognitive processes can be summarized as follows: delta oscillations appear to be implicated in autonomic functions, motivational processes and in cognitive processes related to high emotional involvement and attention (Harmony, 2013); theta oscillations correlate with high cognitive processing, specifically in tasks involving memory (Gevins, 1997; Mitchell et al., 2008) and executive functions (Sauseng et al., 2007); alpha oscillations have been associated with memory (Klimesch, 1997), attentional processes (Pfurtscheller and Stancák, 1996), and motor tasks (Klimesch et al., 2007); beta oscillations have been linked to sensorimotor processing (Pfurtscheller and Silva, 1999), yielding a decrease in power (i.e. desynchronization) during preparation and execution of motor tasks; gamma oscillations may reflect cortical activation upon functions related to attentive processing of information (Fries, 2001) and formation of neural representations of events,

i.e. perception (Singer, 2006).

Importantly, as Lopes da Silva emphasizes (Lopes Da Silva, 2013), a unique alpha rhythm doesn't exist, rather three independent rhythms that can be distinguished based on the brain system where they are dominant: the occipitoparietal alpha associated with the visual system; the central alpha, called mu rhythm, associated with the sensorimotor system; and the temporal alpha, called tau rhythm, which is not commonly identified in EEG studies.

Functional Connectivity of the Electroencephalogram

As mentioned above, neurons do not function as isolated units, in fact large populations of neurons distributed along the cortex can become associated, communicating through synchronous activity, and then dissociated, giving rise to transient assemblies that are thought to play a crucial role in brain function. Each assembly as a whole has emergent properties that do not exist at the level of individual neurons (John, 2002). Cognitive processes may thus be modulated by changing the strength, pattern, or frequency with which neural populations engage in synchronous activity. The goal of brain connectivity analysis is to identify the interactions that make up such assemblies, as well as their strength. In this context, EEG-based connectivity analysis plays a dominant role because, to capture the transients inherent to brain connectivity, a temporal resolution in the order of milliseconds is of special interest.

For the purposes of EEG-based connectivity analysis, a distinction may be done between functional (or non-directed) and effective (or directed) connectivity. While the former describes the strength of the association of neural activation patterns between two brain regions, the latter describes the causal influence that one neural system exerts in another, thus accounting for both strength and direction of the interaction. Computational methods to quantify both types of connectivity may be further divided into model-free and model-based approaches. The latter typically assume linearity regarding the interactions that may take place. In this work, only model-based functional connectivity methods are explored.

The simplest linear method to quantify functional connectivity between two channels is to compute the Pearson correlation coefficient, which measures the linear relationship between two random variables, in this case, the amplitude of the EEG signal at each of the channels. More specifically, in a multiple regression framework, the squared correlation coefficient (R^2) measures the fraction of the variance of one of the variables that can be explained by the other. This measure ignores the temporal structure of the data, since it treats the EEG time-series as a set of independent realizations of random variables. In other words, this property means that estimated connectivity will be the same irrespective of whether the time-series has been randomly shuffled or not. Hence, an alternative method that does account for the temporal structure of the data is to compute the cross-correlation function for that channel pair, which describes the correlation between both time-series as a function of their time lag. Because of their simplicity, the correlation and the cross-correlation were the first methods used to assess EEG-based connectivity (Adey et al., 1961; Brazier and Casby, 1952).

Subsequentially, other metrics were proposed that were based on the frequency domain representation of the EEG signal, in opposition to the previous metrics, computed from its time domain representation. Among the methods that rely on the spectral profile of the EEG, some quantify the power of the

relationships between the spectra of both signals, some quantify phase of these relationships, others are sensitive to both. Metrics that are particularly sensitive to the relation of phase between the oscillatory components of signals may be referred to as synchronization metrics (Bastos and Schoffelen, 2016).

One widely used metric to quantify synchronization of neural processes is coherency (Nunez et al., 1997), which may be obtained from the cross-spectrum of the signals under analysis. The complex Fourier spectra of two signals can be combined to compute the cross-spectrum of the pair. The cross-spectrum is obtained by multiplying, at each frequency, the complex value of one signal by the complex conjugate of the other. This multiplication results as well in a complex number, with amplitude equal to the product of both signals' amplitudes, and phase equal to the phase difference of both signals. The cross-spectrum of the channels i and j, S_{ij}, can then be defined as:

$$S_{ij}(f) = \langle x_i(f) x_i^*(f) \rangle \Leftrightarrow \tag{2.1}$$

$$\Leftrightarrow S_{ij}(f) = \langle r_i(f) r_j(f) e^{i\Delta\Phi} \rangle$$

where x_i and x_j are the complex Fourier transform of the timecourse of channel i and j (respectively), r_i and r_j their amplitude, $\Delta\Phi$ their phase difference, * the complex conjugation operator and $\langle\rangle$ the expected value operator. Coherency, C_{ij}, is now defined as the normalized cross-spectrum:

$$C_{ij}(f) = \frac{S_{ij}(f)}{(S_{ii}(f) S_{jj}(f))^{1/2}} \tag{2.2}$$

In practice, it is only possible to obtain a true estimate of the cross-spectrum's expected value, S_{ij}, by computing an average over a sufficiently large number of trials. For this reason, coherency essentially reflects how the phases in two EEG channels are coupled to each other, and measures to which extent this coupling is consistent across trials. If the signals are independent, their phase lag $\Delta\Phi$ for a given trial is a random value, thus its expected value $\langle\Delta\Phi\rangle$ is zero. Hence, in this scenario coherency is zero as well (Nolte et al., 2004). Further, coherency weights these phase relationships with respect to the amplitudes of the two signals. Here, the question arises of whether phase relationships as given by $\langle\Delta\Phi\rangle$ are statistically correlated with the associated signal amplitudes. For this reason, another measure that does not weight the amplitudes in is often explored in alternative: the phase locking value (PLV), given by $PLV(f) = \langle e^{i\Delta\Phi} \rangle$. Notably, Nolte at al. report that, in their experience, results obtained with coherency are typically more robust than those obtained with the PLV. They suggest that, by assigning a stronger weight to samples with high amplitude product, coherency favors the samples that yield higher quality phase difference estimates. This is to assume that, for weak signals, it is more likely that noise corrupts the phase structure.

Importantly, in EEG-based connectivity analysis it is necessary to account for consequences of volume conduction, by means of which a deep source of electrical potential can reach many regions on the scalp surface, and thus be simultaneously recorded at different channels. This can give rise to spurious instantaneous correlations between scalp EEG signals that may lead to misinterpretations regarding

connectivity. In extreme cases, volume conduction can create purely artifactual coherency (or phase-locking) estimates, in which the apparent connectivity between two channels arises from activity that was generated by the same underlying source. Notably, the effects of volume conduction are instantaneous, at least in the frequency range relevant for neural signals, a property that can be exploited in order to derive connectivity measures that are robust against volume conduction. Indeed, the simplest methods proposed to address this problem rely on applying connectivity measures designed to suppress all instantaneous effects (Bastos and Schoffelen, 2016). Examples of such measures are the imaginary part of coherency and the phase slope index, both derived from coherency. The associated cost is that these measures are not able to detect true interactions occurring at zero phase difference.

Analyzing pairwise connectivity measures between the timecourse of all EEG channels requires an elevated computational cost. Moreover, it may be relevant to investigate the pattern of multiple pairwise interactions, from a complex network perspective. For this, graph theoretic approaches can be employed (Sporns, 2016). Given that the functional connectivity matrix may be regarded as representation of a weighted graph, some authors (Filho et al., 2017) suggest that EEG connectivity can be indirectly assessed by estimating the weighted degree of each channel (i.e., node). In a weighted graph, the weighted degree of a node is the average value of edges connected to that node, which in this context corresponds to the average connectivity values between that channel and all other channels. The weighted degree of channel i, W_i, can then be given by:

$$W_i = \sum_{j=1}^{C} S_{ij} \tag{2.3}$$

where C is the total number of channels, and S_{ij} is the functional connectivity between the channel pair (i, j). The underlying assumption is that local changes in the functional connectivity matrix should be accompanied by changes on weighted degree values.

2.1.2 Functional Magnetic Resonance Imaging (fMRI)

Functional magnetic resonance imaging (fMRI) is a widespread technique for measuring the activation of the human brain in response to specific experimental conditions. fMRI studies use imaging methods that are sensitive to changes in the blood oxygen level dependent (BOLD) signal, which indirectly reflects local neural activation. The aim is to build activation maps that depict the average activation levels of different regions in the brain (Buxton, 2009).

Principles of Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) relies on the nuclear magnetic resonance (NMR) signal from hydrogen (^{1}H) atoms of water molecules in tissue. Protons and neutrons in atomic nuclei possess an intrinsic angular momentum, often referred to as spin, which is only allowed specific quantized values. Nuclei with an odd charge number, such that of (^{1}H), hold a non-zero net spin that creates a non-zero nuclear magnetic moment, generating the NMR signal. In the absence of an external magnetic field, (^{1}H) nuclei

spins are oriented randomly, and the net macroscopic magnetic moment is zero. When subjected to an external magnetic field, the spins begin to rotate about the field axis with a characteristic precession frequency called the Larmor frequency. In doing so, they split into two quantized energy states: a lower energy state in which the nuclear moment is aligned parallel to the applied magnetic field, and a higher energy state in which the nuclear moment is aligned anti-parallel to the applied magnetic field. The energy difference between both states is directly proportional to the Larmor frequency, which in its turn is directly proportional to the strength of the applied magnetic field, B_0 (Tesla, T). At physiological temperatures, thermal agitation phenomena result in only slightly more spins aligning parallel to the field than anti-parallel, creating a macroscopic net magnetization, M, that is measurable for strong external magnetic fields B_0, typically at 1.5T or above.

Importantly, in an MRI experiment, to be able to detect the signal that arises from the macroscopic net magnetization of the sample, it is necessary to disturb the spins into an excited state, so that a magnetization component lies in the transverse plane. This is done by applying an alternating weaker (B_1) magnetic field that is transverse to the main static (B_0) magnetic field and oscillates with the Larmor frequency of the spins: the radiofrequency (RF) pulse. The result is a coherent (in phase) precession of the spins along the transverse plane. When the RF pulse is switched off, the excited spins return to their equilibrium state in a process called relaxation. This process comprises not only the restoring of their initial orientation (parallel to B_0) but also the dephasing of its transverse components. The return of the magnetization along the longitudinal axis is characterized by the time constant T_1, while the time constant characterizing the decay of its transverse component is called T_2. These relaxation time constants, along with the density ρ of ^{1}H atoms, have different values for different tissues of the human body, making up the basis of the contrast mechanism of MRI.

Changes of magnetization in the transverse plane are detected because the precessing magnetization vectors induce an electromotive force in the receiver RF coil (Faraday's law of induction), generating a current that is proportional to the magnitude of M. The time signal picked up is a damped sine wave, called free-induction decay (FID), that decays under an exponential envelope, proportionally to the decay of the transverse magnetization that happens due to the loss of phase coherence. The source for this dephasing phenomenon is the occurrence of variations in the Larmor frequencies of the sample's spins that arise from two main mechanisms: the first, characterized by time constant T_2, is a result of random interactions between neighboring nuclei, called spin-spin interactions, that cause low frequency random fluctuations in the local field; the second, characterized by time constant T_2*, stems from local, non-random, variations in the magnetic field caused by different magnetic susceptibilities of neighboring tissues. These susceptibility variations are commonly present near blood vessels, where the level of deoxyhemoglobin in the blood affects the T_2* in the surrounding tissues. This effect is the basis of the BOLD contrast used in fMRI (Ogawa et al., 1992; Kwong et al., 1992).

In MRI, acquisition and RF sequences can be designed with a range of different timing and amplitude parameters, adjusted to obtain images that are sensitive to specific tissue properties, among which: density ρ of ^{1}H atoms, time constants T_1, T_2 or, in the case of functional imaging, T_2*. MR signals are useful tools for neuroscience applications because they enable the detection of T_2* across the entire

brain with high spatial resolution, which can be achieved by superimposing small gradients upon the main B_0 magnetic field.

Blood Oxygen Level Dependent (BOLD) Signal

The biophysical source of the BOLD signal is the level of oxygenation of the blood. Red blood cells are rich in hemoglobin, a large iron-containing molecule that can bind to oxygen and is responsible for its transportation. This molecule exists in two distinct configurations: oxyhemoglobin (Hb), if it is bound to oxygen, and deoxyhemoglobin (dHb), if it is not. Hence, blood is rich in Hb when oxygenated, and rich in dHb otherwise. The presence of iron atoms in hemoglobin confers high magnetic susceptibility to this molecule, i.e., it is paramagnetic (Pauling and Coryell, 1936). However, when oxygen is bound, the molecule alters from being paramagnetic to being diamagnetic (i.e. low magnetic susceptibility). Paramagnetic agents such as dHb create local changes in the magnetic field that increase local field inhomogeneities (Brooks et al., 1975), thus decreasing transverse relaxation in surrounding tissues, in particular that related to T_2*. Hence, in T_2*-weighted images, if the oxygenation level of the blood rises, there is a consequent decline in the local levels of dHb that causes T_2* of the blood to increase, resulting in an enhanced signal (and vice-versa) (Ogawa et al., 1990).

Neural activity in brain tissue is related to blood oxygenation of blood vessels that perfuse it because neural activation leads to the consumption of Adenosine Triphosphate (ATP), a molecule that requires oxygen for its formation, via the glucose pathway. Thus, local neural activity is reflected by co-localized changes in blood oxygenation, changes that are depicted in the BOLD signal. Within this framework, it seems reasonable to expect that increased neural activity would drive a rise in oxygen consumption and thus a weaker signal, yet the opposite occurs, as first noticed by Ogawa et. al. (Ogawa et al., 1992). The dynamics that dictate this phenomenon are commonly referred to as the neurovascular coupling.

The rise in local oxygen demand that follows neural activation is translated in terms of an increased cerebral metabolic rate of oxygen ($CMRO_2$). To compensate for this, there is an increase in the regional Cerebral Blood Flow (rCBF), which is the rate of delivery of arterial blood to the capillary bed, both directly via signaling molecules and indirectly via astrocytes. Through mechanisms that are currently unclear, there is a mismatch of rCBF and $CMRO_2$ changes, as the augment of the rCBF overcompensates for the increase in the metabolic demand of oxygen, delivering an oversupply of oxygenated blood (Fox and Raichle, 1986). This mismatch means that neural activation is indirectly coupled to a decrease in the levels of dHb, and thus to a stronger BOLD signal.

The rise of rCBF is achieved by dilation of arterioles and consequent increase in arterial Cerebral Blood Volume (CBV). Because capillaries do not expand significantly, the greater blood volume delivered to the capillary bed is thought to be accommodated mostly by an increase in capillary overall flow velocity. Although other hypotheses exist, some authors (Buxton and Frank, 1997) argue that the inefficient oxygen diffusion that occurs at these high flow rates is responsible for the diminished Oxygen Extraction Factor (OEF) that triggers the necessity for excess oxygen in the blood supply, causing the imbalance between rCBF and $CMRO_2$ (Buxton, 2009).

Hemodynamic Response Function (HRF)

For the interpretation of activation studies performed using BOLD fMRI it is important to establish quantitative models that reflect the relationships between blood flow, oxygen metabolism, and electrical activity in the brain. These are generally characterized by the so-called hemodynamic response. The main events that characterize the hemodynamic response are reflected in the timecourse of the BOLD response to a brief stimulus, a temporal impulse response function referred to as the hemodynamic response function (HRF).

Despite scientific consent regarding heterogeneity of the HRF across brain regions, individuals and stimuli, the typical waveform consists of a positive peak roughly 5 s after stimulus onset, return to baseline and a post-stimulus undershoot (Logothetis and Wandell, 2004). Although some interpretations exist, there is still no consensus on a single explanation for the post-stimulus undershoot, as it may even be the result of multiple neural, vascular, and metabolic effects contributing under different circumstances. Several experiments report yet an additional transient of the BOLD response, called the initial dip, that occurs immediately after stimulus onset. There is strong evidence that this transient is likely a result of an early mismatch of the onset of $CMRO_2$ and CBF changes, with $CMRO_2$ responding more quickly. However, because it is a weak signal and often not observed, the initial dip is usually not included in the typical HRF waveform (Buxton, 2009).

Under the assumption that the relationship between a neural stimulus and the associated BOLD response is described by a linear-time invariant (LTI) system, the expected BOLD signal can be derived by linear convolution of the stimulus with the system's impulse response, the HRF. However, multiple authors have proven that the BOLD response is not a linear function of the neural stimulus, which is to say that the response to a sustained stimulus can't been predicted from that of a brief stimulus. Specifically, it has been consistently reported that response to a brief stimulus (e.g., < 4 s) is stronger than what would be expected from the response to a longer stimulus. This nonlinearity may stem from the neural response itself, since recordings show that upon stimulus onset, neural output peaks and then decreases over time until reaching a lower plateau, that lasts for as long as the excitatory input remains. This pattern, observed as well in many other similar systems, is considered to be an adaptation effect to stimuli. Another likely source of nonlinearity is the transformation of the CBF response to the BOLD signal, in the sense there is a maximum signal that could be measured in a BOLD experiment, corresponding to removing all dHb from the voxel under analysis. Hence, as the CBF change increases to infinite values, the BOLD response reaches a ceiling. Still, because the approximation that the BOLD response behaves linearly with stimuli produces reasonably accurate predictions for most applications, this model is commonly used in standard fMRI analysis (Buxton, 2009).

Statistical Analysis of the fMRI

Statistical processing of BOLD data is a critical step of studies concerning functional mapping. Within this framework, statistical analysis is concerned with making inferences about the underlying patterns in data, as it aims to provide estimates of certainty in the experimental hypothesis, given the available

observations. This type of analysis can be guided by building a model for the way in which the BOLD response is expected to depend on the stimulus, i.e., for the experimental hypothesis. Such models must include a component of random error along with a component that characterizes the expected BOLD response. The latter constitutes the experimental paradigm and is commonly designed by convolving the HRF with the stimulus (Worsley, 2000).

In this context, the most common approach is to implement the general linear model (GLM) (Friston et al., 1994), which is an equation that expresses the observed BOLD response in terms of a linear combination of independent variables. These include the explanatory variables (EVs), which are the known paradigms, and an error term. The model is fitted to the data by linear regression, computing the parameter estimates, or effect sizes, which quantify how strongly the associated effects explain the variability of the data. The GLM is usually implemented within a univariate or voxel-wise statistic framework, in which the timecourse of each voxel is assumed to be an independent process. For every voxel in the dataset, the parameter estimates are calculated by representing each time-point by a linear equation.

The next step is to determine whether an effect is statistically significant, and this can be done by building statistical parametric maps (SPMs) (Friston et al., 1991). To obtain such maps, effect sizes are used to compute a statistic at each voxel, usually a t-statistic (or t-value), or alternatively an F or Z statistic. These values are then tested against the null hypothesis, by comparing them to the null distribution of the statistic, which is a known probability density function, respectively the t-Student, F and normal. The null hypothesis is rejected, i.e. the effect is deemed as statistically significant, if the value of the test is below a pre-defined significance level. A common significance level used is $p = 0.05$.

However, because the SPM entails a large number of repeated measures over the entire brain, to reject the null hypothesis a threshold adjustment must be made that accounts for the large number of comparisons made. The most common methods used to resolve this problem, referred to as the multiple comparisons problem, rely on results from the Gaussian random field (GRF) theory (Adler, 2004).

2.1.3 EEG and fMRI in Resting State

Resting state activity comprises the spontaneous fluctuations in human brain activity that occur when subjects are not engaged in a particular task or higher cognitive process. In opposition to the standard task-based paradigm, which aims to build activation maps that depict the level of engagement in response to a specific stimulus or task, in resting state studies there is no controlled stimulus, this is, activation maps are produced during rest.

An important related concept is that of resting state networks (RSNs), which are networks that integrate functionally connected brain regions, i.e. brain regions that share correlated temporal patterns. Since these patterns have first been reported in fMRI studies by Biswal et al. (Biswal et al., 1995), several RSNs have been identified. These networks are highly reproducible in multiple resting state conditions and are spatially consistent across different subjects as well. The most widely studied and robustly detected RSN is the default mode network (DMN) (Raichle et al., 2001). This network was first identified by

Raichle and colleagues, when stumbling upon a consistent pattern of regions that showed higher activation during control state than during task performance, in PET (positron emission tomography) studies. The DMN includes the anterior cingulate cortex (ACC), the posterior cingulate cortex (PCC), the inferior lateral parietal cortex and the medial prefrontal cortex (mPFC) (Greicius et al., 2002). Although the exact cognitive functions of the DMN are still unclear, there is a growing consensus concerning its involvement in introspection, self-referential cognition and environment monitoring processes. Recent findings suggest that the DMN may be divided into a set of subnetworks, that play distinct roles and exhibit slightly different fluctuations, as the resting brain switches between engagement in self-directed thought and representation of current surroundings (Smith et al., 2018).

The study of RSNs has been proven to be of significant clinical value, providing sensitive markers of disease. Specifically, the DMN has been demonstrated to be altered in a number of conditions such as attention deficit hyperactivity disorder (ADHD) (Rubia et al., 2019) and several chronic pain disorders including migraine and neuropathic pain (Edes et al., 2017; Sağ et al., 2018). Regrettably, only very few studies have been employed to date that try to understand the role of resting state neurofeedback in the treatment of such conditions. Notably, Rubia and colleagues reported a significant increase in the functional connectivity of the DMN, accompanied by improvement of clinical symptoms, following NF therapy with ADHD patients. Moreover, some studies report improvement of pain symptoms in patients with migraines and other types of chronic pain after NF therapy (Marzbani et al., 2016). Altogether, these results indicate that RS patterns may be promising new targets in the field of NF based therapy.

fMRI Resting State Networks

Resting state functional magnetic imaging (rsfMRI) (Biswal et al., 1995) is a widespread technique for examining the functional connectivity of the human brain, through the study of spontaneous changes in the BOLD signal. Because the BOLD signal indirectly reflects neural activation, BOLD fluctuations during periods of absence of activity may reflect the neuronal baseline activity of the brain. Hence, resting state fMRI data may reveal how components of large-scale distributed neural systems are coupled together, based on their synchronous activity during rest. The organization, interrelationship and integrated performance of these regions can be globally described by the term functional connectivity. Measurement of functional connectivity in resting state studies is done by evaluating the similarity between different voxels' time-series, on the underlying assumption that the spontaneous time-series from functionally related brain regions are correlated.

Methodologies used to identify the rsfMRI source components corresponding to the RSNs can be broadly divided into two categories: model-free (or data driven) approaches, which attempt to map the connectivity in the entire brain; and model-based approaches, which use prior knowledge to restrict the analysis to a specific set of regions. Regarding the latter, one of the simplest approaches is seed-based correlation, which was the first method adopted by Biswal et al. to identify the resting state networks. This technique requires first selecting a region, often called seed, belonging to the putative network under analysis, and then computing the correlation between its BOLD signal time-series and all voxels in the entire brain. This can be implemented using the GLM to build statistical parametric maps.

EEG Correlates of Resting State Networks

In recent years, several studies have used simultaneous EEG-fMRI measurements to investigate the link between EEG oscillations and RSNs. In a study investigating the temporal correlation of the BOLD signal within several RSNs and the EEG power across the classical frequency bands, Mantini et al. (Mantini et al., 2007) concluded that each RSN may be characterized by a specific electrophysiological signature, reflecting a specific combination of EEG frequency components. More recent studies (e.g., Munck et al., 2009) further support these remarks, reinforcing the idea that distributed neural networks do not display pure frequency components within distinct bands, but instead a coalescence of rhythms. Notably, a growing body of research has revealed that mostly the alpha rhythm, but also the beta and gamma rhythms, are modulated during rest by the activity of the default mode network (DMN). Specifically, it has been reported a positive relationship between beta oscillations and the DMN, a positive relationship between gamma oscillations and the medial prefrontal cortex (mPFC) of the DMN, and a negative relationship between medial theta power changes and some regions of the DMN (Mantini et al., 2007).

However, reports of the relationship between DMN and alpha power are conflicting. In an attempt to clarify these questions, Bowman and colleagues (Bowman et al., 2017) tried to assess this relationship by analyzing concurrent EEG-fMRI data using a hypothesis-free approach, in which voxels of the DMN were identified by ICA (independent component analysis) and cross-correlated with the alpha frequency power of occipital and parietal channels. Results showed that different regions of the DMN exhibit divergent relationships to alpha power. Specifically, a positive correlation was found for one IC (independent component) that included regions within the PCC and the mPFC, and a negative correlation was found for the remaining. Such findings do not contradict the overall results obtained in previous studies, and Bowman et al. hypothesize that this behavior may point out to the dual function that is often ascribed to the DMN, which is thought to play a role both in introspection and in maintaining a level of outward vigilance. In this regard, they state that is then reasonable to look for duality in the behavior of DMN regions. More, they note that regions found to be positively correlated with alpha power have often been linked in literature to introspective processes. Accordingly, alpha power is believed to be inversely correlated with attentional processes, which is consistent with its positive association with regions engaged in introspection, where suppression of sensory information is key. In a final remark, it is proposed that the fact that the DMN can be decomposed into subnetworks with differing correlations with EEG alpha power may explain much of the inconsistency in previous studies looking to examine the link between the DMN and alpha power. More recently, studies have been exploring the potential use of EEG functional connectivity measures for predicting BOLD fluctuations within resting state networks. In particular, Jann et al. (Jann et al., 2009) investigated the relationship between alpha global field synchronization (GFS; a measurement of the phase-locked activity across all EEG channels) and the BOLD timecourse of the DMN, and found a positive correlation between the two.

2.1.4 EEG and fMRI in Motor Imagery

The term motor imagery (MI) (Jeannerod, 1994) is commonly used in the scientific community to address tasks in which the subject imagines moving a specific body part, without physically executing the movement. Research on motor imagery has identified several similarities between imagined and executed actions, thus supporting the idea that motor imagery is executed via many of the brain structures involved in the programming of movements. Indeed, neuroimaging studies report similar spatiotemporal patterns of neural activation in the two conditions, both engaging a set of frontal motor areas (the supplementary motor area, SMA; the premotor cortex, PMC; the primary motor cortex, M1) along with posterior parietal regions (the secondary somatosensory cortex; S2) and subcortical structures such as the cerebellum and the basal ganglia (BG) (Pfurtscheller and Neuper, 1997). Over the years, MI has been adopted as a tool for motor rehabilitation (Ramos-Murguialday et al., 2013), improvement of sport performance (Guillot and Collet, 2008), and musical practice (Lotze and Halsband, 2006). For this, the idea of directly training the central nervous system was promoted by establishing an alternative pathway between the user's brain and a computer system (Wolpaw et al., 2002). This is possible by using EEG-based brain-computer Interfaces (BCIs), since they can provide an alternative non-muscular channel for communication with the external world, while providing also a cost-effective solution for training (Vourvopoulos and Badia, 2016). BCI systems aim to translate features extracted from recorded brain activity into signals able to communicate with external computer devices, whether for assistance or rehabilitative purposes. In this regard, although EEG is the most popular imaging technique for feature extraction in BCI applications, EEG-based BCIs lack from high accuracy due to poor signal-to-noise ratio, low spatial resolution and non-stationarity of the signals (Lotte, 2014). A way to overcome the current limitations in EEG-based BCIs is to understand the modulation of EEG patterns by capturing user-specific correlates of MI. This can be achieved resorting to fMRI, because of its high spatial resolution. Hence, combining EEG and fMRI in this context may allow the identification of specific EEG correlates that best represent the brain activity associated with the execution of the MI task.

EEG Correlates of Motor Imagery

Regarding the association of EEG oscillations with planning and execution of movements, previously to the movement the ongoing EEG activity shows characteristic changes, namely either decreases or increases of power in specific frequency components. Because such phenomena may be interpreted as due to changes in the degree of synchrony of underlying neuronal networks, the former has been called event related desynchronization (ERD) and the latter event related synchronization (ERS) (Pfurtscheller and Neuper, 1997). More specifically, MI induces ERSs in the gamma frequency band in the ipsilateral (to the MI) hemisphere and ERDs in the mu and beta bands in the sensorimotor areas of the contralateral hemisphere, roughly two seconds prior to the movement onset, followed by a post-movement rebound beta ERS. Additionally, voluntary movement of the upper limbs is associated with ERS of the mu rhythm over the area representing the inferior limbs (and vice-versa), a phenomenon referred to as focal ERD, surround ERS (Pfurtscheller and Neuper, 2001; Pfurtscheller and da Silva, 2017). Notably, several

studies report that only a few electrodes show to consistently capture the EEG patterns of MI upper limb tasks, namely the hand representation areas in the primary motor cortex, C3 and C4, and the secondary sensorimotor cortex, Cz (Hamedi et al., 2016).

Although EEG spectral power metrics have been the most popular for MI feature extraction, the use of such features entails several limitations, as outlined by Hamedi (Hamedi et al., 2016) in their review. These include the inconsistency of MI responsive frequency bands across inter and intra-subject trials, as well as the heterogeneity, regarding cortex location and timings, of ERD and ERS occurrences. These challenges motivate the search for alternative features, specifically those that account for the relationship between the temporal patterns of multiple cortical regions, in the form of functional brain connectivity. Indeed, it is well known that the execution of motor tasks requires the participation of interconnected different parts of the brain, thus making it difficult to isolate only a few.

fMRI Correlates of Motor Imagery

In accordance to the results of electrophysiological studies, fMRI studies of motor imagery and motor execution indicate that both tasks share activation of overlapping cortical and subcortical networks. In particular, the SMA, PMC, several regions of the parietal cortex, the cerebellum and the BG show to be robustly activated during motor imagery. Remarkably, overall research reports mixed findings regarding the enhancement of the BOLD signal within the primary motor cortex during motor imagery training (Hardwick et al., 2017). Hence, it remains unclear whether M1 represents an effective target region for motor imagery-based fMRI paradigms.

2.1.5 EEG-fMRI Integration

In recent years, the use of bimodal approaches that combine EEG and fMRI has received a great deal of attention. Although originally motivated by the need to accurately map epileptic networks, it quickly expanded to a wider field of applications. In fact, the reason it has been sought as a promising novel method lies not only the complementary temporal and spatial resolutions of both modalities, but also on the different neural correlates inherent to each technique. Together, these bring the potential to overcome limitations specific to each modality and obtain a deeper understanding on a multiplicity of neuronal processes. Yet, several efforts have been employed over the years into overcoming the technological barriers of simultaneous EEG-fMRI recordings. These include EEG hardware modifications to guarantee compatibility with the MR environment and subject safety, as well as innovative signal processing tools to remove MR-induced EEG artifacts and EEG-induced fMRI artifacts (Abreu et al., 2018). Apart from these, further challenges lie on the data integration itself.

Multimodal EEG-fMRI data integration strategies can be categorized as asymmetrical (EEG-informed fMRI or fMRI-informed EEG) or symmetrical (model-driven or data-driven) (Jorge et al., 2014; Lahat et al., 2015). Whilst the former tries to build a model that maps neural activity as measured by one of the modalities to neural activity as measured by the other, the latter aims to build a single fusion model of neural activity that can explain both the EEG and the fMRI data, thus being often referred

to as data fusion. EEG-fMRI data integration encloses several challenges, particularly due to the fact that each modality depicts neural activity through a different physiological process. While this type of complementary information may indeed constitute one of the main advantages of such integration, it also poses restrictions to its conjoint interpretation, notably the fact that fMRI and EEG hold a different time-frame to a neural event: EEG provides in real time a direct measure of the electric potential that arises from neural events, whereas fMRI indirectly reflects neural activity by changes in the BOLD signal that are modulated by a hemodynamic response pattern that follows the neural event and typically peaks a few seconds after it occurs.

In the literature, asymmetrical approaches are by far the most common, partially due to their relative conceptual and methodological simplicity. They can be applied in either direction, one corresponding to fMRI-informed EEG and the other to EEG-informed fMRI. Studies employing fMRI-informed EEG frequently aim to improve EEG with respect to localization using anatomical and functional constraints derived from the simultaneous fMRI. On the other hand, EEG-informed fMRI studies often describe methods to extract relevant EEG features that share temporal patterns with the BOLD signal within a specific region of interest (Abreu et al., 2018). To that end, a multiplicity of strategies has been investigated, since different solutions can be employed depending on the type of activity under analysis. For instance, some studies select one or a few EEG channels of interest before feature extraction, either by using a priori knowledge or by running a preliminary analysis that identifies the most relevant channels (e.g., the channels that most correlate with the activity of interest); others employ regularization in order to impose spatial sparsity, this is, to learn the most relevant channels for model estimation. In this regard, EEG-informed fMRI methods can be divided into univariate and multivariate: in univariate methods only a limited number of representative EEG timecourses is considered, whereas in multivariate methods features are extracted from multiple channels, in order to capture spatial information. These include methods based on spatial correlation features or functional connectivity measures across different EEG channels, which cannot be assessed by considering only a few isolated channels. Regarding univariate metrics, both temporal and frequency information may be extracted from each timecourse under analysis, in order to compute either temporal or spectral features.

