TÉCNICO LISBOA

Regression of EEG correlates from simultaneous fMRI signals in epilepsy

Ana Carolina Santa Marta da Silva

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors: Prof. Patrícia Margarida Piedade Figueiredo

Dr. Rodolfo Telo Martins de Abreu

Examination Committee

Chairperson: Prof. Mário Jorge Costa Gaspar da Silva

Supervisor: Dr. Rodolfo Telo Martins de Abreu

Member of the Committee: Prof. Mário Alexandre Teles de Figueiredo

November 2018

Acknowlegments

This thesis would not be possible without the help of many people. Starting with the ones that have been constantly up-to-date during this period and ending with the ones that although completely out of field, had a huge contribution by always sharing their support, curiosity and how proud they were of me.

I would like to acknowledge my supervisor Prof. Patrícia Figueiredo for always being completely available, for all her guidance and motivation throughout the whole work. I am also truly grateful to Rodolfo Abreu for all the support and for readily providing anything that was needed along this journey, by finding errors in my code in 5 sec while I had spent the whole day trying to figure it out and by recurrently sharing new ideas and tests when no meeting was scheduled.

My deepest gratitude goes to my family for all the unconditional love, laughs and for making my life so much easier and lighter. Undoubtedly, they are the most valuable sources of inspiration for me. There are not enough words to describe how important my father, Abílio, my sister, Eva and my grandparents, Rosa e António, are important to me, for closely following all the steps in my life.

Finally, a special thanks goes to all my colleagues for going through these difficult 5 years with me, making them amazing and certainly unforgettable, specially to the ones that shared the thesis time with me in LaSEEB. Also to all my friends from high-school for never stop believing that I could do it. And of course to Bruno and his family, Tita, Vitor and Catarina, for the constant presence, patience and understanding during the last five years.

Resumo

As propriedades complementares das técnicas de Electroencefalografia (EEG) e imagem por Ressonância Magnética Functional (fMRI), em particular o contraste *Blood-Oxygen-Level-Dependent* (BOLD), têm sido particularmente importantes no mapeamento da actividade epiléptica, apesar da sua relação ainda não ser completamente clara.

Para investigar esta questão, as flutuações do sinal BOLD-fMRI médio extraídas das redes epiléticas previamente identificadas no fMRI de cinco doentes com epilepsia, baseada na aquisição simultânea das duas técnicas, foram estimadas a partir do EEG -*EEG Fingerprint* (EFP): *fMRI - informed EEG*. EFPs basearam-se em três métricas ponderadas na potência de modo a prever o sinal BOLD: 1) *Linear Combination of EEG power over frequency bands* (LC), 2) *Root Mean Squared Frequency* (RMSF) e 3) *Total Power* (TP). A performance destas usando diferentes métodos de regressão lineares foi comparada por meio de critérios de seleção. Os métodos considerados foram: *Least Angle Regression* (LAR) e *Elastic Net Regularization*, *Stepwise Regression* (SR) e, regressão dos mínimos quadrados (LS), por motivos de comparação. Foi verificado que quando os preditores de todos os canais do EEG eram incorporados, o SR destacava-se positivamente, enquanto que quando o estudo era realizado ao nível de cada canal, o *Ridge Regression* (RR) era escolhido, devido a propriedades que permitem minimizar o problema de *overfitting*. O LC sem *Hemodynamic Response Funtion* (HRF) pré-definida, foi consistemente o melhor modelo.

Concluindo, este trabalho mostrou que o SR e RR do modelo LC sem HRF pré-definida tiveram os melhores resultados a prever o sinal BOLD. Esta metodologia poderá exibir importantes aplicações no uso das EFPs para modular o sinal BOLD nas redes epiléticas com base em técnicas de *Neuro-Feedback* (NF).

Palavras Chave: EEG Fingerprint, EEG-fMRI simultâneo, Tempo-frequência, Epilepsia, Métodos de regressão lineares, Métricas ponderadas na potência

Abstract

The complementary properties of the Electroencephalography (EEG) and the Blood Oxygen Level Dependent (BOLD)-Functional Magnetic Resonance Imaging (fMRI) techniques have been particularly important for mapping epileptic brain networks. However, the relationship between both neuroimaging signals remains an open question.

We addressed this problem by estimating the EEG correlates -EEG Fingerprint (EFP)- of BOLD-fMRI fluctuations measured at specific distributed epileptic networks previously identified on the fMRI, based on simultaneous EEG-fMRI recordings from a group of five epilepsy patients: fMRI-informed EEG. Due to previous findings, EFPs were based on three power-weighted metrics: 1) Linear Combination of EEG power over frequency bands (LC), 2) Root Mean Squared Frequency (RMSF) and 3) Total Power (TP) - in order to predict the BOLD signal changes.

The resulting EFPs were estimated using different linear regression methods and, their prediction performance was compared by different selection criteria. The considered methods were: Elastic Net Regularization approach including Ridge and Lasso, Least Angle Regression (LAR), Stepwise Regression (SR) and Least-squares (LS) for comparison purposes. We found that when the predictors from all channels were integrated, the SR yielded the best BOLD prediction measures, whereas when the investigation was restricted to each channel, no significant differences were found between SR, Ridge and LS. Therefore, Ridge Regression was chosen due to its favourable advantages dealing with the overfitting effect. In relation to the different metrics, LC with a flexible hemodynamic delay was consistently the best model.

In conclusion, this work revealed that SR and Ridge of LC model with no-predefined delay performed best at predicting the BOLD signal of epileptic networks. This methodology may have important applications in the subsequent use of the epileptic EFPs to drive EEG-based Neuro-Feedback (NF) interventions.

Keywords: EEG Fingerprint, Simultaneous EEG-fMRI, Time-frequency, Epilepsy, Linear methods for regression, Power-weighted metrics

Contents

List of Abbreviations

AIC Akaike Information Criterion

ANOVA Analysis of Variance

AP Action Potential

APP Automatic Pre-Processing pipeline

BIC Bayesian Information Criterion

BOLD Blood Oxygen Level Dependent

CBF Cerebral Blood Flow

CBV Cerebral Blood Volume

$CMRO_2$ Cerebral Metabolic Rate of Oxygen

CV Cross-Validation

DOF Degrees of Freedom

ECG Electrocardiogram

ECoG Electrocorticography

EEG Electroencephalography

EFP EEG Fingerprint

EPI Echo-planar imaging

EPSP Excitatory Postsynaptic Potentials

ERP Event-Related Potential

FFT Fast Fourier Transform

FID Free-Induction Decay

fMRI Functional Magnetic Resonance Imaging

FSL FMRIB's Software Library

GFS Global Field Synchronization

GSW Generalized Spike-and-Wave

HRF Hemodynamic Response Function

IC Independent Component

ICA Independent Component Analysis

ICs Independent Components

IED Interictal Epileptiform Discharge

IPSP Inhibitory Postsynaptic Potentials

LAR Least Angle Regression

LAURA Local Autoregressive Average

LC Linear Combination of EEG power over frequency bands

LFP Local Field Potential

LORETA Low Resolution Electromagnetic Tomography

LS Least-squares

LTI Linear Time-Invariant system

MPs Motion Parameters

MR Magnetic Resonance

MRI Magnetic Resonance Imaging

MUA Multi-Unit Activity

NBR Negative Bold Response

NF Neuro-Feedback

NMSE Normalized Mean Squared Error

PSI Phase Synchronization Index

PSP Postsynaptic Potential

RF Radiofrequency

RMSF Root Mean Squared Frequency

ROR Relative Overfitting Rate

RSS Residual Sum-of-Squares

SPM Statistical Parametric Mapping

SR Stepwise Regression

SSW Slow Spike-and-Wave

SURE Stein's Unbiased Risk Estimation

SVD Singular Value Decomposition

TLE Temporal Lobe Epilepsy

TP Total Power

VE Variance Explained

List of Figures

List of Tables

1. Introduction

1.1 Motivation

The Electroencephalography (EEG) is by far the most commonly used technique to study human brain function. The fast electrical dynamics of neuronal populations can be adequately sampled due to a millisecond temporal resolution of this technique. This property allows it to be the gold standard to non-invasively monitor epileptic activity. Particularly in epileptic patients with focal epileptic activity, it would be of great importance to accurately infer about the localization of such abnormal activity in the brain in order to consider a posterior surgical treatment (Niedermeyer et al., 2011).

However, the signals that reach the scalp and are measured by the EEG result from a mixture of electrical potentials that propagate from several brain sources. Therefore, the EEG source reconstruction has no unique solution (Lopes, 2010). Even though some strategies have been implemented, they all require several assumptions and provide limited spatial resolution with a heterogeneous spatial sensitivity particularly for superficial versus deep sources (Liu and He, 2008).

On the other hand, Blood Oxygen Level Dependent (BOLD)-Functional Magnetic Resonance Imaging (fMRI) is a neuroimaging technique that allows the localization of brain activity with an excellent spatial resolution. Its major limitation is regarding the origin of the measured signals: it is an indirect measure of neuronal activity. Moreover, the temporal resolution of fMRI is limited by the relatively slow Hemodynamic Response Function (HRF), with BOLD changes being delayed relative to the onset of neuronal activity (Kilner et al., 2005; Logothetis et al., 2001; Logothetis, 2008).

The EEG and fMRI exhibit highly complementary characteristics and their integration has been widely used, particularly, on epileptic networks mapping. Several integration strategies are available, the most common being the asymmetrical approaches where: (i) the spatial location of focal fMRI activations are used to constrain the space of solution of EEG sources (fMRI-driven EEG) or (ii) information of interest from the EEG is used to estimate BOLD-fMRI fluctuations (EEG-informed fMRI) (Goldman et al., 2002; Laufs et al., 2003b).

In this work, a different approach was investigated, whereby we estimated the EEG correlates of BOLD-fMRI fluctuations measured at specific distributed epileptic networks previously identified on the fMRI, based on simultaneous EEG-fMRI recordings. These network-specific EEG correlates are termed EEG Fingerprints (EFP)s (Meir-Hasson et al., 2014). This complex disease has reported a variety of epileptic foci and networks that considerably differ from patient to patient. Therefore, an individual EFP based on the respective fMRI network was estimated for each patient (Perronnet et al., 2018; Meir-Hasson et al., 2016).

In future, the fMRI-informed EEG features could be used to develop an EEG-based Neuro-Feedback (NF) protocol for targeting those distributed epileptic brain networks. The NF consists in measuring brain activity, using techniques such as EEG or fMRI, and presenting a representation of an activity pattern of interest to an individual in real-time while instructing them to manipulate it through self-regulation. The NF has been successfully used to normalize aberrant activity patterns in various neurological diseases

such as epilepsy (Sitaram et al., 2017). Although fMRI can in principle be used for NF protocols, EEG is by far more practical NF modality, given its portability and much lower cost. Therefore, the construction of a model of EEG features, driven by the fMRI measurement has been used for the identification of specific epileptic networks.

1.1.1 Objectives

The main objective of this thesis was to develop a subject-specific EFP, from the EEG-fMRI data collected from five epilepsy patients in the scope of the DynImag project (Ref:FCT-PTDC/SAU-ENB/112294/2009). The epileptic brain networks were identified on the fMRI, following previous work by the team Abreu et al. (2018b). The specific objectives of this thesis were:

1. To investigate different models comprising several combinations of EEG features in terms of their ability to predict the average BOLD-fMRI signal of each patient's epileptic network.
2. To estimate the EFP from each EEG models using different regression methods (e.g., shrinkage and subset selection) and compare their predictive performance based on selection criteria (e.g., Bayesian Information Criterion (BIC));
3. To select the best regression method and EEG model;

1.1.2 Thesis Outline

The practical part of the thesis is presented in figure 1.1. The thesis is organized into five main parts: Introduction, Methods, Results, Discussion and Conclusion/Future work. The Introduction comprises the general principles of the EEG and BOLD-fMRI, their advantages and disadvantages which motivated their integration into a multimodal technique to study epileptic activity, the electrophysiological metrics that were extracted in this work and the state-of-the-art. Furthermore, the theoretical background of the linear methods for regression and the selection criteria for model assessment that were used, are subsequently introduced. In Methods, the methodology steps applied throughout the work are described and their main results are depicted in the Results chapter. Following the same structure, in Discussion the most important findings in Results are explained and associated to previous works.

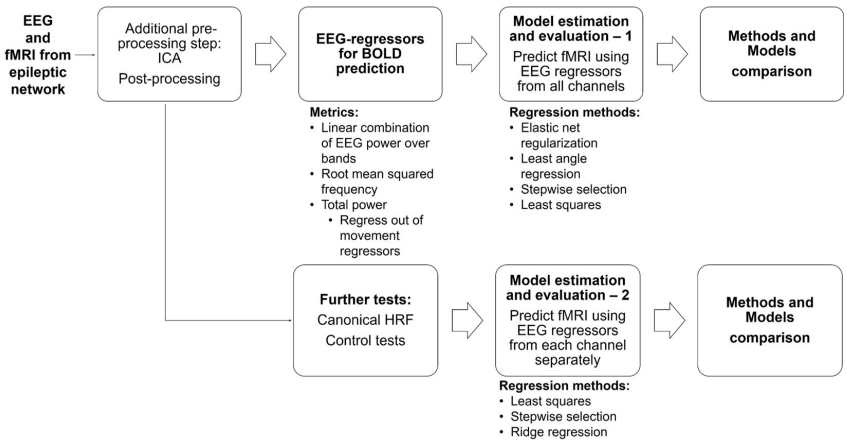

Figure 1.1: Pipeline overview of the thesis, highlighting the main steps that were performed throughout this work.

1.2 EEG-fMRI in epilepsy

1.2.1 Context of the disease

Epilepsy is a common and chronic neurological disease with a prevalence that exceeds more than 50 million people worldwide. This disease affects humans of all ages, geographic location, ethnicity and social background. Although the precise etiology remains unknown in many patients, there are factors proven to be involved in epilepsy onset such as tumours, strokes, dementia, head trauma, genetic factors, meningitis, prenatal injury, oxygen deprivation, among others (van Mierlo et al., 2014).

It is a disease characterized by recurrent and unpredicted seizures, defined as an abnormal, excessive firing of neurons with synchronous electrical activity (hypersyncronization) in the brain. Seizures are patient-specific, therefore may exhibit different characteristics among individuals. The severity of the manifestations intrinsically depends on the seizure's features and its epileptogenic focus, defined as the brain regions that are responsible for the generation and propagation of seizures (Tatum et al., 2018).

The most useful non-invasive instrument in the diagnostic of epilepsy is the EEG, by allowing to retrieve information about the electrical activity with a high temporal resolution in the order of milliseconds (Noachtar and Rémi, 2009; Niedermeyer et al., 2011).

1.2.2 Electroencephalography (EEG)

The EEG is one of the two imaging techniques employed in the present study. This section provides an introduction to this approach, namely a brief description of the cellular basis for signal measurement, the broad division of its rhythms into frequency bands, the Independent Component Analysis (ICA) procedure for time series denoising and finally, discuss its main disadvantage and how it can be in principle mitigated by the simultaneous integration with the fMRI technique.

1.2.2.1 Introduction to EEG

Hans Berger (1873-1941) is seen as the pionner of the modern EEG and started human EEG studies in 1924 (Niedermeyer et al., 2011). It is a method that enables the recording of spontaneous or evoked electrical activity generated by an ensemble of brain cells acting in synchrony in brain. This electrical activity is recorded by means of electrodes, which can be placed: 1) at a short distance from the sources with depth electrodes, 2) on the cortical surface (Electrocorticography (ECoG)) or 3) on the scalp (EEG in the most common sense) (Olejniczak, 2006; Lopes, 2010). A differential amplifier is then used to record the potential difference between each electrode and the reference electrode.