The methodology used in this work falls onto the category of EEG-informed fMRI, because it aims to extract a set of EEG-derived features, as well as learn its coefficient estimates, in order to simulate the simultaneous BOLD signal from a region of interest. According to the definition presented above, this work explored a number of univariate spectral features and multivariate functional connectivity features, specifically described in the following section.

EEG Correlates of fMRI

Although a great amount of work has been dedicated into formulating a model that expresses the transfer function between the EEG and the BOLD signal, this is still an active area of research.

Spectral Features Several transfer functions between EEG and BOLD have been proposed that are based on the time-frequency decomposition of the EEG signal, thus accounting for its temporal and

spectral profiles. Hence, these can be referred to as spectral features. Amongst the most well-studied are the root mean square frequency (Kilner et al., 2005), the total power (Wan et al., 2006), the average power across a specific frequency band, and the linear combination of band-specific average power Goense and Logothetis, 2008).

Root Mean Squared Frequency Kilner and colleagues (Kilner et al., 2005) described a heuristic that expresses hemodynamic changes as a function of the spectral profile of ongoing EEG activity, as opposed to the power across any particular frequency, or to the sum of power across all frequencies. Through a simplistic biophysical model, they proposed that neuronal activation, as reflected by increases in the BOLD signal, is associated with decrease in low frequencies relative to high frequencies or, conversely, an increase in high frequencies relative to low frequencies. In brief, the physiological basis of such model lies on the assumption that high-frequency dynamics, associated with fast changes in potential and high neuronal firing rates, incur an energy cost to which the BOLD signal is sensitive. The model is further supported by a dimensional analysis of the problem, in the sense that both activation (energy dissipation per second) and frequency of EEG oscillations express a rate. More, these predictions are consistent with empirical observations of how changes in the EEG spectrum are expressed hemodynamically. Hence, this heuristic specifies that the BOLD signal can be given by a measure of the average frequency of the EEG spectrum: the root-mean-squared EEG frequency. In this work the normalized root mean squared frequency (RMSF) of the EEG spectrum, as expressed by the following equation, is assessed.

$$EEG_{RMSF}(t) = \sqrt{\sum_{f_{min}}^{f_{max}} f^2 \tilde{P}(f,t)} \tag{2.4}$$

with

$$\tilde{P}(f,t) = \frac{P(f,t)}{\sum_{f_{min}}^{f_{max}} P(f,t)} \tag{2.5}$$

where $EEG_{RMSF}(t)$ is the time-series of the normalized RMSF EEG feature, f_{min} and f_{max} the minimum and maximum frequency under analysis, and $P(f,t)$ the value of the EEG power spectrum at the frequency bin f, at time t. $\tilde{P}(f,t)$ is then the normalized power spectrum (normalized by the total power) of the representative EEG time-series. Hence, this feature describes the changes in the relative spectral power of all the different frequencies considered, weighted by the square of the respective frequency.

Total Power Several studies (e.g., Wan et al., 2006) report a significant spatiotemporal correlation between the BOLD and the total power (TP) of the EEG, suggesting that the latter may comprise a reasonable transfer function between both signals. In this sense, it is stated that the neurovascular coupling could be modeled by a power transducer, transferring the power of the neural activity across all frequencies (i.e., the total spectral power) into the vascular input. The total power of the EEG spectrum

can be formulated as follows.

$$EEG_{TP}(t) = \sum_{f_{min}}^{f_{max}} P(f,t) \tag{2.6}$$

where $EEG_{TP}(t)$ is the time-series of the TP EEG feature, f_{min} and f_{max} the minimum and maximum frequency under analysis, and $P(f,t)$ the value of the EEG power spectrum at the frequency bin f, at time t. Given their experimental results, Wan et al. (Wan et al., 2006) suggest that, while both neural and hemodynamic responses behave nonlinearly with stimulus strength and duration, the relationship between EEG source power and BOLD amplitude remains itself close to linear. As they emphasize, this may indicate that the nonlinearity of the hemodynamic response primarily reflects the nonlinearity of the underlying neural activity, measured by EEG. Moreover, theoretical models (e.g., Babajani and Soltanian-Zadeh, 2006) also exist that assume a relationship between BOLD and neural activity that is independent of its frequency content.

Linear Combination of Band-Specific Total Power Other authors propose a relationship between BOLD and band-specific power of the EEG spectrum. As covered in previous sections, several studies report that changes in the BOLD signal share a significant temporal correlation with the EEG power across specific frequency bands. This has been verified for both resting state and task based experimental paradigms (for a review see Murta et al., 2014). The average power across a specific frequency band k, EEG_k, can be formulated as follows.

$$EEG_k(t) = \frac{1}{|b_k|} \sum_{f_{b_{k_{min}}}}^{f_{b_{k_{max}}}} P(f,t) \tag{2.7}$$

where $[f_{b_{k_{min}}}, f_{b_{k_{max}}}]$ is the frequency range of the frequency band k, and $|b_k|$ the total number of frequency bins within this range. More recently, other studies reported that, when used in a multiple regression framework, all the classical frequency bands explained a significant part of the BOLD variability (Goense and Logothetis, 2008; Munck et al., 2009). Thus, it has been suggested that the average power across each of all frequency bands should be modelled simultaneously in studies trying to predict the BOLD signal through EEG spectral features. This irrespective of any prior knowledge regarding the specific frequency bands of interest for the particular study. Hence, this work explores yet another measure based on the frequency profile of the EEG spectrum, with the potential to model the relationship between EEG and BOLD: the linear combination of the EEG power across each of the classical frequency bands. For the purposes of this work, this can be referred to as the linear combination (LC) metric. The time-series of the LC EEG feature, EEG_{LC}, can then be formulated as follows.

$$EEG_{LC}(t) = \sum_{k=1}^{N} a_k EEG_k = \sum_{k=1}^{N} a_k \left(\frac{1}{|b_k|} \sum_{f_{b_{k_{min}}}}^{f_{b_{k_{max}}}} P(f,t) \right) \tag{2.8}$$

where N is the total number of frequency bands considered and a_k the linear combination coefficient associated with the frequency band k.

A number of studies exist that compare the performance of different models of the transfer function between EEG and BOLD signals in the prediction of fMRI data. In particular, studies addressing this question in a visual task with healthy subjects (Rosa et al., 2010) and during resting state with epileptic patients (Leite et al., 2013) reported that frequency-weighted metrics yield a better performance than power-weighted metrics. However, overall reports in the literature do not provide yet a clear picture regarding the link between EEG and the BOLD signal, and a consensual optimal approach to model such relationship is still lacking.

Connectivity Features Concurrently, a growing body of research has been focused on describing the link between the EEG and the simultaneous BOLD in terms of the functional connectivity of the electroencephalogram. Although they are typically more complex and computational costly, the interest of using such measures lies on the possibility to incorporate in the models the communication patterns across distributed brain regions. Specifically, EEG synchronization measures have been deemed as promising within this context (Abreu et al., 2018), since they reflect aspects of brain activity that are complementary to those captured by spectral power measures. However, the current research on the usage of these measures to predict BOLD activity is rather scarce. As Abreu et al. reviews, both the global field synchronization (GFS) and the phase synchronization index (PSI) have been successfully used to predict simultaneous BOLD changes, and have additionally been found to outperform some of the spectral power measures described above.

Motivated by these promising results, this work explores the potential of another EEG synchronization measure, the imaginary part of coherency (IPC), for the prediction of concurrent bold changes. In this context, the transfer function between the EEG and the BOLD signal may be formulated as follows:

$$EEG_{IPC_{ij}}(t) = \sum_{i=1}^{N} a_k \left(\frac{1}{|b_k|} \sum_{f_{b_{k_{min}}}}^{f_{b_{k_{max}}}} IPC_{ij}(f,t) \right) \tag{2.9}$$

where

$$IPC_{ij}(f,t) = Im\left(\frac{\langle P_i(f,t) P_j(f,t) e^{i\Delta\Phi} \rangle}{(\langle P_i(f,t)^2 \rangle \langle P_j(f,t)^2 \rangle)^{1/2}} \right) \tag{2.10}$$

where $IPC_{ij}(f,t)$ is the IPC between channel i and channel j at frequency f and instant t, $\Delta\Phi$ their phase difference at frequency f and instant t, and the remaining parameters are the ones described above for the spectral measures. Further, the weighted node degree (WND) of the IPC is also explored in this work, in an attempt not only to reduce the computational cost of employing a traditional connectivity measure, but also to investigate the potential of incorporating into the models these patterns of multiple pairwise interactions. The corresponding transfer function may be formulated as follows:

$$EEG_{WND_i}(t) = \sum_{k=1}^{N} a_k \left(\sum_{j=1}^{C} \left(\frac{1}{|b_k|} \sum_{f_{b_{k_{min}}}}^{f_{b_{k_{max}}}} IPC_{ij}(f,t) \right) \right) \tag{2.11}$$

where C is the total number of EEG channels considered and EEG_{WND_i} the WND at channel i.

2.2 Model Estimation

2.2.1 Linear Methods for Regression

Linear regression is the simplest and most extensively studied method for statistical supervised learning. Supervised learning is the machine learning task of using a set of measured input variables, the so-called predictors, features or independent variables, to predict the values of one or more measured output variables, the responses or dependent variables. Supervised learning is so called because the learning process is guided by a set of measured input-output pairs, commonly referred to as the training set. This is in contrast to the unsupervised learning paradigm, in which no measurements of the outputs exist. Hence, in supervised learning methods, a prediction model is inferred from the training data, with the goal of predicting with good enough accuracy the output of new unseen input data.

Linear regression is a simple but powerful prediction method, that aims to model linearly the relationship between the output response, Y, and a set of predictors, $X_1, ..., X_p$. To hypothesize that the data yields a linear structure is to make a very large assumption, hence linear models often reach stable but sometimes inaccurate predictions. Yet, such models still thrive in modern statistical analysis exactly because of their simplicity and straightforward interpretation. More, in some situations they can outperform more complex non-linear models. Additionally, linear models can be applied to non-linear transformations of the original inputs, which considerably expands their scope (Hastie et al., 2017).

According to the framework described, linear regression methods aim to predict the output Y from an input vector of p features $\mathbf{X}^T = X_1, ..., X_p$, via the model:

$$\hat{\mathbf{Y}} = f(\mathbf{X}) = \beta_0 + \sum_{j=1}^{p} \mathbf{X_j} \beta_j \tag{2.12}$$

where $f(\mathbf{X})$ is the regression function, $\hat{\mathbf{Y}}$ is the estimate of output $\mathbf{Y}$, $\boldsymbol{\beta} = \beta_1, ..., \beta_p$ is the vector of coefficient estimates and β_0 is the intercept.

In this work, linear regression is used to learn a model that predicts the BOLD response from a set of p features, extracted from the simultaneous EEG signal. $\mathbf{Y}$ is then a N-length vector that represents the BOLD time-series, with N the total number of time-points of the series. Likewise, each predictor in $\mathbf{X}$ represents a feature of the EEG time-series, so $\mathbf{X_j}$ is a N-length vector and $\mathbf{X}$ a $N \times p$ length matrix.

Least Squares Regression

Many methods exist that fit the linear model described in equation 2.12 to the set of training data. Among these, the most widely used is the least squares regression (LSR). In this approach, the vector of coefficients $\boldsymbol{\beta}$ is estimated so as to minimize the residual sum of squares (RSS) between all the $(\hat{y}_i, y_i)$ pairs within the training set, defined as:

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \sum_{i=1}^{N} (y_i - \hat{y_i})^2 \Leftrightarrow \tag{2.13}$$

$$\Leftrightarrow RSS(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

where x_{ij} and y_i are the ith observations of $\mathbf{X_j}$ and $\mathbf{Y}$, respectively. Because this is a minimization problem, the optimal vector of estimates $\hat{\boldsymbol{\beta}}$ can then be given by:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} RSS(\boldsymbol{\beta}) \Leftrightarrow \tag{2.14}$$

$$\Leftrightarrow \hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

In statistical learning, the right hand side of equation 2.13 is often referred to as the objective function of the optimization problem, since this is the function that the method seeks to optimize (in this case, minimize). Because $RSS(\boldsymbol{\beta})$ is a quadratic function of the parameters, its minimum always exists, but may not be unique. The solution is easiest to characterize in matrix notation. To do so, it is convenient to include the intercept β_0 in the vector of coefficients $\boldsymbol{\beta}$, which can be done by adding a new row of ones at the beginning of $\mathbf{X}$. Equation 2.13 can then be rewritten as follows.

$$RSS(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{2.15}$$

where $\mathbf{X}$ is now an $N \times (p+1)$ matrix and $\boldsymbol{\beta}$ a $(p+1)$-length vector. The minimum of the $RSS(\boldsymbol{\beta})$ function can be obtained by differentiating it with respect to $\boldsymbol{\beta}$ and setting the result to zero. By doing so, the following linear system of equations is obtained:

$$\frac{\partial RSS}{\partial \hat{\boldsymbol{\beta}}} = 0 \Leftrightarrow \tag{2.16}$$

$$\Leftrightarrow \mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0$$

If matrix $\mathbf{X}^T\mathbf{X}$ is invertible (or nonsingular), then a unique solution to equation 2.16 exists, and may be obtained by solving the following equation.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \Rightarrow \tag{2.17}$$

$$\Rightarrow \hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{2.18}$$

The matrix that maps the transformation of the vector of observed output values into the vector of estimated output values, $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, is called the hat matrix, $\mathbf{H}$.

However, in real scenarios, even if a solution is found and the model seems to fit the data, optimization problems such as this one are often ill-posed or prone to overfitting. For instance, if two or more

input features are correlated, the columns of $\mathbf{X}$ are not all linearly independent or orthogonal, i.e., $\mathbf{X}$ is not of full rank. As a consequence, $\mathbf{X}^T\mathbf{X}$ is a singular (non-invertible) matrix. In this case, the least squares solution $\hat{\boldsymbol{\beta}}$ is not uniquely defined, which is to say that exists more than one optimal vector of coefficients $\hat{\boldsymbol{\beta}}$ that minimizes 2.13. As a practical example, if two of the input features $\mathbf{X_j}$ and $\mathbf{X_k}$ are inversely correlated, then an infinite number of symmetric pairs (β_j, β_k) could result in canceling the effect of such features in the estimated response. As a result, it might even be the case that neither of these features has a significant relation to the output variable, but both its coefficients hold high magnitudes. Hence, these features would be falsely deemed as statistically significant. Altogether, the effects of applying this model to a set of correlated inputs could largely confound the interpretation of results.

Further rank deficiencies can occur if the number of input features p approximates or exceeds the number of training samples N. This is often the case within signal analysis frameworks, either because a large number of features is included in the model, the number of observations is limited, or both. Here it is convenient to speak in terms of degrees of freedom (DOF). In classical statistics, the number of degrees of freedom of a model may be defined as the number of linearly independent parameters in it and is often used to quantify model complexity. In Least Squares Regression, this corresponds to the number of linearly independent input features. Depending on the number of DOF of the problem, the model fit may range from a straight line fit to an interpolating fit. Importantly, the latter is only possible if the DOF approximates or exceeds the number of samples. If this is the case, then LSR on a set of training samples cannot be used to find a set of coefficients $\hat{\boldsymbol{\beta}}$ that fit the entire population from which the samples are drawn. This is because when given the possibility, in its attempt to minimize the RSS on the training samples, LSR chooses an interpolating fit that ensures zero residuals. Such an overadjustment is made at the expense of assigning high effects to features that randomly explain small details in the output, but do not truly correlate with it. In other words, LSR coefficients become poorly determined and exhibit high variance. The model learnt is then perfectly fitted to the structure of the training data, but will likely fail to predict unseen data. In statistical learning, this effect is referred to as overfitting.

The problems described above illustrate the two main reasons why LSR often produces models that are unsatisfactory for most of its applications. The first reason has to do with the interpretability of results. In statistical learning, it is often desirable to identify the subset of features that exhibit the strongest effects on the fit. This is important to get a clearer view of the relationships being modelled, which will further allow to develop more cost-efficient solutions. The second reason has to do with the predictive performance of the estimated models. Indeed, the least squares estimates of the parameters $\boldsymbol{\beta}$ often have low bias but large variance (as discussed in more detail in section 2.3.1). This large variance translates into poor prediction accuracy on unseen data, as explained above.

Both of these problems can typically be alleviated by regularization strategies that impose smoothness restrictions so as to remove some DOF and reduce the complexity of the LS model. This means that model bias is sacrificed in order to reduce the variance of its coefficients, and thus improve the overall prediction accuracy and interpretability of results. In the following section, a set regularization approaches for linear models, termed shrinkage regression methods, is reviewed.

Shrinkage Regression Methods

Shrinkage regression methods aim to shrink the sizes of coefficient estimates towards zero, in order to reduce their variance. They do so by imposing a smoothness constraint on the least squares objective function (equation 2.13). As Hastie et al. emphasizes, shrinkage methods (as with all regularization strategies) often result in biased estimates. This is because any methods that impose restrictions on the optimization function do not remove the ambiguity caused by the multiplicity of solutions, rather transfer this ambiguity to the choice of constraint.

Ridge Regression. Ridge regression (RR) (Hoerl and Kennard, 1970) shrinks the coefficients $\hat{\boldsymbol{\beta}}$ by adding a new term to the LS objective function that imposes a quadratic penalty (given by an L_2 norm) on the coefficients size. The reformulated objective function is then given by the expression inside curly brackets on right hand side of the equation:

$$\hat{\boldsymbol{\beta}}^{ridge} = \underset{\boldsymbol{\beta}}{\arg\min} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{2.19}$$

where $\lambda > 0$ is a complexity parameter, termed ridge regularization parameter, that allows to control the amount of shrinkage. The greater the complexity parameter, the greater the amount of shrinkage, the lower the complexity of the model.

Notice that coefficient β_0 is left out of the penalty term, since penalizing the intercept would make the solution depend on the origin of $\mathbf{Y}$. Additionally, because the solutions of RR are not invariant to the scale of its inputs, it is standard procedure to perform normalization of the feature input vectors (columns of $\mathbf{X}$), before fitting the model. Normalization is accomplished by centering (subtracting the mean of) each feature vector, and then normalize the result by its standard deviation. Hence, the resulting vectors hold zero mean and standard deviation one. Using centered inputs means that each x_{ij} gets replaced by $(x_{ij} - \overline{x}_{ij})$, and thus equation 2.19 can be separated in two: β_0 may be estimated by equation 2.20, whereas the remaining p coefficients may be estimated by RR without the intercept, as expressed in equation 2.21 (in matrix notation).

$$\hat{\beta}_0 = \overline{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{2.20}$$

$$\hat{\boldsymbol{\beta}}^{ridge} = \underset{\boldsymbol{\beta}}{\arg\min}\, RSS(\boldsymbol{\beta}, \lambda) \Leftrightarrow \tag{2.21}$$

$$\Leftrightarrow \hat{\boldsymbol{\beta}}^{ridge} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}$$

Differentiating the objective function with respect to $\boldsymbol{\beta}$ and setting the result to zero, the solution is given by:

$$\hat{\boldsymbol{\beta}}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \tag{2.22}$$

Because the penalty $\boldsymbol{\beta}^T\boldsymbol{\beta}$ is a quadratic expression, the RR solution is, similarly to the LSR solution, a linear function of $\mathbf{Y}$. More, the ridge solution adds a positive constant ($\lambda\mathbf{I}$) to the diagonal of $\mathbf{X}^T\mathbf{X}$ before inversion, hence matrix $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})$ is nonsingular, even if $\mathbf{X}^T\mathbf{X}$ is not.

An important new concept related to linear regularization methods is that of effective degrees of freedom. Although ridge regularization does not actually set any component of $\boldsymbol{\beta}$ to zero, the coefficients are shrunk, and thus the classical definition of DOF in terms of dimensionality is no longer useful. Instead, in RR models, the number of effective degrees of freedom $df(\lambda)$ is a parameter that depends on the model's complexity, as defined by λ, instead of on the number of input parameters p. Yet, the effective degrees of freedom of the ridge fit are defined not only as function of λ, but as a function of the singular values of $\mathbf{X}$ as well. These can be obtained through singular value decomposition (SVD) of the input matrix $\mathbf{X}$, which provides important insights on the nature of the ridge fit.

Singular value decomposition (Golub and Reinsch, 1971) is essentially a means of decomposing any matrix into a form that is extremely useful for numerous applications in signal processing and statistics. The SVD of the $(N \times p)$ matrix $\mathbf{X}$ is a factorization of the form $\mathbf{UDV}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are $(N \times N)$ and $(p \times p)$ orthogonal matrices, with the columns of $\mathbf{U}$ spanning the column space of $\mathbf{X}$ and the rows of $\mathbf{V}$ spanning the row space of $\mathbf{X}$. $\mathbf{D}$ is an $(N \times p)$ diagonal matrix, with entries $d_1 \geqslant d_2 \geqslant ... \geqslant d_p \geqslant 0$ called the singular values of $\mathbf{X}$. If one or more values $d_j = 0$, then $\mathbf{X}$ is singular (non-invertible). Using the orthogonality of $\mathbf{V}$, it may also be written in the form $\mathbf{XV} = \mathbf{UD}$. Through SVD of $\mathbf{X}$, the matrix form solution of RR (equation 2.22) can be decomposed as follows:

$$\hat{\mathbf{Y}}^{ridge} = \mathbf{X}\hat{\boldsymbol{\beta}}^{ridge} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \Leftrightarrow$$

$$\Leftrightarrow \hat{\mathbf{Y}}^{ridge} = \mathbf{UD}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{DU}^T\mathbf{Y} \tag{2.23}$$

$$\Leftrightarrow \hat{\mathbf{Y}}^{ridge} = \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T\mathbf{Y}$$

where u_j are the columns of $\mathbf{U}$. Hence, in estimating the output $\hat{\mathbf{Y}}$, the coordinates of $\mathbf{Y}$ with respect to the orthonormal basis $\mathbf{U}$ are shrunk by the factor $d_j^2/(d_j^2 + \lambda)$. The shrinkage factor is inversely proportional to d_j, which means that a greater amount of shrinkage is applied to the directions with smaller d_j. This is an important observation, because the order of the singular values $d_1, d_2, ..., d_p$ reflects the order of the principal component directions of $\mathbf{X}$. Indeed, SVD of $\mathbf{X}$ discloses its principal components (PCs).

As a side note, principal component analysis (PCA) (Hotelling, 1933; Jolliffe and Cadima, 2016) is a statistical procedure whereby a set of observations of correlated variables (in this case, the input features or columns of $\mathbf{X}$) are decomposed into a set of linearly independent variables called the principal components of the data (of $\mathbf{X}$). This decomposition is defined in such a way that the first principal component maximizes the variance of the projected data, and the last principal component minimizes that variance.

Through SVD of $\mathbf{X}$, its principal component directions are decomposed into the columns of $\mathbf{V}$. The first column of $\mathbf{V}$, $\mathbf{v_1}$, is the first principal component direction of $\mathbf{X}$, and thus the first principal component of $\mathbf{X}$, $\mathbf{z_1}$, is given by $\mathbf{z_1} = \mathbf{X}\mathbf{v_1}$. Hence, because $\mathbf{z_1} = \mathbf{X}\mathbf{v_1} = \mathbf{u_1}d_1$ (given that $\mathbf{XV} = \mathbf{UD}$), d_1 reflects the amount of variance of the first principal component. Therefore, RR projects $\mathbf{Y}$ onto the principal components of $\mathbf{X}$ and shrinks the coefficients of the low-variance components more than those of the high-variance components. The implicit assumption is that the expected output response $\hat{\mathbf{Y}}$ varies more in the directions of higher variance of the inputs.

Returning to the matter of the effective degrees of freedom, in RR, $df(\lambda)$ may be defined by the trace of the hat matrix, $\mathbf{H}$. Hence, as mentioned above, it is defined as a function of λ as well of d_j:

$$df(\lambda) = \mathbf{H} = \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T \tag{2.24}$$

Notice that $df(\lambda) \to p$ when $\lambda \to 0$, as the objective function approximates that of the LSR and no regularization is applied. Conversely, as λ increases towards infinite, and an infinite amount of regularization is applied, $df(\lambda)$ decreases towards zero.

Lasso Regression. Similarly to RR, lasso regression (LR) (Tibshirani, 1996) shrinks the coefficients in $\hat{\boldsymbol{\beta}}$ by adding a penalty term to the LS objective function. However, the lasso penalty is a linear function of $\boldsymbol{\beta}$ (given by an L_1 norm), in contrast to the quadratic penalty imposed by ridge. The reformulated objective function is given by the expression inside curtly brackets on the right hand side of the equation:

$$\hat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2.25}$$

As before, λ is the complexity parameter, here called the lasso regularization parameter. In matrix notation, the problem may be rewritten as follows:

$$\hat{\boldsymbol{\beta}}^{lasso} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \tag{2.26}$$

Just like in RR, LR continuously shrinks the coefficients towards zero as λ increases. However, owing to the nature of the L_1 penalty, some coefficients are shrunk to exactly zero if λ is large enough. Hence, lasso does both continuous shrinkage and automatic variable selection simultaneously.

Another important consideration is that the L_1 constraint makes the lasso solution a nonlinear function of $\mathbf{Y}$, and thus there is no closed form expression as there is in ridge regression. Instead, computing the lasso solution is a quadratic programming problem. For the majority of nonlinear modeling procedures such as this one, it is difficult to derive the analytical expression of the effective degrees of freedom of the fit. However, for lasso in particular, the number of nonzero components of $\hat{\boldsymbol{\beta}}$ has been proven to be an exact unbiased estimate of the effective degrees of freedom (Zou and Hastie, 2005). This was held as a surprising result, given that such a clean relationship between the number of nonzero coefficients and the effective DOF is not found in other regularization methods.

Elastic Net Regularization. As a continuous shrinkage method, RR achieves a better prediction performance than LSR via a bias-variance trade-off. Yet it produces rather unintelligible models, for it keeps all the original features. LR comes then as a suitable alternative, since it produces significantly sparse models. Although none of these methods conclusively dominates the other in terms of prediction performance (Tibshirani, 1996), LR is becoming more appealing in modern data analysis, as variable selection is increasingly desirable. However, it encloses some pitfalls, as illustrated by the following scenarios (Zou et al., 2007): (a) if $p > N$, lasso selects at most N variables out of p candidates before it saturates; (b) if a set of highly correlated input features exist, lasso selects only one, regardless of which one (this is a concern especially for large p, small N problems); (c) if $N > p$, and there are high correlations among input features, ridge typically outperforms lasso.

Within this context, Zou and Hastie (Zou et al., 2007) proposed a novel regularization technique, which they called the elastic net regularization (ENR), that aimed to keep the best features of lasso, whilst fixing the problems highlighted above. The naive ENR is a penalized least squares method that uses a novel elastic net penalty, given by a linear combination of the ridge and lasso penalties. The objective function is then defined by the expression inside curly brackets on the right hand side of the equation:

$$\hat{\boldsymbol{\beta}}^{naive} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \right\} \tag{2.27}$$

This means that the elastic net solution is not a linear function of the output $\mathbf{Y}$, and thus solving equation 2.27 is equivalent to solving the lasso optimization problem.

The naive elastic net properties overcome the limitations of lasso described in scenario (a), as it can potentially select all p features in all situations. More, the method exhibits a so-called grouping effect for correlated variables, a property not shared by lasso, as described in scenario (b). This grouping effect means that the method assigns identical coefficients for highly correlated variables (up to a change of sign if they are negatively correlated), instead of blindly choosing one and rejecting the rest (Zou et al., 2007). However, this naive approach appears to incur a double amount of shrinkage, which introduces extra bias without removing extra variance. To address this issue, the authors proposed a corrected method, called simply ENR, which coefficients are a rescaled version of the former. Such a scaling transformation preserves the variable selection property of the naive ENR, while undoing the excess shrinkage. The ENR coefficients are then determined by the following equation:

$$\hat{\boldsymbol{\beta}}^{enr} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda P_\alpha(\boldsymbol{\beta}) \right\} \tag{2.28}$$

where λ is the elastic net complexity parameter, and P_α is the elastic net penalty term, defined as:

$$P_\alpha(\boldsymbol{\beta}) = \sum_{j=1}^{p} \left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \right] \tag{2.29}$$

where α is computed from the ratio between λ_2 and λ_1, thus defining the "amount" of each regular-

ization incurred.

$$\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2} \tag{2.30}$$

In the extreme cases of $\alpha = 0$ and $\alpha = 1$, the elastic net penalty corresponds to the ridge and lasso penalty, respectively. Therefore, the sparsity of the elastic net solution increases as α rises from 0 to 1. Importantly, Zou and Hastie demonstrated that this latter method yields higher prediction performance when compared with ridge and lasso alone. Hence, in this work, only ENR is employed.

Regarding the effective DOF of the elastic net fit, Zou and Hastie (Zou and Hastie, 2005) found the number of nonzero model coefficients to be a biased estimate. An unbiased estimate of the elastic net DOF may be obtained as follows:

$$df(\hat{\mathbf{Y}}) = tr(\mathbf{X}_A(\mathbf{X}_A^T\mathbf{X}_A + \lambda_2\mathbf{I})^{-1}\mathbf{X}_A^T \tag{2.31}$$

where $\mathbf{X}_A$ is the active set of $\mathbf{X}$, defined as the subset of parameters with nonzero coefficients.

2.3 Model Selection and Evaluation

All regularization methods have a smoothing or complexity parameter that has to be determined. For the shrinkage linear regression methods described above, this corresponds to the multiplier of the penalty function, λ. Model selection is the process whereby the best of several possible models is chosen so as to optimize a predefined criterion, whilst model evaluation (or assessment) is the procedure of estimating the prediction error of the selected model on unseen data. The specifics of model selection and evaluation are discussed hereafter.

2.3.1 Bias-Variance Trade-off

In general terms, as harsher regularization is applied and model complexity increases, model variance tends to decrease and model bias tends to increase. As briefly discussed in section 2.2.1, criteria employed to choose model complexity often seek to achieve an optimal bias-variance trade-off so as to minimize the expected prediction error or other performance criterion. Because training an algorithm and evaluating its performance on the same data yields overoptimistic results, measures of performance are typically assessed in an independent dataset, termed the test set. For models with high variance, the training error is usually low but test error tends to be large, as models are overfitted and do not generalize well. In contrast, models with very low complexity have large bias and may be underfitted, again resulting in poor generalization. This is the empirical essence of the bias-variance trade-off.

Further insights on the nature of the bias-variance trade-off can be obtained through the bias-variance decomposition of the data. For all regression models described, the output variable $\mathbf{Y}$ may be expressed by a linear function of the input $\mathbf{X}$ plus an error term which the model seeks to minimize.

$$\mathbf{Y} = \hat{f}(\mathbf{X}) + \boldsymbol{\epsilon} = \hat{\boldsymbol{\beta}}\mathbf{X} + \boldsymbol{\epsilon} = \hat{\mathbf{Y}} + \boldsymbol{\epsilon} \tag{2.32}$$

where the error ϵ follows a normal distribution $\mathcal{N}(\mu, \sigma_\epsilon^2)$, with expected value $E(\epsilon) = \mu = 0$ and variance $Var(\epsilon) = \sigma_\epsilon^2$. The expected prediction error of the regression fit $f(\mathbf{X})$ for a given set of input features $\mathbf{X}_0$ is determined as:

$$Err(\mathbf{X}_0) = L(\mathbf{Y}_0, f(\mathbf{X}_0)) \tag{2.33}$$

where L is the loss function, typically defined for linear regression problems as the squared error, $(\mathbf{Y}_0 - f(\mathbf{X}_0))^2$. The expected error may also be decomposed as:

$$\Leftrightarrow Err(\mathbf{X}_0) = \sigma_\epsilon^2 + Bias(f(\mathbf{X}_0))^2 + Var(f(\mathbf{X}_0)) \tag{2.34}$$

The first term on the right hand side corresponds to the error between the the best-fitting linear approximation $\hat{f}(\mathbf{X}_0)$ and the true function $\mathbf{Y}_0$, as defined by equation 2.32. The second term is the average squared estimation bias, which is the error between the best-fitting linear approximation $\hat{f}(\mathbf{X}_0)$ and average estimate $E[f(\mathbf{X}_0)]$. The third term is the error between each estimate $f(\mathbf{X}_0)$ and the average estimate. While the first term does not depend on the method and constraints chosen, the last two do and thus can be optimized.

As mentioned above, the loss function for regression methods usually corresponds to a mean squared error. However, when it comes to model selection and evaluation, other loss functions are typically applied as criteria to assess model performance. These include information criteria such as the Bayesian information criterion (BIC), the Akaike information criterion (AIC), or less often the normalized mean squared error (NMSE). Criteria for model selection is discussed further ahead in section 2.3.3. Although the expected prediction error is defined as the error across an independent test set, this value can be estimated from the training data alone, using cross-validation. Cross-validation procedures are addressed in the following section.

2.3.2 Cross-Validation

General Methods for Cross-Validation

Cross validation (CV) (Stone, 1974) is a well-established method for both model selection and evaluation. For a scenario in which there is a large amount of data, the well-accepted approach for both model selection and evaluation is to divide the dataset into three parts: a training set, a validation set and a test set. The training set is used to fit the model, the validation set is used to estimate the prediction error of several models for model selection, and the test set is used to estimate the prediction error of the selected model for model evaluation. However, in most real scenarios the amount of data available is not sufficient for being split into three parts. CV comes then as a convenient alternative for this type of situation, for it performs efficient data re-usage. There is a wide range of CV approaches, but they all share a common structure: the available data is split into a number of independent partitions; some of

these are set aside for testing, others for validation, and the remaining are used to fit the model. Then, another iteration of the method is performed, and the partitions assigned to test, validation and train set rotate. The result is that after a number iterations, a single and scarce dataset has been used used to simulate several training, test and validation sets.

The major interest of CV lies in the universality of the underlying assumptions: that data are identically distributed and that training, validation and test sets are independent. Owing not only to this universality but also to its simplicity, CV has become arguably the most commonly used method for tuning model parameters and producing robust measurements of model performance. Regarding model selection, CV procedures are recognized for their ability to control for overfitting effects, especially in datasets with limited number of observations (Arlot and Celisse, 2010).

Among the most popular types of cross-validation are the holdout and the k-Fold CV. Holdout CV (Devroye and Wagner, 1979) is the simplest CV as it relies on a single split of data. In each iteration, data is randomly divided into training and test set according to a specified ratio. On the other hand, k-Fold Cross-Validation (Geisser, 1975) splits the data into k roughly equal sized subsets. One of the subsets is retained for testing and the remaining k-1 make up the training set. In the traditional approach, a total of k iterations or folds is performed: the model is fitted k times, and k prediction errors are determined. Hence, every observation is used for both training and testing, and every observation is used for testing exactly once. The prediction error computed for the test set can then be averaged across folds in order to obtain an unbiased prediction error estimate for the model. Typical values of k range from 5 to 10. For small values of k, the training sets in each fold are quite different from the original training set and thus the cross-validation estimator usually yields low variance. However, as k decreases so does the number of samples within the training set of each fold, and thus bias may be a problem. Whether this bias is a drawback depends on how the performance of the estimator varies with the size of the training set: if there is a given size bellow which the learning performance decreases, sample size should be kept above that threshold. Overall, 5- to 10-fold procedures are recommended as a good compromise (Hastie et al., 2017).

The two types of CV fall onto the category of non-exhaustive CV. Exhaustive CV methods are so called because they train and test on all possible ways to divide the original dataset. The most common example is the leave-one-out CV (LOOCV) method, in which the test set consists of only one sample, and every sample of the data is assigned to the test set exacly once. However, because these methods are computationally very costly, these are not viable for most applications.

If the model being assessed requires the specification of prior parameters, an additional inner CV procedure may be applied. Here it is convenient to distinguish between model parameters and hyperparameters. Model parameters in the strict sense may be defined as the parameters estimated for the underlying system under analysis. For the linear regression and shrinkage methods addressed in this work, these correspond to the components of β. On the other hand, model hyperparameters may refer to parameters that must be defined prior to fitting the model, thus corresponding to the complexity parameters of the shrinkage methods discussed. Methods that, for each outer CV iteration, implement an additional inner CV procedure to tune model hyperparameters are often referred to as nested CV tech-

niques. For each inner iteration of the nested procedure, the training data is divided into two partitions, the learning set and the validation set. Then a number of hyperparameters within a range of interest is used to fit the model in the learning set, and model performance is estimated in the validation set. After several iterations, an optimal hyperparameter is selected and used to fit the model in the entire training set of the current k^{th} iteration of the outer CV procedure.

In this work a nested cross-validation procedure is implemented. For the outer cycle, a few variants of the the traditional k-fold CV procedure are explored, whereas for the inner cycle a holdout CV procedure is employed. The particulars of the CV procedures used are detailed further ahead in chapter 3.

Cross-Validation for Time-series

Important considerations must be accounted for when cross-validation is applied to time-series data, which is the framework explored in this work. Time-series data contains temporal dependencies, which means that if samples are randomly assigned to training and test sets, these subsets are no longer independent. More, some time-series are generated by non-stationary processes, in which time-evolving effects may occur, and in such cases samples are not identically distributed. These issues raise theoretical concerns, as the previous assumptions of i.i.d (independent and identically distributed) data are broken down. The question arises of whether these theoretical shortcomings are relevant in practice and lead indeed to misleading results, particularly regarding model performance estimation. Some authors (Bergmeir and Benítez, 2012; Arlot and Celisse, 2010) address such issues, and review the methods proposed in the literature to deal with CV on dependent data.

Non-dependent CV methods attempt to remove from the training set all data that is not independent from the data that is used for testing. One assumption that is common to almost every one of these methods is that the auto-correlation of two values only depends on the relative position of the values (i.e., on their lag). Under these circumstances, it is possible to assume that there is a constant h such that the samples x_i and x_j are approximately independent, if $|i - j| > h$. Hence, data correlated to the test set can be omitted by removing the previous and following h samples of all test samples. However, although such assumptions are valid for stationary processes, that is sometimes not the case for non-stationary processes. Notice that using a loose definition of stationarity, a series may be defined as stationary if its basic statistics (most importantly, the mean) are kept constant over time.

Depending on the amount of lags considered and on the size of each test set, omission of dependent values can lead to significant losses on the amount of training data. Hence, for non-dependent CV methods to be viable, the test set should only contain a low percentage of the overall data, or the amount of relevant lags must be small. Because of the former, non-dependent CV methods are often applied to LOOCV procedures. However, these are computationally costly, and hence not viable for most applications, as mentioned above. To implement non-dependent k-fold CV, the number of folds must be large enough so there remains sufficient amount of data in the training set. Although this procedure has been found to yield robust results, it requires a large waste of data, and thus is not viable for some applications.

Other methods tackle this problem by using, in each iteration, a non-interrupted portion of the time-

series as the test set. This reduces tremendously the amount of training data that has to be removed in order to ensure independence from the test set. The methods that engage in this type of data splitting may be referred to as blocked cross-validation methods, as each partition consists of a block of samples. However, for non-stationary processes this poses an additional problem, given that the variability of the continuous signal may not be captured by the test subsets, and hence training and test data are not identically distributed.

2.3.3 Criteria for Model Selection

Whatever fitting procedure is employed, a crucial step is to select the most appropriate model from a set of candidates. Model selection is key for reliable and reproducible statistical inference, and it relies a great deal in the criterion chosen for optimization. Hence wide variety of model selection criteria have been proposed and compared in the past few decades.

One of the most simple criteria to evaluate model performance is the normalized mean squared error (NMSE) (equation 2.35), which may be used as a similarity measure between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$.

$$NMSE = \frac{MSE}{Var} = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{\mathbf{Y}})^2} \tag{2.35}$$

The rationale is that if the NMSE is less than 1, then the prediction is doing better than the simple mean of the output itself. However, for model selection purposes, likelihood alone cannot be used to decide on the most appropriate model. If that was to be the case, more complex models would always be preferred over simpler models, and thus methods would be prone to overfitting effects and yield poor generalization. In traditional regression and model selection theory, one popular way to consider model complexity is to use information criteria that balance simplicity and goodness of fit, the latter measured by maximized likelihood (Aho et al., 2014). The most common among these are the AIC (Akaike, 1973) and the BIC (Schwarz, 1978), defined as:

$$AIC = -2ln(\hat{L}) + 2p \tag{2.36}$$

$$BIC = -2ln(\hat{L}) + ln(N)p \tag{2.37}$$

where p is the total number of parameters estimated in the model, defined for the shrinkage methods under study as the effective degrees of freedom, and $\hat{L}$ is the maximized value of the likelihood function of the estimated model. In the context of regression models, this corresponds to the likelihood of the parameters $\hat{\boldsymbol{\beta}}$ given the response $\boldsymbol{Y}$. Under the Gaussian model, assuming that the model residuals are i.i.d. according to a normal distribution $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, the maximum likelihood may be given by the estimated variance of the residuals, σ_ϵ^2. The latter may be expressed in terms of the residual sum of squares, $\sigma^2 = RSS/N$, and hence the previous expressions can be simplified into:

$$AIC = Nln\left(\frac{RSS}{N}\right) + 2p \tag{2.38}$$

$$BIC = Nln\left(\frac{RSS}{N}\right) + ln(N)p \tag{2.39}$$

Methods that rely on such criteria for model selection aim to pick the model that minimizes AIC or BIC. Notably, there is no clear universal choice between using AIC or BIC. To decide on which to use, the particular circumstances of the problem under analysis must be taken into account. BIC penalizes more heavily model complexity, and thus it may be more appropriate for situations where models are expected to be relatively simple and sparsity is desirable. Additionally, because the complexity penalty in BIC grows with the sample size, it is better suited for situations in which likelihood is expected to increase with the amount of data used for fitting (Aho et al., 2014).

2.4 State of the Art

The approach employed in this work was first inspired by the work of Meir-Hasson and colleagues (Meir-Hasson et al., 2014), who introduced a general framework for the prediction of patterns of electrophysiological signatures that could simulate the BOLD response. In their paper, they referred to these patterns as EEG Fingerprints (EFPs), a term further adopted in this work. To build such models, signal processing and machine learning algorithms were used to learn, from a broad set of EEG-derived features, the best parameters to predict robust estimates of the simultaneous BOLD response. This approach further extends the scope of the large body of studies that attempt to derive the best EEG correlates of BOLD response. The long-term goal of the study conducted by Meir-Hasson was that in the future, these estimated EFPs could be applied during real-time neurofeedback training, in order to attain the quality of the results obtained with a NF-fMRI session, without the need to use fMRI. Therefore, they focused on the brain activity of sub-cortical structures, under the argument that for these structures training is only possible if guided by fMRI, since the EEG alone offers little information on deep neural regions.

The framework introduced by Meir-Hasson enclosed a specific set of solutions that together aimed to build robust models with high generalization ability, so as to better predict unseen fMRI data. It modeled within a linear regression context the BOLD response of interest through a set of band-specific EEG power timecourses from a set of channels, where a range of delays was introduced at each timecourse. Hence, an optimization problem was used to search the feature space ([band] $\times$ [channel] $\times$ [delay]), instead of making any prior assumptions regarding the best frequencies, channels or HRF delays. More, because such models were likely prone to overfit the data, they used a machine learning approach that included robust statistics methods such as regularization and cross-validation for model selection and evaluation. Specifically, they performed ridge (L_2) regularization and used a nested CV procedure (outer 2-5-fold CV, inner holdout CV with n = 30 number of cycles) to control for overfitting and assess model performance. The NMSE was used as a criterion for model performance assessment. Importantly, their main conclusions were that the band-specific power across several frequency bands contributed altogether for the prediction of BOLD activity and that different brain regions were best modeled by

different weight patterns regarding frequency, channels and HRF delays. The latter result points out to the variability of the optimal EFP across brain regions, and potentially experimental conditions, which is supported by the overall literature on EEG correlated of the BOLD response.

More recenlty, Cury et al. (Cury et al., 2019) implemented a similar approach in a motor imagery study with data from simultaneous EEG-fMRI recordings, acquired during bimodal NF-EEG-fMRI sessions. Their goal was too to learn activation patterns of EEG-derived features so as to improve the quality of neurofeedback scores when EEG is used alone. Their approach was also based on a machine learning mechanism within a linear regression framework. However, in contrast to the approach of Meir-Hasson et al., neurofeedback scores were fed to the learning algorithm, instead of the EEG and fMRI signals alone. Regularization was applied by adding a mixed L_{12} penalty along with a L_1 penalty to the LSR objective function. The rationale was that because only a few brain regions are expected to be activated by a given cognitive task, it is reasonable to impose spatial sparsity through the use of a L_1 penalty. On the contrary, because engagement of brain rhythms is not expected to be sparse across frequencies, and might even be smooth, it would be appropriate to use an additional mixed L_{12} penalty. Here, a mixed norm was used instead of a simple L_2 norm to further allow non-relevant frequency bands to hold null coefficients. While the weight of the L_1 norm was empirically chosen and fixed, the weight of the L_{12} norm was estimated through an optimization procedure. Both model selection (optimization of the weight parameter) and evaluation were performed using a nested cross-validation procedure (outer 3-fold CV, inner holdout CV with n = 50 number of cycles). Here, the criterion used to assess model performance was the NMSE. Importantly, to select among the many possible models, the method sought to minimize the combined NMSE on the validation and learning set. This was justified on the grounds of whilst using the NMSE from the learning set alone would introduce bias, using the NMSE from the test set alone would introduce variance. Remarkably, Cury et al. showed that only specific channels and frequency bands held high activation, over the several different subjects for whom models were derived. This led them to believe that there may exist a common, more general, model for the population, even if the best possible models are subject-specific.

Chapter 3

Methods

This chapter introduces the approach implemented to build the EEG Fingerprints (EFPs), for both resting state and motor imagery data. For each condition, the pipeline of the method employed may be divided in the following steps: 1. data acquisition; 2. standard pre-processing of raw data to remove artifacts related to data acquisition; 3. post-processing procedures to build the feature space from the EEG data and the output response from the BOLD data; 4. model estimation (i.e., EFP estimation) and evaluation. Both data acquisition and pre-processing cleanup steps (step 1. and 2.) were performed prior to this work. Data acquisition and pre-processing operations are gathered in section 3.1 (Data Characterization), extraction of the BOLD time-series of interest is described in section 3.2 (Extraction of the Representative BOLD Response), computation of the EEG-derived features is described in section 3.3 (Extraction of the EEG Features) and the methods for model estimation and evaluation are depicted in section 3.4 (Models Estimation and Evaluation).

3.1 Data Characterization

3.1.1 Resting State Dataset

The EEG-fMRI data was acquired during rest and eyes open, for a run of 10 minutes. This dataset was acquired and pre-processed in the scope of a previous project (Abreu et al., 2017). The subject was a patient selected from a group of drug-refractory focal epilepsy, undergoing presurgical evaluation from the Program of Surgery for Epilepsy of the Hospital Center of West Lisbon. Epileptic activity was recorded on the EEG for a brief period towards the end of the acquisition. The samples corresponding to the epileptic event were removed from the data analysed in this work, so as to minimize any confounds of healthy resting state activity.

EEG-fMRI Data Acquisition

Imaging was performed on a 3T Siemens Verio scanner (Siemens, Erlanger) using a 12-channel RF coil. Whole brain functional images were acquired using a 2D multi-slice gradient-echo Echo-Planar

Imaging (EPI) sequence with the following parameters: repetition time (TR) / echo time (TE) = 2500/30 ms, voxel size = $3.5 \times 3.5 \times 3.0$ mm^3, 37 or 40 slices (interleaved acquisition). EPI (Delapaz, 1994) is a fast imaging technique, in which k-space is traversed in one or a small number of excitations, allowing to obtain individual MR slices in short time frames, thus minimizing the effects of patient motion. The 1 mm isotropic whole-brain structural image was acquired on the same scanner, using an MPRAGE sequence. The MPRAGE sequence (Brant-Zawadzki et al., 1992) is a 3D magnetization-prepared rapid gradient-echo pulse, fast to acquire and T1-weighted.

The EEG data was acquired on a MR-compatible 32-channel BrainAmp MR amplifier (Brain Products, Germany), with a standard montage according to the 10-20 system. Two additional electrodes were used: a reference electrode and an electrode placed on the back for ECG recording. The EEG acquisition was synchronized with the fMRI scanner, and sampled at 5 kHz.

EEG and fMRI Data Pre-Processing

EEG and fMRI data pre-processing were previously performed as described in Abreu et al (Abreu et al., 2017). Correction for the most common fMRI-induced EEG artifacts was performed. Specifically, EEG data underwent gradient and ballistiocardiogram (BCG) artifact correction. Subsequently, data was down-sampled at 250 Hz and band-pass filtered (1-45 Hz).

Regarding the fMRI data, pre-processing of both structural and functional images, as well as the majority of post-stats carried out, was performed using tools from the FMRIB's Software Library (FSL) (Smith et al., 2004). Non-brain tissue was removed from the structural image followed by registration to the MNI standard space (MNI152 standard-space T1-weighted average structural template image). The functional data was then registered to the structural image, and the first three volumes were discarded. Subsequently, pulsatility and respiratory induced artifacts were regressed out of the data and motion and slice timing correction was also performed. Spatial smoothing was achieved by applying a Gaussian kernel with FWHM of 5 mm (1.4 - 1.7 $\times$ voxel size). Finally, the data was subjected to high-pass temporal filtering with cut-off frequency 0.01 Hz.

3.1.2 Motor Imagery Dataset

The EEG-fMRI data was acquired from a healthy volunteer during a single motor imagery session of 120 seconds, schematically illustrated in figure 3.1. The session consisted of 6 blocks, alternating between rest, eyes open, and motor imagery of the right arm.

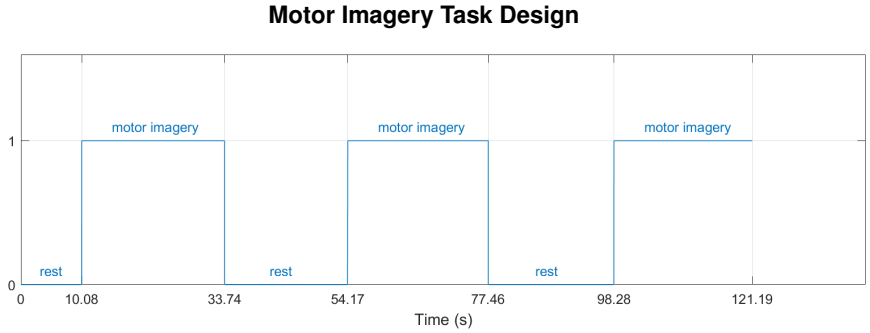


Figure 3.1: Representation of the motor imagery task block design and the corresponding timings.

EEG-fMRI Data Acquisition

Imaging was performed on a Siemens Vida at 3T with a 64-channel head coil. The fMRI data was collected using a 2DEPI sequence, with the following parameters: TR/TE = 1260/30ms, 2.2 mm isotropic, 60 axial slices, with simultaneous multi-slice acceleration factor 3, and GRAPPA in-plane acceleration factor 2. The EEG data was acquired on a MR-compatible 32-channel BrainAmp MR amplifier (Brain Products, Germany), with a standard montage according to the 10-20 system. Two additional electrodes were used: a reference electrode and an electrode placed on the back for ECG recording. The EEG acquisition was synchronized with the fMRI scanner, and sampled at 5 kHz.

EEG and fMRI Data Pre-Processing

EEG pre-processing steps were implemented in MATLAB ®. Data was filtered with a high-pass filter at 1 Hz and down-sampled at 250 Hz. Regarding the fMRI data, standard pre-processing of the functional and structural images was performed as described for the resting state fMRI data (section 3.1.1).

3.2 Extraction of the Representative BOLD Response

3.2.1 Resting State Dataset

The BOLD resting state network of interest, the default mode network (DMN), was mapped through seed-based GLM analysis. The seed (or region of interest, ROI) selected for the analysis was a well-defined area within the DMN, the PCC. A mask of the seed was created in the FSL's program FSLeyes, using the Harvard-Oxford Cortical Structural Atlas (Desikan et al., 2006). The mask was then binarized and converted from standard to functional space. The average pre-processed BOLD time-series within the PCC was thereby extracted from the functional data and used as an explanatory variable (EV) for GLM analysis. GLM was performed using the FSL's tool FEAT (Woolrich et al., 2001), and the z-statistic images returned were thresholded. Notice that high z-scores in the z-statistic images belong to brain regions that share strong temporal correlation with the PCC, and therefore are assumed to belong to the DMN. The DMN map (figure 3.2) was visually confirmed, and its average time-series extracted as the BOLD signal of interest to be further used in the analysis.

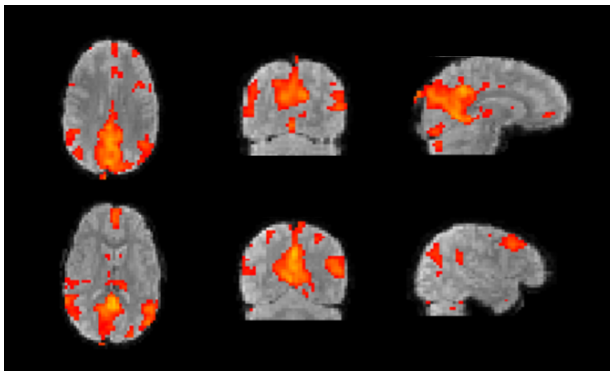


Figure 3.2: ROI from which the BOLD time-series of interest was extracted (resting state dataset). Spatial maps obtained by GLM analysis using the PCC as an explanatory variable. The red-to-yellow color images are the z-statistic images obtained, overlaid on an example fMRI volume image. This figure shows, at each column, the two most representative transverse (on the left), coronal (in the middle) and sagittal (on the right) slices, all obtained with the FSL's Fsleyes software.

The remaining processing steps were implemented in MATLAB ®. The BOLD signal was first up-sampled to 4 Hz using cubic spline interpolation. This interpolation inserts zeros between the original data values and then applies a low-pass interpolating filter to the expanded sequence. The signal was then normalized to have zero mean and one standard deviation. The resulting BOLD time-series was used as the output response $Y(t)$ of the learning algorithms.

3.2.2 Motor Imagery Dataset

The BOLD signal of interest was obtained through GLM analysis with the FSL's tool FEAT, using the square waveform of the task paradigm (figure 3.1) as an explanatory variable (EV) in the design matrix. The square wave was first convolved with the canonical HRF, as given by a double gamma function with overshoot at 6 seconds relative to onset (see details of HRF modelling in section 3.3.4). Its first temporal derivative was also included as a second regressor in the design matrix. All previous steps were implemented in the FEAT GUI. The z-statistic images returned for the EV_1 were thresholded to obtain the regions that shared high temporal correlation with the input task paradigm. The resulting map was binarized and then multiplied by a binary mask of the motor cortex to remove voxels from non-motor areas. The motor mask was created in the FSL's program FSLeyes, using the Harvard-Oxford Cortical Structural Atlas (Desikan et al., 2006) and the Jüelich Histological Atlas (Eickhoff et al., 2007). From the latter, the following cortical regions were included: the premotor cortex (PMC; Brodmann area 6), the primary motor cortex (M1; Brodmann area 4), the primary somatosensory cortex (Brodmann areas 1,2 and 3). Additionally, it was also included the supplementary motor area (SMA), from the Harvard-Oxford Cortical Structural Atlas. The mask was then binarized and converted from standard to functional space. The average BOLD time-series of the resulting map (figure 3.3) was then extracted to be further used in the analysis.

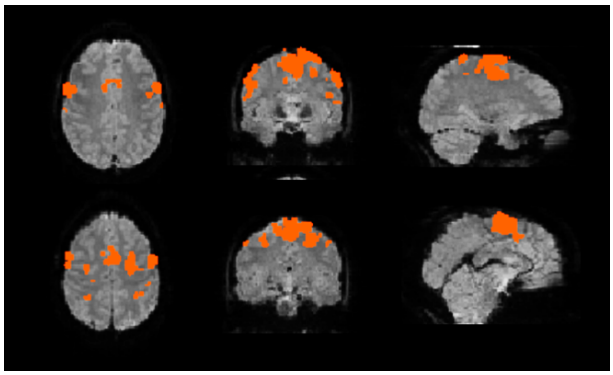


Figure 3.3: ROI from which the BOLD time-series of interest was extracted (motor imagery dataset). Spatial maps obtained by GLM analysis using the task paradigm, convolved with the HRF (double gamma with overshoot at 6 sec), as an explanatory variable. The red-to-yellow color images are the z-statistic images obtained, masked with a mask of the motor cortex, overlaid on an example fMRI volume image. This figure shows, at each column, the two most representative transverse (on the left), coronal (in the middle) and sagittal (on the right) slices, all obtained with the FSL's Fsleyes software.

The remaining processing steps, implemented in MATLAB ®, were the steps already described for the resting state BOLD signal (section 3.2.1). The resulting BOLD time-series was used as the output response $Y(t)$ of the learning algorithms.

3.3 Extraction of the EEG Features

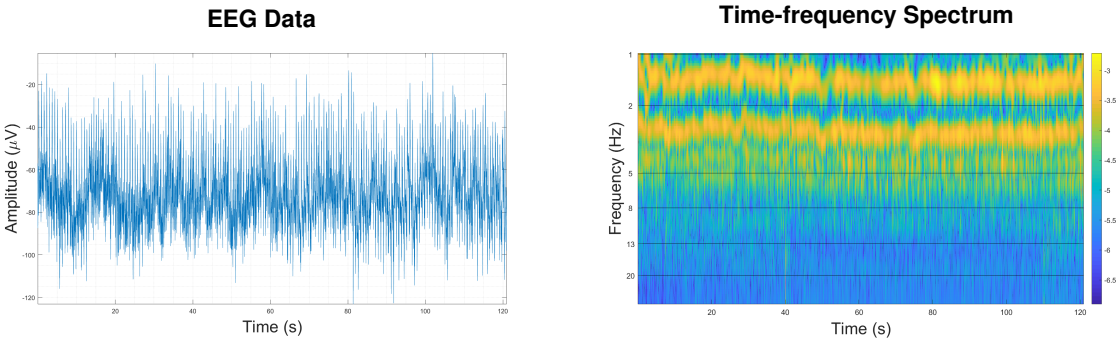
In this section, the post-processing pipeline employed to build the feature space from the EEG data is described. The input data fed to the pipeline corresponds to the post-processed EEG 2D data matrix, within which each row represents the time-series of one EEG channel. Although the original data contained information from 32 channels, the channel corresponding to the ECG time-series was not to be integrated in the learning algorithms, and hence was removed from the data matrix henceforward.

3.3.1 EEG Time-frequency decomposition

The first step was to decompose the 2D EEG data into the time-frequency domain. This time-frequency decomposition was accomplished by Morlet wavelet convolution in the time-domain. The general time-series of a Morlet wavelet, w, is described by:

$$w(t,f) = \left(\sigma_t\sqrt{\pi}\right)^{1/2} exp\left(\frac{-t^2}{2\sigma_t^2}\right) exp(i\pi ft) \tag{3.1}$$

where σ_t is the time sampling resolution of the wavelet, and thus $\sigma_t = 1/2\pi\sigma_f$. The frequency resolution of the wavelet, σ_f, may be given by f/R, where R is a constant often referred to as the wavelet factor. Notice that w may be decomposed as the product of a Gaussian wave with standard deviation σ_t and zero mean, with a sine wave with frequency f. Hence, the general shape of the Morlet wavelet is given by a rapidly decaying sine-like wave with zero mean, as illustrated in figure 3.4.



(a) (b)

Figure 3.5: Result of the time-frequency decomposition performed on the signal from channel 5 (C3) of the motor imagery dataset. (a) EEG raw data. (b) EEG time-frequency spectrum.

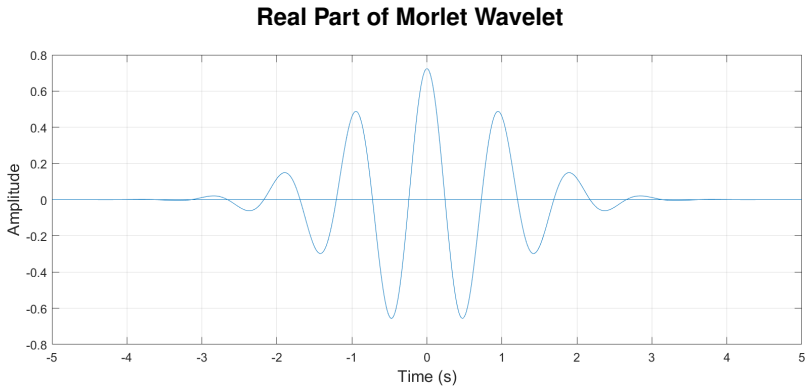


Figure 3.4: Time-domain representation of the real component of a complex Morlet wavelet with wavelet factor R=7, at frequency 1.0346 Hz.

The wavelet parameter R specifies the wavelet's number of cycles in the time-domain, thus defining the trade-off between temporal and frequency resolution. When R increases so does the number of cycles, resulting in a less steep wavelet shape in the time-domain, and hence in worse temporal resolution (because more neighbouring time-points are weighted into the central time-point upon convolution). $R = 7$ is deemed as good compromise, typically used for most applications. The frequency-domain representation of the wavelet is then a Gaussian shape (notice that the frequency domain representation of a Gaussian is a Gaussian as well), with mean corresponding to frequency of the sine wave, f. Hence, the Morlet wavelet works as a band-pass filter, with central frequency f.

Each of the EEG time-series $X_j(t)$ was then convolved with a Morlet wavelet $w(t, f)$ with wavelet factor $R = 7$, to obtain the time-series of the spectral power at frequency f, given by $P_j(t, f) = |X_j(t) * w(t, f)|^2$. This process was repeated for a 100 discrete frequency values, logarithmically distributed from 1 Hz to 30 Hz. The result was a 3D matrix of spectral power values $\mathbf{X} \in \mathbb{R}^{C \times F \times T}$ (C the number of channels, F the number of frequency bins, T the number of time-points), as illustrated in figure 3.5.

From the spectrum of the EEG data, both spectral and functional connectivity features were derived. These represent the range of possible functional links between EEG activity and the BOLD response explored in this work.

3.3.2 EEG Spectral Features

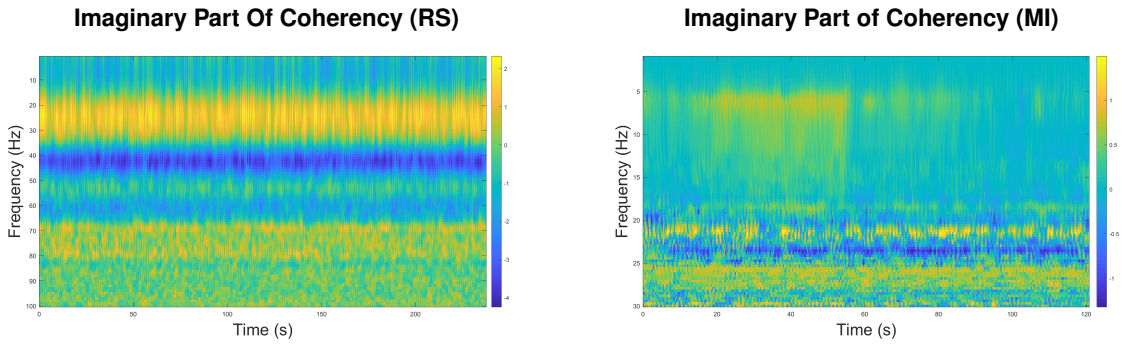
Based on the literature reviewed, the following features were explored for modeling the linear relationship between the EEG and the simultaneous BOLD within the ROI: the root mean squared frequency (RMSF), the total power (TP) and the linear combination of band-specific power (LC) (Abreu et al., 2018). The EEG-derived feature matrix, $\boldsymbol{X}$, built for each of these models, is referred to in this work as $\boldsymbol{X}_{RMSF}$, $\boldsymbol{X}_{TP}$, $\boldsymbol{X}_{LC}$ (respectively). These features were derived through the expressions in equations 2.4 to 2.8 (of the chapter Background). Because the latter two rely on the average power across specific ranges of frequencies, the frequency bands of interest had to be defined. Four frequency bands were considered in this work, defined as follows: delta (1-4 Hz), theta (4-8 Hz), alpha (8-13 Hz) and beta (13-30 Hz).

3.3.3 EEG Functional Connectivity Features

Results in the literature reviewed suggest that modeling the linear relationship between EEG functional connectivity features and the simultaneous BOLD signal may comprise a promising new field within the framework of this work. Hence, some functional connectivity measures were also explored as potential features. To avoid the necessity for source reconstruction of the EEG signals, the priority was to select features robust against volume conduction effects. Here, imaginary part of coherency (IPC) was chosen as an adequate measure to fulfill the requirements stated. As covered in previous sections, this measure suppresses all instantaneous effects ($\Delta(\Phi) = 0$), since these are only accounted for in the real part of coherency.