1.2.2.2 The cellular basis of EEG signals

The main parts of a typical neuron are: the dendrites, the body cell, the myelinized tube called axon for brain signals propagation, and the presynaptic terminals.

The pyramidal neurons are the principal type of neurons in the cortex. It is a type of multipolar neuron because it possesses a single axon and a branch of dendrites. Like the other types of cells in the human

body, they present a negative charge in the intracellular space of -60 mV, named of resting potential, and a positive charge in the extracellular space. The selective permeability of the cell's membrane to ions such as K^+, Na^+ and Ca^{2+} is the reason behind this unbalanced distribution of charges (Guyton and Hall, 2006).

The communication of two neurons is mediated by the release of neurotransmitters from a presynaptic neuron to the synaptic cleft. Then the neurotransmitters can couple with specific receptors on the membrane of the postsynaptic neurons and trigger an alteration on its permeability to ions. This leads to the generation of subthreshold electrical potential called Postsynaptic Potential (PSP). Afterwards, if the transmembrane potential is depolarized to values approximately above -50 mV, voltage-gated ion channels in the membrane open, rapidly increasing its permeability to Na^+ ions. Their influx can change the polarity of the membrane from its equilibrium potential to approximately +30 mv leading to a phenomenon called Action Potential (AP), which is the exchange of ions through the membrane of the cell (dentrites through the axon until it reaches the presynaptic terminal). The PSPs on a dendrite are more likely to contribute to the EEG than the action potentials because there are too fast to contribute much to EEG (Guyton and Hall, 2006; Akay, 2006; Olejniczak, 2006).

The PSP can either induce a membrane depolarization called Excitatory Postsynaptic Potentials (EPSP), or a membrane hyperpolarization called Inhibitory Postsynaptic Potentials (IPSP) (Akay, 2006). An EPSP results from the inward flow of positive ions (i.e. calcium - Ca^{2+} and sodium - Na^+ ions), creating a negative extracellular voltage at the dendrites, called an active sink. Due to volume conduction in the extracellular medium, this creates a positive extracellular voltage at the soma, called a passive source. This difference of potential along the neuron generates a current dipole. The reverse effect occurs during an IPSP due to the outward flow of positive ions (K^+) or the inward flow of negative ions (Cl^-) mediated by Ca^{2+} and GABA, respectively, creating an active source at the level of the synapse and distributed passive sinks along the soma-dendrite membrane (figure 1.2). Thus, the positive electric current is directed to the intracellular medium in the case of an EPSP and is directed from the inside of the neuron to the outside in the case of an IPSP (Lopes, 2010).

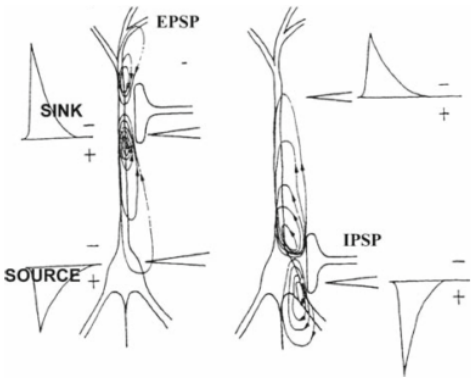

Figure 1.2: Postsynaptic extracellular potentials. Representation of a cortical pyramidal cell showing the current flow patterns from an excitatory synaptic activation, EPSP, and a inhibitory synaptic activation, IPSP. Both cases show a dipolar source-sink configuration. The extracellular recorded EPSP and IPSP is drawn on the left and on the right, respectively (adapted from Lopes (2010)).

Also the existence of a spatially synchronized neurons population (approximately 10^8 neurons) is a

requirement for their electric and magnetic field to be picked up at a distance. In order to reach the scalp, the neuronal electric currents generated at the level of the cortex are propagated through several layers with different geometry and electrical properties. The magnitude of the electric current reached at the scalp reflects the complex summation of AP, EPSP and IPSP from a pool of neurons (Lopes da Silva, 2013).

At the macroscopic level, the measured electrical potentials that result from the postsynaptic activity from a large synchronous population of pyramidal cells is called a Local Field Potential (LFP). This can sum up constructively due to the open-field geometry of the pyramidal cells which allows a laminar current along the main axes. This type of cells plays a major role in the generation of the synaptic potentials captured in the EEG, due to their unique orientation with their long apical dendrites perpendicular to the cortical surface.

Although APs are not the main substrate underlying EEG measures, if they overlap in time, the sum of their fluctuations can be locally measured, originating a Multi-Unit Activity (MUA) (Nunez and Silberstein, 2000).

Compromising the imbalance between the inflow/outflow of neurotransmitters can result in an excessive excitation and a relative reduction of inhibition. This unmanageable situation is the essence of epileptic activity as well as the so-called Interictal Epileptiform Discharge (IED), which occur between seizures. It is triggered by a violent outburst of neuronal depolarization that creates a large EPSP and is continuously excited by many simultaneous neurons belonging to the same population. This synchronous neuronal depolarization starts from calcium and sodium influx and is identified as an IED by the EEG. An influx of negative charges at the axon will follow afterwards, leading to negative slow-waves (Tatum et al., 2018; Niedermeyer et al., 2011).

The electric activity measured on the scalp exhibits characteristic rhythms which have been traditionally separated into the following frequency bands: delta (0.4-4Hz), theta (4-8Hz), alpha (8-13Hz), beta (13-30Hz) and gamma (above 30 Hz) (Akay, 2006). The power spectra of the EEG follows a power law distribution (relation in 1.1). Specific peaks in the power spectra measured in a human brain are associated with the mentioned different rhythms.

$$P \propto \frac{1}{f^{\beta}}, \tag{1.1}$$

where P is the power of the EEG, f is the frequency and β is a spectral parameter experimentally determined.

1.2.2.3 Epileptic Networks

A brain network is defined as a graph of interconnected neuronal populations with temporally correlated activity. The concept of graph theory has been currently used to understand and simulate the neuronal networks: a group of neuronal populations is called a node, and highly connected nodes are called hubs which are linked by a specific path that allows the flow of information.

The notion of networks interactions and connectivity at short and long range provide a better un-

derstanding of the pathophysiology of epilepsy. The concept is that seizures occur in strongly coupled networks due to a shift in the dynamic balance between excitatory and inhibitory processes at the microscopic level (as explained before in the subsection 1.2.2.2). These networks recruit more distant brain regions, recognized by the EEG and fMRI recordings, such as the thalamocortical system involved in Absence epilepsies (have short period of "blanking out") or the recuitment of the parahippocampal cortices involved in some models of Temporal Lobe Epilepsy (TLE).

This imbalance can be triggered by several reasons such as structural lesions like tumors. These lesions can be synaptically connected with other brain regions, such that the epileptic activity may propagate along the connecting pathways constituting an "epileptogenic network", or otherwise stay confined to a region around the lesion (Gotman, 2008; Stefan and da Silva, 2013). This difference is the core of seizure's classification which is addressed hereafter.

1.2.2.4 EEG in patients with Focal and Generalized Epilespy

The epileptic seizures can be classified into three sub-types: focal, generalized and unknown source. This classification is based on the location where the seizure starts from and how it propagates. This is closely linked to the severity of seizures' manifestations (Tatum et al., 2018).

1.2.2.4.1 Focal Epilepsy

The International League Against Epilepsy (ILAE) defined focal seizures as an electrical discharge in a limited area of the brain (without contra-lateral propagation) which can be further classified depending on the involved brain lobe: temporal, frontal, parietal and occipital lobe (mentioned in descending order of frequency). In adults, the temporal lobe is the most epileptogenic site of focal seizure onset, associated with TLE (Tatum et al., 2018).

Patients with focal and drug resistant epilepsy are usually the best candidates for surgery because a local resection of the brain region responsible for the seizures can be performed. In order to improve the delineation of the epileptic focus, the placement of invasive electrodes such as in intra-cranical EEG is generally done to get a higher sensitivity of the measured activity and of its possible sources (Niedermeyer et al., 2011).

The capacity of seizure activity to propagate through brain networks, sometimes leads to doubts about an event is a unitary seizure or a sum of multiple seizures starting from different networks (Fisher et al., 2017).

1.2.2.4.2 Generalized Epilepsy

A generalized onset seizure happens when both hemispheres are activated (contra-lateral manifestation), occurring a widespread epileptic discharge captured across a large number of electrodes (Tatum et al., 2018).

A seizure should be limited in time, however the following two main parts are sometimes indistinct: the ictal and the inter-ictal (between seizures) epileptiform activity. The differentiation of the transitions between states is usually difficult because a standard EEG can only detect some of the IED events,

providing limited information of the cortical epileptiform activity (Fisher and Scharfman, 2015). In the majority of patients, seizures are seen as infrequent events in an outpatient setting while inter-ictal activity is the most dominant.

The epileptiform activity between seizures, the IEDs, are distinctive waveforms or series of waveforms which may have several morphologies depending on the type of epilepsy (Bagheri et al., 2017). Despite intra- and inter-patient variability, the inter-ictal period may contain high amplitude spikes or sharp waves. The former is overall characterized by a short duration around 20-70 ms and in the case of the latter, a longer duration around 70-200 ms. A slow wave lasting 200-500 ms is usually present after spike, called the spike and wave complex. It is thought to be a consequence of postsynaptic activity.

A generalized genetic epilepsy (GGE) is characterized by Generalized Spike-and-Wave (GSW) patterns within repeating bursts at 3 Hz or faster during the ictal period, whereas other types with an aggressive ictal activity can be characterized by the presence of Slow Spike-and-Wave (SSW) at least of 3 Hz (Niedermeyer et al., 2011).

The quality of EEG data, and consequently the ability to detect and discriminate the different types of epileptic events, can be severely hampered by the presence of several sources of artefacts (e.g., eye blinks, eye movements, bad channels). Because some of these artefacts are mainly stationary, one of the most common and effective approaches to tackle them is to use Independent Component Analysis (ICA), which is described next.

1.2.2.5 Independent Component Analysis (ICA) of EEG signals

Signals of non-neuronal origin can easily contaminate the EEG data. Therefore, the interpretation of the EEG signal recorded at each electrode becomes complicated because it reflects a mixture of several activity sources generated inside and outside the brain. The power line noise at 50/60 Hz caused by the electrodes' grounding or the interference of other electric devices are examples of non-physiological noise.

Although approaches such as the segmentation of data periods that exceeds a certain level of contamination or a frequency band filtering appear like simple solutions to deal with the artefacts, they may lead to a great loss of artefact-free information. One clear example of this issue is to use the filtering approach to clean-up data that is overlapped with artefacts in the spectral domain (Chaumon et al., 2015; Ramos et al., 2018). The ICA comes as a widely used method that allows to identify and isolate independent sources of activity, called Independent Components (ICs), by maximizing their statistical independence from each other (Formaggio et al., 2011). Subsequently, the contribution of the artefactual ones can be removed from the EEG signals. This is possible because this technique assumes that the EEG signal is a linear summation of the neuronal and artefactual ICs recorded at the scalp.

Selection of the components

The established relation between sources and the measured signals by the electrodes can be formulated as:

$$EEG = A \times s, \tag{1.2}$$

where EEG represents the EEG signals recorded at the scalp, $EEG = [EEG(1), \ldots, EEG(N)] = [EEG_i(j)]_{nN}$, s is the matrix of the unknown sources, $s = [s(1), \ldots, s(N)] = [s_i(j)]_{mN}$, and A is the unknown mixing matrix with $n \times m$ dimension, where n represents the number of channels, N the number of time samples and m represents the brain sources and artefacts. If we assume that the number of sources is equal to the number of recorded EEG signals, then the ICs, $y = [y(1), \ldots, y(N)] = [y_i(j)]_{mN}$, can also be defined as:

$$Y = W \times EEG, \tag{1.3}$$

$W = A^{-1}$. Each IC is associated to a unique topography (column of A) which describes its projection onto each channel. It is also characterized by a time course, which can be thought as the signal that would have been measured with a channel placed at the correspondent source (Chaumon et al., 2015).

Reconstruction of EEG signal

After the ICs of interest have been assessed, they are back projected in order to obtain the reconstruction of the EEG signal. To perform this step, an identity matrix Z is constructed, having z_{ii} elements equal 0 if the ICs positioned at y_i are aimed to be eliminated from data. It can be represented by:

$$EEG_{rec} = W^{-1} \times Z \times y \tag{1.4}$$

As previously mentioned, the EEG cleaned data are not sufficient to accurately determine the sources of activity due to the the nature of the measured signals.

1.2.2.6 The ill-posed problem and the source reconstruction of EEG

As previously reported, the EEG has a significant disadvantage in terms of localizing the electrical activity sources. The estimation of the sources by measuring the distribution of the electrical activity at the scalp is called the inverse problem (Lopes da Silva, 2013).

This problem is ill-posed because there is an infinite number of possible configurations of the sources that lead to the same recordings. In other words, there is no unique solution. Therefore, some assumptions (physical, geometric, anatomical) regarding the nature of the sources, conductive media and recording electrodes must be set in order to estimate the EEG sources.

The current dipole is the simplest source model, suggesting that the scalp distribution is better represented by an equivalent dipolar current source which can be thought as the centroid of the dipole layers activated at a certain moment, useful in the case of Event-Related Potentials (ERPs) and focal epileptic spikes.

Other methods have been developed in order to handle multiple dipoles. However, they must be assumed to be located at the surface of the cortex. Besides the constraint on the location, it is essential to formulate the conduction problem to determine any source estimation, because it allows to describe the electrical potential at the scalp as a function of the postulated current dipole sources. This step is called the forward part of the inverse problem. By changing the parameters - orientation and position -

of the “constructed” equivalent dipole, “forward solutions” can be calculated in order to obtain a possible distribution of the potentials at the scalp. If the difference between this solution and the measured (original) potentials at the scalp is small, then the solution is generally found. Finding this solution means that is possible but does not mean that is the same as the real generators of brain activity. Having further information about possible sources location (e.g., evoked potential with activity in specific areas) helps to shrink the space of solutions.

Moreover, structural Magnetic Resonance Imaging (MRI) images improve the accuracy on the location of the sources by helping to get realistically shaped volume conductor models.

Among different strategies for solving the inverse problem, Low Resolution Electromagnetic Tomography (LORETA) and Local Autoregressive Average (LAURA) methods stand as two important examples (more details in Niedermeyer et al. (2011)).

However, the limited number of electrodes and, most importantly, the volume conduction from the sources to the surface of the brain (including the lack of information regarding individual geometry and inhomogeneous tissue conductivity) continue to impose a severe limitation on the EEG source localization. This issue is exacerbated with regard to the possibility to measure activity from sub-cortical regions. In order to overcome this problem, the use of simultaneous EEG-fMRI recording allows to increase overall spatio-temporal resolution by combining the high temporal resolution of the EEG with the high spatial resolution of the fMRI (Liu and He, 2008).

1.2.3 Functional Magnetic Resonance Imaging (fMRI)

In this section, the basic principles of the MRI signal and specially the BOLD signal will be briefly introduced. Hereafter, the Hemodynamic Response Function (HRF) shape and its variability in epilepsy will be also addressed. The HRF is involved in the conversion of the EEG regressors into the BOLD information, by convolution, as is going to be also explained in more details throughout this work.