However, an important remark must be made regarding the computation of such features for the data under analysis in this work. Coherency (and thus, imaginary part of coherency as well) requires to determine the expected value of the cross-spectrum, $\langle S_{ij} \rangle$ (as detailed in section 2.1.1 of the chapter Background). In theory, this can only be done by averaging the cross-spectrum acquired over a sufficiently large number of trials. However, in this work, a single session was performed for both resting state and motor imagery conditions. Therefore, a different solution was employed, that relied on the Welch overlapped-segment averaging method, commonly used for spectral density estimation (Carter, 1987). Accordingly, to compute the cross-spectrum of two signals X_i and X_j throughout time, the following steps were performed: for both X_j and X_i, time intervals of 2 seconds, centered in time instant t, were divided into T overlapping segments of 250 ms, and each of the segments was windowed with a Hanning window; at each segment, the complex Fourier spectrum was obtained for both signals, and from the two Fourier spectra the cross-spectrum at that segment was computed (equation 2.1); finally, the cross-spectrum obtained for all T segments was averaged to obtain the expected value of the cross-spectrum, $\langle S_{ij} \rangle$, at time instant t. However, it is important to emphasize that the procedure described assumes stationarity within each time interval for which coherence is calculated. While this may be a fair assumption for the most part of the motor imagery time-series, it may not be the case for the resting state data, in which the time interval considered may be sufficient for a spurious process to change the connectivity patterns recorded.

The entirety of this procedure was implemented with the MATLAB ® function `cpsd`. From the cross-



(a) (b)

Figure 3.6: Imaginary Part of Coherency (IPC) estimated by the Welch method (Hanning window of 250 ms). (a) IPC between the signals of channels 10 (O2) and 18 (Cz) of the resting state dataset. (b) IPC between the signals of channels 5 (C3) and 21 (FC1) of the motor imagery dataset.

spectra obtained, coherency was computed (equation 2.2), and its imaginary part determined to derive the IPC. Pairwise IPC estimates were averaged across the four frequency bands of interest (delta, theta, alpha and beta) to obtain the feature matrix $\boldsymbol{X_{IPC}}$. Lastly, from this IPC matrix, the weighted node degree of each channel was estimated (equation 2.11), to obtain the feature matrix $\boldsymbol{X_{WND}}$.

3.3.4 EEG HRF Convolution

To better match the time-series of the simultaneous BOLD signal, each EEG feature $X_j(t)$ was non-linearly transformed through convolution in the temporal-domain with the HRF. This processing step was performed using the MATLAB ® toolbox SPM12 (Ashburner et al., 2019), designed for the analysis of brain imaging data. The time-series of the HRF was obtained using the function `spm_hrf`, which approximates the HRF to a combination of two gamma functions. Notice that this is in principle the same HRF function as the one used for the the GLM analysis of the MI fMRI data in FSL, described in section (3.2.2). Indeed, the double gamma is considered to be a standard function to model the hemodynamic response: one of the gamma functions models the response peak and the other models the post-stimulus undershoot. Notice that this function lacks the initial dip, sometimes observed in the hemodynamic response. The double gamma implemented by `spm_hrf` is characterized by a set of parameters that dictate its shape and scale, p_1 to p_5, where: p_1 is the delay of the response and p_2 the delay of the undershoot (both relative to onset), p_3 is the dispersion of the response, p_4 is the dispersion of the undershoot and p_5 is the ratio of the response to undershoot. More, `spm_hrf` requires the specification of two additional parameters, p_6 and p_7, which are the onset and length of the kernel (in seconds). The function `spm_hrf` considers the canonical (default) HRF to be defined by parameters: $p_1 = 6$, $p_2 = 16$, $p_3 = 1$, $p_4 = 1$, $p_5 = 6$, $p_6 = 0$, $p_7 = 32$. However, since the HRF is known to vary considerably across subjects, brain regions, and even cognitive tasks, a set of different HRF shapes was considered in this work, characterized by the following range of overshoot delays: 10, 8, 6, 5, 4 and 2 seconds. To maintain a linear relation between the values of the shape parameters (p_1 to p_5), the canonical values of each of these parameters were multiplied by the scale factor $s = (p_1/6)$. The family

of HRF functions used in this work is illustrated in figure 3.7.

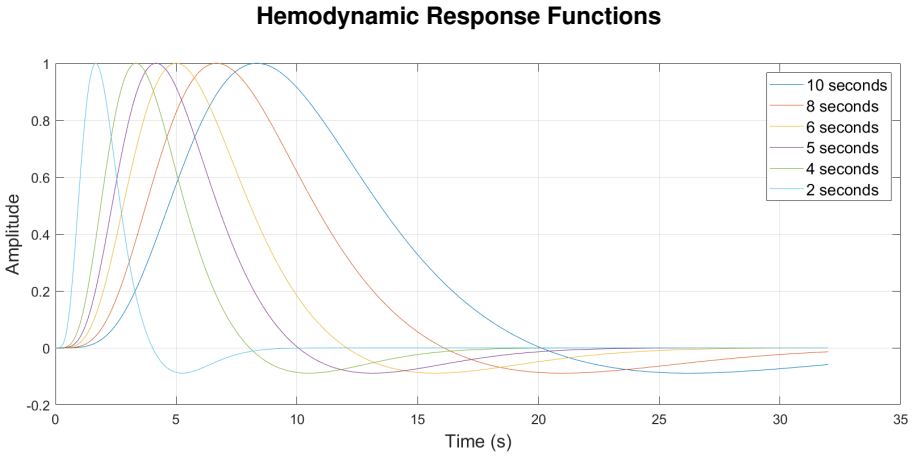


Figure 3.7: Family of hemodynamic response functions (HRFs) used in this work to perform temporal convolution of each of the EEG-derived features. Overshoot delays relative to onset of 10, 8, 6, 5, 4 and 2 seconds.

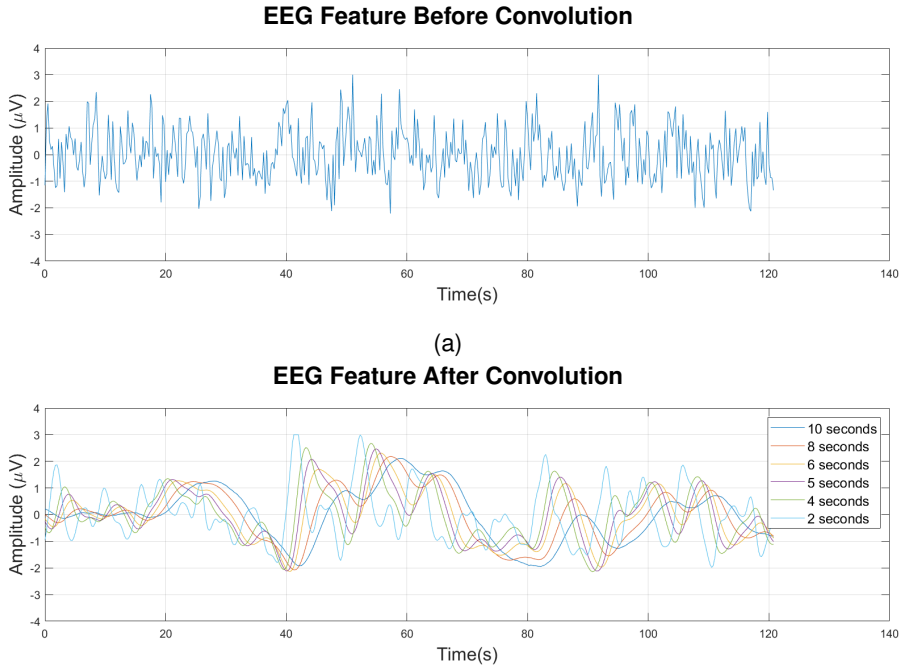


Figure 3.8: Effect of the temporal convolution between a family of HRFs and the EEG X_{TP} feature from channel 5 (C3) of the motor imagery dataset. (a) Original EEG feature. (b) EEG feature convolved with a family of HRFs with overshoot delays relative to onset of 10, 8, 6, 5, 4 and 2 seconds.

Convolution between the input feature $X_j(t)$ and the HRF was performed with the function `spm_Volterra` from the same toolbox. This convolution temporally smooths and gives a BOLD-like shape to the input features of the model, so as to increase their linear relationship with the output response. Figure 3.8 illustrates an example of the temporal convolution between one EEG feature and the family of six HRF

functions used.

3.3.5 Further Processing Steps

The resulting EEG features were down-sampled at 4 Hz, and normalized to have zero mean and one standard deviation, so as to facilitate model computation and the interpretability of coefficient estimates.

Hence, the final feature space for each of the main models considered may be decomposed into:

$$\mathbf{X}_{LC},\ \mathbf{X}_{WND}\ \in\ \mathbb{R}^{\,C\times B\times D\times T} \tag{3.2}$$

$$\mathbf{X}_{RMSF},\ \mathbf{X}_{TP}\ \in\ \mathbb{R}^{\,C\times D\times T} \tag{3.3}$$

$$\mathbf{X}_{IPC}\ \in\ \mathbb{R}^{\,C\times C\times B\times D\times T} \tag{3.4}$$

where C is the total number of channels, 31, B is the total number of frequency bands, 4, D is the total number of HRF delays, 6, and T is the total number of time points. Because the functional connectivity matrix is a symmetric matrix, redundant combinations of channels were removed from the X_{IPC} feature matrix, thus only a total of $(31 * 31 - 31)/2 = 465$ channel pairs remained. This corresponds to a total number of features p of: $31 \times 4 \times 6 = 744$ (for the $\mathbf{X}_{LC}$ and $\mathbf{X}_{WND}$ matrices); $31 \times 6 = 186$ (for the $\mathbf{X}_{RMSF}$ and $\mathbf{X}_{TP}$ matrix); $465 \times 4 \times 6 = 11160$ (for the $\mathbf{X}_{IPC}$ matrix).

After the computation of the models above, a further analysis was conducted only for the LC features, that aimed to evaluate the cost-effectiveness trade-off of computing models with a reduced feature space. Specifically, new feature matrices were derived for each of the frequency bands, $\boldsymbol{X}_{\delta}$, $\boldsymbol{X}_{\theta}$, $\boldsymbol{X}_{\alpha}$, $\boldsymbol{X}_{\beta}$, and for each of the EEG channels, $\boldsymbol{X}_{Fp1}$, ..., $\boldsymbol{X}_{POz}$. For each dataset, the one-band model and one-channel model that yielded higher performances were used to derive a matrix that only contained features from that frequency band and that channel combined (e.g., $\boldsymbol{X}_{\alpha,O2}$). Finally, another matrix was derived for the best band, the best model, and the combination of best band and model, that only contained features convolved with the canonical HRF (overshoot delay at 6 seconds), further reducing the feature space (e.g., $\boldsymbol{X}_{\alpha,can}$, $\boldsymbol{X}_{O2,can}$, $\boldsymbol{X}_{Alpha,O2,can}$).

3.4 Models Estimation and Evaluation

3.4.1 Prediction Model

In this work, an independent EFP was learnt for a range of different EEG-BOLD transfer functions, here referred to as: the LC, the TP, the RMSF, the IPC and the WND. Each of the previous transfer functions was used to integrate a prediction model defined by an unique feature space, as described in section 3.3. Regardless of the transfer function considered, the prediction model was estimated by linear regression, using the ENR (elastic net regularization) method. Hence, ENR was used to fit the training data, so as to predict the BOLD response of interest $Y(t)$, from the set of EEG features $\boldsymbol{X(t)}$. The usage of a regularization strategy such as this one was motivated by the need to reduce model complexity so as to improve model interpretability and control for overfitting effects (and hence improve overall prediction accuracy). The choice of this particular method was supported by the theoretical and empirical knowledge presented in section 2.2.1 of the chapter Background. Indeed, the ENR combines the better properties of Ridge and Lasso regression methods, yielding better performance than either of them when used alone. Specifically, it enables to produce sparse models, by setting some of the coefficients to zero, whilst producing a grouping effect, in which highly correlated features are assigned similar weights.

As already covered, the ENR incurs a penalty, imposed by a mixed L_{12} norm, on the LS objective function. Hence, it requires the specification of two hyperparameters prior to model estimation: the α parameter, which defines ratio between the weights of the L_1 and the L_2 norm, and the λ parameter, which is the complexity parameter of the model. Model selection (as defined by the tuning of its hyper-parameters) was performed through a nested CV procedure, in which the inner CV cycle was used to choose among a set of hyperparameters within a range of interest (the nested CV procedure is detailed in sections 3.4.2 to 3.4.4). Because the ENR requires the specification of two hyperparameters, two popular approaches may be employed. One possible way is to test for all the possible combinations between the range of parameters α and the range of parameters λ. Instead, α may be fixed (by defining prior assumptions or empirically testing), and hence only λ needs to be optimized. The latter approach significantly reduces the amount of computation and has the advantage of creating a common ground whereby comparison between different models is easier. Therefore, this was the one implemented in this work, as α was set to 0.5 in all models. The choice of this value was twofold: first, to impose sufficient sparsity on the models, small values of α were discarded (notice that sparsity increases as α rises from 0 to 1); additionally, a simplistic parameter search was employed, in which model performance, as measured by the BIC, was assessed for a range of α values. Figure 3.9 illustrates a representative run of results.

The pattern represented in figure 3.9 was somewhat consistent for all models and both datasets. The lower BIC was typically easier to identify with $\alpha = 0.5$ and would often correspond to lower model complexities (higher λ values). The parameter λ was deemed as more important to the optimization problem, and hence estimated automatically. The list of parameters to test $\Lambda = \{\lambda_1, ..., \lambda_l\}$ was obtained by the MATLAB ® function `lasso`. This function performs elastic net regularization regression on the

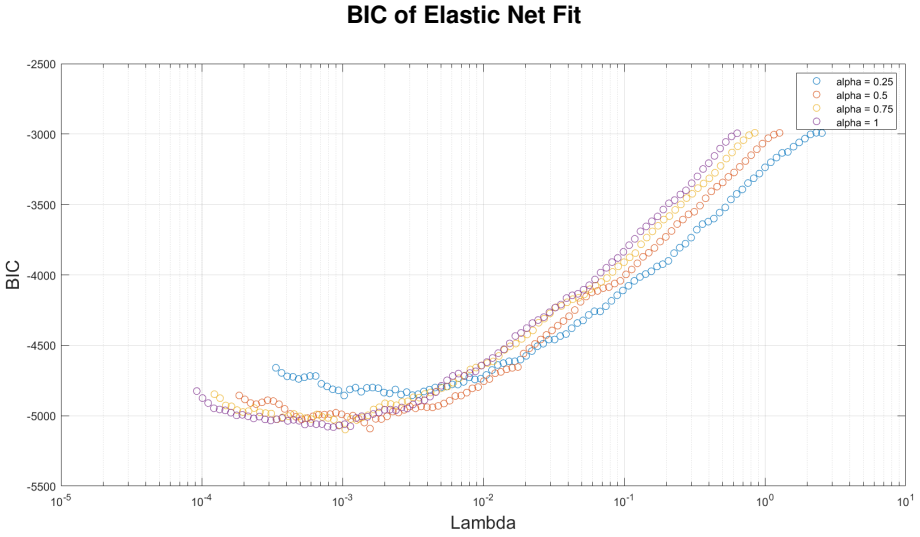


Figure 3.9: Cross validated Bayesian Information Criterion (BIC) of the elastic net fit vs λ, for a set of α parameters. Example taken from the modelling of input features X_{LC} of the motor imagery dataset.

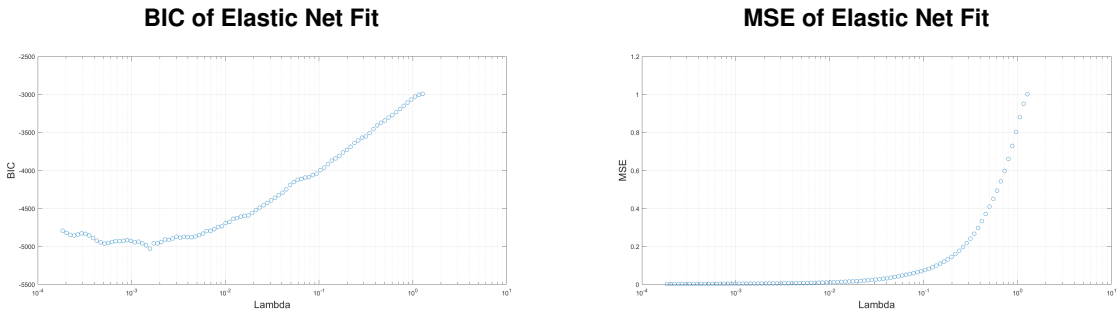
input variables in $\boldsymbol{X}$ and Y, for a specified $\alpha \in]0, 1]$. By default, the function calculates the largest value of λ that gives a non-null model (one effective DOF), and assigns this as the upper limit of the list Λ. Then, it assigns the lower limit of the list through an established ratio between the upper and lower limit, which was set to $1\mathrm{e}{-2}$. The total length of the list was set to 20 ($\Lambda = \{\lambda_1, ..., \lambda_{20}\}$), and the distribution of values between the lower and upper limit was defined to be logarithmic.

3.4.2 k-Fold Cross Validation

For model selection and evaluation, a nested k-fold CV procedure was implemented, with $k = 15$ outer cycles and $n = 20$ inner cycles. For the outer procedure, a few variants of the traditional k-fold CV procedure were explored, whereas for the inner procedure a holdout CV procedure was employed, with learning/validation ratio of $70/30$.

The first version of k-fold CV explored, which is also the most simple, is described in this section. The original dataset was split into k roughly equal sized, random partitions. In other words, each sample of the original set $\{1, ..., N\}$ was randomly allocated into a specific partition $\{1, ..., k\}$. Then, a total of k outer iterations was performed, and at each iteration one partition was retained as the test set, while the remaining $k - 1$ partitions were used to fit the model. After all iterations, each partition was used as the test set exactly once. For each of the k training sets, an inner holdout CV procedure with n iterations was performed to estimate the optimal hyperparameter λ for that particular set. In each iteration of the inner procedure, training data was randomly assigned to the learning and validation sets, according to a ratio of $70/30$. Data within the learning set was used to compute a family of 20 models, one for each of the 20 λ values belonging to the list $\Lambda = \{\lambda_1, ..., \lambda_{20}\}$.

The criterion used for model selection was the BIC value. It was chosen over the AIC criterion because it was deemed as more appropriate for the framework of this work, given that it penalizes more heavily model complexity, and that the size of the penalty grows with the total number of samples



(a) (b)

Figure 3.10: Cross Validated (a) BIC and (b) MSE of the elastic net fit for a range of λ values. Example took from the modelling of input features X_{LC} of the motor imagery dataset.

(as discussed in section 2.3.3 of the Background). Accordingly, the optimal $\hat{\lambda}$ selected was the one to minimize the combined BIC on the validation and learning set, summed across all N iterations of the inner CV procedure, according to the expression:

$$\hat{\lambda} = \arg\min \left\{ \sum_{n=1}^{N} (BIC_\lambda)_n|_{learn} + \sum_{n=1}^{N} (BIC_\lambda)_n|_{val} \right\} \tag{3.5}$$

This solution was inspired by the procedure described in Cury et al. (Cury et al., 2019), and sought to avoid the bias that would be introduced if using the BIC from the learning set alone, as well as the variance that would be introduced if using the BIC from the validation set alone.

However, because the data being modeled in this work is time-series data, considerations regarding temporal dependencies should be accounted for. Importantly, samples within the test set are not independent from those of the training set, and thus estimates of model performance may be overoptimistic. Since this is a critical aspect for model assessment and comparison, two modified CV methods to deal with dependent data were also explored in replacement of the outer k-fold CV procedure. These two methods, here referred to as non-dependent k-fold CV and blocked k-fold CV, are described in the following sections (3.4.3 and 3.4.4).

Additionally, temporal dependencies may also exist between the learning and validation sets of the inner CV procedure, which may be suboptimal for hyperparameter selection. Specifically, simplicity may not be favored enough due to good apparent performance, in terms of the estimated prediction error, of models that are significantly overfitted to the learning set. However, CV procedures that remove temporal dependencies lead to significant data waste, and the data available in this work was insufficient to apply such procedures in the inner CV cycles as well. Removing temporal dependencies from the outer cycles was prioritized, since the first concern was to obtain reliable estimates of the final model's prediction error. More, notice that despite the inner CV procedure may not promote model simplicity, the BIC is used as a criterion for model selection, and this acts as a safeguard to avoid too complex models that may be prone to overfit. This effect is illustrated in figure 3.10.

As figure 3.10 illustrates, although the minimum values of the mean squared prediction error (MSE) occur at the higher model complexities (lower λ values), BIC weights in the effective DOF of the models,

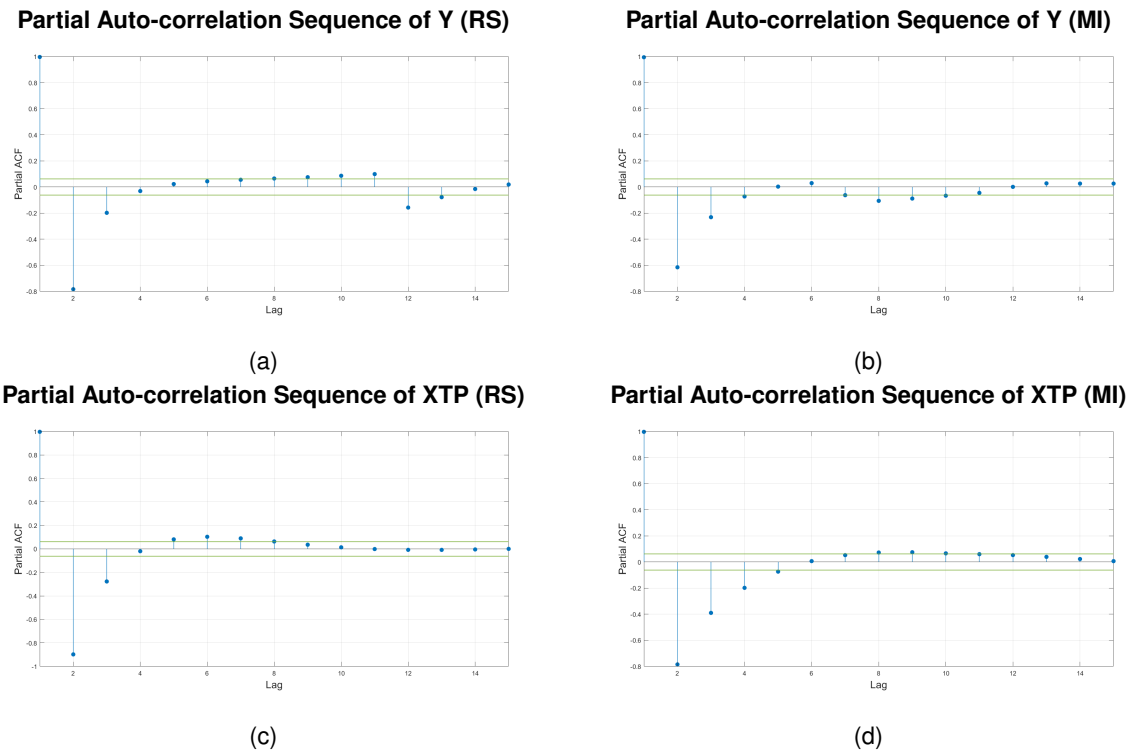
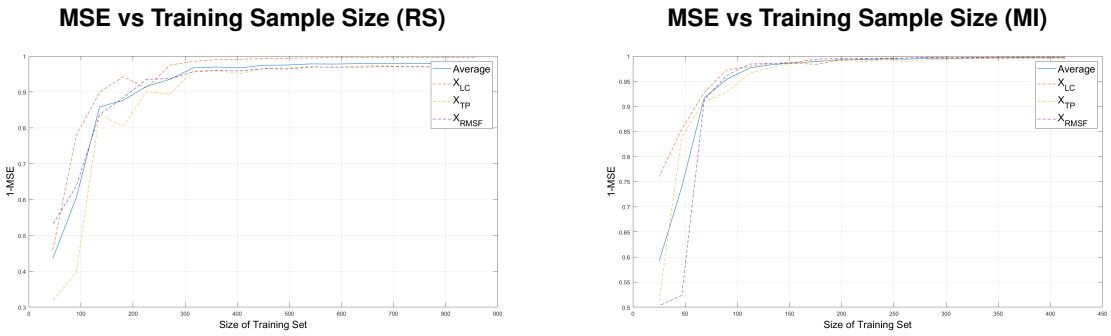


Figure 3.11: Auto-Correlation Function (ACF), out to lag 15, of four representative time-series. The two horizontal green lines represent the 95% confidence bounds. (a) ACF of the BOLD response Y of the resting state dataset. (b) ACF of the BOLD response Y of the motor imagery dataset. (c) ACF of EEG feature X_{TP}, from channel 10 and with HRF delay of 6 seconds, of the resting state dataset. (d) ACF of EEG feature X_{TP}, from channel 5 and with HRF delay of 6 seconds, of the motor imagery dataset.

thus being minimized at more moderate model complexities.

3.4.3 Non-dependent k-Fold Cross Validation

A nested non-dependent k-fold CV procedure was implemented, with $k = 15$ outer cycles and $n = 20$ inner cycles. The inner CV remained absolutely unchanged. In each iteration of the outer procedure, each of the samples of the original dataset $\{1, ..., N\}$ was randomly allocated into one of k equal sized partitions. However, right after allocation, samples within the training set that shared dependencies with any sample within the test set were removed. Importantly, the temporal correlation between two samples was assumed to only dependent on their lag, which is to assume that there is a constant h such that samples x_i and x_j are approximately independent, if $|i - j| > h$. For a given time-series, the value of this constant may be determined by analysis of its partial auto-correlation sequence. To obtain an approximate measure of h, the partial auto-correlation sequence, out to lag 15, was determined for the output of the model Y and for several representative input features within $\boldsymbol{X}$. Some of the results obtained are displayed in figure 3.11.



(a) (b)

Figure 3.12: Change of learning performance, in terms of model Mean Squared Error (MSE) (estimated by cross-validation), with the size of the training set. (a) Evaluation performed for the input features $\boldsymbol{X}_{LC}$, $\boldsymbol{X}_{TP}$ and $\boldsymbol{X}_{RMSF}$ of the resting state dataset. (b) Evaluation performed for the input features $\boldsymbol{X}_{LC}$, $\boldsymbol{X}_{TP}$ and $\boldsymbol{X}_{RMSF}$ of the motor imagery dataset.

The lag values of the auto-correlation sequence outside the 95% confidence bounds indicate that data may be generated by an auto-regressive process of order equal to those lags. The analysis of the partial ACF plots suggests that the assumption that the auto-correlation of two samples only depends on their lag may be a fair approximation, since the auto-correlation decreases with increasing lag. More, it may be extrapolated that each sample is significantly correlated with the 3 previous samples and the 3 following samples. However, to minimize data waste, $h = 2$ was chosen, since it should allow to remove almost all of the significant dependencies. Hence, the 2 neighboring samples, in both axis directions, of each test sample were removed from the training set. It was also confirmed that, after removal of dependent data, the size of the training set remained within the acceptable limits in terms of model performance (figure 3.12).

As observable in the figure, the learning performance has a considerable slope at training size $N = 300$ (for the resting state dataset) and $N = 100$ (for the motor imagery dataset) and a very high slope at $N = 100$ (for the resting state dataset) and $N = 50$ (for the motor imagery dataset). If the training size is bellow the these values, then cross validation will overestimate the true prediction error. This effect is particularly worrying for training sizes bellow the second limit ($N = 100$ for the RS and $N = 50$ for the MI), which is why it is important to guarantee that these limits are not trespassed after removing dependent data.

To conclude, with this modified CV procedure, the estimated prediction errors obtained is then expected to be a reliable estimate of the true prediction error.

3.4.4 Blocked k-fold Cross Validation

Finally, another modified k-fold CV procedure, referred to as blocked CV, was also explored. Similarly to the non-dependent CV, this procedure too works by removing the $h = 2$ neighboring samples of the test samples. The difference relies on the way data is allocated to each partition: instead of randomly allocating samples to one of the k partitions, partitions are created by chronological order, i.e., they

consist of uninterrupted blocks of samples. This tremendously minimizes data waste, since only a total of 4 samples need to be removed from the training set in each iteration: the 2 samples before the first sample of the test set, and the 2 samples that follow the last sample of the test set.

However, it should be considered that this procedure may be suboptimal for highly non-stationary processes, since the variability of the series throughout time may not be captured at each partition. Hence, if a specific process occurs at a specific moment in time, it is probable that the unique structure that it incurs on the data will not be considered for training and for testing simultaneously. For the particular case of the motor imagery data, a 15-fold blocked CV procedure will certainly not capture resting and task periods on the same subset. As so, k was changed to 5. Indeed, by dividing the data in only 5 blocks, each block is guaranteed to capture one trial of rest and imagery. The same procedure was applied to the resting state dataset, to keep the pipeline the most similar as possible.

3.4.5 Estimation of the EFPs

Because models were fitted by k-fold CV, a total of k sets of coefficient estimate vectors β were obtained for each method. However, one of the goals of this work was to analyse the clusters of channels, HRF delays and frequency bands that yielded higher weights in the final models, so as to compare them to the known paradigms from both RS and MI studies. Hence, a final EEG Fingerprint (EFP) was estimated for each method, by computing the average of non-zero values of the components in β across the k iterations. Then, two final representations of the EFP, illustrated in figure 3.13, were always obtained.

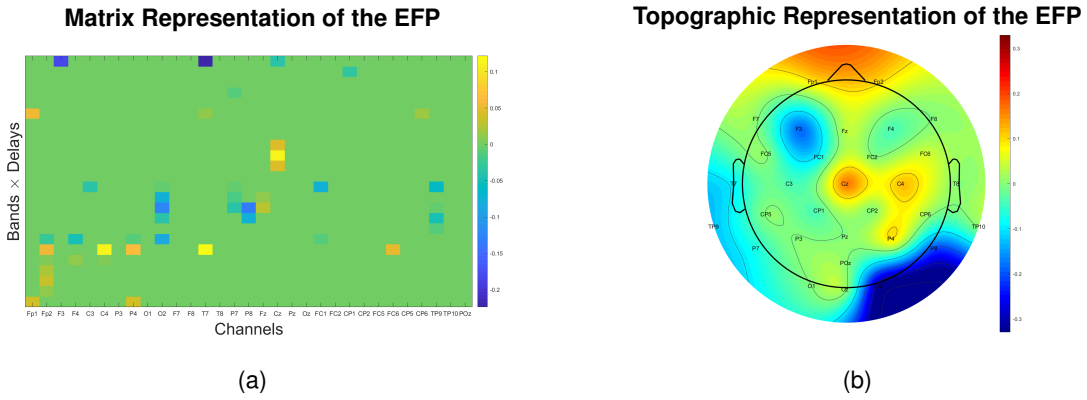


Figure 3.13: Example of one EEG Fingerprint (EFP), estimated with the X_{LC} model for the resting state data. The sizes of the coefficient estimates are color-coded. (a) Matrix representation of the EFP: rows represent the HRF delays (in decreasing order from top to bottom) for each of the frequency bands; columns represent the EEG channels. (b) Topographic representation of the EFP: coefficient estimates are averaged across delays and frequency bands, for each EEG channel (equivalent to averaging each column of the matrix representation).

3.4.6 Statistical Analysis

The main models for each CV procedure were compared via Analysis of Variances (ANOVAs), in order to evaluate whether significant differences (p-value < 0.05) regarding model performance were obtained. For this evaluation, model performance was assessed independently through NMSE and BIC results. This analysis was performed using the MATLAB ® function `anova1`. The statistics returned by `anova1` further allowed to run multiple comparison tests, which enabled to determine which pairs of models held significantly different predictive performances. The latter were performed with the MATLAB ® function `multcompare`.

Chapter 4

Results

In this chapter, the elastic net fitting results are summarized. The chapter is divided in two main sections: section 4.1 introduces the results obtained for the resting state (RS) dataset and section 4.2 introduces those obtained for the motor imagery (MI) dataset. Both these sections are divided into three subsections: first, the fitting results of all methods and models employed are described; then, the estimated EEG fingerprints (EFPs) are displayed in the form of topographic maps; finally, the predictive performance of all models is compared via analysis of variance (ANOVA).

4.1 Resting State

4.1.1 Results of Model Fitting

In this section, the average results of the elastic net fitting are reported. The first part of this section presents results concerning the five main models assessed in this work, here termed as: the X_{RMSF} model, the X_{TP} model, the X_{LC} model, the X_{IPC} model and the X_{WND} model. The second part of this section presents the results regarding models derived from feature space reduction of the X_{LC} model (one-band models, one-channel models, etc).

Models with All Features

The tables bellow summarize the average results obtained with each of the cross validation procedures implemented: tables 4.1, 4.2 and 4.3 show the results for 15-fold CV, non-dependent 15-fold CV and blocked 5-fold CV, respectively. The measures reported are the λ parameter and the NMSE and BIC values, estimated by averaging across all folds. Notice that the average λ cannot be regarded as an actual parameter of the model, since in each fold the λ used to fit the data was different. However, it is shown in these tables in order to provide a sense of the amount of regularization incurred. More, the estimated effective DOF of each model is also presented. This was determined from the final EFP of the model, derived by averaging the non-zero values of the coefficient estimates across all folds. An important observation drawn from all the results in this section is that the size of the λ parameter

was significantly high, regardless of the CV procedure, and certainly high enough to greatly reduce model complexity, especially for the models with a large feature space. This effect may be observed by comparing the DOF of the final models with the original number of features. Additionally, the amount of regularization was particularly high for the blocked CV procedure with the X_{LC} and X_{WND} models. Since the λ parameter was determined through the inner CV procedure, which was the same for the three methods employed, this result possibly arises from the increase in the size of the training data with 5-fold CV. Indeed, BIC penalizes model's complexity more heavily for large training subsets.

Table 4.1: Cross validated (15-fold CV) λ parameter, NMSE and BIC values of the elastic net fit of the main models explored: X_{RMSF}, X_{TP}, X_{LC}, X_{IPC} and X_{WND}. Effective degrees of freedom (DOF) of the final model, derived by averaging the nonzero values of the fitted coefficient estimates across all folds. Results respective to the RS data.

	Average λ $(\times 10^{-2})$	DOF	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
RMSF	6.68	26.0	4.56	1.07
TP	5.02	46.0	6.24	1.71
LC	8.77	47.0	2.93	2.56
IPC	7.72	124.0	0.797	8.92
WND	5.56	80.0	2.26	4.16

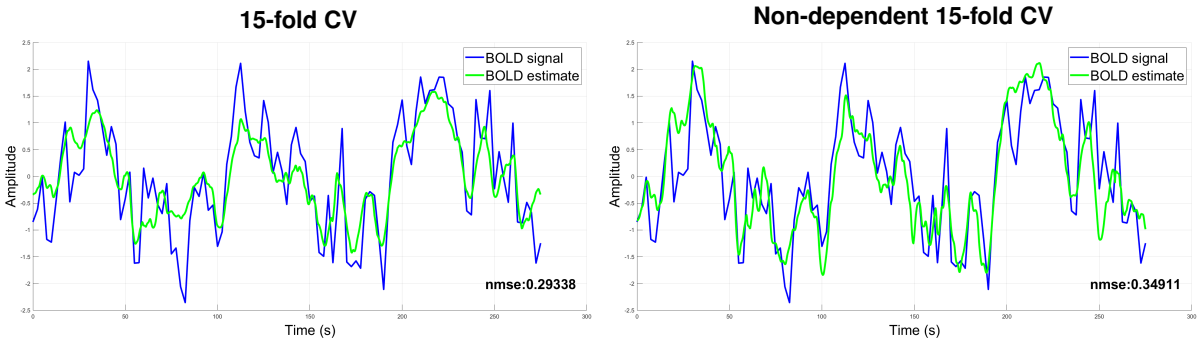
Table 4.2: Cross validated (non-dependent 15-fold CV) λ parameter, NMSE and BIC values of the elastic net fit of the main models explored: X_{RMSF}, X_{TP}, X_{LC}, X_{IPC} and X_{WND}. Effective degrees of freedom (DOF) of the final model, derived by averaging the nonzero values of the fitted coefficient estimates across all folds. Results respective to the RS data.

	Average λ $(\times 10^{-2})$	DOF	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
RMSF	5.07	74.0	4.77	1.49
TP	8.16	99.0	6.82	2.23
LC	7.10	125.0	3.49	3.02
IPC	9.62	234.0	1.80	7.74
WND	6.60	148.0	3.72	4.01

Figure 4.1 displays an example of the prediction of the BOLD response from the region of interest. The particular example shown concerns the prediction performed by fitting the X_{LC} model, using regular 15-fold CV (figure (a)) and non-dependent 15-fold CV (figure (b)). The BOLD estimate represented was computed from the final EFP derived in each of the procedures.

Table 4.3: Cross validated (blocked 5-fold CV) λ parameter, NMSE and BIC values of the elastic net fit of the main models explored: X_{RMSF}, X_{TP}, X_{LC}, X_{IPC} and X_{WND}. Effective degrees of freedom (DOF) of the final model, derived by averaging the nonzero values of the fitted coefficient estimates across all folds. Results respective to the RS data.

	Average λ $(\times 10^{-1})$	DOF	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
RMSF	0.550	26.0	5.50	1.22
TP	0.497	42.0	7.91	1.79
LC	5.55	41.0	6.35	1.57
IPC	0.710	104.0	0.743	9.18
WND	1.61	71.0	3.72	3.58



(a) (b)

Figure 4.1: Example of prediction of the BOLD response. BOLD signal (Y, in blue) and respective BOLD estimate ($\hat{Y}$, in green), obtained with the EFP estimated for the X_{LC} model. (a) EFP estimated by 15-fold cross Validation. (b) EFP estimated by non-dependent 15-fold cross validation. Average NMSE of the fitting displayed on the lower right corner. Results respective to the resting state data.

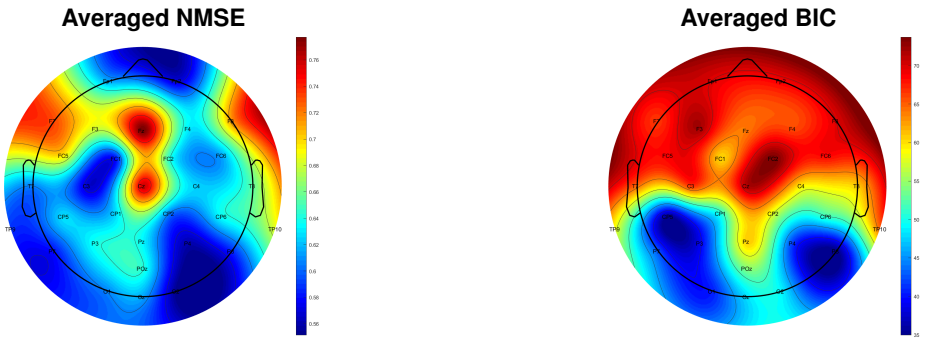
Models with Reduced Feature Space

The following tables report the results obtained in the scope of reducing the feature space of the X_{LC} model. The parameters referent to the X_{LC} model are always shown as well, for comparison purposes. Because the relationship between models remained similar across all CV procedures explored, only the results of regular 15-fold CV procedure are presented in this section (the results obtained with the other two procedures are presented in appendix A.1). Table 4.4 summarizes the results of the one-band models. This analysis revealed that the X_{α} model yielded higher predictive performances than the remaining one-band models, both in terms of NMSE and BIC. This result was consistent across all CV procedures employed.

In this scope, one-channel models were also fitted. Figure 4.2 displays the topography of the predictive performance, in terms of NMSE and BIC, estimated for the model fitted at each channel with 15-fold CV.

Table 4.4: Cross validated (15-fold CV) λ parameter, NMSE and BIC values of the elastic net fit of the models X_{LC}, X_{δ}, X_{θ}, X_{α} and X_{β}. Effective degrees of freedom (DOF) of the final model, derived by averaging the nonzero values of the fitted coefficient estimates across all folds. Results respective to the RS data.

	Average λ $(\times 10^{-2})$	DOF	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
LC	8.77	47.0	2.93	2.56
Delta	5.55	43.0	6.49	1.70
Theta	4.87	30.0	5.23	1.31
Alpha	5.40	30.0	3.56	1.11
Beta	18.17	45.0	6.03	1.37



(a) (b)

Figure 4.2: Topography of the predictive performance of one-channel models, as measured by the NMSE and BIC, estimated through 15-fold CV. (a) NMSE, averaged across all folds, of the model fitted at each channel. (b) BIC, averaged across all folds, of the model fitted at each channel. Results respective to the resting state data.

A rough analysis of the topographies shown reveals that the best BIC performances were achieved by the models derived with posterior EEG channels (particularly in occipital areas), whereas the best NMSE performances were achieved with a more distributed range of channels. A more detailed analysis was performed for the four one-channel models that yielded better NMSE performance. The parameters estimated for such models are reported in table 4.5.

Model X_{O2} yielded better predictive performance, in terms of NMSE, in all of the CV procedures. However, in terms of model BIC, neither one of the models was consistently better across all CV procedures. The channel selected to perform further analyses was then the channel O2. Hence, features with respect to the alpha frequency band and channel O2 combined were used to build the model $X_{\alpha,O2}$. More, the feature space of models X_{α}, X_{O2} and $X_{\alpha,O2}$ was further reduced by solely integrating features that were convolved with the canonical HRF (overshoot delay at 6 seconds). The results referent to this analysis are summarized in table 4.6.

Table 4.5: Cross validated (15-fold CV) λ parameter, NMSE and BIC values of the elastic net fit of the models X_{LC}, X_{O2}, X_{P4}, X_{C3} and X_{Fp2}. Effective degrees of freedom (DOF) of the final model, derived by averaging the nonzero values of the fitted coefficient estimates across all folds. Results respective to the RS data.

	Average λ $(\times 10^{-1})$	DOF	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
LC	0.877	47.0	2.93	2.56
O2	1.10	13.0	5.51	0.489
P4	1.76	12.0	5.69	0.412
C3	0.584	19.0	5.58	0.666
Fp2	0.067	19.0	5.69	0.690

Table 4.6: Cross validated (15-fold CV) λ parameter, NMSE and BIC values of the elastic net fit of the models X_{LC}, X_{α}, $X_{\alpha,\,can}$, X_{O2} and $X_{O2,can}$, $X_{\alpha,O2}$ and $X_{\alpha,O2,can}$. Effective degrees of freedom (DOF) of the final model, derived by averaging the nonzero values of the fitted coefficient estimates across all folds. Results respective to the RS data.

	Average λ $(\times 10^{-1})$	DOF	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{1})$
LC	0.877	47.0	2.93	25.60
Alpha	0.540	30.0	3.56	11.12
Alpha,can	0.672	11.0	5.20	3.68
O2	1.10	13.0	5.51	4.89
O2,can	3.64	2.0	6.99	0.392
Alpha,O2	0.629	6.0	5.99	1.78
Alpha,can	3.64	2.0	6.99	0.392

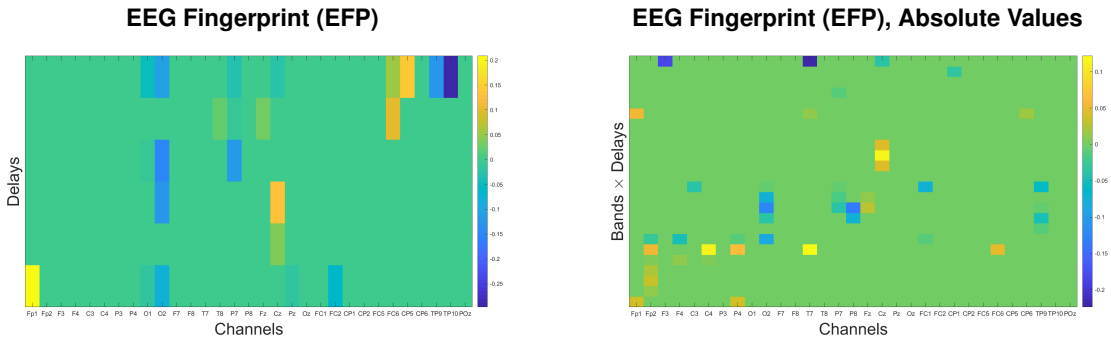
4.1.2 EEG Fingerprints

This section presents the most relevant topographic representations of the EEG Fingerprints (EFPs) estimated for each model. Because maps were consistently similar across CV procedures, only those referent to the EFPs estimated by regular 15-fold CV are displayed.

Models with All Features

Here the EFPs obtained for the main models, the X_{RMSF}, X_{TP}, X_{LC}, X_{IPC} and X_{WND}, are shown. Two examples of the matrix representation of the EFP are shown in figure 4.3.

From these figures it is possible to verify that the models estimated were significantly sparse, and that the number of parameters was successfully decreased by the regularization procedures. These results validate the decision of imposing regularization strategies on the models assessed, since the estimated models were able to perform rather successfully using a way smaller cluster of features than



(a) (b)

Figure 4.3: Example of one EEG Fingerprint (EFP) in matrix representation. The sizes of the coefficient estimates are color-coded. (a) EFP estimated for the X_{RMSF} model: rows represent each of the HRF delays, 10, 8, 6, 5, 4 and 2 seconds (top to bottom); columns represent the EEG channels. (b) EFP estimated for the X_{LC} model: rows represent the HRF delays for each of the frequency bands, δ, θ, α, β (top to bottom); columns represent the EEG channels. Results respective to the resting state data.

those that were originally considered. The following figure 4.4 displays a topographic representation of the EFPs, obtained by averaging the coefficient estimates across all frequency bands and HRF delays.

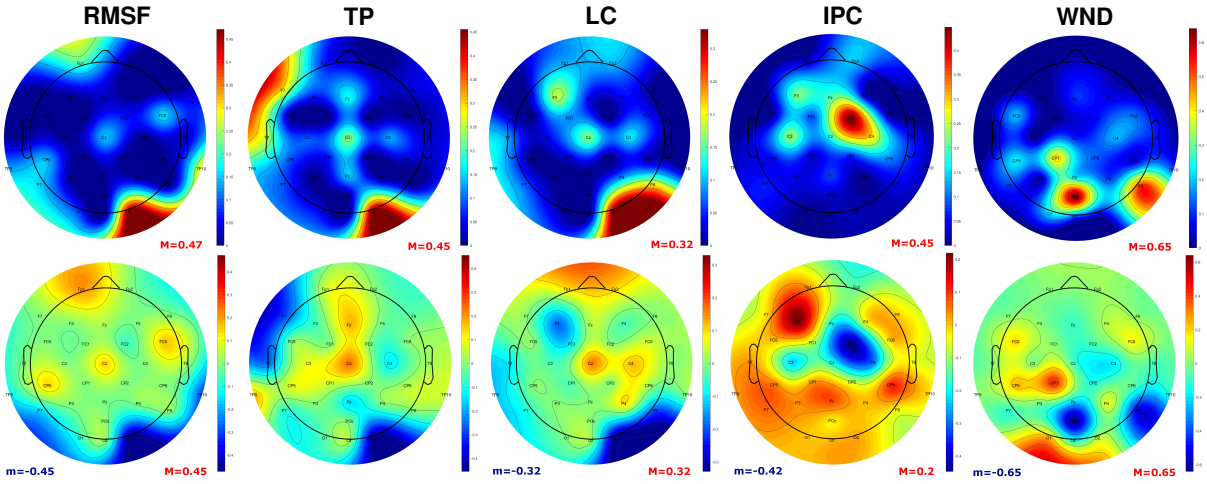


Figure 4.4: Topographic representation of the coefficient estimates (EFP), averaged across all folds, obtained for each of the main models: X_{RMSF}, X_{TP}, X_{LC}, X_{IPC} and X_{WND}. Bottom line: coefficient estimates averaged across all HRF delays and (in the case of the X_{LC}, X_{IPC} and X_{WND} models) across all frequency bands. Top line: heat map representing the absolute values of those coefficient estimates. Minimum and maximum values are indicated for each map with m (blue) and M (red). Results respective to the resting state data.

One important observation considering all the topographies shown, is that only a limited set of EEG channels yield high coefficient sizes. Such sparsity across channels has been previously reported within a similar framework (Cury et al., 2019). From the EFPs of the models that included spectral EEG features (X_{RMSF}, X_{TP} and X_{LC}), the topographies displayed are considerably consistent, regardless of the model:

the highest weights are found within occipital regions, in particular in the channel O2, which holds a large negative coefficient. Regarding the models that included connectivity features (X_{IPC} and X_{WND}), such clean and interpretable maps were not achieved. The topographies are not similar across the two, although it must be considered that each of these models measures a distinct aspect of functional connectivity: while the former weights each pairwise connectivity individually, the latter weights the overall connectivity at each channel.

To obtain more insights on the frequency profile of the EFPs derived for the X_{LC} model, these results were also organized in another topographic representation (figure 4.5).

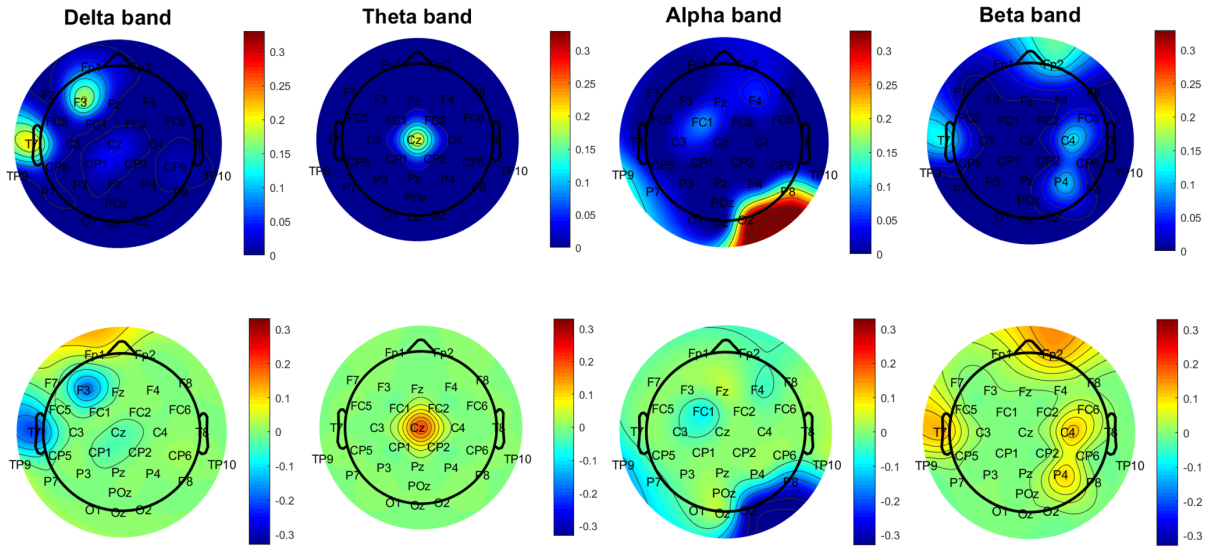


Figure 4.5: Topographic representation of the coefficient estimates (EFP), averaged across all folds, obtained for the X_{LC} model. Bottom line: coefficient estimates corresponding to each of the frequency bands, averaged across all HRF delays. Top line: heat map representing the absolute values of those coefficient estimates. Results respective to the resting state data.

The representations in this figure were all plotted following an unique color-code. Comparison of the absolute topographies in the upper figure shows that the higher coefficient sizes were estimated for features corresponding to the alpha frequency band. Most importantly, it is possible to conclude that alpha band features were responsible for the high coefficient sizes within the occipital regions observed in the overall topography of figure 4.4. Notably, the negative value of such coefficients points out to to a negative correlation between the alpha occipital power and the BOLD changes within the DMN, which is in line with several previous findings (Bowman et al., 2017). Similar conclusions may be drawn by observing the frequency profiles of the EFP obtained for the X_{WND} model (figure 4.6).

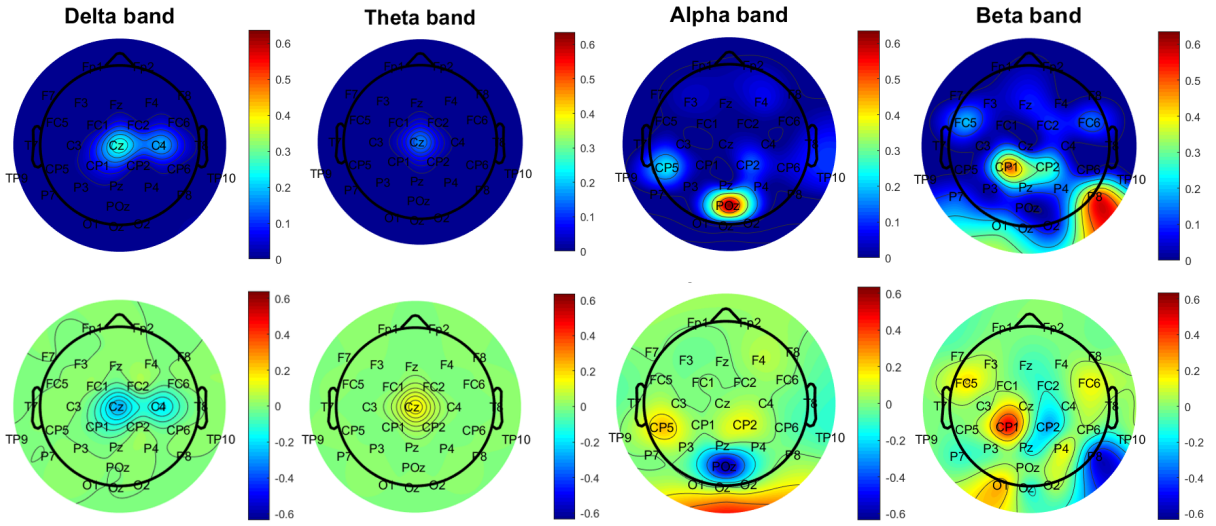


Figure 4.6: Topographic representation of the coefficient estimates (EFP), averaged across all folds, obtained for the X_{WND} model. Bottom line: coefficient estimates corresponding to each of the frequency bands, averaged across all HRF delays. Top line: heat map representing the absolute values of those coefficient estimates. Results respective to the resting state data.

Here too, some of the higher coefficients estimated belong to the alpha band within occipital regions (in this case, the channel POz), and these coefficients hold negative values as well. However in this model, negative coefficients with high absolute value were also found for the beta band within parietal regions.

Models with Reduced Feature Space

Further relevant conclusions may be drawn by analysing the topographies of the EFP obtained for the X_{α} model. The results obtained for this model should provide more detailed insights on the temporal structure of the alpha band features, since all other features are removed, and thus prevented from confounding the interpretation of the alpha coefficients. Figure 4.7 shows that the negative relationship between the occipital channels and the output response of the system is most prominent in the feature that was convolved with the canonical HRF delay.

4.1.3 Comparison of Methods and Models

The main models for each CV procedure were compared via analysis of variance (ANOVA), in order to evaluate whether significant differences (p-value < 0.05) regarding model performance were obtained. For this evaluation, model performance was assessed independently through the NMSE and BIC results. The results of the pairwise comparison tests among all models and for all methods are presented in appendix B.1. In figures 4.8 to 4.11, results of model comparison are displayed in the form of boxplots. The central mark within each box indicates the median of the corresponding group (i.e., of the NMSE or BIC values of that model across all folds), and the bottom and top edges indicate the 25th and 75th

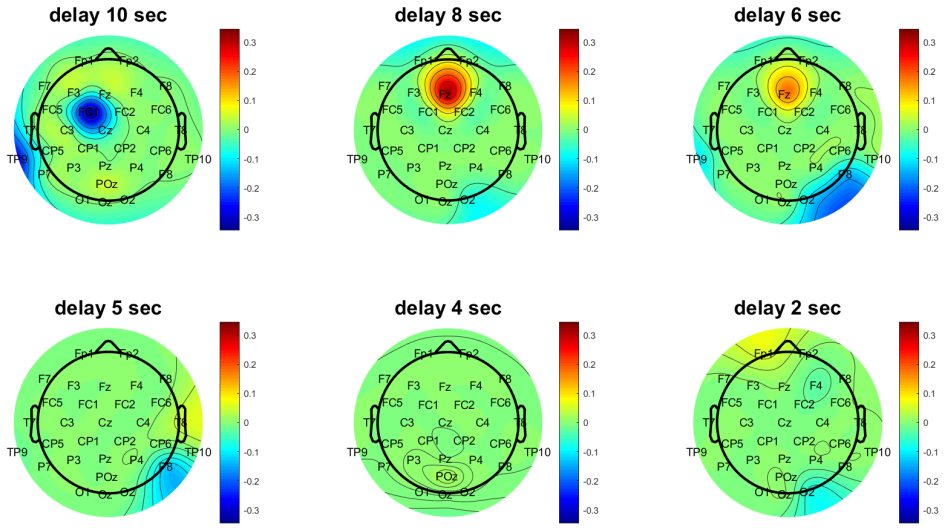


Figure 4.7: Topographic representation of the coefficient estimates (EFP), averaged across all folds, obtained for the X_{α} model. Coefficient estimates relative to each HRF delay displayed for each EEG channel. Results respective to the resting state data.

percentiles (respectively). The whiskers extend to the most extreme values of the groups (outliers not considered), and outliers are plotted individually using the '+' symbol.

Both NMSE and BIC values showed similar patterns across the three cross validation methods employed. This indicates that, event if temporal dependencies are not dealt with (regular CV), comparison between models is not impaired. NMSE values showed a significant higher variance across folds than the BIC values. Additionally, both NMSE and BIC values estimated by 5-fold blocked CV showed much higher variance than those estimated by 15-fold dependent and regular CV. This was somewhat expected, since not only did this procedure held a lower number of folds, but also because, since partitions were assigned in chronological order, specific processes that may have occurred were certainly not captured simultaneously in all CV partitions. Although the BIC values estimated were of the same order of magnitude for all procedures explored, the NMSE was slightly higher for the non-dependent CV than for the regular CV, and considerably higher for the blocked CV than for the remaining two. The fact that the modified procedures yielded higher NMSE values may indeed indicate an underestimation of the error with the regular CV, commonly verified with datasets with a high number of inter-sample dependencies such as this one (Bergmeir and Benítez, 2012). Additionally, the fact that the blocked procedure reported lower predictive performance than the other two may partially arise from the fact that, since the dependent samples were only removed up to a lag of $h = 2$, this was the only procedure that really ensured the removal of all stationary temporal dependencies.

Regarding the performance of the main models assessed, the pairwise comparisons indicate that all models held significantly different (p-value < 0.05) NMSE and BIC values, which means that the BIC of the X_{RMSF} model was deemed as significantly better than that of the X_{TP} model, and so on. These

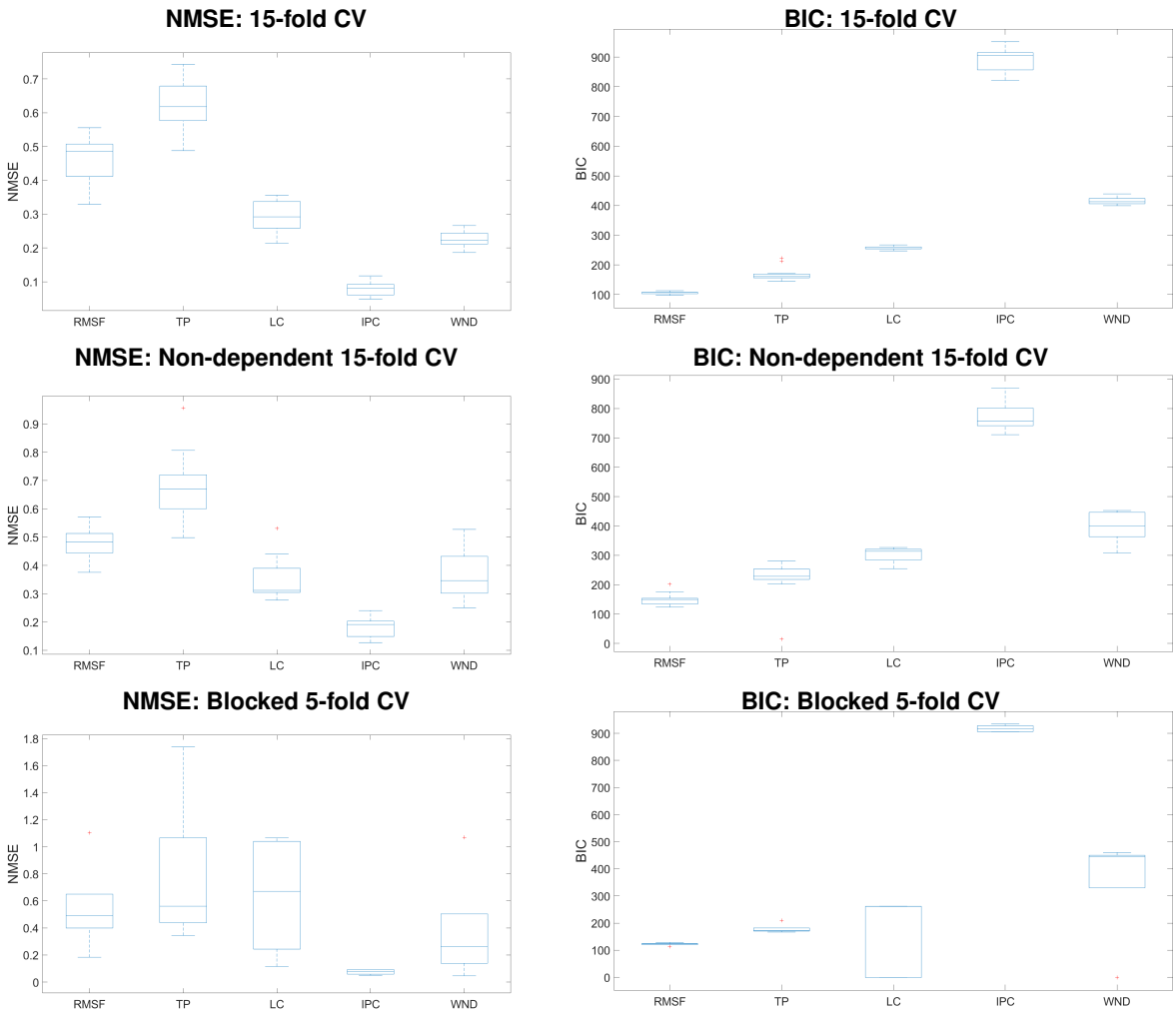


Figure 4.8: Boxplots of the NMSE and BIC values across folds, for each of the main models: X_{RMSF}, X_{TP}, X_{LC}, X_{IPC} and X_{WND}, obtained through 15-fold CV. Results respective to the resting state data.

results allow to conclude that, although the NMSE of the X_{IPC} model was significantly better than those of the remaining models, the increase in model complexity that it enclosed did not compensate at all for the higher prediction accuracy. Conversely, although yielding the second higher NMSE, the X_{RMSF} model was deemed as the one to enclose the best likelihood-complexity trade-off among all the main models assessed.

For the comparison of models with a reduced feature space, only the results for the regular CV procedure are presented herein, since the patterns of BIC and NMSE values across models remained similar for the other two procedures. The boxplots for the non-dependent and blocked CV are presented in Appendix A.1.

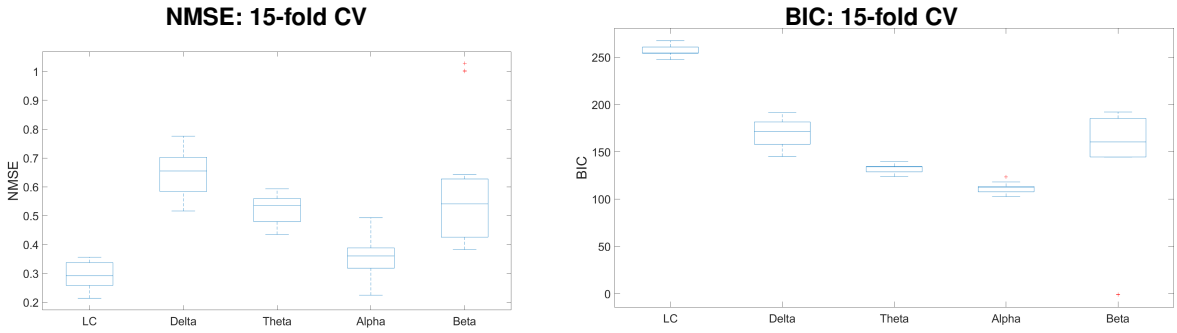


Figure 4.9: Boxplots of the NMSE and BIC values across folds, for the models X_{LC}, X_{δ}, X_{θ}, X_{α} and X_{β}, obtained through 15-fold CV. Results respective to the resting state data.

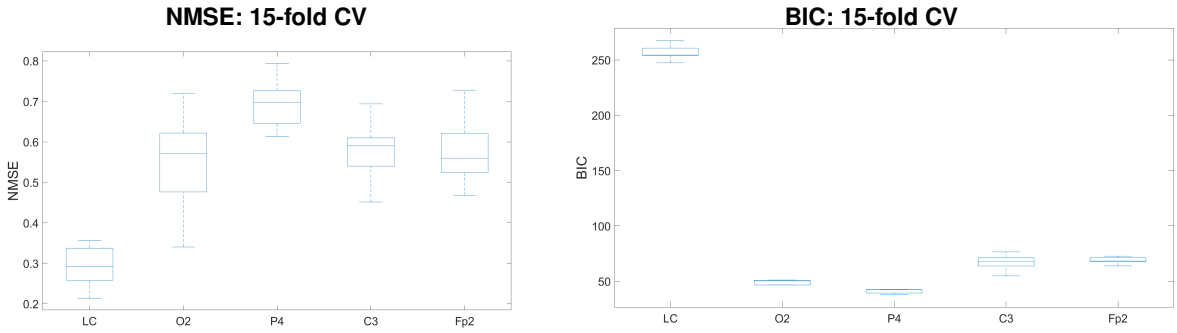


Figure 4.10: Boxplots of the NMSE and BIC values across folds, for the models X_{LC}, X_{O2}, X_{P4}, X_{C3} and X_{Fp2}, obtained through 15-fold CV. Results respective to the resting state data.