1.2.3.1 Principles of Magnetic Resonance Imaging (MRI)

In MRI, the measured signal usually arises from the energy transitions of the protons in hydrogen atoms, due to their significant density in the human body. A proton possesses a physical property called the spin that has a small magnetic dipole moment that aligns when submitted to an external and static magnetic field (B_0). If the tissue is submitted to a strong magnetic field, as 1.5 T or above, an macroscopic magnetization arises, M, because a tiny majority of spins assume a parallel alignment and their magnetic moments add up. If the protons inside the magnetic field B_0 are exposed to a Radiofrequency (RF) electromagnetic wave with a specific frequency, the so-called Larmor Frequency (rate of precession of the magnetic moment of the proton around the external magnetic field: $\omega \propto \gamma B_0$, with γ being the gyromagnetic ratio), the magnetization is tilted, allowing the nuclei to get exited from the resting-state to a higher energy state.

After the exposure to the RF electromagnetic wave ceases, the magnetization returns to be parallel to B_0 which is called the relaxation process. The excited magnetization has a longitudinal component that

is parallel to the magnetic field and a transverse component that is perpendicular to the magnetic field. The T_1 time constant (longitudinal relaxation time) is related with the longitudinal relaxation, whereas the T_2 time constant (transverse relaxation time) is related with the vanishing of the perpendicular component. The transverse relaxation signal is described by an exponential decay, Free-Induction Decay (FID), in a magnetic field ideally homogeneous. However, the tissues have susceptibility differences that induce field inhomogeneities, resulting in an accelerated signal decay.

The energy transitions between states, high- and low-energy states, are influenced by the properties of the nearby tissue and physiological state of the brain and, the standard MRI signal is a measure of how these transitions spatially differ.

Thus, the signal decays with the effective (apparent) transverse relaxation time, T_2^*, which is shorter than T_2 and depends on the degree of field inhomogeneities. In particular, this degree depends on the composition of the local blood supply, which is related to the neuronal activity. Therefore, the value of T_2^* can be seen as an indirect measure of the neuronal activity (Deichmann et al., 2010; Logothetis and Wandell, 2004).

1.2.3.2 Principles of Blood Oxygen Level Dependent signal (BOLD)

The hemoglobin is a protein present in blood, responsible for oxygen transportation. Blood is oxygenated when rich in oxyheamoglobin (Hb) and deoxygenated when rich in deoxyheamoglobin (dHb) (Guyton and Hall, 2006).

The relative concentration of oxygenated and deoxygenated blood is changed by neuronal activity, which is reflected into T_2^*. The BOLD is the contrast mechanism behind these events (Logothetis and Wandell, 2004; Logothetis et al., 2001).

The deoxyhaemoglobin has paramagnetic properties which means that will interact with the external magnetic fields, increasing the field inhomogeneities locally. The paramagnetic properties result from the free/unpaired electron in the oxygen molecule. On the other hand, the brain tissue around vessels has diamagnetic properties which means that will not attractively interact with the external fields. Therefore at their interfaces, there are magnetic field inhomogeneities that make the T_2^* shorter and lead to a signal reduction in the images (Deichmann et al., 2010).

After neuronal activation, the site of the brain tissue requires a higher demand of oxygen, which is supported by an increase of Cerebral Blood Flow (CBF). CBF or brain perfusion quantifies the arterial blood that is delivery to the brain tissue, and thus of oxygen (Ogawa et al., 1990).

This rapidly leads to a higher concentration of oxyhaemoglobin (comparing to resting state) and a washout of deoxyhaemoglobin. Since oxyhaemoglobin is diamagnetic, this change results in similar magnetic properties of blood and the surrounding tissues, consequently a drop in the field inhomogeneities and finally an increase on the local image intensity, when weighted on the T_2^* constant (Deichmann et al., 2010).

To summarize, an increase of the Magnetic Resonance (MR) signal (BOLD contrast) is an indirect measure of the increase in neuronal activity because between these occurrences a series of changes take place, overall underlined by neurovascular coupling mechanisms.

1.2.3.3 Hemodynamic Response Function (HRF)

The use of the information provided by the BOLD signal to make conclusions about the underlying unobservable neuronal activation is the primary goal of fMRI research. The ability to model an evoked hemodynamic response triggered by a neuronal event plays a crucial role in the EEG-fMRI understanding.

The concept

The HRF is referred as the BOLD signal that is obtained in response to a infinitely short stimulus, as represented in figure 1.3.

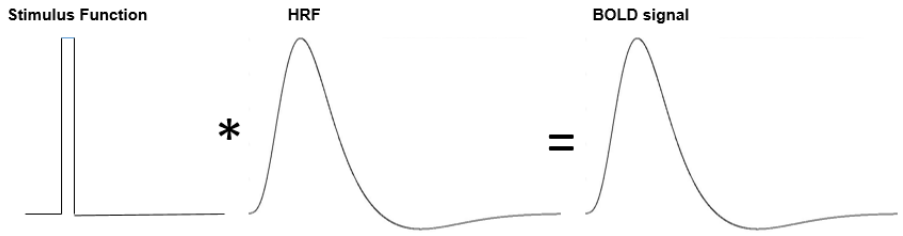

Figure 1.3: The BOLD signal modelling as the convolution of the experimental stimulus and the HRF.

Assuming under certain conditions that the BOLD response is linearly dependent on neural signals, i.e., that it is a Linear Time-Invariant system (LTI), the impulse responses to individual stimulus add linearly to yield the total BOLD response. The HRF acts as an impulse response function of the LTI, i.e., between a mathematical representation of the stimulus over time and the measured BOLD signal.

A canonical (standard) HRF can be defined if it is assumed to be valid across the whole brain and across subjects. One popular choice is the gamma function, which has been shown to provide a reasonably good fit to the impulse response, although it lacks an undershoot. The canonical HRF is usually set by the combination of two gamma functions (double-gamma function) which results in a peak approximately 5 sec after stimulation and a subsequent undershoot: one of the gamma functions models the peak and the other models the undershoot (Friston et al., 1998). A gamma function is characterized by a set of parameters that dictates the basis function shape: delay of response, delay of undershoot, dispersion of response, dispersion of undershoot, ratio of response to undershoot, onset and the length of the kernel (Grouiller et al., 2010). Formally, the canonical HRF is defined as:

$$h(t) = A\left(\frac{t^{p_1-1}p_3^{p_1}e^{-p_3 t}}{\Gamma(p_1)} - c\frac{t^{p_2-1}p_4^{p_2}e^{-p_4 t}}{\Gamma(p_2)}\right), \tag{1.5}$$

where p_1=6, p_2=16, p_3=p_4=1 and c =1/6. Parameters A, p_1/p_2 and p_3/p_4 control the amplitude, shape and scale, respectively. Constant c determines the ratio of the response to the undershoot and Γ acts as a normalizing parameter (Lindquist et al., 2009).

The finite impulse response (FIR) and Fourier basis set are popular basis sets that can be applied directly in a linear regression framework. However, the gamma functions are a more parsimonious solution, by allowing a reduced number of Degrees of Freedom (DOF) to modulate the variablity, when compared with the last two (Murta et al., 2016a). However if the canonical HRF assumption is not true,

then the approach of using a pre-defined canonical HRF may be sub-optimal or inappropriate to analyze BOLD data from different subjects (Aguirre et al., 1998; Grouiller et al., 2010).

Therefore, in order to increase the flexibility of the BOLD response model, the HRF, $h(t)$, may be modeled by a linear combination of basis functions, $g_i(t)$, $h(t) = \sum_{i=1}^{B} \beta_i g_i(t)$, where B represents the number of basis functions in the model, with $i = 1, ..., B$ and β_i are the model parameters to be estimated (Lindquist et al., 2009).

The HRF shape

The figure 1.4 shows a typical design of a HRF, which was the one used at the experimental part of this work. Overall its shape presents a strong positive BOLD response and by last a negative undershoot. In a more general case, an initial dip can also occur before the strong positive peak. The observed phenomena on each of the three stages can be explained by the variation over time of the incoming physiological parameters: the Cerebral Metabolic Rate of Oxygen ($CMRO_2$), the CBF and the Cerebral Blood Volume (CBV). The time course of these variables are also displayed in the same figure.

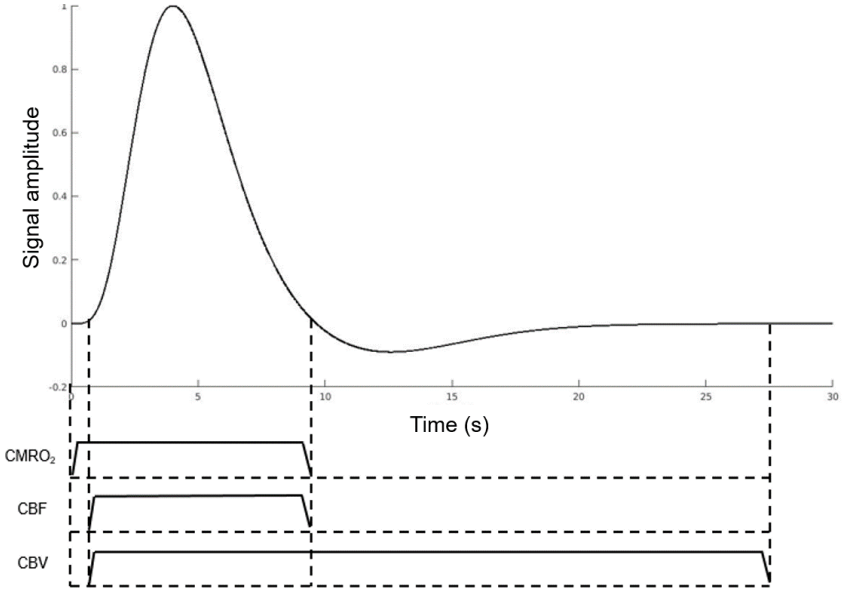

Figure 1.4: The HRF shape. The shape of the HRF explained by the variability of three physiological parameters over time: the $CMRO_2$, the CBF and the CBV. Adapted from Deichmann et al. (2010).

Immediately after a neuronal onset, the initial dip may in some cases be observed due to the $CMRO_2$ raise. This happens because the higher consumption of $CMRO_2$ leads to an increase of the deoxy-heamoglobin concentration, which results in a reduction of the signal. As in most cases, the initial dip is not observable in figure 1.4, unless at high field strengths.

After this short period of time, there is an increase of the CBF and CBV to hold the need for oxygen by the brain tissue. The CBV is the definition for the blood volume per tissue mass or volume (ml/100g or ml/100ml). Due to the increase of CBF, the flow of oxygen goes up and thus the concentration of deoxyheamoglobin goes down. Therefore, this change will induce an increase in the BOLD signal. On the other hand, the CBV increase brings together a higher concentration of deoxyheamoglobin, decreasing the BOLD signal. During this 5-10 sec the BOLD response remains positive because the CBF effect overtakes the increase of the $CMRO_2$ and CBV values.

Regarding the last stage, although the values of $CMRO_2$ and CBF return to baseline, the CBV value does not due to a slower relaxation. So the rest of the time, up to 30 sec after the stimulus onset, the CBV is responsible for the resulting undershoot because the increase of the deoxyheamoglobin concentration is linked to the signal reduction (Buxton et al., 2004; Hoge et al., 1999; Deichmann et al., 2010).

Although the explanation was focused on a strong positive response during the second stage, negative BOLD responses have also been observed. In particular, Moosmann et al. (2003) and Laufs et al. (2003a) revealed an inverse relationship between spontaneous rhythms in the EEG (more specifically alpha activity) and BOLD signal in the occipital cortex.

As the positive BOLD response was shown to be correlated to an increase of the LFP activity, the negative BOLD response has been associated to a decrease in neuronal activity compared to the baseline. Although the mechanism underlying this phenomenon is not completely understood, it is consistent with a decrease in blood flow due to a lower oxygen demand. The Carmichael et al. (2008) investigated the coupling between the time courses of Negative Bold Response (NBR) and CBF in epilepsy patients based on EEG-fMRI simultaneous acquisition.

1.2.3.4 The variability of HRF for epileptic activity

The simultaneous acquisition using EEG-fMRI has a great potential as a clinical tool to study epileptic activity, but the HRF of epileptiform activity remains incompletely characterized. Furthermore, a high intra- and inter-subject variability of the HRF to epileptic discharges has been observed in adults (Bagshaw et al., 2005).

The shape of the HRF also varies as a function of both task and brain region. Therefore, any fixed model would be misspecified for much of the brain. Regarding other factors that may induce variability, the impact of age has been addressed (Huettel et al., 2001; Jacobs et al., 2008; Esposito et al., 1999). Huettel et al. (2001) compared the characteristics of the visually evoked HRF in groups of young and elderly adults based on the fact that age-related change in brain anatomy, neuronal density, vasculature, metabolism, or neuronal responsivity would influenced the fMRI-measured HRF. More recent studies documented that age is considered as a factor of HRF variability associated with changes in brain maturation in children with epilepsy. In the case of Jacobs et al. (2008) was found that youngest children had longer peak times of the HRF.

To increase its ability to fit responses that are shifted in time or have extended activation duration, models using a mixture of gamma functions (canonical HRF) plus its temporal derivative and its dispersion derivatives have been used (Lindquist et al., 2009; Rosa et al., 2010). Other type is the Fourier basis functions: a linear combination of sines/cosines. The principal advantage of Fourier functions is that they are insensitive to artefactual differences in the timing of the stimulus and the acquisition of data (Thornton et al., 2010; Friston et al., 1998).

The use of a flexible HRF has proven to improve results by allowing the detection of additional BOLD responses. The relevance of using different HRF models on the event-related fMRI design has been investigated in inter-ictal studies. The variability of gamma-based HRFs has been assessed by

specifically changing the dispersion and delay (time-to-peak) parameters in a brute force fashion on a subject basis. In the case of epilepsy, found “new” optimal HRF parameters but significantly different from the ones of the canonical HRF (Leite, Marco; Figueiredo, 2012; Grouiller et al., 2010).

Interestingly, some studies already showed that when the HRF is estimated directly from the data (subject- and pathological- specific HRF), brings increased sensitivity to the analysis of epileptic spikes by EEG-fMRI and the results are in better agreement with hypothesized epileptic focus for those patients (Bagshaw et al., 2005; Kang et al., 2003).

1.2.4 EEG and fMRI integration

There are two main types of integration strategies: the symmetrical and the asymmetrical approaches. The first tries to construct a forward model that goes from neuronal activity to both electromagnetic and hemodynamic changes, exploring the association of some features of the EEG and some features of the fMRI. On the other hand, the asymmetrical approach can be applied in both directions, one of them being to improve the EEG source localization based on fMRI-BOLD signal constraints. Other example, which is more commonly applied, is to identify brain regions in which an EEG representative time-course of the activity of interest is correlated with the BOLD-fMRI signal, called the EEG-informed fMRI. In order to do this, an EEG predictor is built based on EEG analysis. For example, by identifying an IED, it is now possible to extract relevant EEG features to localize the brain networks associated with the epileptic discharges (Abreu et al., 2018a; Murta et al., 2015).

The framework used at this work falls into the category of fMRI-informed EEG because the average BOLD signal from a region of interest (e.g., epileptic network area) is first assigned as a constraint in order to investigate the EEG features that better predict it. However, the integration is not trivial because of the strong artefacts that arise but also because the two modalities have different neural correlates (Abreu et al., 2018a). Therefore, the two modalities have a different time frame to a neuronal event: while EEG exhibits an immediate response, fMRI corresponds to a change in BOLD signal which is modulated by the HRF pattern that typically peaks around 5 sec after the appearance of a neuronal event (Herrmann and Debener, 2008; Friston et al., 1994).