Regarding the pairwise comparison of one-band models (figure 4.9), the NMSE of X_{α} was found to be significantly ($p < 0.05$) better than those of the remaining models, and the BIC of X_{δ} to be significantly worse than those of the remaining. Notably, the X_{LC} model did not yield a significantly better NMSE than the X_{α}, but did yield a significantly worse BIC. These results suggest that, for this dataset, models that only include features within the alpha band compensate not only in terms of cost-effectiveness but also in terms of likelihood-complexity trade-off. Regarding one-channel models (figure 4.10), no significant differences were found in terms of BIC, while almost all NMSE values were found to be significantly different. More, although the X_{LC} performed significantly better than the X_{O2} in terms of NMSE, it performed significantly worse in terms of BIC. Finally, the results from the comparison of the

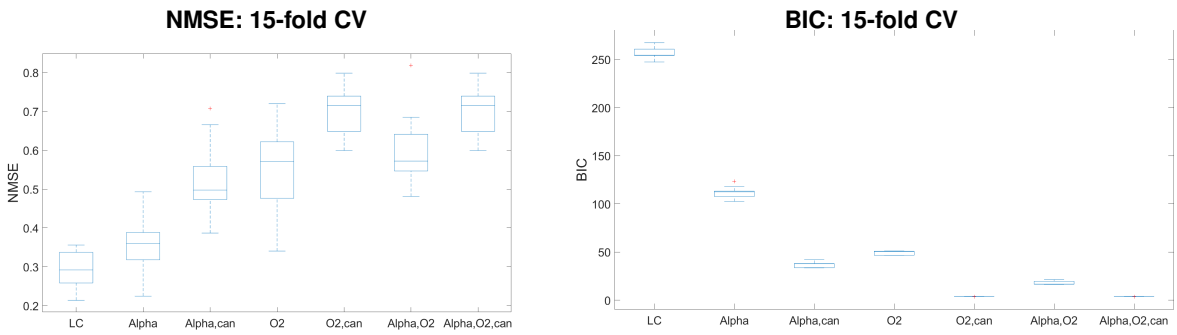


Figure 4.11: Boxplots of the NMSE and BIC values across folds, for the models X_{LC}, X_{α}, $X_{\alpha,\,can}$, X_{O2} and $X_{O2,can}$, $X_{\alpha,O2}$ and $X_{\alpha,O2,can}$, obtained through 15-fold CV. Results respective to the resting state data.

models in figure 4.11 indicate that all models exhibited significant differences regarding both BIC and NMSE (except for the X_{LC}- X_{α} pair, as mentioned above). Importantly, none of the models yielded unacceptably high NMSE, not even the model fitted with only one feature ($X_{\alpha,O2,can}$). This suggests that modelling resting state BOLD fluctuations within the DMN may be done successfully with a small number of features, and that both α activity and the activity from channel O2 may comprise promising features within this scope.

4.2 Motor Imagery

4.2.1 Results of Model Fitting

In this section, the average results of the elastic net fitting, obtained for the motor imagery data, are reported. The section is organized similarly to that of the resting state data: first the results concerning the five main models assessed are presented, and then the results derived in the scope of feature space reduction are shown.

Models with All Features

The tables bellow summarize the average results obtained with each of the cross validation procedures implemented: tables 4.7, 4.8 and 4.9 show the results for 15-fold CV, non-dependent 15-fold CV and blocked 5-fold CV, respectively. The measures reported are the average λ parameter and NMSE and BIC values, as well as the effective DOF of the final models. From the analysis of these results it is possible to conclude that the overall regularization across all models was significantly high, as reflected on the size of the λ parameter and on the effective DOF of the estimated models.

Table 4.7: Cross validated (15-fold CV) λ parameter, NMSE and BIC values of the elastic net fit of the main models explored: X_{RMSF}, X_{TP}, X_{LC}, X_{IPC} and X_{WND}. Effective degrees of freedom (DOF) of the final model, which was derived by averaging the nonzero values of the fitted coefficient estimates across folds. Results respective to the MI data.

	Average λ $(\times 10^{-1})$	DOF	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2}$
RMSF	1.09	39.0	2.25	1.72
TP	1.04	40.0	2.00	1.84
LC	1.11	49.0	0.843	2.55
IPC	1.76	62.0	0.564	4.20
WND	1.30	57.0	0.925	3.03

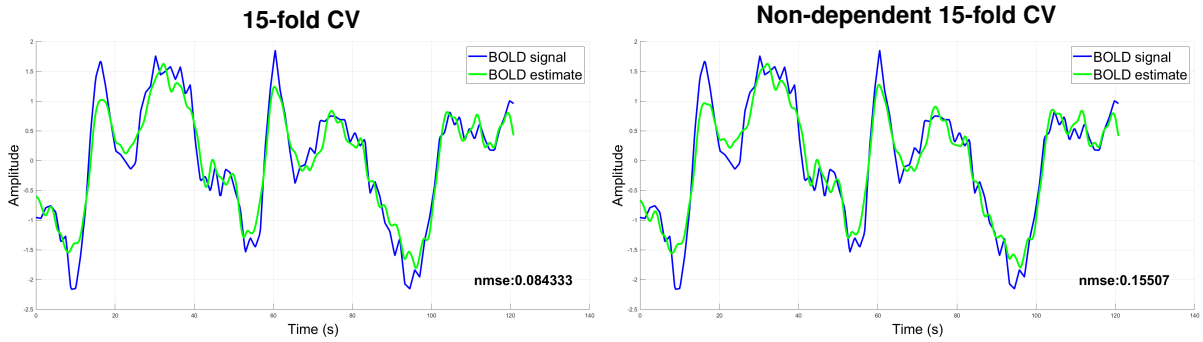
Table 4.8: Cross validated (non-dependent 15-fold CV) λ parameter, NMSE and BIC values of the elastic net fit of the main models explored: X_{RMSF}, X_{TP}, X_{LC}, X_{IPC} and X_{WND}. Effective degrees of freedom (DOF) of the final model, derived by averaging the nonzero values of the fitted coefficient estimates across all folds. Results respective to the MI data.

	Average λ $(\times 10^{-1})$	DOF	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
RMSF	1.03	54.0	2.65	1.71
TP	1.44	48.0	3.24	1.58
LC	1.65	64.0	1.55	2.15
IPC	2.06	92.0	0.863	4.02
WND	1.34	69.0	1.33	2.87

Table 4.9: Cross validated (blocked 5-fold CV) λ parameter, NMSE and BIC values of the elastic net fit of the main models explored: X_{RMSF}, X_{TP}, X_{LC}, X_{IPC} and X_{WND}. Effective degrees of freedom (DOF) of the final model, derived by averaging the nonzero values of the fitted coefficient estimates across all folds. Results respective to the MI data.

	Average λ $(\times 10^{-1})$	DOF	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
RMSF	0.719	46.0	1.66	1.94
TP	0.966	40.0	1.95	1.99
LC	0.884	45.0	0.713	2.69
IPC	1.23	54.0	0.420	4.54
WND	1.28	54.0	0.975	3.09

Figure 4.12 displays an example of the prediction of the BOLD response of interest. The particular example shown concerns the prediction performed by fitting the X_{LC} model, using regular 15-fold CV (figure (a)) and non-dependent 15-fold CV (figure (b)). The BOLD estimate represented was computed from the final EFP derived in each of the procedures.



(a) (b)

Figure 4.12: Example of prediction of the BOLD response. BOLD signal (Y, in blue) and respective BOLD estimate ($\hat{Y}$, in green), obtained with the EFP estimated for the X_{LC} model. (a) EFP estimated by 15-fold cross validation.(b) EFP estimated by non-dependent 15-fold cross Validation. Average NMSE of the fitting displayed on the lower right corner. Results respective to the motor imagery data.

Models with Reduced Feature Space

The following tables report the results obtained in the scope of feature space reduction of the X_{LC} model. Here too, only the results of the regular 15-fold CV procedure are shown, whilst the results obtained with the other two procedures are presented in appendix A.2.

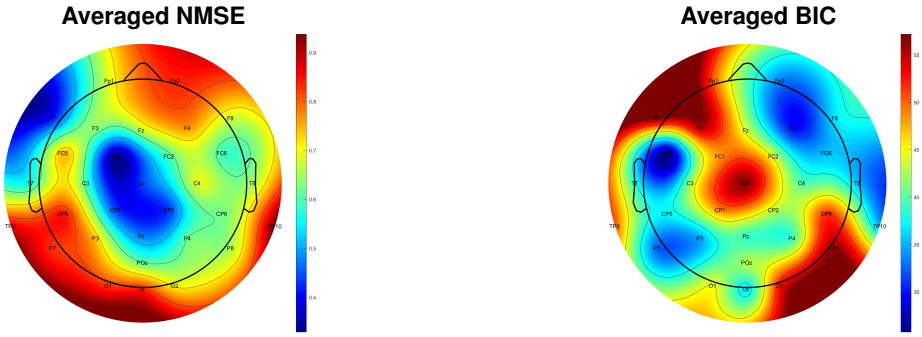
Table 4.10: Cross validated (15-fold CV) λ parameter, NMSE and BIC values of the elastic net fit of the models X_{LC}, X_{δ}, X_{θ}, X_{α} and X_{β}. Effective degrees of freedom (DOF) of the final model, derived by averaging the nonzero values of the fitted coefficient estimates across all folds. Results respective to the MI data.

	Average λ $(\times 10^{-1})$	DOF	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
LC	1.11	49.0	0.843	2.55
Delta	2.07	44.0	3.26	1.53
Theta	0.972	39.0	4.00	1.49
Alpha	0.851	39.0	1.33	1.57
Beta	0.723	37.0	1.29	1.36

Table 4.10 summarizes the results of the one-band models. Both X_{α} and X_{β} models yielded significantly better NMSE than the remaining models, although no significant differences for the BIC values were found (this is further detailed in section 4.2.3). However, between the X_{α} and the X_{β} model, the NMSE was not deemed as significantly different. Hence, despite the X_{β} showing a slightly better NMSE,

the X_α was selected to be used the subsequent analyses.

Regarding the analyses with one-channel models, figure 4.13 displays the topography of the predictive performance, in terms of NMSE and BIC, estimated for the model fitted at each channel with 15-fold CV.



(a) (b)

Figure 4.13: Topography of the predictive performance of one-channel models, as measured by the NMSE and BIC, estimated through 15-fold CV. (a) NMSE, averaged across all folds, of the model fitted at each channel. (b) BIC, averaged across all folds, of the model fitted at each channel. Results respective to the motor imagery data.

A rough analysis of the topographies shown reveals that the best NMSE performances were achieved by the models derived with central EEG channels, many of them corresponding to regions within the motor cortex. However, the best BIC performances were achieved with a more distributed range of channels. A more detailed analysis was performed for the four one-channel models that yielded better NMSE performance. The elastic net fitting parameters of such models are reported in table 4.11.

Table 4.11: Cross validated (15-fold CV) λ parameter, NMSE and BIC values of the elastic net fit of the models X_{LC}, X_{FC1}, X_{F7}, X_{CP1} and X_{CP2}. Effective degrees of freedom (DOF) of the final model, derived by averaging the nonzero values of the fitted coefficient estimates across all folds. Results respective to the MI data.

	Average λ $(\times 10^{-1})$	DOF	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
LC	1.11	49.0	0.843	2.55
FC1	0.378	20.0	3.29	0.486
F7	0.225	23.0	4.18	0.572
CP1	0.919	21.0	4.35	0.489
CP2	0.698	22.0	4.36	0.433

Regarding model NMSE, model X_{FC1} yielded better predictive performance in either of the CV procedures, and thus it was the one selected to perform further analyses. This result is consistent with the overall findings in literature, which describe that motor channels within the contralateral hemisphere to the MI task typically report high engagement in such tasks (Hamedi et al., 2016). Indeed, channel FC1 is

located within the left areas of the premotor cortex. Hence, features with respect to the alpha frequency band and channel FC1 combined were used to build the model $X_{\alpha,FC1}$. More, the feature space of models X_{α}, X_{FC1} and $X_{\alpha,FC1}$ was further reduced by solely integrating features that were convolved with the canonical HRF (overshoot delay at 6 seconds). The results referent to this analysis are summarized in table 4.12.

Table 4.12: Cross validated (15-fold CV) λ parameter, NMSE and BIC values of the elastic net fit of the models X_{LC}, X_{α}, $X_{\alpha,\,can}$, X_{FC1} and $X_{FC1,can}$, $X_{\alpha,FC1}$ and $X_{\alpha,FC1,can}$. Effective degrees of freedom (DOF) of the final model, derived by averaging the nonzero values of the fitted coefficient estimates across all folds. Results respective to the MI data.

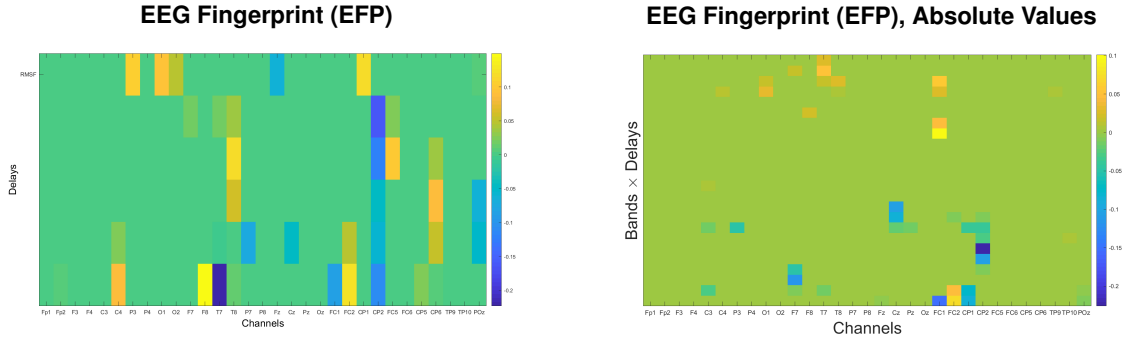
	Average λ $(\times 10^{-2})$	DOF	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{1})$
LC	11.1	49.0	0.843	25.48
Alpha	8.51	39.0	1.33	15.71
Alpha,can	5.35	25.0	4.04	4.93
FC1	3.78	20.0	3.29	4.86
FC1,can	0.77	2.0	10.46	0.346
Alpha,FC1	4.36	7.0	10.15	1.32
Alpha,FC1,can	0.77	2.0	10.46	0.346

4.2.2 EEG Fingerprints

This section presents the most relevant topographic representations of the EFPs estimated for each model. Here too, only maps referent to the EFPs estimated by regular 15-fold CV are displayed.

Models with All Features

In this section the EFPs obtained for the five main models assessed are shown. Two examples of the matrix representation of the EFP are displayed in figure 4.14. These figures reveal that the models estimated were significantly sparse, which further validates the decision of implementing regularization strategies to fit the models. Figure 4.15 displays a topographic representation of the EFPs, obtained by averaging the coefficient estimates across all frequency bands and HRF delays.



(a) (b)

Figure 4.14: Example of one EEG Fingerprint (EFP) in matrix representation. The sizes of the coefficient estimates are color-coded. (a) EFP estimated for the X_{RMSF} model: rows represent each of the frequency bands δ, θ, α, β (top to bottom); columns represent the EEG channels. (b) EFP estimated for the X_{LC} model: rows represent the HRF delays, 10, 8, 6, 5, 4 and 2 seconds (top two bottom), for each of the frequency bands; columns represent the EEG channels. Results respective to the motor imagery data.

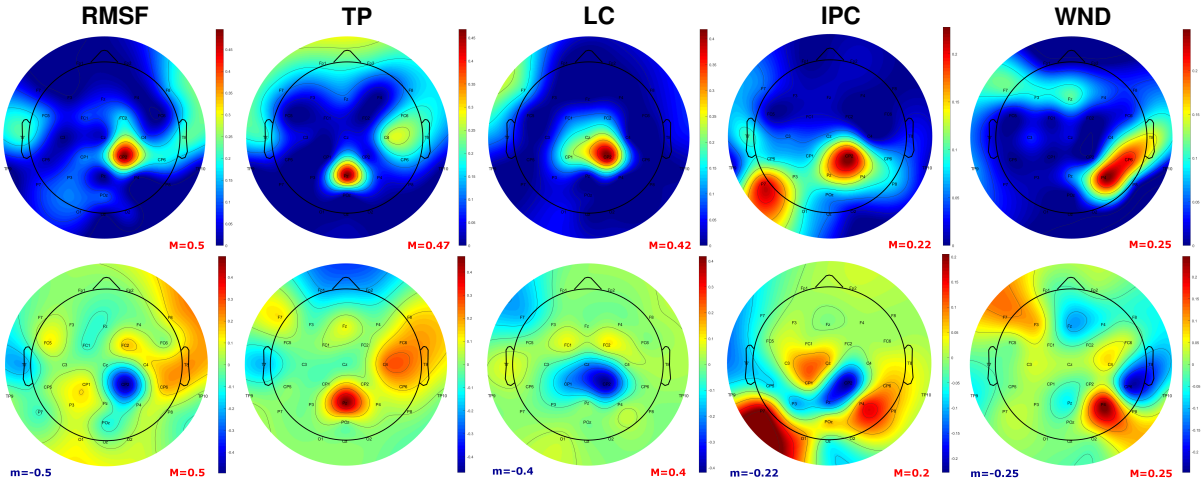


Figure 4.15: Topographic representation of the coefficient estimates (EFP), averaged across all folds, obtained for each of the main models: X_{RMSF}, X_{TP}, X_{LC}, X_{IPC} and X_{WND}. Bottom line: coefficient estimates averaged across all HRF delays and (in the case of the X_{LC}, X_{IPC} and X_{WND} models) across all frequency bands. Top line: heat map representing the absolute values of those coefficient estimates. Minimum and maximum values are indicated for each map with m (blue) and M (red). Results respective to the motor imagery data.

In all topographies displayed, only a limited set of electrodes is shown to yield high coefficient sizes, which is in line with the theoretical expectations that only a few brain regions are highly engaged in motor tasks such as this one. Notably, all maps show engagement of regions within the motor cortex, although not exclusively. Specifically, in the EFP of the X_{RMSF}, X_{LC} and X_{IPC} models, the highest weights appear to belong to channels Cz and CP2, both within the secondary sensorimotor cortex. Notably, the CP2 is located on the ipsilateral hemisphere to the MI task, instead of being on the contralateral hemisphere as most commonly observed (Hamedi et al., 2016).

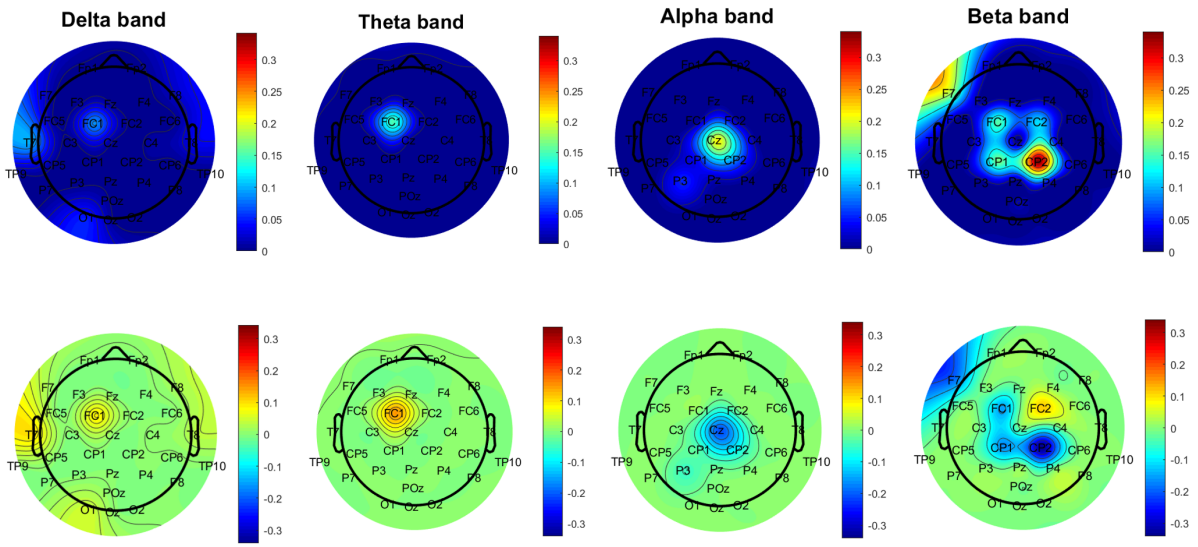


Figure 4.16: Topographic representation of the coefficient estimates (EFP), averaged across all folds, obtained for the X_{LC} model. Bottom line: coefficient estimates corresponding to each of the frequency bands, averaged across all HRF delays. Top line: heat map representing the absolute values of those coefficient estimates. Results respective to the motor imagery data.

To obtain more insights on the frequency profile of the EFPs derived for the X_{LC} model, these results were also organized in another topographic representation (figure 4.16).

The representations in figure 4.16 were all plotted following an unique color-code. In the upper figure, it is possible to verify that the higher coefficient sizes were estimated for features corresponding to the alpha and beta frequency bands. Most importantly, the negative values of such coefficients suggest an ERD of these frequency bands with the motor task, which is consistent with the results most frequently reported in literature (Hamedi et al., 2016).

Models with Reduced Feature Space

Further relevant conclusions may be drawn by analysing the topographies of the EFP obtained for the X_{α} model. The results obtained for this model should provide more detailed insights on the temporal structure of the alpha band features, since all other features are removed, and thus prevented from confounding the interpretation of the alpha coefficients. Figure 4.17 shows that the negative relationship between the motor channel Cz and output response of the system was most prominent in the features that were convolved with the HRFs of delays 8, 6 and 5 seconds. A remarkable observation that may be drawn from this figure is that channel C3, located on the primary motor cortex of the contralateral hemisphere, yielded a positive coefficient at 10 second delay, and a negative coefficient at 4 second delay, both with high absolute values. This may explain why the coefficients of this channel did not appear to be significant in the topographic EFP representations of figures 4.15 and 4.16, in which coefficients were averaged across delays.

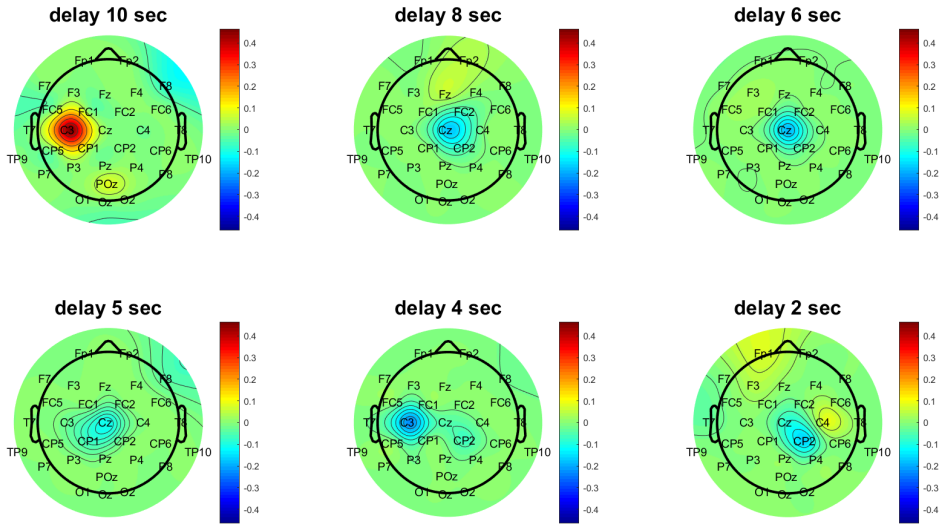


Figure 4.17: Topographic representation of the coefficient estimates (EFP), averaged across all folds, obtained for the X_{α} model. Coefficient estimates relative to each HRF delay displayed for each EEG channel. Results respective to the motor imagery data.

4.2.3 Comparison of Methods and Models

The main models for each CV procedure were compared via ANOVA, in order to evaluate whether significant differences (p-value < 0.05) regarding model performance, in terms of NMSE and BIC, were obtained. The results of the pairwise comparison tests among all models and for all methods are presented in appendix B.2. In figures 4.18 to 4.21, results of model comparison are displayed in the form of boxplots.

Here too, both NMSE and BIC values showed similar patterns across the three cross validation methods employed, which suggests that the comparison between models is not impaired by dependencies within the data. NMSE values showed a significant higher variance across folds than the BIC values, similarly to what was observed for the RS data. However, in contrast to the results obtained for the RS data, NMSE and BIC values estimated by 5-fold blocked CV did not show significantly higher variance than those estimated by 15-fold dependent and regular CV. This is an interesting result, suggesting that the higher variance observed for the blocked procedure on the RS data may indeed arise from the occurrence of specific spurious processes that were not captured simultaneously in all partitions. On the contrary, for the MI data, dividing the data in 5 blocks was sufficient to capture one trial of rest and imagery at each block. Moreover, in this dataset too the BIC values were of the same order of magnitude for all the procedures explored, whereas the overall NMSE values were sightly higher for the two modified CV procedures, which may indicate an underestimation of the true NMSE with the regular CV procedure.

Regarding the performance of the main models assessed, the pairwise comparisons indicate that all models held significantly different ($p < 0.05$) NMSE and BIC values, which means that the BIC of the

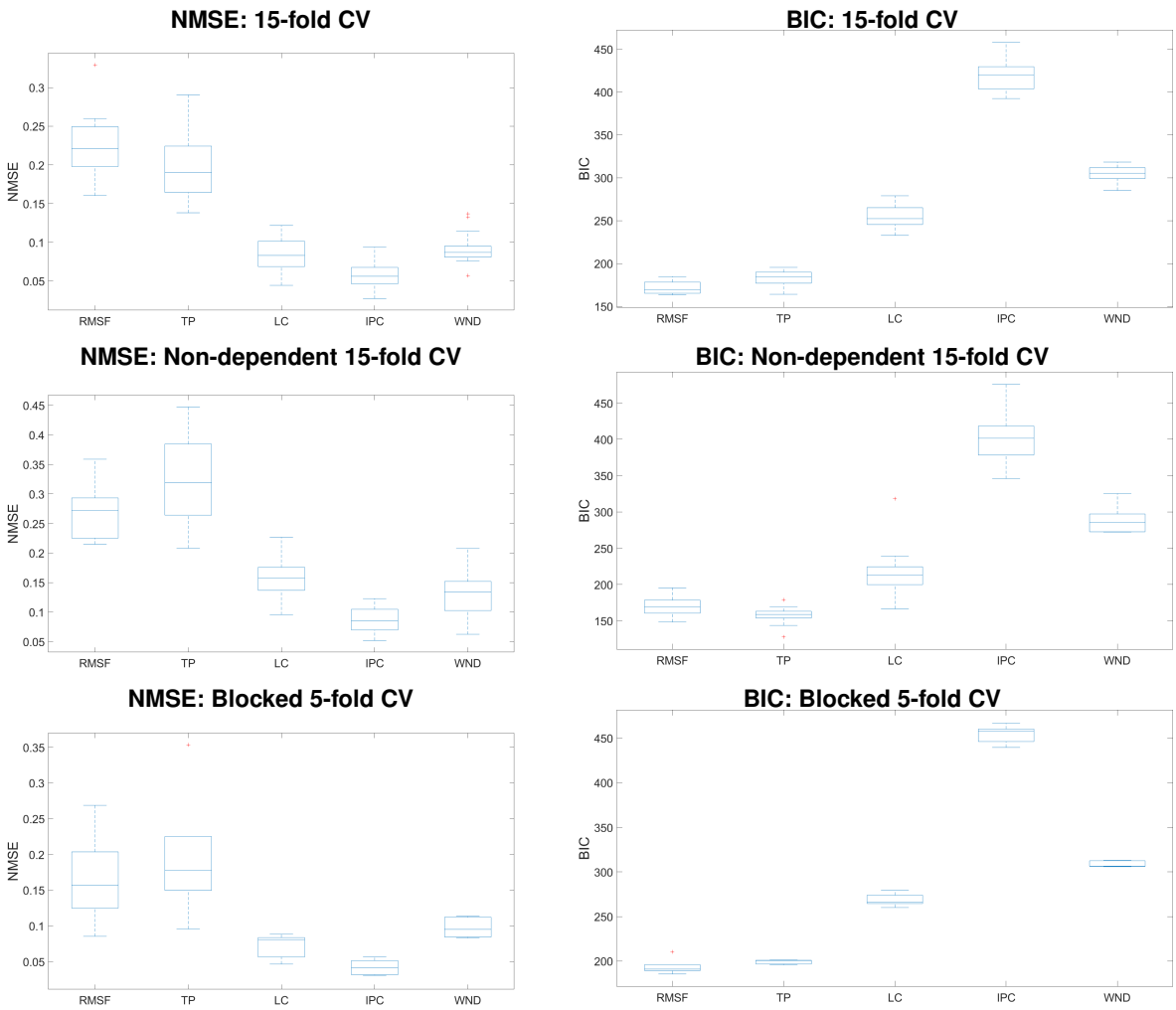


Figure 4.18: Boxplots of the NMSE and BIC values across folds, for each of the main models: X_{RMSF}, X_{TP}, X_{LC}, X_{IPC} and X_{WND}. Results respective to the motor imagery data.

X_{RMSF} model was significantly better than that of the X_{TP} model, and so on. Hence, although the NMSE of the X_{IPC} model was significantly better than those of the remaining models, the X_{RMSF} model was deemed as the one to enclose the best likelihood-complexity trade-off.

For the comparison of models with a reduced feature space, only the results for the regular CV procedure are presented herein, since the patterns of BIC and NMSE values across models remained similar for the other two procedures. The boxplots for the non-dependent and blocked CV are presented in Appendix A.2.

Regarding the pairwise comparison of one-band models (figure 4.19), the NMSE of both the X_{α} and the X_{β} model was found to be significantly ($p < 0.05$) better than those of the remaining models, however no significant differences were observed in terms of BIC. Moreover, although the X_{LC} performed significantly better than both X_{α} and X_{β} in terms of NMSE, it performed significantly worse in terms of BIC. Regarding one-channel models (figure 4.20), no significant differences were found in terms of BIC, however model X_{FC1} significantly outperformed the remaining in terms of NMSE. Finally, the results from the comparison of the models in figure 4.21 indicate that all models exhibited significant differences

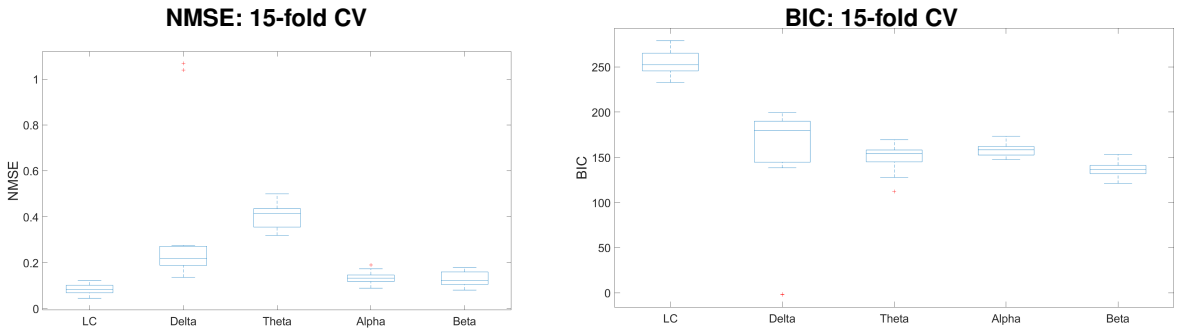


Figure 4.19: Boxplots of the NMSE and BIC values across folds, for the models X_{LC}, X_{δ}, X_{θ}, X_{α} and X_{β}. Results respective to the motor imagery data.

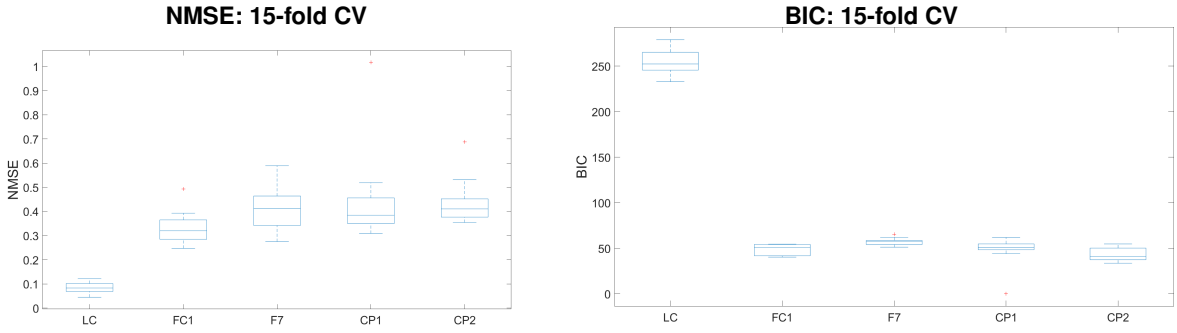


Figure 4.20: Boxplots of the NMSE and BIC values across folds, for the models X_{LC}, X_{FC1}, X_{F7}, X_{CP1} and X_{CP1}. Results respective to the motor imagery data.

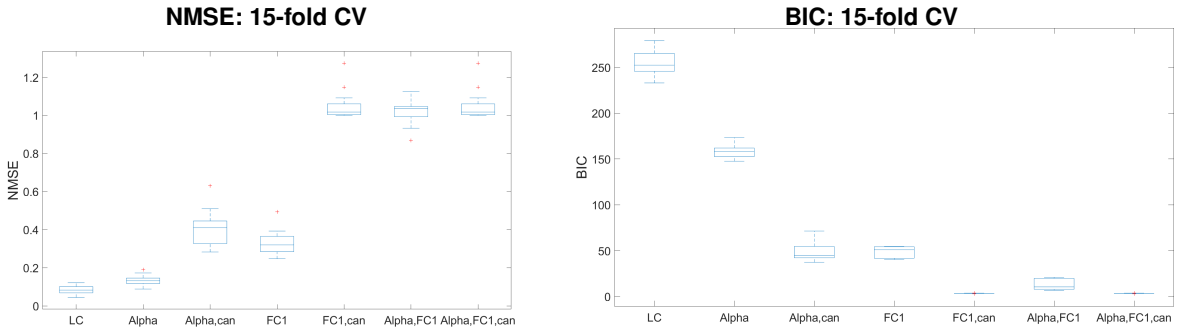


Figure 4.21: Boxplots of the NMSE and BIC values across folds, for the models X_{LC}, X_{α}, $X_{\alpha,\text{can}}$, X_{FC1} and $X_{FC1,\text{can}}$, $X_{\alpha,FC1}$ and $X_{\alpha,FC1,\text{can}}$. Results respective to the motor imagery data.

regarding BIC, except for the last three ($X_{FC1,can}$, $X_{\alpha,FC1}$ and $X_{\alpha,FC1,can}$). However, these three models all yielded unacceptably high NMSE (greater than 1), which indicates that these features alone could not have been used to derived good enough BOLD predictions. More, the best one-band model significantly outperformed the best one-channel model in terms of NMSE, but the one-channel model yielded a significantly better BIC. Altogether, these results suggest that modelling the BOLD fluctuations activated during MI tasks may be done successfully with a considerably small number of features, and that both α activity and the activity from channel FC1 may comprise promising features within this scope.

Chapter 5

Conclusions

In this chapter, the main conclusions drawn from the results obtained are discussed, and possible future developments to optimize the performance of the methods assessed are outlined.

5.1 Achievements

This work described an integration strategy whereby relevant EEG-features were extracted and their coefficient estimates learnt so as to build EEG fingerprints (EFPs) able to predict the simultaneous BOLD signal from a specific region of interest. Within this scope, analyses were performed with two EEG-fMRI datasets: one acquired during resting state (RS), and another during a motor imagery (MI) task. Notably, several of the conclusions drawn from the results obtained were shared among the two independent analyses. Specifically, for both cases, the regularization incurred through elastic net fitting was successful in producing models with relative low complexities and good enough predictive performances, in which only the most relevant EEG features were kept. This effect reflected on the effective degrees of freedom (DOF) estimated for the final models, which was consistently lower than the number of features originally considered. Moreover, the topographic representations of the EFPs suggest that the models obtained were rather sparse across channels, which is in line with the theoretical expectation that only a few brain regions are highly engaged in a specific cognitive process. These results validate not only the choice of elastic net regularization to fit the models, but also the adequacy of the cross validation procedures used for model selection, as well as the criterion adopted to choose amongst competing models, which was the BIC value.

Regarding the performance of the several models tested, as measured by the estimated BIC values, results were also consistent across the two datasets. Specifically, the X_{RMSF} model was deemed to yield the best predictive performance, followed by the X_{TP}, the X_{LC}, the X_{WND} and finally, the X_{IPC}. Regarding the analyses performed in the scope of reducing the feature space of the X_{LC} model, results from both datasets suggest that the BOLD fluctuations of interest may be successfully modelled through a considerably small number of features, particularly features arising from either one single frequency band or one single EEG channel.

Results across the two analysed datasets were also consistent in terms of the evaluation patterns obtained through each of the three cross validation procedures implemented: 15-fold cross validation, 15-fold non-dependent cross validation, and 5-fold blocked cross validation. Specifically, the two modified CV procedures, which attempted to remove temporal dependencies from the time-series data in order to obtain reliable predictions of model performance, estimated overall higher prediction errors than the traditional CV procedure. This may indicate an underestimation of the true NMSE with the regular CV procedure, further validating the need to employ modified CV procedures for model evaluation. Notably, in all the CV methods explored, both the NMSE and the BIC showed to vary similarly amongst the different models. This suggests that the comparison between the models here assessed was not impaired by dependencies within the data.

Finally, for both datasets analysed, the topographic representation of the overall EFPs estimated was in line with previous findings reported for each of the experimental conditions. Specifically, for the resting state data, the topographic representation of the EFPs estimated for the spectral feature models (X_{RMSF}, X_{TP} and X_{LC}) showed negative coefficients of high absolute value for channels within occipital regions, and in the case of the X_{LC} model these were associated with the alpha frequency band. This suggests a negative correlation between the alpha activity in occipital channels and the BOLD fluctuations within the DMN, a result that has been previously reported. On the other hand, for the motor imagery data, all maps obtained show engagement of regions within the motor cortex (although not exclusively). Specifically, in the EFPs estimated for the X_{RMSF}, X_{LC}, X_{IPC} models, the highest weights were associated with channels within the secondary sensorimotor cortex. More, for the X_{LC} model, higher coefficient sizes were estimated for features associated with the alpha and beta frequency bands. More, these held negative values, suggesting an ERD of these rhythms with the motor task, which is consistent with the results most frequently reported in the literature. Moreover, by analysing the temporal structure of the coefficient estimates related with the alpha frequency band, it was revealed that the channel C3, located on the primary motor cortex of the contralateral hemisphere to the MI task, was also associated with high coefficient sizes. However, these weights switched between positive and negative values depending on the HRF delay considered. This poses the question of weather it is redundant to consider such a wide range of HRF delays for the purposes of the methodology here studied.

5.2 Limitations and Future Work

The methodology employed entailed several limitations. An important consideration has to do with the convolution of the EEG-derived features with the family of HRF waveforms, which certainly increased the presumably already high correlation amongst the cluster of features originally fed to the models. This is a concern especially for problems with high number of features and small sample size, which was the case in this work. The redundancy that arises from this phenomenon significantly hinders the interpretation of the models estimated, and thus the identification of the features with the strongest fitting effects. This compromises the search for a clearer view on the relationships being modelled, and thus

the future development of more cost-effective solutions. Notably, the elastic net regression algorithm exhibits a grouping effect, whereby similar weights are assigned to highly correlated features. While this effect is often desirable so as to not discard relevant features, it may also be suboptimal when the correlation amongst features is introduced by the processing of the EEG data. In the context of these suppositions, a different pipeline could and should be assessed, in which EEG-derived features were not to be convolved with a family of HRFs, but instead delayed to match the time-frame of the simultaneous BOLD signal.

To improve the predictive performance of the models assessed, increasing the training sample size by using more than one session (of resting state or of motor imagery) could be a possible solution. More, data from separate sessions would not share temporal dependencies, thus being more suitable for evaluating model performance. In another scope, by recording multiple sessions, the motor imagery cross-spectrum could have been averaged across several trials, thus increasing the quality of the functional connectivity estimates. However, this procedure could not be done for the resting state data, which does not yield trial stationarity. To improve the quality of the resting states connectivity estimates, alternative methods for deriving the expected value of the cross-spectrum could have been tested. A possible approach could have been to wavelet-based methods, which are particularly suited to quantifying time varying coherency (Lachaux et al., 2002).

Finally, as a future prospect, the methodology employed could be extended to an EEG-fMRI-NF training framework, in which the NF scores could be directly used to learn the optimal EFPs. This as been recently done on a study that assessed a similar approach to this one (Cury et al., 2019) and the results were rather promising. More, another future hope would be to apply the pipeline here described on data recorded from multiple subjects, in order to search for a common model, able to account for the differences between subjects.

Bibliography

(1961). The ten twenty electrode system: International federation of societies for electroencephalography and clinical neurophysiology. *American Journal of EEG Technology*, 1(1):13–19. doi:10.1080/00029238.1961.11080571.

Abreu, R., Leal, A., and Figueiredo, P. (2018). Eeg-informed fmri: A review of data analysis methods. *Frontiers in Human Neuroscience*, 12:29. doi:10.3389/fnhum.2018.00029.

Abreu, R., Nunes, S., Leal, A., and Figueiredo, P. (2017). Physiological noise correction using ecg-derived respiratory signals for enhanced mapping of spontaneous neuronal activity with simultaneous eeg-fmri. *NeuroImage*, 154:115–127. doi:10.1016/j.neuroimage.2016.08.008.

Adey, W., Walter, D., and Hendrix, C. (1961). Computer techniques in correlation and spectral analyses of cerebral slow waves during discriminative behavior. *Experimental Neurology*, 3(6):501–524. doi:10.1016/s0014-4886(61)80002-2.

Adler, R. J. (2004). Gaussian random fields on manifolds. In Dalang, R. C., Dozzi, M., and Russo, F., editors, *Seminar on Stochastic Analysis, Random Fields and Applications IV*, pages 3–19, Basel. Birkhäuser Basel.

Aho, K., Derryberry, D., and Peterson, T. (2014). Model selection for ecologists: the worldviews of aic and bic. *Ecology*, 95(3):631–636. doi:10.1890/13-1452.1.

Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60(2):255. doi: 10.2307/2334537.

Amzica, F. and da Silva, F. H. L. (2017). Cellular substrates of brain rhythms. In Schomer, D. L. and da Silva, F. H. L., editors, *Niedermeyer's ElectroencephalographyBasic Principles, Clinical Applications, and Related Fields*, Oxford, UK. Oxford University Press.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4(0):40–79. doi:10.1214/09-ss054.

Ashburner, J., Barnes, G., Chen, C.-C., Daunizeau, J., Flandin, G., Friston, K., Kiebel, S., Kilner, J., Litvak, V., Moran, R., Penny, W., Razi, A., Stephan, K., Tak, S., and Zeidman, P. (2019). *SPM12 Manual*. Institute of Neurology, ULC. http://www.fil.ion.ucl.ac.uk/spm/.

Babajani, A. and Soltanian-Zadeh, H. (2006). Integrated meg/eeg and fmri model based on neural masses. *IEEE Transactions on Biomedical Engineering*, 53(9):1794–1801. doi:10.1109/tbme.2006.873748.

Bastos, A. M. and Schoffelen, J.-M. (2016). A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Frontiers in Systems Neuroscience*, 9. doi:10.3389/fnsys.2015.00175.

Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213. doi:10.1016/j.ins.2011.12.028.

Biswal, B., Yetkin, F. Z., Haughton, V. M., and Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic Resonance in Medicine*, 34(4):537–541.

Bowman, A. D., Griffis, J. C., Visscher, K. M., Dobbins, A. C., Gawne, T. J., Difrancesco, M. W., and Szaflarski, J. P. (2017). Relationship between alpha rhythm and the default mode network. *Journal of Clinical Neurophysiology*, 34(6):527–533. doi:10.1097/wnp.0000000000000411.

Brant-Zawadzki, M., Gillan, G. D., and Nitz, W. R. (1992). Mp rage: a three-dimensional, t1-weighted, gradient-echo sequence–initial experience in the brain. *Radiology*, 182(3):769–775. doi:10.1148/radiology.182.3.1535892.

Brazier, M. A. and Casby, J. U. (1952). Crosscorrelation and autocorrelation studies of electroencephalographic potentials. *Electroencephalography and Clinical Neurophysiology*, 4(2):201–211. doi:10.1016/0013-4694(52)90010-2.

Brooks, R. A., Battocletti, J. H., Sances, A., Larson, S. J., Bowman, R. L., and Kudravcev, V. (1975). Nuclear magnetic relaxation in blood. *IEEE Transactions on Biomedical Engineering*, BME-22(1):12–18. doi:10.1109/tbme.1975.324533.

Buxton, R. B. (2009). *Introduction to functional magnetic resonance imaging: principles and techniques*. Cambridge University Press.

Buxton, R. B. and Frank, L. R. (1997). A model for the coupling between cerebral blood flow and oxygen metabolism during neural stimulation. *Journal of Cerebral Blood Flow Metabolism*, 17(1):64–72. doi:10.1097/00004647-199701000-00009.

Buzsáki, G. (2009). *Rhythms of The Brain*. doi:10.1093/acprof:oso/9780195301069.001.0001.

Carter, G. (1987). Coherence and time delay estimation. *Proceedings of the IEEE*, 75(2):236–255. doi:10.1109/proc.1987.13723.

Cury, C., Maurel, P., Gribonval, R., and Barillot, C. (2019). A sparse eeg-informed fmri model for hybrid eeg-fmri neurofeedback prediction. doi:10.1101/599589.

de Nó, R. L. (1995). *Action potential of the motoneurons of the hypoglossus nucleus*.

Delapaz, R. L. (1994). Echo-planar imaging. *RadioGraphics*, 14(5):1045–1058. doi:10.1148/radiographics.14.5.7991813.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., and et al. (2006). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *NeuroImage*, 31(3):968–980. doi:10.1016/j.neuroimage.2006.01.021.

Devroye, L. and Wagner, T. (1979). Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604. doi:10.1109/tit.1979.1056087.

Edes, A. E., Kozak, L. R., Magyar, M., Zsombok, T., Kokonyei, G., Bagdy, G., and Juhasz, G. (2017). Spontaneous migraine attack causes alterations in default mode network connectivity: a resting-state fmri case report. *BMC Research notes*, 10(1):165. doi:10.1186/s13104-017-2484-1.

Eickhoff, S. B., Paus, T., Caspers, S., Grosbras, M.-H., Evans, A. C., Zilles, K., and Amunts, K. (2007). Assignment of functional activations to probabilistic cytoarchitectonic areas revisited. *NeuroImage*, 36(3):511–521. doi:10.1016/j.neuroimage.2007.03.060.

Filho, C. A. S., Attux, R., and Castellano, G. (2017). Eeg sensorimotor rhythms' variation and functional connectivity measures during motor imagery: linear relations and classification approaches. *PeerJ*, 5. doi:10.7717/peerj.3983.

Fox, P. T. and Raichle, M. E. (1986). Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects. *Proceedings of the National Academy of Sciences*, 83(4):1140–1144. doi:10.1073/pnas.83.4.1140.

Fries, P. (2001). Modulation of oscillatory neuronal synchronization by selective visual attention. *Science*, 291(5508):1560–1563. doi:10.1126/science.1055465.

Friston, K. J., Frith, C. D., Liddle, P. F., and Frackowiak, R. S. J. (1991). Comparing functional (pet) images: The assessment of significant change. *Journal of Cerebral Blood Flow & Metabolism*, 11(4):690–699. doi:10.1038/jcbfm.1991.122.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4):189–210. doi:10.1002/hbm.460020402.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320. doi:10.2307/2285815.

Gevins, A. (1997). High-resolution eeg mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. *Cerebral Cortex*, 7(4):374–385. doi:10.1093/cercor/7.4.374.

Goense, J. B. and Logothetis, N. K. (2008). Neurophysiology of the bold fmri signal in awake monkeys. *Current Biology*, 18(9):631–640. doi:10.1016/j.cub.2008.03.054.

Golub, G. H. and Reinsch, C. (1971). Singular value decomposition and least squares solutions. *Linear Algebra*, page 134–151. doi:10.1007/978-3-662-39778-7_10.

Greicius, M. D., Krasnow, B., Reiss, A. L., and Menon, V. (2002). Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*, 100(1):253–258. doi:10.1073/pnas.0135058100.

Guillot, A. and Collet, C. (2008). Construction of the motor imagery integrative model in sport: a review and theoretical investigation of motor imagery use. *International Review of Sport and Exercise Psychology*, 1(1):31–44. doi:10.1080/17509840701823139.

Hamedi, M., Salleh, S.-H., and Noor, A. M. (2016). Electroencephalographic motor imagery brain connectivity analysis for bci: A review. *Neural Computation*, 28(6):999–1041. doi:10.1162/neco_a_00838.

Hardwick, R. M., Caspers, S., Eickhoff, S. B., and Swinnen, S. P. (2017). Neural correlates of motor imagery, action observation, and movement execution: A comparison across quantitative meta-analyses. doi:10.1101/198432.

Harmony, T. (2013). The functional significance of delta oscillations in cognitive processing. *Frontiers in Integrative Neuroscience*, 7. doi:10.3389/fnint.2013.00083.

Hastie, T., Friedman, J., and Tisbshirani, R. (2017). *The Elements of statistical learning: data mining, inference, and prediction*. Springer, 2^{nd} edition.

Herrmann, C. S., Strüber, D., Helfrich, R. F., and Engel, A. K. (2016). Eeg oscillations: From correlation to causality. *International Journal of Psychophysiology*, 103:12–21. doi:10.1016/j.ijpsycho.2015.02.003.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67. doi:10.1080/00401706.1970.10488634.

Hotelling, H. (1933). *Analysis of a complex of statistical variables into principal components*. Warwick & York.

Jann, K., Dierks, T., Boesch, C., Kottlow, M., Strik, W., and Koenig, T. (2009). Bold correlates of eeg alpha phase-locking and the fmri default mode network. *NeuroImage*, 45(3):903–916. doi:10.1016/j.neuroimage.2009.01.001.

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences*, 17(2):187–202. doi:10.1017/s0140525x00034026.

John, E. (2002). The neurophysics of consciousness. *Brain Research Reviews*, 39(1):1–28. doi:10.1016/s0165-0173(02)00142-x.

Jolliffe, I. and Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374:20150202. doi:10.1098/rsta.2015.0202.

Jorge, J., van der Zwaag, W., and Figueiredo, P. (2014). Eeg–fmri integration for the study of human brain function. *NeuroImage*, 102(1):24–34. doi:10.1016/j.neuroimage.2013.05.114.

Kilner, J., Mattout, J., Henson, R., and Friston, K. (2005). Hemodynamic correlates of eeg: A heuristic. *NeuroImage*, 28(1):280–286. doi:10.1016/j.neuroimage.2005.06.008.

Klimesch, W. (1997). Eeg-alpha rhythms and memory processes. *International Journal of Psychophysiology*, 26(1-3):319–340. doi:10.1016/s0167-8760(97)00773-3.

Klimesch, W., Sauseng, P., and Hanslmayr, S. (2007). Eeg alpha oscillations: The inhibition–timing hypothesis. *Brain Research Reviews*, 53(1):63–88. doi:10.1016/j.brainresrev.2006.06.003.

Krishnan, V., Chang, B. S., and Schomer, D. L. (2017). Normal eeg in wakefulness and sleep: Adults and elderly. In Schomer, D. L. and da Silva, F. H. L., editors, *Niedermeyer's ElectroencephalographyBasic Principles, Clinical Applications, and Related Fields*, Oxford, UK. Oxford University Press.

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., Kennedy, D. N., Hoppel, B. E., Cohen, M. S., Turner, R., and et al. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences*, 89(12):5675–5679. doi:10.1073/pnas.89.12.5675.

Lachaux, J.-P., Lutz, A., Rudrauf, D., Cosmelli, D., Quyen, M. L. V., Martinerie, J., and Varela, F. (2002). Estimating the time-course of coherence between single-trial brain signals: an introduction to wavelet coherence. *Neurophysiologie Clinique/Clinical Neurophysiology*, 32(3):157–174. doi:10.1016/s0987-7053(02)00301-5.

Lahat, D., Adali, T., and Jutten, C. (2015). Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477. doi:10.1109/jproc.2015.2460697.

Leite, M., Leal, A., and Figueiredo, P. (2013). Transfer function between eeg and bold signals of epileptic activity. *Frontiers in Neurology*, 4. doi:10.3389/fneur.2013.00001.

Logothetis, N. K. and Wandell, B. A. (2004). Interpreting the bold signal. *Annual Review of Physiology*, 66(1):735–769. doi:10.1146/annurev.physiol.66.082602.092845.

Lopes Da Silva, F. (2013). Eeg and meg: Relevance to neuroscience. *Neuron*, 80(5):1112–1128. doi:10.1016/j.neuron.2013.10.017.

Lotte, F. (2014). A tutorial on eeg signal-processing techniques for mental-state recognition in brain–computer interfaces. *Guide to Brain-Computer Music Interfacing*, page 133–161. doi:10.1007/978-1-4471-6584-2_7.

Lotze, M. and Halsband, U. (2006). Motor imagery. *Journal of Physiology-Paris*, 99(4-6):386–395. doi:10.1016/j.jphysparis.2006.03.012.

Mantini, D., Perrucci, M. G., Gratta, C. D., Romani, G. L., and Corbetta, M. (2007). Electrophysiological signatures of resting state networks in the human brain. *Proceedings of the National Academy of Sciences*, 104(32):13170–13175. doi:10.1073/pnas.0700668104.

Marzbani, H., Marateb, H. R., and Mansourian, M. a. (2016). Methodological note: Neurofeedback: A comprehensive review on system design, methodology and clinical applications. *Basic and Clinical Neuroscience journal*, 7(2). doi:10.15412/J.BCN.03070208.

Meir-Hasson, Y., Kinreich, S., Podlipsky, I., Hendler, T., and Intrator, N. (2014). An eeg finger-print of fmri deep regional activation. *NeuroImage*, 102:128–141. doi:10.1016/j.neuroimage.2013.11.004.

Mitchell, D. J., Mcnaughton, N., Flanagan, D., and Kirk, I. J. (2008). Frontal-midline theta from the perspective of hippocampal "theta". *Progress in Neurobiology*, 86(3):156–185. doi:10.1016/j.pneurobio.2008.09.005.

Munck, J. D., Goncalves, S., Mammoliti, R., Heethaar, R., and Silva, F. L. D. (2009). Interactions between different eeg frequency bands and their effect on alpha - fmri correlations. *NeuroImage*, 47. doi:10.1016/s1053-8119(09)70268-x.

Murta, T., Leite, M., Carmichael, D. W., Figueiredo, P., and Lemieux, L. (2014). Electrophysiological correlates of the bold signal for eeg-informed fmri. *Human Brain Mapping*, 36(1):391–414. doi:10.1002/hbm.22623.

Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S., and Hallett, M. (2004). Identifying true brain interaction from eeg data using the imaginary part of coherency. *Clinical Neurophysiology*, 115(10):2292–2307. doi:10.1016/j.clinph.2004.04.029.

Nunez, P. and Cutillo, B. A. (1995). *Neocortical dynamics and human EEG rhythms*. Oxford University Press.

Nunez, P., Srinivasan, R., F.Westdorp, A., Wijesinghe, R. S., Tucker, D. M., B.Silberstein, R., and J.Cadusch, P. (1997). Eeg coherency. i: statistics, reference electrode, volume conduction, laplacians, cortical imaging, and interpretation at multiple scales. *Electroencephalography and Clinical Neurophysiology*, 103(5):499–515. doi:10.1016/S0013-4694(97)00066-7.

Ogawa, S., Lee, T. M., Kay, A. R., and Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87(24):9868–9872. doi:10.1073/pnas.87.24.9868.

Ogawa, S., Tank, D. W., Menon, R., Ellermann, J. M., Kim, S. G., Merkle, H., and Ugurbil, K. (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 89(13):5951–5955. doi:10.1073/pnas.89.13.5951.

Pauling, L. and Coryell, C. D. (1936). The magnetic properties and structure of hemoglobin, oxyhemoglobin and carbonmonoxyhemoglobin. *Proceedings of the National Academy of Sciences*, 22(4):210–216. doi:10.1073/pnas.22.4.210.

Perronnet, L., Lécuyer, A., Mano, M., Clerc, M., Lotte, F., and Barillot, C. (2018). Learning 2-in-1: towards integrated eeg-fmri-neurofeedback. *bioRxiv*. doi:10.1101/397729.

Perronnet, L., Lécuyer, A., Mano, M., Bannier, E., Lotte, F., Clerc, M., and Barillot, C. (2017). Unimodal versus bimodal eeg-fmri neurofeedback of a motor imagery task. *Frontiers in Human Neuroscience*, 11:193. doi:10.3389/fnhum.2017.00193.

Pfurtscheller, G. and da Silva, F. L. (2017). Eeg event-related desynchronization and event-related synchronization. In Schomer, D. L. and da Silva, F. H. L., editors, *Niedermeyer's Electroencephalographyand Related Fields, Clinical Applications, Basic Principles*, Oxford, UK. Oxford University Press.

Pfurtscheller, G. and Neuper, C. (1997). Motor imagery activates primary sensorimotor area in humans. *Neuroscience Letters*, 239(2-3):65–68. doi:10.1016/s0304-3940(97)00889-6.

Pfurtscheller, G. and Neuper, C. (2001). Motor imagery and direct brain-computer communication. *Proceedings of the IEEE*, 89(7):1123–1134. doi:10.1109/5.939829.

Pfurtscheller, G. and Silva, F. L. D. (1999). Event-related eeg/meg synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, 110(11):1842–1857. doi:10.1016/s1388-2457(99)00141-8.

Raichle, M. E., Macleod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(2):676–682. doi:10.1073/pnas.98.2.676.

Ramos-Murguialday, A., Broetz, D., Rea, M., Läer, L., Yilmaz, , Brasil, F. L., Liberati, G., Curado, M. R., Garcia-Cossio, E., Vyziotis, A., and et al. (2013). Brain-machine interface in chronic stroke rehabilitation: A controlled study. *Annals of Neurology*, 74(1):100–108. doi:10.1002/ana.23879.

Rosa, M., Kilner, J., Blankenburg, F., Josephs, O., and Penny, W. (2010). Estimating the transfer function from neuronal activity to bold using simultaneous eeg-fmri. *NeuroImage*, 49(2):1496–1509. doi:10.1016/j.neuroimage.2009.09.011.

Rubia, K., Criaud, M., Wulff, M., Alegria, A., Brinson, H., Barker, G., Stahl, D., and Giampietro, V. (2019). Functional connectivity changes associated with fmri neurofeedback of right inferior frontal cortex in adolescents with adhd. *NeuroImage*, 188:43–58. doi:10.1016/j.neuroimage.2018.11.055.

Sauseng, P., Hoppe, J., Klimesch, W., Gerloff, C., and Hummel, F. C. (2007). Dissociation of sustained attention from central executive functions: local activity and interregional connectivity in the theta range. *European Journal of Neuroscience*, 25(2):587–593. doi:10.1111/j.1460-9568.2006.05286.x.

Sağ, A. T., Has, A. C., Öztekin, N., Çağrı Mesut Temuçin, and Oğuz, K. K. (2018). Tracking pain in resting state networks in patients with hereditary and diabetic neuropathy. *Noro psikiyatri arsivi*, 56(2):92–98. doi:10.5152/npa.2017.22660.

Schomer, D. L., Epstein, C. M., Herman, S. T., Maus, D., and Fisch, B. J. (2017). Recording principles:analog and digital principles; polarity and field determinations; multimodal monitoring; polygraphy (eog, emg, ecg, sao). In Schomer, D. L. and da Silva, F. H. L., editors, *Niedermeyer's ElectroencephalographyBasic Principles, Clinical Applications, and Related Fields*, Oxford, UK. Oxford University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464. doi:10.1214/aos/1176344136.

Siegel, M., Donner, T. H., and Engel, A. K. (2012). Spectral fingerprints of large-scale neuronal interactions. *Nature Reviews Neuroscience*, 13(2):121–134. doi:10.1038/nrn3137.

Singer, W. (1993). Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology*, 55(1):349–374. doi:10.1146/annurev.ph.55.030193.002025.

Singer, W. (2006). Consciousness and the binding problem. *Annals of the New York Academy of Sciences*, 929(1):123–146. doi:10.1111/j.1749-6632.2001.tb05712.x.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., Luca, M. D., Drobnjak, I., Flitney, D. E., and et al. (2004). Advances in functional and structural mr image analysis and implementation as fsl. *NeuroImage*, 23. doi:10.1016/j.neuroimage.2004.07.051.

Smith, V., Mitchell, D. J., and Duncan, J. (2018). Role of the default mode network in cognitive transitions. *Cerebral Cortex*, 28(10):3685–3696. doi:10.1093/cercor/bhy167.

Sporns, O. (2016). *Networks of the brain*. The MIT Press.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133. doi:10.1111/j.2517-6161.1974.tb00994.x.

Sutter, R., Kaplan, P. W., and Schomer, D. L. (2017). Historical aspects of eeg. In Schomer, D. L. and da Silva, F. H. L., editors, *Niedermeyer's ElectroencephalographyBasic Principles, Clinical Applications, and Related Fields*, pages 1–16, Oxford, UK. Oxford University Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288. doi:10.1111/j.2517-6161.1996.tb02080.x.

Vourvopoulos, A. and Badia, S. B. I. (2016). Usability and cost-effectiveness in brain-computer interaction. *Proceedings of the 7th Augmented Human International Conference 2016 on - AH 16*. doi:10.1145/2875194.2875244.

Vourvopoulos, A., Jorge, C., Abreu, R., Figueiredo, P., Fernandes, J.-C., and Badia, S. B. I. (2019). Efficacy and brain imaging correlates of an immersive motor imagery bci-driven vr system for upper limb motor rehabilitation: A clinical case report. *Frontiers in Human Neuroscience*, 13. doi:10.3389/fnhum.2019.00244.

Wadman, W. J. and da Silva, F. H. L. (2017). Biophysical aspects of eeg and meg generation. In Schomer, D. L. and da Silva, F. H. L., editors, *Niedermeyer's ElectroencephalographyBasic Principles, Clinical Applications, and Related Fields*, Oxford, UK. Oxford University Press.

Wan, X., Riera, J., Iwata, K., Takahashi, M., Wakabayashi, T., and Kawashima, R. (2006). The neural basis of the hemodynamic response nonlinearity in human primary visual cortex: Implications for neurovascular coupling mechanism. *NeuroImage*, 32(2):616–625. doi:10.1016/j.neuroimage.2006.03.040.

Wolpaw, J. R., Birbaumer, N., Mcfarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791. doi:10.1016/s1388-2457(02)00057-3.

Woolrich, M. W., Ripley, B. D., Brady, M., and Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of fmri data. *NeuroImage*, 14(6):1370–1386. doi: 10.1006/nimg.2001.0931.

Worsley, K. (2000). Statistical analysis of activation images. *Functional MRI: An introduction to methods*, 14. doi:10.1093/acprof:oso/9780192630711.003.0014.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net (vol b 67, pg 301, 2005). *Journal of the Royal Statistical Society Series B*, 67:768–768. doi:10.1111/j.1467-9868.2005.00527.x.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192. doi:10.1214/009053607000000127.

Appendix A

Model Fitting Results

The tables here presented summarize the fitting results obtained for the feature space reduction models through non-dependent 15-fold and blocked 5-fold CV. Additionally, the boxplots referent to the NMSE and BIC values estimated at all folds, for each model and procedure, are also displayed.

A.1 Resting State Dataset

Table A.1: Cross validated (blocked 5-fold CV) λ parameter, (NMSE) and (BIC) values, and effective DOF of the models X_{LC}, X_{δ}, X_{θ}, X_{α} and X_{β}. Results respective to the RS data.

	Average λ $(\times 10^{-2})$	**DOF**	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
LC	7.10	125.0	3.49	3.02
Delta	3.57	110.0	7.51	2.35
Theta	2.70	88.0	5.49	1.87
Alpha	4.46	70.0	3.82	1.41
Beta	3.65	82.0	5.59	1.87

Table A.2: Cross validated (blocked 5-fold CV) λ parameter, (NMSE) and (BIC) values, and effective DOF of the elastic net fit of the models X_{LC}, X_{δ}, X_{θ}, X_{α} and X_{β}. Results respective to the RS data.