1.2.4.1 Electrophysiological correlates of BOLD signal: EEG power-weighted metrics

Although the simultaneous acquisition of EEG and fMRI signals provides a strong neurophysiological basis for deeper understanding of the link between the two modalities, the sensitivity and specificity of EEG-relevant characteristics to predict BOLD fluctuations remains still limited due to lack of information about the basis of the neurovascular coupling (Rainer and Fabrizio, 2010). Concerning epilepsy, literature has revealed an extensive number of EEG-derived metrics to predict BOLD signal changes (Goldman et al., 2002; Laufs et al., 2006; Moosmann et al., 2003; Scheeringa et al., 2008). Indeed a more general question remains about which are the electrophysiological derived quantities that best correlates with the BOLD signal over the epileptic networks (Murta et al., 2015).

From the time-frequency analysis of the EEG signal, several transfer functions and their respective regressors have been derived, exploring its temporal and frequency information (Abreu et al., 2018b).

The Linear Combination of EEG power over frequency bands (LC), the Root Mean Squared Frequency (RMSF) and the Total Power (TP) are the most commonly used power-weighted metrics which were also used in this work. Their formulation is specified below.

Linear Combination of EEG Power over bands

Studies as the one performed in Meir-Hasson et al. (2014) assume a relationship between a linear combination of the EEG power averaged across each frequency band and the BOLD signal. The different frequency bands that were tested are divided into the following variants: delta [1,4] Hz, theta [4,8] Hz, alpha [8,13] Hz, beta [13,30] Hz and gamma [30,45] Hz frequency. The LC EEG metric, q_{LC}, is therefore computed for each of the depicted frequency bands, $b = [fb_{min}, fb_{max}]$, as:

$$q_{LC_b}(t) = \frac{1}{|b|} \sum_{fb_{min}}^{fb_{max}} P(f,t), \tag{1.6}$$

where b is defined as one of the frequency ranges mentioned before, $|b|$ is the number of frequency bins within the specific range and $P(f,t)$ is the power at the frequency bin f at the time point t. This model is going to be continuously called along the work as the LC for simplicity's sake.

Total Power

The model covered at this point which is based on the work of Wan et al. (2006) assumes that the neurovascular coupling is a power transducer. For this purpose, the TP of the EEG time series captured on the scalp was derived. This metric was obtained by summing the power over all frequency bins within the range under analysis, $[f_{min}, f_{max}] = [1, 45]$ Hz:

$$q_{TP}(t) = \sum_{f_{min}}^{f_{max}} P(f,t), \tag{1.7}$$

where $P(f,t)$ represents the EEG power (Rosa et al., 2010; Murta et al., 2015).

Root Mean Squared Frequency

The RMSF model motivated by the result of Kilner et al. (2005), is known as the "Heurist" approach which assumes that the BOLD signal is best explained by variations in the relative power of the different frequencies in the EEG spectrum, instead of changes only in a particular frequency band.

There are two variants of this metric: the un-normalized and the normalized power spectrum P by the TP. The first is referred as the "u-Heuristic". The latest studies have found that the normalized one outperforms the non-normalized version in relation to prediction of BOLD fluctuations (Leite et al., 2013). Therefore, the following version was the one investigated in this work.

$$q_{RMSFN}(t) = \sqrt{\sum_{f_{min}}^{f_{max}} f^2 \tilde{P}(f,t)} \tag{1.8}$$

$$\tilde{P}(f,t) = \frac{P(f,t)}{\sum_{f_{min}}^{f_{max}} P(f,t)}, \tag{1.9}$$

where $\tilde{P}$ corresponds to the power spectrum normalized by the TP.

1.3 Model estimation and evaluation

In this section, different regression methods, procedures for model selection and assessment, and selection criteria will be adressed.

1.3.1 Linear methods for regression

1.3.1.1 Least-squares Regression (LS)

The Least-squares (LS) is the most popular estimation method. Assuming that the input is a vector given by $X^T = (X_1, X_2, \ldots, X_p)$, where p with $j = 1, \ldots, p$, is the number of features and X_j is a vector containing N time samples with $i = 1, \ldots, N$, the aim is to predict a real-valued output designated Y. The ε represents the random errors, in other words, the differences between the original and the predicted output, $\hat{Y}$. The linear regression model takes the following form:

$$Y = f(X) = \varepsilon + \beta_0 + \sum_{j=1}^{p} X_j \beta_j, \tag{1.10}$$

where β_j, plus an initial constant β_0, are the unknown coefficients that allow to combine the different inputs to predict an output as much correlated as possible to the Y. These coefficients are estimated from a training set as those that minimize the Residual Sum-of-Squares (RSS) of the training samples. The RSS is defined as:

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2, \tag{1.11}$$

where x_i and y_i are the observations in the time sample i of the matrices X and Y, respectively.

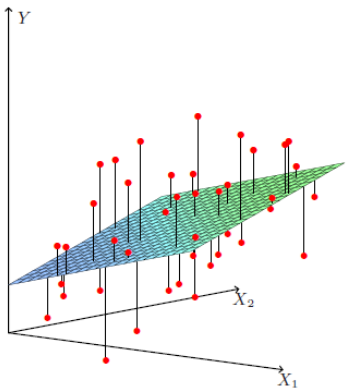

Figure 1.5: Illustration of the geometry of a LS fitting with with $p = 2$ X matrix. The vertical black lines represent the "distance" between the training set and the plane which is the best linear fit to the data (adapted from Hastie et al. (2001)).

Then the question lies on how to minimize the equation above. It is more convenient to deal with multiple regression (estimation of more than β_0-intercept and β_1-slope) when expressed in matrix notation. Defining X as the matrix with dimension $[N \times (p+1)]$, the Y the output vector with dimension $[N \times 1]$, β

as the vector with dimension $[(p+1) \times 1]$ and the ε as the vector with dimension $[N \times 1]$. The additional feature is not omitted when the initial β_0 is considered. The RSS can be rewritten as:

$$RSS(\beta) = \sum_{i=1}^{N} \varepsilon_i^2 = \varepsilon^T \varepsilon = (Y - X\beta)^T (Y - X\beta) \tag{1.12}$$

Thus, the optimal β is given by:

$$\hat{\beta} = \underset{\beta}{argmin} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\} = \underset{\beta}{argmin} \{RSS(\beta)\} \tag{1.13}$$

The RSS is a quadratic function in the $p+1$ parameters, and thus this represents a convex optimization problem. Its unique solution can be easily obtained by setting to zero the first derivative with respect to β:

$$\frac{\partial RSS}{\partial \beta} = -2X^T(Y - X\beta) = 0 \tag{1.14}$$

Finally, the β_j is obtained:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{1.15}$$

The equation 1.15 is called the ordinary LS solution or OLS solution. Therefore, the predicted value at an input vector is given by:

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y \tag{1.16}$$

The $H = X(X^T X)^{-1} X^T$ is named the Hat matrix and maps the transformation of the vector of observed values into a vector of fitted values. The difference between the $\hat{y}_i$ and the y_i is also named residual, $\varepsilon_i = \hat{y}_i - y_i$ (Hastie et al., 2001; Murphy, 2012; Montgomery et al., 2006).

It is also important to cover the question of the DOF of the regression models because they are often used to quantify the model complexity. However there is no direct relation between the number of parameters in the model and the DOF. One clear example is the fact that by finding the predictor, X_j, that most correlates with the response vector, Y, amongst the p different ones and after performing the fit using LS, the DOF would be greater than one by taking into account the stochastic search of that predictor. In the LS method with no prior selection of the predictors, the number of predictors (p) is the number of DOF (Zou et al., 2007).

One problem with LS is that it can easily result in overfitting because when the number of predictors is sufficiently high, it allows to model data, almost perfectly like an interpolation. However if the data are noisy, such parameters often result in complex functions with large negative and positive numbers. In other words, small changes in the data induced by the presence of outliers will strongly influence the estimation of the regression parameters. The variance of the LS estimates of the regression parameters may also be considered inflated when the method is applied to correlated variables, giving rise to parameters with high magnitudes and different signs (Montgomery et al., 2006).

The subset selection methods are defined by keeping only a subset of predictors that exhibit the strongest fitting effects and eliminating the rest from the model, even though some details get sacrificed. Hereafter, the subset selection method used in this work will be introduced: Stepwise Regression (SR).

1.3.1.2 Subset Selection: Stepwise Regression (SR)

The best subset selection is the simplest subset selection method but with a high computational cost that aims to find for each subset size, $k = 1, \ldots, p$, the model with the smallest error (or other criterion that accounts for model complexity). However, this method is only feasible when $p \leq 40$ (Hastie et al., 2001; Draper and Smith, 1998).

The forward-stepwise selection starts with the intercept, β_0, in case of the predictors are not de-meaned (zero mean), and centered predictors with coefficients initially all 0. Then sequentially selects and adds into the model the predictor that most improves the fit. This method is a greedy algorithm that produces a nested sequence of models. It intends to find the optimal choice at each stage in a reasonable number of steps, which is not the same as the optimal solution reached in the best subset selection. Although it might be sub-optimal when compared to best subset selection, when $p >> N$ is always preferable over the second method.

The stepwise-selection can also be performed backwards, the backward-stepwise selection. In contrast to forward-selection, this method can only be used when $N > p$. It starts with the full model and then sequentially selects and removes from the model the predictor that less contributes to the fitting (Hastie et al., 2001).

Other variant is a mixture of these two mentioned methods, since at each step a decision must be taken: whether forward (add a predictor) or backward (removing a predictor) is the best approach. This decision can rely upon criteria (e.g., Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC)) that account for the number of parameters fit or based on the statistical significance of the parameters in a regression. At each step, the p-value on a F-statistic is computed to test models with and without a potential regressor. If a term is not currently in the model, the null hypothesis is that the term would have a zero coefficient if added to the model. If there is sufficient evidence to reject the null hypothesis, the term is added to the model. Conversely, if a term is currently in the model, the null hypothesis is that the term has a zero coefficient. If there is insufficient evidence to reject the null hypothesis, the term is removed from the model.

Basically, the procedure can be summarized into two steps:

1. If any terms not in the model have p-values less than an entrance tolerance (which means, if it is unlikely that they would have zero coefficient if added to the model), add the one with the smallest p-value and repeat this step. Otherwise, the algorithm goes to the next step.
2. If any terms in the model have p-values greater than an exit tolerance (which means, if it is unlikely that the hypothesis of a zero coefficient can be rejected), remove the one with the largest p-value and go to the previous step. Otherwise, it ends.

The method terminates when no single step improves the model (Draper and Smith, 1998).

1.3.1.3 Shrinkage Methods

Contrary to the LS that makes predictions using all the predictors, shrinkage and subset selection methods are a discrete process because predictors can either be retained or discarded, by the use of a constraint. These methods do not suffer as much from high variability as is going to be addressed below.

1.3.1.3.1 Ridge Regression

The Ridge appears as a possible method to minimize this problem, encouraging the parameters to be small by imposing a penalty on their size, thus resulting in a smoother curve. As the LS, the Ridge coefficients are the ones that minimize the RSS with an additional penalty which is the sum of squares of the parameters (second term of the right side in equation 1.17):

$$\hat{\beta}^{ridge} = \underset{\beta}{argmin} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{1.17}$$

where $\lambda > 0$ is a complexity parameter, also called ridge regularization parameter, that allows to control the amount of shrinkage in the regression: greater amount of shrinkage when the λ is larger. In other words, parameters are shrunk towards zero. The L_2 ridge penalty in the equation 1.17 does not have the intercept, β_0, because penalization of the intercept would make the procedure depend on the origin chosen for Y.

The normalization of the feature vectors (columns of X), namely the scaling step, was set as a standard procedure before applying each fitting method. By centering (removing the mean) of each feature vector, $x_{ij} - \bar{x}_j$, the intercept β_0 is no longer needed. Then by using the centered x_{ij}, the p parameters (rather than $p + 1$) are estimated by a ridge regression without intercept.

The equation 1.17 can be rewritten as:

$$\begin{aligned} \hat{\beta}^{ridge} = \underset{\beta}{argmin} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2, \\ \sum_{j=1}^{p} \beta_j^2 \leq t \end{aligned} \tag{1.18}$$

There is a direct relation between the λ from equation 1.17 and the variable t from equation 1.18. As done in the LS problem, the equation 1.17 can be written in the matrix form:

$$RSS(\beta) = (Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta \tag{1.19}$$

With a quadratic penalty, $\beta^T \beta$, the ridge solution is a linear function of Y. By differentiating the equation 1.19 in order to β and making the result equals zero, the ridge solutions are determined by:

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T Y, \tag{1.20}$$

where I is the identity matrix. The constant λI stands out in equation 1.20. A matrix is singular (or non-

invertible) if and only if its determinant is 0, which can only be possible if at least two columns are linearly dependent. In other words, the rank which gives the highest number of linearly independent columns, is not full (less than p). By adding the constant to the diagonal of X^TX before inversion, makes the problem non-singular, even if X^TX is not of full rank. Note that when $\lambda = 0$, the ridge estimator is the LS estimator. The ridge solution is just a scaled version of the LS estimates, $\hat{\beta}^{ridge} = \frac{\hat{\beta}}{(1+\lambda)}$, if the X input matrix is orthonormal ($X^TX = XX^T = I$ and $det(X) = 1$).

Therefore as in 1.16, the predicted value at an input vector is given by (Hastie et al., 2001):

$$\hat{Y} = X\hat{\beta}^{ridge} = X(X^TX + \lambda I)^{-1}X^TY \tag{1.21}$$

A general definition of the DOF for any fitting method is given by Stein's Unbiased Risk Estimation (SURE) theory. Considering any model fitting method, the $\hat{Y}$ the fitted data, the $Y \sim (\mu, \sigma^2 I)$, where μ is the mean and σ^2 the corresponding variance. The DOF formulation is defined as (Zou et al., 2007):

$$df(\hat{Y}) = \sum_{i=1}^{N} \frac{cov(\hat{y}_i, y_i)}{\sigma^2} \tag{1.22}$$

Intuitively, this relation makes sense because the harder the data are fitted, the larger the covariance and the $df(\hat{Y})$ (Hastie et al., 2001). Assuming that $\hat{Y} = HY$, then $cov(\hat{Y}, Y) = cov(HY, Y) = Hcov(Y, Y) = Hvar(Y) = \sigma^2 H$. Therefore the equation 1.22 can be simplified into $df(\hat{Y}) = tr(H)$ (Zou et al., 2007).

The DOF in Ridge can also be determined by the Singular Value Decomposition (SVD) of the predictors matrix X. This decomposition gives extremely useful information about the nature of Ridge. The X matrix is equivalent to UDV^T decomposition, where $U_{N\times p}$ and $V_{p\times p}$ are orthogonal matrices ($U^TU = UU^T$) and $D_{p\times p}$ is a diagonal matrix whose entries are the singular values ($d_j, \ldots, d_p$) of X by a decreasing order.

By integrating the SVD decomposition of the matrix X ($X = UDV^T$), the predicted value $\hat{Y}$ equals to 1.23 after simplification.

$$\hat{Y} = UD(D^2 + \lambda I)^{-1}DU^TY = \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T Y, \tag{1.23}$$

where u_j is the columns of U. Then since $\lambda \geq 0$, the constant $\frac{d_j^2}{d_j^2+\lambda}$ will be always less than 1, which acts as an shrinkage factor of the Y: the larger the d_j^2, the smaller the amount of shrinkage.

The columns of the V matrix are also called *principal components* directions of X, which reflects the directions of all possible linear combinations of the columns of X. The first column of V, v_1, ends up to the the first principal component direction. It has associated the first principal component of X, $z_1 = Xv_1$, with the property of having the largest sample variance of the data. By the SVD decomposition, the z_1 is the same as $z_1 = Xv_1 = u_1 d_1$, where d_1 reflects the amount of variance of the first principal component.

Then, to conclude, the smaller the singular values, d_j, the smaller the variance of the directions that they correspond. As mentioned before, the Ridge skrinks these directions the most by shrinking the associated coefficients, β_j. In this way, Ridge minimizes the high variance that may occur along the

short directions.

After this explanation, the definition of the *effective DOF*, $df(\lambda)$, in Ridge, can be established by being the trace of the hat matrix, H, which is the same as, $\sum_{j=1}^{p} \frac{d_j^2}{d_j^2+\lambda}$. It means that its value decreases as the λ value increases. Although all the parameters in Ridge are non-zero, the effect of the penalty term results in a number of $df(\lambda)$ that is no longer equal to p. It can be noticed that when λ tends to zero or to infinite, the $df(\lambda)$ tends to p (in the case of LS) or to 0, respectively (Hastie et al., 2001).

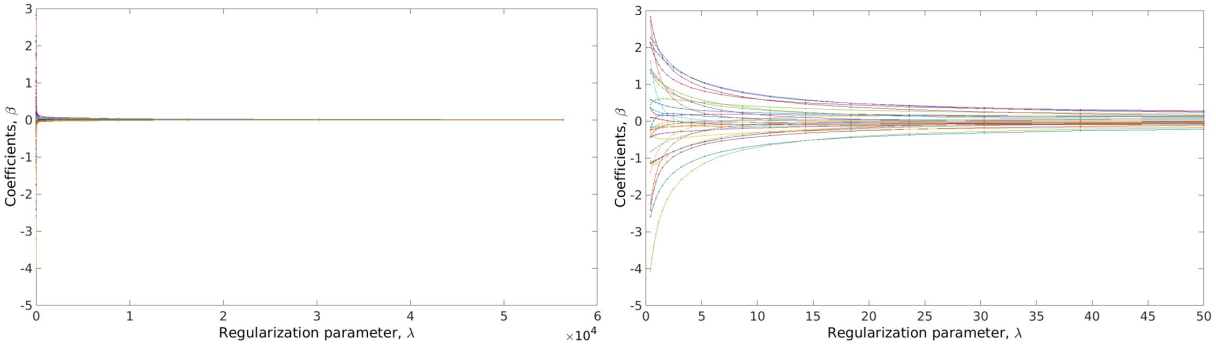

Figure 1.6: Ridge coefficients for an example model (the TP convolved with the canonical HRF) plotted versus the regularization parameter λ. The range of λ values in the left figure goes from 0 (no regularization) towards the value that generates only one DOF. The right figure is a zoom of the smallest λ values.

1.3.1.3.2 Lasso Regression

Like Ridge, Lasso which stands for Least Absolute Shrinkage and Selection Operator, is also a shrinkage method but with some underlying differences. Its formulation is:

$$\hat{\beta}^{lasso} = \underset{\beta}{argmin} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2, \quad \sum_{j=1}^{p} |\beta_j| \leq t \tag{1.24}$$

The Lasso problem can also be formulated in the *Lagrangian form*:

$$\hat{\beta}^{lasso} = \underset{\beta}{argmin} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{1.25}$$

The λ is again a complexity parameter but here called of lasso regularization parameter. The main difference of Lasso compared to Ridge expression (equation 1.17) is regarding the penalty term. In Lasso, a L_1 penalty (instead of L_2) is established, therefore the solutions are no longer linear with Y. One consequence of this difference is the fact that when t is sufficiently small it can lead to some parameters to be exactly zero, similar to subset selection (or feature selection). Then, as λ increases, the number of non-zero parameters also increase.

The t_0 is defined as $\sum_1^p \left|\hat{\beta}_j\right|$, being the parameters β_j equal to the ones from LS estimation. If t in the penalty term is smaller than t_0, then the $\hat{\beta}_j$ parameters are estimated by Lasso. Indeed if $t = \frac{t_0}{2}$ means that on average the LS parameters suffered a shrunk by about 50%. Other parameter usually used in the interpretation of Lasso is the shrinkage parameter, $s = \frac{t}{\sum_{j=1}^{p} |\hat{\beta}_j|}$. When $s = 1$, the result equals the

LS estimation (Hastie et al., 2001). As s decreases to zero, the Lasso coefficients are shrunk.

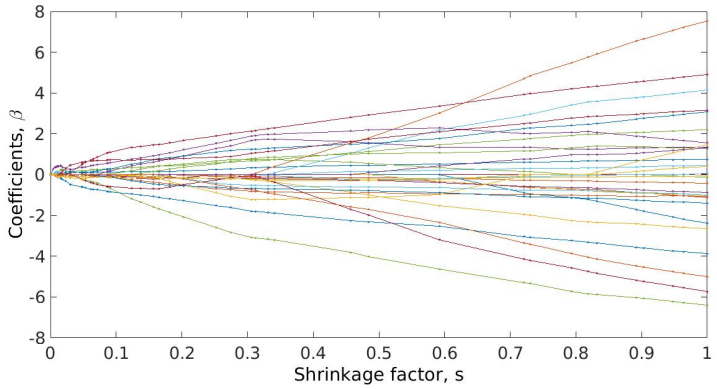

Figure 1.7: Lasso coefficients for an example model (the TP convolved with the canonical HRF) plotted versus the shrinkage parameter s. The range of s values goes from 0 (just one DOF) towards the value that generates the LS coefficients (no regularization).

Zou et al. (2007) showed that the number of non-zero predictors in the model, for any given λ, is an unbiased estimation of the DOF.

1.3.1.3.3 Elastic Net Regularization (ENR)

The naive elastic net is a regularized regression method that results from the linear combination of the L_1 and L_2 penalties from the Lasso and Ridge, respectively (Zou and Hastie, 2005). Its formulation is given in equation 1.26.

$$\hat{\beta}^{naive} = \underset{\beta}{argmin} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda_2 \left\|\beta\right\|_2^2 + \lambda_1 \left\|\beta\right\|_1 \right\} \tag{1.26}$$

The elastic net formulation introduced hereafter is nothing more than a rescaled version of the naive elastic net (Friedman et al., 2010). The estimates from the elastic net method are defined by:

$$\hat{\beta}^{elastic} = \underset{\beta}{argmin} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda P_\alpha(\beta) \right\}, \tag{1.27}$$

where $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$ and the elastic-net penalty term, P_α, is defined as:

$$P_\alpha(\beta) = (1-\alpha)\frac{1}{2}\left\|\beta\right\|_{l_2}^2 + \alpha\left\|\beta\right\|_{l_1} = \sum_{j=1}^{p} \left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \right] \tag{1.28}$$

The α parameter allows for intermediate penalty terms between the two types of regularization: in an extreme $\alpha = 0$ correspondents to Ridge penalty and for $\alpha = 1$ to Lasso penalty. Therefore, when α goes from 0 to 1, for a given λ, the sparsity of the solution, which means the number of parameters equal to zero, increases monotonically from 0 to the sparsity of the lasso solution (Friedman et al., 2010). By comparing the two equations (1.26 and 1.27), the following relations, $\lambda_2 = \lambda(1-\alpha)$ and $\lambda_1 = \lambda(\alpha)$, can be established (Zou and Hastie, 2005).

It is known that the Ridge penalty shrinks the coefficients of correlated predictors towards each other

while Lasso tends to pick one of them and discard the others. In the extreme case of having p identical predictors, Ridge attribute identical coefficients with $1/p$ the size that any single one would get if fit alone. On the other hand, Lasso would break down. One use of α is for numerical stability; for example, the elastic net with $\alpha = 1-\varepsilon$ for some small $\varepsilon > 0$ performs almost like Lasso, but removes any degeneracies caused by extreme correlations (Friedman et al., 2010).

The number of non-zero coefficients, contrary to Lasso, is a biased estimate of the DOF of the elastic net. In order to get an unbiased estimate of the DOF, they should be calculated by the general equation in 1.22. For the elastic net, the DOF are defined as (Zou et al., 2007):

$$df(\hat{Y}) = tr(S_{\lambda_2}(A)) = tr(X_A(X_A^T X_A + \lambda_2 I)^{-1} X_A^T), \tag{1.29}$$

where λ_2 is Ridge parameter and A is the active set, defined as the subset of predictors whose parameters are non-zero.

1.3.1.3.4 Least Angle Regression (LAR)

The Least Angle Regression (LAR) uses a sequential strategy to add parameters similar to the one implemented in the forward-stepwise (previously introduced), however with some differences as subsequently described. As in the above-mentioned regression methods, it is assumed that the predictors are normalized (Hastie et al., 2001). Its algorithm can be summarized into the five steps described afterwards:

1. First, the residual is computed $r = Y - \bar{Y}$ with all the β_j set to zero, $j = 1, \ldots, p$.
2. The predictor X_j that most correlates with r is selected.
3. The value of the corresponding parameter, β_j, is moved from zero towards its LS value, until other predictor X_i, with $i \neq j$, has as much correlation with the current residual, $r_1 = Y - X_j\beta_j$, as does X_j.
4. Then both parameters, β_j and β_i are moved together towards their LS coefficient $r_2 = Y - [X_j X_i][\beta_j \beta_i]^T$
5. The same procedure continues until the p predictors are introduced into the model. At the last step, the full LS takes place.

The name of “least angle” arises from the geometrical interpretation of this method. Consider the explanatory example where $p = 2$, and the predictor that has the most correlation with Y is X_1. The next step is to find the value of β_1 such that the difference between Y and $X_1\beta_1$ (which is equal to r_1) bisects the angle between X_1 and X_2. Since there are no more remaining predictors to search for, the updated residual will simply be $r_2 = Y - [X_1 X_2][\beta_1 \beta_2]^T = 0$. All the details concerning this method can be found in Efron et al. (2004).

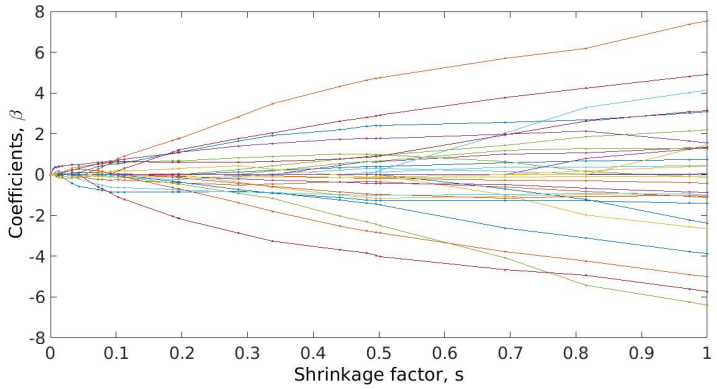

Figure 1.8: LAR coefficients for an example model (the TP convolved with the canonical HRF) plotted versus the shrinkage parameter s. The range of s values goes from 0 (just one DOF) towards the value that generates the LS coefficients (no regularization).

Efron et al. (2004) based on the equation 1.22, showed that the DOF of the fitting at each step k are approximately equal to k, $df(\hat{Y}_k) = k$.

1.3.2 Model selection and assessment

Some of the models described above have complexity parameters that have to be determined (e.g., shrinkage methods). In model selection, the performance of various competing models is estimated with the hope of choosing the best one, given the result of some selection criterion. Having chosen the final model, the model is assessed by estimating measures such as the prediction error on new data. This theme will be addressed hereafter.

1.3.2.1 Bias-variance-complexity trade-off

Supposing that the target data, Y, arises from the sum of the optimal prediction, $\hat{Y} = \hat{f}(X) = X\beta$, with an associated training error that follows a normal distribution, (μ, σ^2). The training error is formulated in equation 1.30:

$$\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i)), \tag{1.30}$$

The loss function, L, can be defined as the squared error, $(Y - \hat{f}(X))^2$ or as the absolute error, $|Y - \hat{f}(X)|$.

The prediction error, also known as the generalization error, is defined as the average of the errors measured on the test sets (sets not used for fitting the model). The Cross-Validation (CV) procedure, namely the data division into sets, is discussed in section 1.3.2.2. It can be divided into the bias term, the variance term and the irreducible error: $Err = bias^2(\hat{Y}) + variance(\hat{Y}) + error$.

The bias component is a measure of the difference between the average of the predictions and the optimal prediction. The variance component is related with the difference between each of the predictions (across the test sets) and the average of the predictions. The noise is the difference between the optimal prediction and the true function, which is beyond control.

The bias and the variance terms vary in a different way with the complexity of the model (i.e., the number of parameters). The bias decreases as model complexity increases, whereas variance increases with model complexity (Arlot and Celisse, 2009). Instead of the average of the prediction errors, the average of the training errors (equation in 1.30) shows a descent trend as the complexity increases.

In the scope of Ridge, the complexity of the model decreases with λ, which means that in figure 1.9, as λ decreases (beginning of the x-axis), the variance gets larger and the bias gets smaller. This effect on the variance can be easily noticed since the range of the predicted values across the ten test folds is larger (box with a larger length) (Hutchison and Mitchell, 2006).

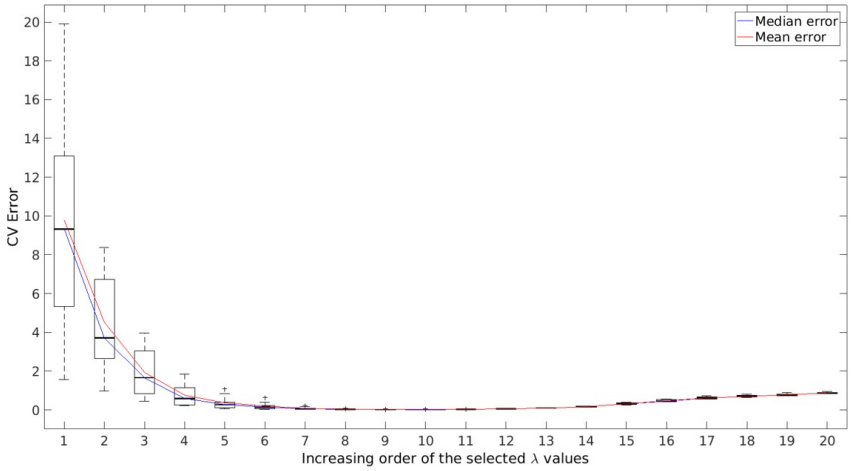

Figure 1.9: Prediction error (test error) estimation obtained by a 10-fold CV using Ridge. The values in the x-axis do not correspond to λ but instead to the order they were selected - from the smaller towards the larger. The blue and the red curves refers to the median and the mean values of the prediction error along different λs, respectively.

In figure 1.9, the NMSE is used as the similarity measure between Y (original BOLD signal) and $\hat{Y}$ (predicted BOLD signal) throughout this work. If NMSE is less than 1, then the prediction is doing better than the simple mean of the signal. Its formula (equation 1.31) is defined below.

$$NMSE = \frac{MSE(Y)}{VAR(Y)} = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{Y})^2} \tag{1.31}$$

In order to find the optimal complexity and therefore the optimal value of λ, a trade-off on the relationship among bias, variance and complexity must be taken into account. The best λ corresponds to the minimum of the sum of the bias and variance values (minimum of the red curve in the figure above) (Hastie et al., 2001; Zou and Hastie, 2005). Although the bias-variance-complexity trade-off using the NMSE as a criterion can be applied in many situations, sometimes the minimum of the curve is not so easily observed. Therefore new criteria (e.g., BIC, AIC) discussed in the next sections are needed to determine the optimal model complexity and minimize the overfitting (Murphy, 2012).