	Average λ $(\times 10^{-2})$	**DOF**	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
LC	55.45	41.0	6.35	1.57
Delta	5.14	37.0	7.88	1.89
Theta	4.70	31.0	6.18	1.34
Alpha	4.22	30.0	4.09	1.47
Beta	17.96	39.0	7.21	1.51

Table A.3: Cross validated (blocked 5-fold CV) λ parameter, (NMSE) and (BIC) values, and effective DOF of the elastic net fit of the models X_{LC}, X_{δ}, X_{θ}, X_{α} and X_{β}. Results respective to the RS data.

	Average λ $(\times 10^{-2})$	**DOF**	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
LC	7.10	125.0	3.49	3.02
O2	3.45	24.0	5.52	0.629
P4	9.35	21.0	4.95	0.539
C3	24.35	20.0	6.38	0.479
FP2	9.02	24.0	6.79	0.588

Table A.4: Cross validated (blocked 5-fold CV) λ parameter, (NMSE) and (BIC) values, and effective DOF of the elastic net fit of the models X_{LC}, X_{δ}, X_{θ}, X_{α} and X_{β}. Results respective to the RS data.

	Average λ $(\times 10^{-2})$	**DOF**	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
LC	55.45	41.0	6.35	1.57
O2	7.21	14.0	6.53	0.673
P4	4.20	17.0	5.31	0.681
C3	2.45	19.0	5.95	0.844
FP2	1.71	17.0	7.68	0.684

Table A.5: Cross validated (blocked 5-fold CV) λ parameter, (NMSE) and (BIC) values, and effective DOF of the elastic net fit of the models X_{LC}, X_{α}, $X_{\alpha,\,can}$, X_{O2} and $X_{O2,can}$, $X_{\alpha,O2}$ and $X_{\alpha,O2,can}$. Results respective to the RS data.

	Average λ $(\times 10^{-1})$	**DOF**	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
LC	0.710	125.0	3.49	3.01
Alpha	0.446	70.0	3.82	1.41
Alpha,can	0.221	31.0	5.00	0.726
O2	0.345	24.0	5.52	0.629
O2,can	9.58	2.0	9.42	0.0420
Alpha,O2	2.34	7.0	6.25	0.236
Alpha,O2,can	9.58	2.0	9.42	0.0420

Table A.6: Cross validated (blocked 5-fold CV) λ parameter, (NMSE) and (BIC) values, and effective DOF of the elastic net fit of the models X_{LC}, X_{α}, $X_{\alpha,\,can}$, X_{O2} and $X_{O2,can}$, $X_{\alpha,O2}$ and $X_{\alpha,O2,can}$. Results respective to the RS data.

	Average λ $(\times 10^{-2})$	**DOF**	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
LC	55.45	41.0	6.35	1.57
Alpha	4.22	30.0	4.09	1.47
Alpha,can	4.79	15.0	5.87	0.618
O2	7.21	14.0	6.53	0.673
O2,can	95.65	2.0	10.62	0.0526
Alpha,O2	2.17	7.0	7.34	0.318
Alpha,O2,can	95.65	2.0	10.62	0.0526

Figure A.1: Boxplots of the NMSE and BIC values across folds, for the models X_{LC}, X_{δ}, X_{θ}, X_{α} and X_{β}. Results respective to the resting state data.

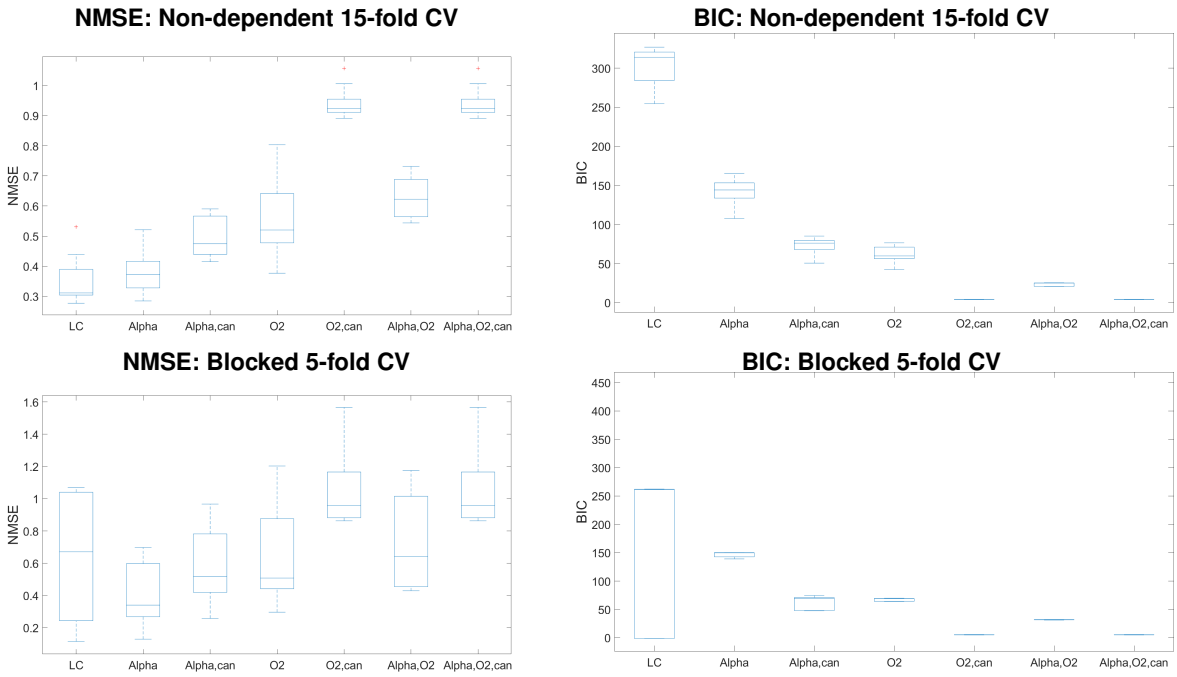


Figure A.3: Boxplots of the NMSE and BIC values across folds, for the models X_{LC}, X_{α}, $X_{\alpha,can}$, X_{O2} and $X_{O2,can}$, $X_{\alpha,O2}$ and $X_{\alpha,O2,can}$. Results respective to the resting state data.

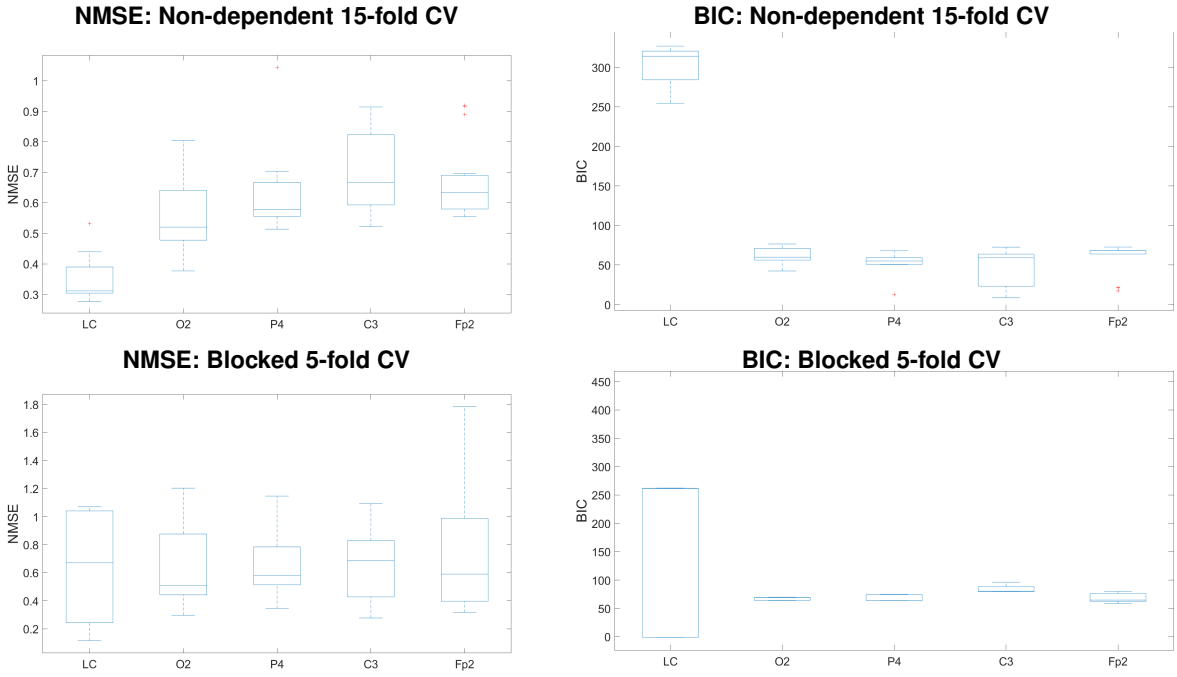


Figure A.2: Boxplots of the NMSE and BIC values across folds, for the models X_{LC}, X_{O2}, X_{P4}, X_{C3} and X_{Fp2}. Results respective to the resting state data.

A.2 Motor Imagery Dataset

Table A.7: Cross validated (blocked 5-fold CV) λ parameter, (NMSE) and (BIC) values, and effective DOF of the elastic net fit of the models X_{LC}, X_{δ}, X_{θ}, X_{α} and X_{β}. Results respective to the MI data.

	Average λ $(\times 10^{-1})$	**DOF**	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
LC	1.65	64.0	1.55	2.15
Delta	1.10	49.0	2.97	1.57
Theta	1.12	65.0	5.24	1.33
Alpha	0.725	60.0	1.62	1.67
Beta	1.73	41.0	2.38	1.16

Table A.8: Cross validated (blocked 5-fold CV) λ parameter, (NMSE) and (BIC) values, and effective DOF of the elastic net fit of the models X_{LC}, X_{δ}, X_{θ}, X_{α} and X_{β}. Results respective to the MI data.

	Average λ $(\times 10^{-2})$	**DOF**	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
LC	8.84	45.0	0.713	2.69
Delta	8.71	37.0	2.10	1.86
Theta	8.89	31.0	3.90	1.54
Alpha	8.34	33.0	1.54	1.53
Beta	8.15	34.0	1.51	1.38

Table A.9: Cross validated (blocked 5-fold CV) λ parameter, (NMSE) and (BIC) values, and effective DOF of the elastic net fit of the models X_{LC}, X_{δ}, X_{θ}, X_{α} and X_{β}. Results respective to the MI data.

	Average λ $(\times 10^{-1})$	**DOF**	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
LC	1.65	64.0	1.55	2.15
FC1	1.34	19.0	4.36	0.401
F7	0.476	23.0	5.30	0.523
CP1	0.816	21.0	5.42	0.447
CP2	2.09	18.0	5.18	0.349

Table A.10: Cross validated (blocked 5-fold CV) λ parameter, (NMSE) and (BIC) values, and effective DOF of the elastic net fit of the models X_{LC}, X_{δ}, X_{θ}, X_{α} and X_{β}. Results respective to the MI data.

	Average λ $(\times 10^{-2})$	**DOF**	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
LC	8.84	45.0	0.713	2.69
FC1	2.55	17.0	3.59	0.691
F7	1.44	22.0	4.22	0.775
CP1	1.96	19.0	4.44	0.729
CP2	2.44	19.0	4.44	0.656

Table A.11: Cross validated (blocked 5-fold CV) λ parameter, (NMSE) and (BIC) values, and effective DOF of the elastic net fit of the models X_{LC}, X_{α}, $X_{\alpha,\,can}$, X_{O2} and $X_{O2,can}$, $X_{\alpha,O2}$ and $X_{\alpha,O2,can}$. Results respective to the MI data.

	Average λ $(\times 10^{-1})$	**DOF**	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
LC	1.65	64.0	1.55	2.15
Alpha	0.725	60.0	1.62	1.67
Alpha,can	1.91	26.0	5.34	0.440
FC1	1.34	19.0	4.36	0.401
FC1,can	0.140	2.0	10.48	0.0347
Alpha,FC1	0.505	7.0	10.56	0.139
Alpha,FC1,can	0.140	2.0	10.48	0.0347

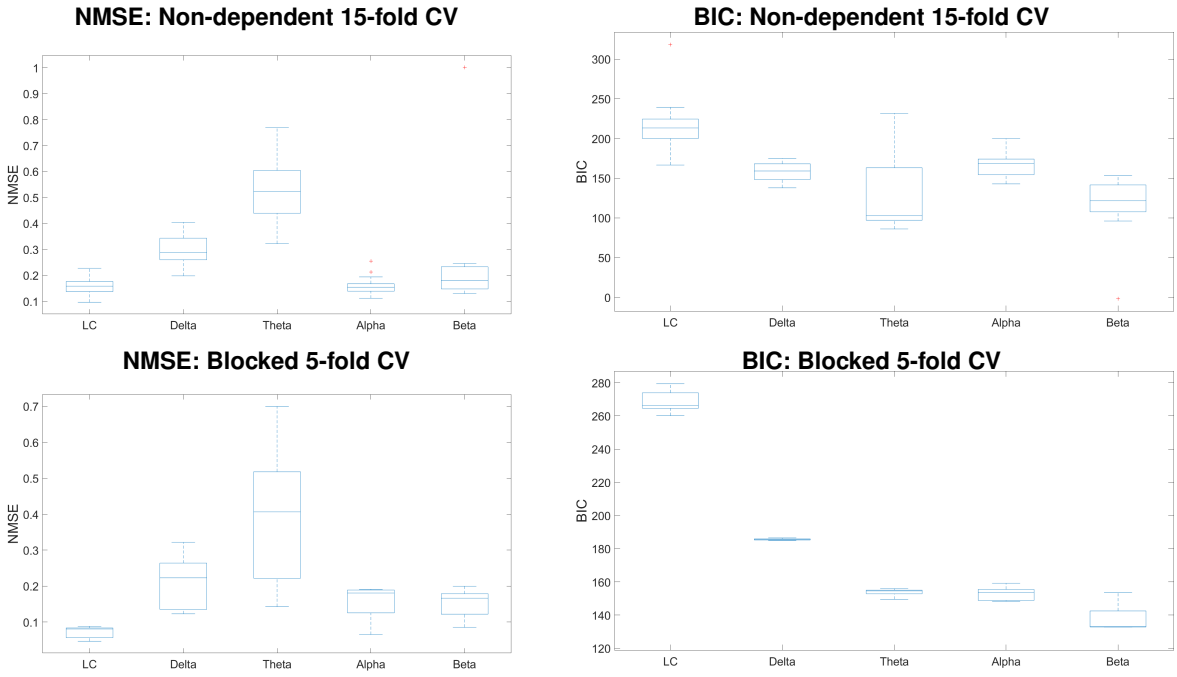


Figure A.4: Boxplots of the NMSE and BIC values across folds, for the models X_{LC}, X_{δ}, X_{θ}, X_{α} and X_{β}. Results respective to the motor imagery data.

Table A.12: Cross validated (blocked 5-fold CV) λ parameter, (NMSE) and (BIC) values, and effective DOF of the elastic net fit of the models X_{LC}, X_{α}, $X_{\alpha,\,can}$, X_{O2} and $X_{O2,can}$, $X_{\alpha,O2}$ and $X_{\alpha,O2,can}$. Results respective to the MI data.

	Average λ $(\times 10^{-2})$	**DOF**	Average NMSE $(\times 10^{-1})$	Average BIC $(\times 10^{2})$
LC	8.84	45.0	0.713	2.69
Alpha	8.34	33.0	1.54	1.53
Alpha,can	4.28	19.0	4.47	0.748
FC1	2.55	17.0	3.59	0.691
FC1,can	0.734	2.0	12.34	0.0452
Alpha,FC1	0.151	7.0	11.52	0.273
Alpha,FC1,can	0.734	2.0	12.34	0.0452

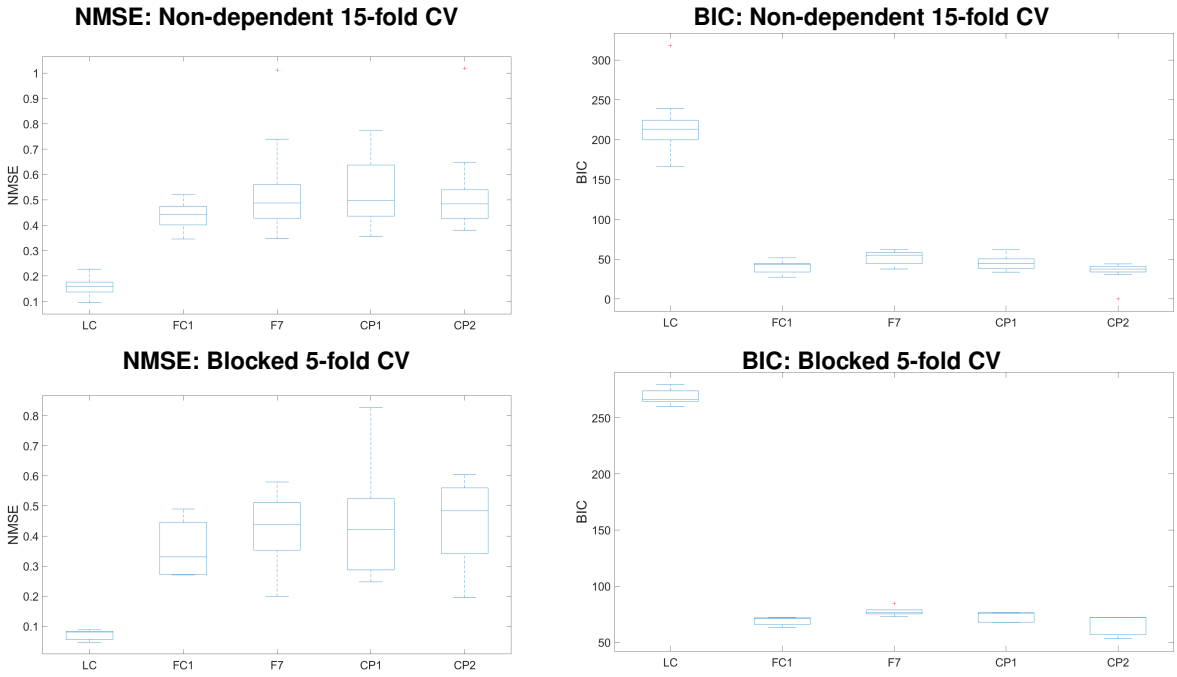


Figure A.5: Boxplots of the NMSE and BIC values across folds, for the models X_{LC}, X_{FC1}, X_{F7}, X_{CP1} and X_{CP1}. Results respective to the motor imagery data.

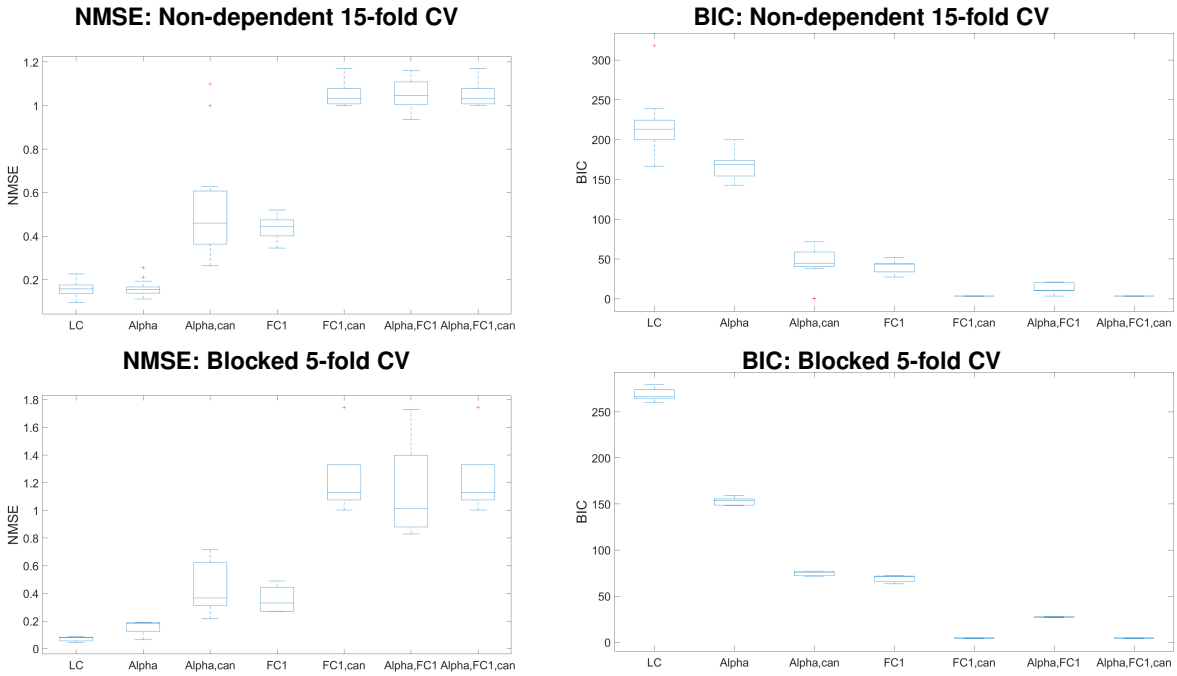


Figure A.6: Boxplots of the NMSE and BIC values across folds, for the models X_{LC}, X_{α}, $X_{\alpha,\,can}$, X_{FC1} and $X_{FC1,can}$, $X_{\alpha,FC1}$ and $X_{\alpha,FC1,can}$. Results respective to the motor imagery data.

Appendix B

Multiple Comparison Results

The following tables summarize the results of the pairwise multiple comparison tests performed for all the models assessed. To compare between models, an independent test was run for the NMSE and another for the BIC. NMSE and BIC values estimated through 15-fold CV.

B.1 Resting State Dataset

Table B.1: Pairwise comparison of 15-fold CV results, in terms of NMSE, of the models: X_{RMSF}, X_{TP}, X_{LC}, X_{IPC} and X_{WND}. Results for the RS data.

	RMSF	TP	LC	IPC	WND
RMSF	-	9.92 $\times 10^{-9}$	9.92 $\times 10^{-9}$	9.92 $\times 10^{-9}$	9.92 $\times 10^{-9}$
TP	-	-	9.92 $\times 10^{-9}$	9.92 $\times 10^{-9}$	9.92 $\times 10^{-9}$
LC	-	-	-	9.92 $\times 10^{-9}$	4.32 $\times 10^{-3}$
IPC	-	-	-	-	1.01 $\times 10^{-8}$
WND	-	-	-	-	-

Table B.2: p-values obtained for the pairwise comparison of 15-fold CV results, in terms of BIC, of the models: X_{RMSF}, X_{TP}, X_{LC}, X_{IPC} and X_{WND}. Results for the RS data.

	RMSF	TP	LC	IPC	WND
RMSF	-	1.03×10^{-8}	9.92×10^{-9}	9.92×10^{-9}	9.92×10^{-9}
TP	-	-	9.92×10^{-9}	9.92×10^{-9}	9.92×10^{-9}
LC	-	-	-	9.92×10^{-9}	9.92×10^{-9}
IPC	-	-	-	-	9.92×10^{-9}
WND	-	-	-	-	-

Table B.3: p-values obtained for the pairwise comparison of 15-fold CV results, in terms of NMSE, of the models: X_{LC}, X_{δ}, X_{θ}, X_{α} and X_{β}. Results for the RS data.

	LC	Delta	Theta	Alpha	Beta
LC	-	9.94×10^{-9}	4.94×10^{-6}	5.55×10^{-1}	1.17×10^{-8}
Delta	-	-	2.78×10^{-2}	2.02×10^{-8}	8.05×10^{-1}
Theta	-	-	-	1.33×10^{-3}	3.15×10^{-1}
Alpha	-	-	-	-	9.81×10^{-7}
Beta	-	-	-	-	-

Table B.4: p-values obtained for the pairwise comparison of 15-fold CV results, in terms of BIC, of the models: X_{LC}, X_{δ}, X_{θ}, X_{α} and X_{β}. Results for the RS data.

	LC	Delta	Theta	Alpha	Beta
LC	-	2.09×10^{-8}	9.92×10^{-9}	9.92×10^{-9}	9.92×10^{-9}
Delta	-	-	1.98×10^{-2}	8.31×10^{-5}	6.21×10^{-2}
Theta	-	-	-	4.69×10^{-1}	9.92×10^{-1}
Alpha	-	-	-	-	2.37×10^{-1}
Beta	-	-	-	-	-

Table B.5: p-values obtained for the pairwise comparison of 15-fold CV results, in terms of NMSE, of the models: X_{LC}, X_{O2}, X_{P4}, X_{C3} and X_{Fp2}. Results for the RS data.

	LC	O2	C3	P4	Fp2
LC	-	9.92×10^{-9}	9.92×10^{-9}	9.92×10^{-9}	9.92×10^{-9}
O2	-	-	9.99×10^{-1}	9.99×10^{-1}	9.53×10^{-1}
C3	-	-	-	9.99×10^{-1}	9.96×10^{-1}
P4	-	-	-	-	2.52×10^{-1}
Fp2	-	-	-	-	-

Table B.6: p-values obtained for the pairwise comparison of 15-fold CV results, in terms of BIC, of the models: X_{LC}, X_{O2}, X_{P4}, X_{C3} and X_{Fp2}. Results for the RS data.

	LC	O2	C3	P4	Fp2
LC	-	9.92×10^{-9}	9.92×10^{-9}	9.92×10^{-9}	9.92×10^{-9}
O2	-	-	1.64×10^{-5}	9.92×10^{-9}	9.92×10^{-9}
C3	-	-	-	9.92×10^{-9}	9.92×10^{-9}
P4	-	-	-	-	4.84×10^{-1}
Fp2	-	-	-	-	-

Table B.7: p-values obtained for the pairwise comparison of 15-fold CV results, in terms of NMSE, of the models: X_{LC}, X_{α}, $X_{\alpha,\,can}$, X_{O2} and $X_{O2,can}$, $X_{\alpha,O2}$ and $X_{\alpha,O2,can}$. Results for the RS data.

	LC	Alpha	Alpha,can	O2	O2,can	Alpha,O2	Alpha,O2,can
LC	-	2.39×10^{-1}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}
Alpha	-	-	5.61×10^{-7}	3.92×10^{-8}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}
Alpha,can	-	-	-	9.03×10^{-1}	7.51×10^{-8}	6.05×10^{-2}	7.51×10^{-8}
O2	-	-	-	-	7.47×10^{-6}	5.73×10^{-1}	7.47×10^{-6}
O2,can	-	-	-	-	-	6.54×10^{-3}	1.00
Alpha,O2	-	-	-	-	-	-	6.54×10^{-3}
Alpha,O2,can	-	-	-	-	-	-	-

Table B.8: p-values obtained for the pairwise comparison of 15-fold CV results, in terms of BIC, of the models: X_{LC}, X_{α}, $X_{\alpha,\,can}$, X_{O2} and $X_{O2,can}$, $X_{\alpha,O2}$ and $X_{\alpha,O2,can}$. Results for the RS data.

	LC	Alpha	Alpha,can	O2	O2,can	Alpha,O2	Alpha,O2,can
LC	-	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}
Alpha	-	-	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}
Alpha,can	-	-	-	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}
O2	-	-	-	-	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}
O2,can	-	-	-	-	-	3.71×10^{-8}	1
Alpha,O2	-	-	-	-	-	-	3.71×10^{-8}
Alpha,O2,can	-	-	-	-	-	-	-

Table B.10: p-values obtained for the pairwise comparison of 15-fold CV results, in terms of BIC, of the models: X_{RMSF}, X_{TP}, X_{LC}, X_{IPC} and X_{WND}. Results for the MI data.

	RMSF	TP	LC	IPC	WND
RMSF	-	7.85×10^{-2}	9.92×10^{-9}	9.92×10^{-9}	9.92×10^{-9}
TP	-	-	9.92×10^{-9}	9.92×10^{-9}	9.92×10^{-9}
LC	-	-	-	9.92×10^{-9}	9.92×10^{-9}
IPC	-	-	-	-	9.92×10^{-9}
WND	-	-	-	-	-

Table B.11: p-values obtained for the pairwise comparison of 15-fold CV results, in terms of NMSE, of the models: X_{LC}, X_{δ}, X_{θ}, X_{α} and X_{β}. Results for the MI data.

	LC	Delta	Theta	Alpha	Beta
LC	-	7.00×10^{-5}	2.17×10^{-7}	8.66×10^{-1}	8.96×10^{-1}
Delta	-	-	5.82×10^{-1}	2.17×10^{-3}	1.70×10^{-3}
Theta	-	-	-	1.04×10^{-5}	7.78×10^{-6}
Alpha	-	-	-	-	1
Beta	-	-	-	-	-

B.2 Motor Imagery Dataset

Table B.9: p-values obtained for the pairwise comparison of 15-fold CV results, in terms of NMSE, of the models: X_{RMSF}, X_{TP}, X_{LC}, X_{IPC} and X_{WND}. Results for the MI data.

	RMSF	TP	LC	IPC	WND
RMSF	-	1.94×10^{-1}	9.92×10^{-9}	9.92×10^{-9}	9.92×10^{-9}
TP	-	-	9.92×10^{-9}	9.92×10^{-9}	9.92×10^{-9}
LC	-	-	-	1.13×10^{-1}	9.52×10^{-1}
IPC	-	-	-	-	1.85×10^{-2}
WND	-	-	-	-	-

Table B.12: p-values obtained for the pairwise comparison of 15-fold CV results, in terms of BIC, of the models: X_{LC}, X_{δ}, X_{θ}, X_{α} and X_{β}. Results for the MI data.

	LC	Delta	Theta	Alpha	Beta
LC	-	9.92×10^{-9}	9.92×10^{-9}	9.92×10^{-9}	9.92×10^{-9}
Delta	-	-	9.98×10^{-1}	9.95×10^{-1}	6.01×10^{-1}
Theta	-	-	-	9.56×10^{-1}	7.82×10^{-1}
Alpha	-	-	-	-	3.56×10^{-1}
Beta	-	-	-	-	-

Table B.13: p-values obtained for the pairwise comparison of 15-fold CV results, in terms of NMSE, of the models: X_{LC}, X_{FC1}, X_{F7}, X_{CP1} and X_{CP2}. Results for the MI data.

	LC	FC1	F7	FC1	FC2
LC	-	5.88×10^{-8}	9.92×10^{-9}	9.92×10^{-9}	9.92×10^{-9}
FC1	-	-	1.18×10^{-1}	3.81×10^{-2}	3.49×10^{-2}
F7	-	-	-	9.90×10^{-1}	9.87×10^{-1}
CP1	-	-	-	-	1.00
CP2	-	-	-	-	-

Table B.14: p-values obtained for the pairwise comparison of 15-fold CV results, in terms of BIC, of the models: X_{LC}, X_{FC1}, X_{F7}, X_{CP1} and X_{CP2}. Results for the MI data.

	LC	FC1	F7	FC1	FC2
LC	-	9.92×10^{-9}	9.92×10^{-9}	9.92×10^{-9}	9.92×10^{-9}
FC1	-	-	1.35×10^{-1}	1.00	5.92×10^{-1}
F7	-	-	-	1.59×10^{-1}	2.48×10^{-3}
CP1	-	-	-	-	5.43×10^{-1}
CP2	-	-	-	-	-

Table B.15: p-values obtained for the pairwise comparison of 15-fold CV results, in terms of NMSE, of the models: X_{LC}, X_{α}, $X_{\alpha,\,can}$, X_{FC1} and $X_{FC1,can}$, $X_{\alpha,FC1}$ and $X_{\alpha,FC1,can}$. Results for the MI data.

	LC	Alpha	Alpha,can	O2	O2,can	Alpha,O2	Alpha,O2,can
LC	-	3.77×10^{-1}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}
Alpha	-	-	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}
Alpha,can	-	-	-	2.79×10^{-2}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}
FC1	-	-	-	-	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}
FC1,can	-	-	-	-	-	8.48×10^{-1}	1.00
Alpha,FC1	-	-	-	-	-	-	8.48×10^{-1}
Alpha,FC1,can	-	-	-	-	-	-	-

Table B.16: p-values obtained for the pairwise comparison of 15-fold CV results, in terms of BIC, of the models: X_{LC}, X_{α}, $X_{\alpha,\,can}$, X_{FC1} and $X_{FC1,can}$, $X_{\alpha,FC1}$ and $X_{\alpha,FC1,can}$. Results for the MI data.

	LC	Alpha	Alpha,can	O2	O2,can	Alpha,O2	Alpha,O2,can
LC	-	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}
Alpha	-	-	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}
Alpha,can	-	-	-	1.00	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}
FC1	-	-	-	-	3.71×10^{-8}	3.71×10^{-8}	3.71×10^{-8}
FC1,can	-	-	-	-	-	1.23×10^{-2}	1.00
Alpha,FC1	-	-	-	-	-	-	1.23×10^{-2}
Alpha,FC1,can	-	-	-	-	-	-	-