The CV comes also as a procedure to minimize overfitting, mainly if the number of observations is not sufficiently higher than the number of parameters. In addition, it is the most widely method used for determining the prediction error. How the method performs is subsequently explained.

1.3.2.2 K-Fold Cross-Validation

The K-fold CV method uses part of the data to fit the model and the remaining part to test it. In order to assign these two different tasks, the data are firstly divided into K roughly equal-sized parts (also named folds). Two common values for K are 5 and 10 (Hastie et al., 2001). In the context of this work, K=10 was choosen.

The data belonging to K-1 folds is used for fitting the model and the K^{th} fold is used for calculating the prediction error using the fitted model previously determined. The same is repeated until the test fold has been in all the k positions, k= $1, 2, \ldots$,K, and consequently K prediction error values are retained (Hastie et al., 2001).

The CV method assumes that the data are identically distributed, and training and test samples are independent (i.i.d.). However, when it comes to time-series prediction, this assumption is proven to fail due to the inherent auto-correlation and potential non-stationarity of the data. Therefore, the CV application is not straightforward and it is safer to adopt different strategies to ensure the i.i.d. of the samples. Moreover, the K-fold CV has been empirically shown to be favourable compared with other time-series specific techniques such as the ones that aim to remove the dependent samples (non-dependent CV) (Bergmeir and Benítez, 2012; Bergmeir et al., 2018).

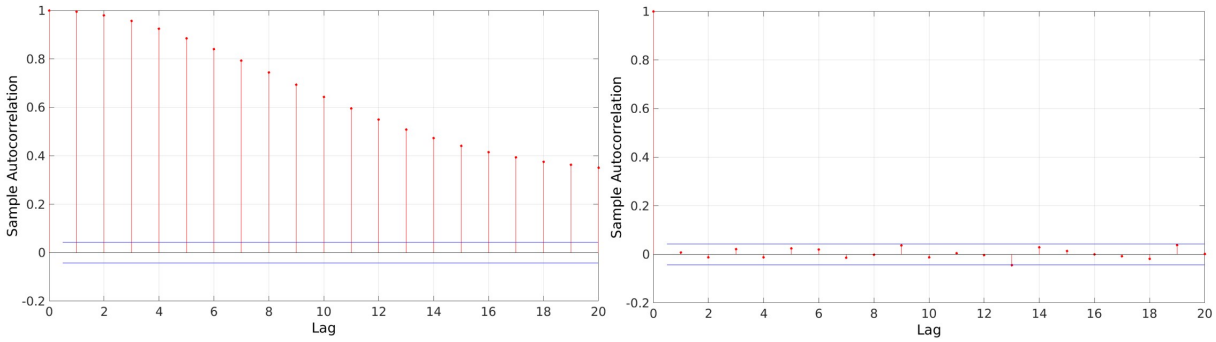

Figure 1.10: The effect of the data randomization on the ten auto-correlation values with 1 sec lags. The left and right figures show an example of before and after sample randomization of the time series, respectively, and the x-axis represents the lag in seconds.

The attribution of each sample to test/training sets can be seen as an indexed function that allocates each observation i to a specific partition k, $\{1, \ldots, N\} \rightarrow \{1, \ldots, K\}$. Since this allocation was performed by a temporal order, the test/training sets resulted into downsampled versions of the original signal. In order to remove the auto-correlation of the time-series, the adopted strategy was to shuffle the data before allocation to sets. This allows to lose the intrinsic auto-correlation (reflected in figure 1.10). Thus, each test set can capture the variability of the continuous signal. Otherwise if an event of interest was restricted to a small portion of the signal and was not used for training, its prediction would be worse.

Ideally, if the amount of data are enough, a validation part is set aside in the training stage. This is called a two stage (nested) M-K fold CV procedure. Inside the training set, a second, inner M-fold CV procedure is used to determine the optimal parameter/s (e.g., λ in the case of Lasso and Ridge). As previously explained in the outer CV, here the training set (also called learning set) is again randomly split into training and validation sets, M times. Then the model is fitted on the training set m, m= $1, 2, \ldots$,M,

for various values of the parameter within a range of interest and selects the one that yields the best results, brings the NMSE or other criteria such as BIC, to minimum, on the correspondent validation set. The same is determined for the remaining validation sets with a fixed test set. The λ^* represented in figure 1.11 is the optimal parameter for a specific k and was determined as the most frequent parameter across the M validation sets. Finally, for a k, the model is fitted in the learning set and the prediction error is calculated in the test set. The average of the prediction errors is then determined across the K test sets (Podlipsky et al., 2012).

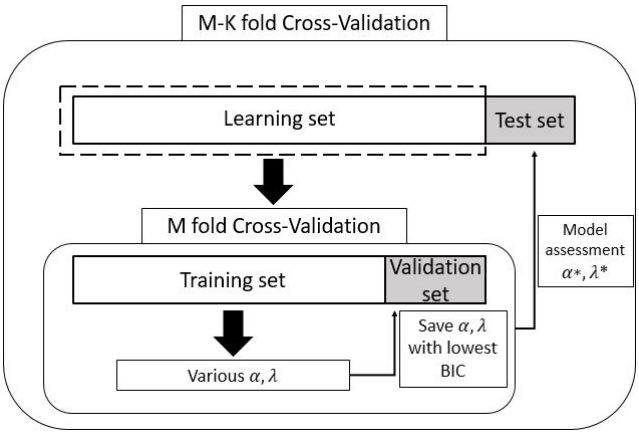

Figure 1.11: The nested CV strategy adopted for this work. The CV is applied twice: the inner M-fold CV is used for choosing the optimal model parameters (α^*, λ^*) and the external K-fold CV is used to evaluate the models on the test sets. This example uses the BIC as the selection criterion of the parameters α, λ at each validation set, but others can be used.

1.3.2.3 Bootstrap Methods

Besides CV method, bootstrap arises as a different approach that can be used to estimate prediction error. This method was essentially used throughout the work to compute an overfitting measure that is explained in the overfitting criteria section and an alternative strategy to considerably increase the sample size (resample) used under the scope of the nested model approach (described in section 2.2.2.1.3).

1.3.2.3.1 The general description

Firstly, bootstrap definition depends on the notion of a bootstrap sample. Assuming that there is a fitting method to the training set of the observed value $z_i = (x_i, y_i), i = 1, ..., N$, a bootstrap sample is defined as a random draw dataset of the training set (z_i^{*b}), each with the same size of the original training set. This index randomization is performed B times in order to obtain B bootstrap samples. A reasonable number is $B = 100$ (Hastie et al., 2001). Then, the bootstrap sample consists of members of the original training set, some appearing zeros times, some appearing one, some appearing twice and so on. It is called bootstrap replication, $S(Z^{*b})$, to the result of computing any quantity S over the bootstrap sample Z^{*b} (figure 1.12). One example is $S(Z)$ is the sample mean $\bar{Z}$ then $S(Z^{*b})$ is the mean of the bootstrap sample, $\bar{Z}^* = \sum_{i=1}^{N} \frac{Z_i^*}{N}$ (Hastie et al., 2001).

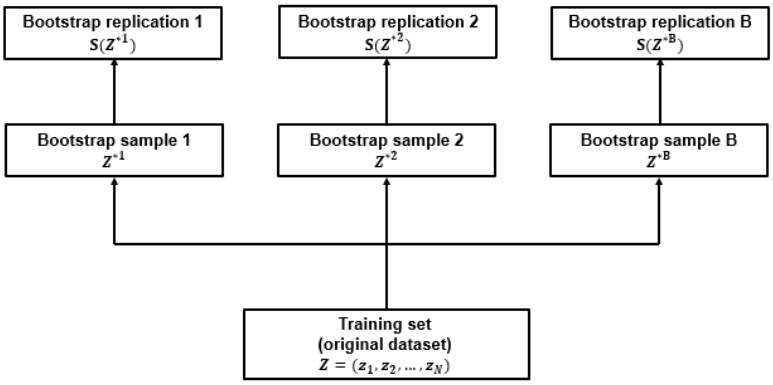

Figure 1.12: General schematic of the bootstrap method: the training set into the B bootstrap samples.

1.3.2.3.2 Bootstrap estimates of prediction error

There are two main variants of this method (a simple and a refined) for this purpose and both are going to be briefly explained hereafter.

Considering the simplest bootstrap approach, the main idea is to generate B bootstrap samples, perform the fitting on each one and by last apply the different fitted model on the original data. Therefore, the final prediction error will be the average of the B prediction errors, $\overline{err}(x^{*b}, \hat{f})$.

The improved bootstrap estimate takes into account what is called the "the optimism", which is the difference between the prediction error obtained by the simplest version and the prediction error when the model estimated at each bootstrap sample is applied to the bootstrap sample itself ("apparent error"), $\overline{err}(x^{*b}, \hat{f}^{*b})$. This error, since it results from training and testing on the same samples, its prediction is unrealistically low. Then, the "optimism" ($\overline{err}_{opt} = \overline{err}(x^{*b}, \hat{f}) - \overline{err}(x^{*b}, \hat{f}^{*b})$) comes as a bias correction that measures the amount by which the averaged residual squared rate decreases the true prediction error.

Finally, the improved prediction error is calculated by adding the optimism value to the average loss over the original training sample, also named "apparent error rate", ($\overline{err}(x, \hat{f})$ described in 1.30) as is defined in Bradley and Robert J. (1993). So, the following is obtained $\overline{err} + \overline{err}_{opt}$.

1.3.2.4 Criteria for Model Selection

1.3.2.4.1 Bayesian Information Criterion (BIC)

The BIC and AIC are two used methods to select between two models for which a likelihood can be calculated. Using the likelihood as a metric to directly compare the goodness-of-fit is inappropriate, because this would unjustly favor models that have many parameters. A compromise between the goodness-of-fit and the number of parameters should be considered in order to avoid overfitting. Both methods introduce a penalty term for the number of parameters in the model (Hastie et al., 2001).

The BIC is given (Aho et al., 2014):

$$BIC = ln(N)p - 2ln(\hat{L}) \tag{1.32}$$

It is aimed to maximize the likelihood of a set of parameters, θ, taking certain values based on the

observed data Y for a specific model M, which is defined as $L(\theta|Y,M) = p(Y|\theta,M)$. The $p(Y|\theta,M)$ is the total (joint) probability density. After maximization, the previous equation can be rewritten as $\hat{L} = L(\hat{\theta}|Y,M)$ (Kuha, 2004).

The penalization for the addition of parameters increases with the sample size, as can be noticed by the penalty term, $ln(N)p$.

The BIC expression can be changed when under the assumption that the residuals of the model, $\varepsilon = Y - \hat{Y}$, are distributed according to independent and identical normal distribution (with zero mean). The maximum likelihood estimate of the residuals' variance is $\hat{\sigma}^2 = RSS/N$. By changing $\hat{L}$ for $\hat{\sigma}^2$, after some simplification, the BIC is defined in terms of the RSS (Burnham and Anderson, 2004):

$$BIC = ln(N)p + Nln\left(\frac{RSS}{N}\right), \tag{1.33}$$

which becomes a simple expression to calculate in regression problems. The BIC is an increasing function of the RSS and an increasing function of the effective degrees of freedom, p. In other words, lower values of BIC implies either fewer variables, better fit or even both. Therefore, given two models, the one with lower value of BIC is the one to the preferred.

1.3.2.4.2 Akaike Information Criterion (AIC)

The AIC is defined as (Aho et al., 2014):

$$AIC = 2p - 2ln(\hat{L}) \tag{1.34}$$

The AIC generally penalizes the parameters less strongly than does BIC ($2p$ instead of $ln(N)p$ penalty), though it depends on the size of N. Under the Gaussian model and assuming that the variance, $\hat{\sigma^2}$, is known, the AIC's formula can be rewriten as done in BIC (Burnham and Anderson, 2004):

$$AIC = 2p + Nln\left(\frac{RSS}{N}\right) \tag{1.35}$$

As in BIC, the value of AIC increases with p and RSS, so a lower AIC is preferable during model comparison. As previously mentioned, the error criterion that was currently used throughput this work to estimate the similarity of two signals was the NMSE instead of the RSS. By analyzing the equations 1.33 and 1.35, the replacement of RSS by NMSE would lead to the addition of a constant (due to logarithmic properties) that only depends on the data points. Since the principal aim of these two approaches was to simply compare different models when applied on the same partition of the signal, the constant can be ignored.

1.3.2.4.3 Coefficient of Determination

The following quantity:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{Y})^2} \tag{1.36}$$

is called the coefficient of determination. Since the Total Sum-of-Squares, $TSS = \sum_{i=1}^{N}(y_i - \overline{Y})^2$, with $\overline{Y} = \frac{\sum_{i=1}^{N} y_i}{N}$, is a measure of the variability in Y without considering the effect of the regressor variable X and RSS (top part of the fraction) is a measure of the variability in Y remaining after the X has been considered, the R^2 is also called proportion of variation explained by the regressor X. Considering that $f(X)$ is a better estimation of Y than its own average, then $0 \leq RSS \leq TSS$, it follows that $0 \leq R^2 \leq 1$. It is aimed to have R^2 values as close to 1 as possible because it means that most of the variability in Y is explained by the regression model. The R^2 should be used with caution because by adding enough explanatory variables to model, it is possible to make R^2 larger, getting closer to what is called a "perfect fit" ($R^2 \simeq 1$). Therefore, this does not consider the overfitting problem.

This leads to the alternative approach which is called the adjusted R^2 (equation 1.37). The explanation of this is almost the same as R^2 however it penalizes the extra variables that are included in the model.

$$R^2_{adj} = 1 - (1 - R^2)\frac{N-1}{N-p-1} = 1 - \frac{N-1}{N-p-1}\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{Y})^2}, \tag{1.37}$$

where p is the number of explanatory variables in the model and N is the sample size, as before. The R^2_{adj} can be negative a given regressor does not explain additional variance to that it would expect by change and will always be $R^2_{adj} \leq R^2$. If a number of regressors is being added to the model one by one with a determined order of importance (set of nested regression models), with R^2_{adj} computed each time, the regression with the ideal combination of having best fit/number of explanatory variables will correspond to the maximum value of R^2_{adj} right before it starts decreasing afterwards (Montgomery et al., 2006).

Then, the percentage of Variance Explained (VE) is usually computed at each step and it is defined as the difference between the R^2_{adj} of two successive models, multiplied by 100:

$$VE(X_i) = 100 \times (R^2_{adj}(X_i) - R^2_{adj}(X_{i-1})), \tag{1.38}$$

where X_i is the model with one or more regressors than X_{i-1} (Bianciardi et al., 2009).

1.3.2.5 Overfitting criteria: Relative Overfitting Rate (ROR)

The ROR formula gives information about the amount of overfitting related to the prediction error that is obtained. This amount can be quantified because the result ranges from 0 if there is no overfitting to 1 if the prediction error is completely corroborated by it. The ROR is written as:

$$\hat{R} = \frac{\widehat{Err}^{(1)} - \overline{err}}{\widehat{\gamma} - \overline{err}}, \tag{1.39}$$

where $\widehat{Err}^{(1)}$ is the leave-one-out bootstrap estimate of prediction error, the $\overline{err}$ is the training error (1.30) and the $\widehat{\gamma}$ is defined as the *no-information error rate*. As mentioned before, one crucial difference between CV and bootstrap for assessing statistical accuracy is the fact that the first method explicitly uses non-overlapping data for training and test sets. On the other hand, the second method uses

bootstrap sample datasets as training set and the original training data as test set. Since these two sets may have samples in common, this way bootstrap method can lead better but unrealistic predictions. By following the CV principle, a better estimation of the prediction can be obtained when compared to bootstrapping. In order to avoid training and testing on the same samples, for each fixed observation, the prediction error is only track on bootstrap samples that did not contain that observation. Then, this different approach called leave-one-out bootstrap estimate of prediction error is specified by:

$$\widehat{Err}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, \hat{f}^{*b}(x_i)) \tag{1.40}$$

The N refers to the size of each bootstrap sample (same as the original training sample), C^{-i} refers to the indices of the bootstrap samples subset b that do not contain the observation indexed by i and $\left|C^{-i}\right|$ is the number of bootstrap samples under that criterion. L represents the loss function between the original y_i value and the predicted value at x_i, using the coefficients derived from the model fitting of the b^{th} bootstrap sample. The value of B (chosen number of bootstrap samples to be computed) should be large enough to avoid $\left|C^{-i}\right|$ equals 0, otherwise the terms for such observation i should be left out.

$$\hat{\gamma} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N} L(y_i, \hat{f}(x_{i'})) \tag{1.41}$$

By using the original training set, $\widehat{\gamma}$ is the average loss between all possible combinations of targets y_i and predictors $\hat{f}(x_i)$ values. Concerning the two different types of loss function addressed in this work, for the determination of both $\overline{err}$ and the $\widehat{Err}^{(1)}$ the sum of squared difference was used. In other words, the result of the ROR equals 0 if $\overline{err} = \widehat{Err}^{(1)}$ and equals 1 if the *no-information value* ($\hat{\gamma} - \overline{err}$) is the same as the overfitting (Hastie et al., 2001).

1.4 State-of-the-art

The first EEG-fMRI studies have used events of interest contained in the EEG temporal information to predict BOLD fluctuations. In most epilepsy studies, by EEG tracing, ictal and IED events can be identified. In order to find correlation between this two types of activity and BOLD changes, both are used as regressors of interest. The ictal activity can be further divided into three phases: early ictal EEG (before a seizure), clinical onset (during seizure) and late ictal EEG (after seizure), presenting a finer representation of the seizure dynamics. The inter-ictal spikes and ictal activity are described by stick functions (zero-duration events) and boxcar functions with the duration of a seizure, respectively (Leite et al., 2013). The Bagshaw et al. (2005); Thornton et al. (2010) are examples of studies that have used this strategy to estimate IED-related BOLD signal changes. As in Levan et al. (2010); Bénar et al. (2002), some features of the IED such as amplitude, energy or width, can be used to modulate the amplitude of the stick function. This has been referenced to provide improvements on the correlation with BOLD signal changes. Murta et al. (2016b) used intracranial EEG-fMRI recordings to investigate the effect of including the following features: amplitude, width or duration, slope of the rising phase, energy, or

spatial extent (number of contacts over which the sharp wave was observed) of the fast wave of the IED, into the model. This study observed that the width was the only one that explained a significant amount of additional variance relatively to the conventional IED regressor. This finding suggests that BOLD amplitude depends more on the duration of the correspondent field potential than the degree of neuronal activity synchrony linked to the amplitude of the sharp wave. Other EEG feature predictive of BOLD signal fluctuations of interest is the Event-Related Potential (ERP) amplitude and the response latency in a stimuli/task-based EEG-fMRI studies, as used in Debener et al. (2005).

The transfer functions mentioned until now only used the rich temporal profiles of EEG. However, in order to account for both temporal and spectral information based on the time-frequency decomposition of EEG, more complex transfer functions have been developed (Moosmann et al., 2003; Goldman et al., 2002; Laufs et al., 2006; Scheeringa et al., 2008), as presented next.

Several studies have used the EEG spectral components as predictors of BOLD fluctuations of interest, from simultaneous EEG-fMRI acquisition. Namely, the methods in Goldman et al. (2002) are equally applicable to study other electrical phenomena such as IED. They demonstrated that increased alpha power was correlated with decreased MRI signal in multiple regions of occipital, superior temporal, inferior frontal, and cingulate cortex, and with increased signal in the thalamus and insula. Moreover, de Munck et al. (2007) showed that a data-driven estimation of the HRF to the alpha power in specific voxel can improve the prediction of the BOLD signal taken from the exact same voxel.

The previous study have motivated the investigation of other frequency bands and address whether they may affect the correlation between alpha and BOLD. Furthermore, the heart beat variations and respiration are correlated to both alpha power and fMRI signal. Then models with more parameters were required. de Munck et al. (2009) showed that the inclusion of other EEG frequency bands as confounders in the fMRI-alpha correlation resulted in a large effect on the results, specially when the HRF used at each frequency band was extracted from the data. Moreover, the power fluctuations of the other frequency bands were mutually highly correlated. In order to determine the frequency of interest from the EEG-fMRI data, all frequency bands were added as regressors into a regression problem. However, the data in other similar study, Tyvaert et al. (2008), exhibited high inter-subject variability, suggesting models even more complex than a multi-frequency model convolved with the canonical HRF.

Mantini et al. (2007) also investigated the relationship between hemodynamic and electric oscillations during resting state from simultaneous EEG-fMRI, by adopting a different approach from the ones previously mentioned. They used ICA analysis to decompose the fMRI data and identified six different Resting State Networks (RSNs). Concerning EEG data, from the averaged spectrogram across the channels, the power spectral density was calculated across the five frequency bands ($\delta, \theta, \alpha, \beta$ and γ). The group searched the correlation between each time-course of the RSNs and the different neuronal rhythms. They found that each RSN was better characterized by a specific combination of the neuronal rhythms which they called a electrophysiological signature. Also noticed that the power fluctuations in different frequency bands were positively cross-correlated, suggesting that neuronal signals are dynamically coupled in different frequency bands. Moreover, each of the fMRI networks were generally associated with more than one neuronal rhythm, reinforcing that the brain display a coalescence

of rhythms instead of pure rhythms within distinct frequency bands.

Goense and Logothetis (2008) used simultaneous intra-cortical LFP-BOLD recordings to explore the relationship between different LFP frequency-bands and the hemodynamic fluctuations. They performed a multiple regression model in which the activity in different frequency-bands (covering the entire LFP range of frequencies) were used as BOLD predictors in alert behaving monkeys. Interestingly, they demonstrated that all bands explained a significant part of the BOLD variations, being preferentially sensitive to LFPs as opposed to APs.

Other types of transfer functions derived from the EEG power spectrum, have also been studied. In Wan et al. (2006), the correlation between the electric and vascular indices was assessed during the performance of a series of visual stimulation. Although both the ERPs and hemodynamic responses showed nonlinear behaviours and the neurovascular coupling mechanism is still not completely understood, this study indicated that the transfer function of the neurovascular coupling is likely a power transducer, which integrates the fast dynamics of neuronal activity into the vascular input of slow hemodynamics. This study motivated the use of TP of the EEG spectrum as BOLD predictors.

However, when compared to other power-weighted metrics, the RMSF of the EEG spectrum, which comes from the model proposed by Kilner et al. (2005), showed superiority in predicting BOLD signal in epilepsy (Leite et al., 2013). In Leite et al. (2013) was aimed to extract metrics from the EEG to predict fMRI data and also to identify brain areas associated with the EEG epileptic activity. They tested this method on both ictal and interictal EEG-fMRI from a patient with hypothalamic hamartoma (benign tumors found in the hypothalamus associated to seizures).

The mentioned Kilner et al. (2005) suggested that the increase in hemodynamic signals, should be associated with a loss of power in lower EEG frequencies, relative to higher frequencies. This statement was consistent with empirical observations of how changes in the EEG spectrum are hemodynamically expressed.

Rosa et al. (2010) also explored frequency-dependent and -independent transfer functions in healthy subjects, suggesting that changes in BOLD are indeed associated with changes in the spectral profile of neuronal activity and that these changes do not arise from one specific spectral band, in particular, from the relative power between high and low frequencies.

Murta et al. (2017) used not only a combination of the power of multiple EEG frequency bands (α, β and γ) but also the strength of the coupling between the phase of low- and the amplitude of high-frequency (PAC) EEG activities, to model the BOLD changes. By intracranial EEG recording when performing a motor task, they found that the phase amplitude coupling strength explained an additional BOLD variance in relation to the combination of the three frequency bands that were considered.

In general, the referenced metrics are intimately related with the EEG signal amplitude and therefore largely susceptible to artifacts, which may incorrectly lead to study correlations of non-neuronal origin with the BOLD signal. Moreover, since these are derived from one channel or a combination of channels (i.e. univariate metrics), do not take into account the amount of synchronization between signals measured at different locations at the scalp (Abreu et al., 2018a). A suggestion to overcome these problems is to use multivariate methods which capture spatial information from several EEG channels across the

scalp. Spatial correlation features and functional connectivity measures are examples of information extracted from these methods.

Regarding spatial correlation methods, the EEG microstates have been used as predictors of BOLD fluctuations, in resting-state studies. This is performed by spatially correlating the concurrent EEG topographies at each time point with the previously identified EEG microstates, which are EEG topographies that are stable during short periods of time ($\approx$ 80-120 ms). The EEG microstates can be derived by ICA, topographic time-frequency decomposition, or clustering. Britz et al. (2010) used a clustering approach to determine the four EEG topographies that dominated across all subjects. Then the time signal of these maps were convolved with a HRF in order to fit a linear model to the BOLD-fMRI responses, revealing four distinct distributed networks, that were correlated by the RSNs derived by ICA of the fMRI data.

Epilepsy-specific spatial templates from separate and long EEG recordings of epileptic activity was already used in Grouiller et al. (2010) study to provide a BOLD predictor by measuring the spatial correlation between these and the EEG scalp topographies at each time point.

Moreover, Vulliemoz et al. (2010) used continuous Electrical Source Imaging (cESI) to obtain a continuous estimate of the activity of an IED source, responsible for generating a given topography at the scalp. The cESI solution of a topography associated to epileptic activity can be used to predict the respective BOLD fluctuations.

The phase locking or synchronous oscillations in EEG have been extensively searched. The EEG oscillations not only occur in a wide frequency range but also have demonstrated to display a large amount of time and space synchronization, captured at different scalp sites. Since the BOLD-fMRI signal also displays synchronous oscillations across different brain regions, the use of functional connectivity methods to study the EEG-fMRI correlation has been sought. Synchronization has become a complementary aspect of brain electric activity and a potential mechanism for organizing brain function (Jann et al., 2009). There are several synchronization metrics used to construct predictors of epilepsy related BOLD fluctuations. Abreu et al. (2018b) comes as the first study proposing the use of the two synchronization metrics hereafter introduced, as an example of the great capacity of those in comparison with the metrics previously mentioned.

The absolute value of the Global Field Synchronization (GFS) does not offer the location of synchronized activity, rather an overview of the EEG synchronization signature, such as the hypersynchronization associated to epileptic activity. For a specific time window, over all channels across the scalp, the Fast Fourier Transform (FFT) is computed. For each frequency bin, the real-imaginary pair is determined and placed in a 2D-plot. The same is performed for all channels. At the end, the shape of the resulting cloud reflects common phase across channels. The range of values goes from 0 to 1, as the absolute difference of the two main directions of the cloud is higher (more elongated shape). The Kottlow et al. (2012) study used this metric to predict BOLD changes in a EEG-fMRI visual binding task.

By contrast, Phase Synchronization Index (PSI) provides information about the location of synchronized activity. To calculate the PSI between two signals (y_n and y_m) recorded at different channels, at first their instantaneous phase at a given time point t and frequency bin f must be determined

(using the Hilbert transform or Wavelet convolution). Then their difference of instantaneous phase, $\Delta\Phi_{nm}(t,f) = \Phi_n(t,f) - \Phi_m(t,f)$, is defined. As was done in Global Field Synchronization (GFS), the PSI at a time window j, $PSI_{nm}(j,f)$, is described as the mean vector length of the angular dispersion of the calculated phase difference. Namely, Mizuhara et al. (2005) study used this metric to map task-dependent BOLD changes for a specific frequency and channel pair.

In Abreu et al. (2018b), both synchronization metrics were computed across two frequency bands: a narrower frequency band characteristic of the epileptic activity (3-10 Hz) and a broadband (1-45 Hz). Then, the epileptic networks obtained by EEG-correlated fMRI analysis were compared with the results obtained by more common metrics such as the ones previously discussed: RMSF, TP and unitary regressors. Its results showed that mainly the average of the PSI within the narrower band, had the best performance comparing with the other metrics, exhibiting a greater ability to map epileptic networks.

Meir-Hasson et al. (2014) used a different approach, which is similar to the one developed in this thesis, by producing an EFP aimed to predict the fMRI activity at the amygdala. This study compared three approaches, all using Ridge: time/frequency representation, linear combination of frequencies with a pre-defined delay and the theta/alpha activity from the electrode. The three approaches were compared using the results from the electrode that retrieved the BOLD prediction with the highest correlation with the target signal (averaged BOLD signal) from the amygdala. Overall, the first model showed significantly better prediction results than the remaining ones. This study also demonstrated that the BOLD-fMRI signal can be better predicted without making any assumption on the HRF shape, but instead using different time-delays for different frequencies. Furthermore, they conclude that θ, α, β and γ frequency bands contributed all together to the prediction of the BOLD-fMRI signal. Meir-Hasson et al. (2014) had firstly tested it in a simpler problem regarding the activity of medial prefrontal cortex (mPFC), because previous fMRI studies demonstrated that Post-Traumatic Stress Disorder (PTSD) patients exhibited exaggerated amygdala responses and diminished mPFC responses to fearful facial expressions. They found that the best correlation with the amygdala was obtained with a different electrode and different weights of frequencies and time-delays than the best correlation with the dmPFC. This indicates that the amygdala and the dmPFC may derive different type and origin of electrical activity and the variability of the EFPs across the brain. The main goal of Meir-Hasson et al. (2014) was to reduce the need for fMRI scanning in regions such as the amygdala, by adopting NF procedures based on the determined general EFP. This study was the only one trying to estimate EEG features from a predetermined average BOLD signal from a region of interest, by means of fMRI-informed EEG strategy, which is used in this thesis but in the epilepsy scope.

2. Methods

2.1 Data characterization

The data from the patients under study was acquired during resting state, because the purpose was to analyze the ongoing spontaneous brain activity, in other words, the neuronal ensemble oscillations ("rhythms"), in particular those related with epilepsy, rather than averaged or induced brain activity.

2.1.1 Patient characterization

The following patients were selected from a group with drug-refractory focal epilepsy undergoing pre-surgical evaluation from the Program of Surgery for Epilepsy of the Hospital Center of West Lisbon. One of the criteria was to present clear inter-ictal (patients P2 to P5) or ictal (patient P1) epileptic activity recorded on the EEG.

Table 2.1: Characterization of the five patients studies and the respective EEG-fMRI datasets. For each patient, it includes the age, the number of datasets and the respective duration in minutes, the number of IEDs and the mean head motion in mm. It is also provided a clinical description of each patient.

Patient	Age	Dataset	Duration [min]	# IEDs	Mean head motion [mm]	Clinical Condition
P1	11	1	10	1 (seizure)	0.14	Childhood absence epilepsy (CAE), with IEDs restricted to the left hemisphere.
P2	9	1	10	596	0.04	CSWS[1], with IEDs over the left temporal lobe, and verbal agnosia (Wernicke type) and impaired ability to sustain attention.
		2	20	738	0.06	
P3	33	1	10	288	0.16	Continuous partial epilepsy, with large left-temporal cortical dysplasia, accompanied by continuous myoclonias of the right hand.
		2	10	342	0.15	
P4	27	1	10	15	0.19	Refractory focal epilepsy, with IEDs over the posterior occipital-temporal lobe, and frontal propagation.
		2	5	7	0.21	
P5	8	1	20	754	0.16	CSWS[1], with right neonatal thalamic hemorrhage and IEDs over the posterior right quadrant epileptogenic focus and frontal propagation.
		2	10	292	0.12	

[1]The CSWS refers to Continuous Spike Wave discharges in slow wave Sleep.

2.1.2 Data acquisition

2.1.2.1 fMRI acquisition

The fMRI data was acquired by a 3 Tesla Siemens Verio scanner (Siemens, Erlanger) with a 12-channel RF receive coil. A 2D multi-slice gradient-echo Echo-planar imaging (EPI) was the sequenced adopted to perform a whole brain functional images, during rest. Its parameters were the following: TR/TE=2500/30 ms, #slices=37 or 40 (interleaved acquisition) with voxel size=$3.5\times3.5\times3.0$ mm^3. Regarding the 1 mm isotropic whole-brain structural images, a 3D, T_1-weighted, gradient-echo MPRAGE sequence was used.

2.1.2.2 EEG acquisition

The EEG data was acquired by an MR-compatible 32-channel BrainAmp MR plus amplifier (Brain Products, Germany). The cap was a standard montage (10-20 system) and it included: the reference electrode and an Electrocardiogram (ECG) electrode placed on the back. The EEG acquisition had to be

synchronized with the fMRI scanner, and was sampled at 5 kHz. For posterior analysis, only some simultaneous EEG-fMRI runs from Abreu et al. (2018a), with duration of 10 and 20 min, were used (shown in table 2.1).

2.1.3 Data pre-processing

2.1.3.1 fMRI pre-processing and extraction of representative BOLD signal

As described in Abreu et al. (2018b), the following pre-processing steps were previously applied to the fMRI data prior to any other analysis. The fMRI data was analyzed using the functions from FMRIB's Software Library (FSL) `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki` and the main steps are summarized below. The first three volumes were discarded and the non-brain tissue was removed. Then, quasi-periodic and aperiodic blood flow pulsatility and respiratory-induced magnetic field changes were regressed from data. Subsequently was performed a motion and slice timing correction. Afterwards, six Motion Parameters (MPs) from the fMRI data were estimated. Lastly, it was applied a spatial smoothing using a Gaussian kernel with a full width at half-maximum (FWHM) of 5 mm and a high-pass temporal filtering with a cut-off period of 100 s. The BOLD epileptic networks were mapped based on the EEG PSI metric within an IED-specific frequency band. The metric choice was due to its better performance in comparison to commonly used EEG metrics. Then, the BOLD signal time courses were averaged across all the voxels within the identified epileptic network in each patient. The average BOLD time course was then used as the response variable throughout this work. After the mentioned pre-processing steps, the BOLD signal was upsampled to an intermediate sampling frequency of 4 Hz. This interpolation inserts zeros between the original data values and then applies a lowpass interpolating filter to the expanded sequence. Then, the signal was normalized: the mean was removed and was divided by the standard deviation (zero mean and one standard deviation).

2.1.3.2 Standard EEG pre-processing and additional clean-up steps

The EEG pre-processing steps were previously performed by routine functions from Matlab. The gradient and pulse artefacts are the two most important MR-induced EEG artefacts. The gradient artefact was corrected by the procedure explained in Allen et al. (2000). The pulse artefact (commonly referred as the ballistocardiogram artifact) is particularly challenging due to the temporal variability of the cardiac pulse. It was reduced according to the optimized strategy proposed by Abreu et al. (2016a). After both correction, the EEG data was downsampled to 250 Hz.

2.1.3.2.1 Additional EEG data clean-up using ICA

Additional EEG data clean-up was performed in this thesis using ICA. It was performed by the formulation explained in 1.2.2.5 with 31 ICs (the same number as the EEG channels). It was used at the final step to ascertain whether this extra clean-up would influence the performance of the entitled "best" model and method for BOLD prediction.

The supervised inspection of each Independent Component (IC) may become very time consuming

when dealing with a great amount of data, which may advent from large population studies or high-density EEG arrays. In order to cope with this problem and minimize subjectivity, automatic methods have been implemented. The Matlab based Automatic Pre-Processing pipeline (APP) that was used in this work is presented in Ramos et al. (2018). However, the authors of Ramos et al. (2018) recommend to use it only as an auxiliary tool when recordings have less than 32 channels. So after applying it, the selected ICs were visually inspected and others were further discarded (by being considered additional artefactual ICs), aided by a practical guide explained in Chaumon et al. (2015).

The APP steps that were based on the ICA consisted of detecting and removing of eye movement-, muscular-, and bad channel-related artefacts. The eye movement-related artefacts are usually detected based on the correlation with the recorded signals from the EOG (electrooculography) electrodes. Since those were not used, the APP was focused on helping the detection of the other two artefacts. In the example for Patient 1 shown in figure 2.1, the APP did not determine any outliers.

Furthermore, other ICs were classified as artefactual based on visual inspection of their topographies and time-courses. The blink components show a topography essentially flat at all except at frontal electrodes, which exhibit a large amplitude time-course and present no peak at physiological frequencies. On the other hand, the principal differences in relation to the horizontal eye movement components are: an opposite bilateral frontal topography (as can be seen by the ICs numbered by 1 and 2 in figure 2.1) and a step-like events time-course with lower amplitude. The muscle components exhibit a focal topography with a steady noisy time-courses and power at high frequencies. Finally, the bad channel components have an even more focal topography constrained to an unique channel with a noisy time course and a high correlation with the marked bad channel. On the other hand, the neuronal components are evidenced by power peak at the physiological frequencies and present a smooth/dipolar topography (Chaumon et al., 2015).

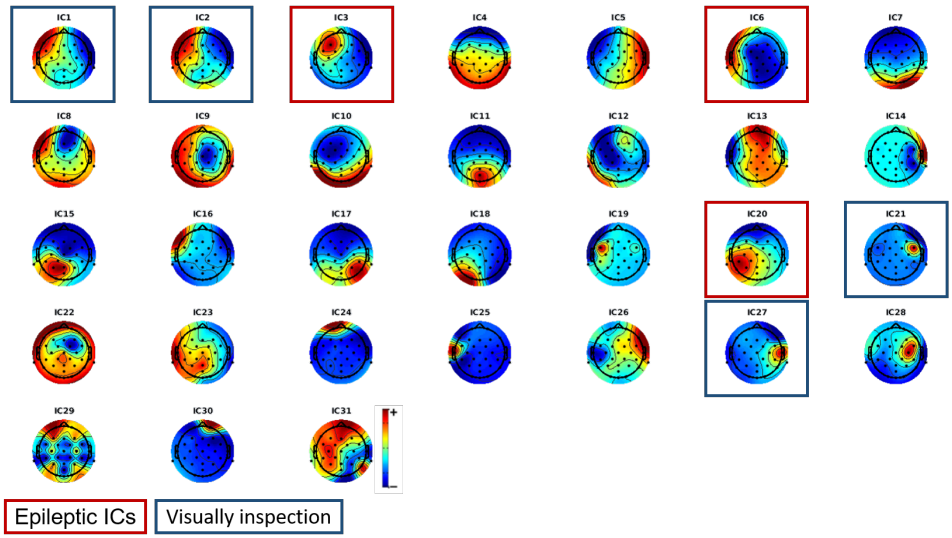

Figure 2.1: The 31 ICs shown for Patient 1, derived from the ICA method. The epilepsy-related ICs determined by the previous work in Abreu et al. (2016b) are highlighted by the red squares around the associated topographies. The ones highlighted by the blue squares were considered artifactual ICs by visual inspection (details in Chaumon et al. (2015)).

Although some topographies complied with the requirements, if the focus was near/similar to the ones shown in the epilepsy-related topographies, the correspondent ICs were not rejected.

2.1.3.2.2 Regression of motion parameters (MPs)

The R^2_{adj} was the chosen criterion to measure the amount of variance of the EEG metric explained by the six MPs (their extraction is referenced in the fMRI pre-processing). In other words, it evaluates the specificity of the EEG metrics to predict the epilespsy-related BOLD changes, rather than motion-related.

Except for Patient 1 (just one dataset), the MPs of both datasets were concatenated after normalization. Then, the six MPs were upsampled to a sampling frequency of 4Hz.

According to equation 1.37, p would be the six MPs, N the sample size after resampling, y_i the value of the EEG regressor at sample i, $\overline{Y}$ the mean of the EEG regressor and $\hat{y}_i = x_i\beta$ would be the predicted value, where x_i represents the value at sample i of the matrix X whose columns are the MPs and β denotes the vector of parameters from the LS. The $\varepsilon = y_i - \hat{y}_i$ represents the EEG regressors that are no longer explained by the MPs.

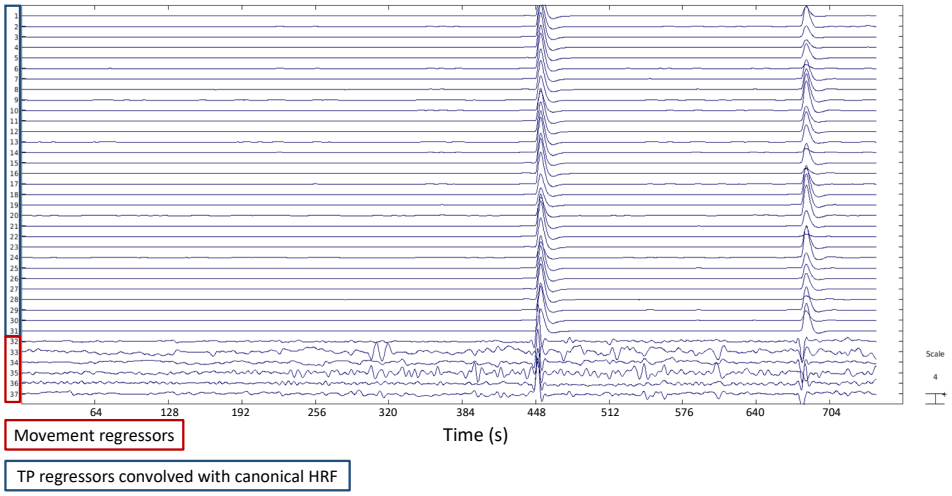

Figure 2.2: The influence of the MPs on the EEG predictors, namely in the TP metric. In the y-axis, the top 31 signals correspond to the TP predictors only convolved with the canonical HRF and the bottom 6 correspond to the MPs. The resemblance of the signals is clearly observable. Since for Patient 1, this similarity was not evident, this figure is representative of Patient 4.

This additional correction was not performed on all EEG models (later described in table 2.2). In each method's description, the models that were submitted to this correction are mentioned.

2.2 Models Estimation and Evaluation

2.2.1 Models of fMRI-based EEG

This section describes the post-processing steps to construct the feature/predictor space from the EEG data before the integration through linear regression.

2.2.1.1 EEG Time-Frequency Decomposition

Apart from Patient 1, the EEG data of the two sessions were concatenated in the time dimension. In order to prevent any discontinuity, a low-pass filter with a cut-off frequency of 250 Hz (equal to the sampling frequency) was after applied.

Each row of the EEG time series matrix was decomposed into time-frequency (Time-Frequency decomposition-TF) using Morlet wavelets. The Morlet wavelet, $w(t, f)$, with t and f representing time and frequency, respectively, is defined as:

$$w(t, f) = (\sigma_t \sqrt{\pi})^{-1/2} exp\left(-\frac{t^2}{2\sigma_t^2}\right) exp(i2\pi f t), \tag{2.1}$$

where $\sigma_t = \frac{1}{2\pi\sigma_f}$ is the time sampling resolution, $\sigma_f = \frac{f}{R}$ is the frequency sampling resolution with a wavelet factor $R = 7$. The $w(t, f)$ is then convolved with each of the EEG time series, $y(t)$. They are decomposed into time-varying power, $P(f, t) = |w(t, f) * y(t)|^2$, within 100 frequency bins whose values are logarithmically distributed from 1 Hz to 45 Hz. The transformed signals constitute a 3D power matrix of $[100$ frequency bins $\times N$ time points $\times 31$ channels$]$.

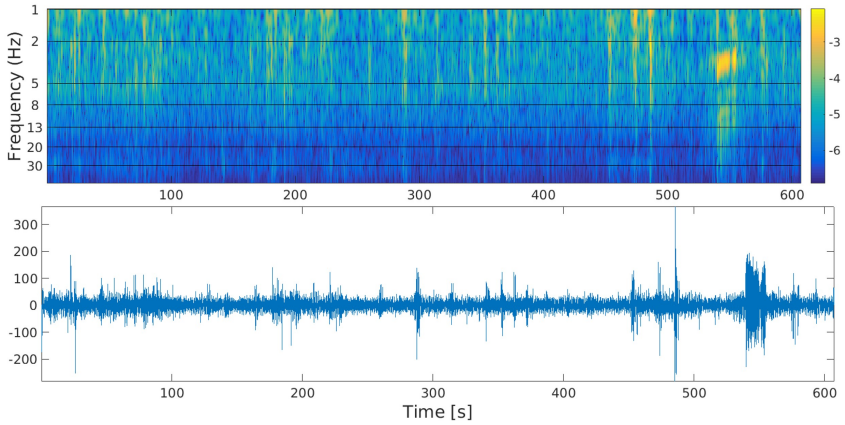

Figure 2.3: Example of the time-frequency spectogram plotted next to the respective time-series. Time-frequency decomposition (upper figure) and the time series (lower figure) of the EEG signal measured at channel 3, from patient 1.

The next step depends upon the type of metric to be applied. As described in section 1.2.4.1, three metrics were applied in this work by the following order: LC, RMSF and TP.

In the case of LC metric, the spectrogram was divided into 5 frequency bands described in section of the EEG rhythms: delta (δ), theta (θ), alpha (α), beta (β) and gamma (γ) with the respective frequency bands of 1-4 Hz, 4-8 Hz, 8-13 Hz, 13-30 Hz and 30-45 Hz. Then, each power band was averaged in order to have one single signal for each frequency band over time. Therefore, five frequency bands were associated to each channel. In the case of RMSF metric, equation 1.8 was applied, where for each time point, an weighted average of the power across all frequency bins was obtained. Therefore, just one RMSF signal was associated to each channel. In the case of TP metric, equation 1.7 was applied, where each time point was obtained by the sum of the power values across all frequency bins. Therefore, just one TP signal was associated to each channel.

Then, for each metric, all regressors were convolved with a number of different HRFs in order to account for the different time-delay between the two imaging modalities and the potential variability of the hemodynamic response between electrodes, frequencies (in case of the LC metric) and subjects.

2.2.1.2 HRF convolution

As previously mentioned, each of the regressors in the model was uniquely convolved with the canonical HRF or with the canonical plus five more variants hereinafter described.

The six HRFs were derived using the Statistical Parametric Mapping (SPM) `https://www.fil.ion.ucl.ac.uk/spm/` function (named spm_hrf) implemented in Matlab. In this function, the four parameters (p_1 to p_4) presented in equation 1.5 were varied in order to create the six double-gamma functions. The principal aim was to obtain six HRFs whose overshoot peaked at $[10, 8, 6, 5, 4, 2]$ sec (where the forth corresponds to the canonical HRF), by maintaining a linear relation between the parameters' values of the five variants HRFs and the canonical HRF (e.g., if $p_1(5) = 6$, then $p_1(10) = 6 * 10/5$ and then successively for the remaining parameters). Furthermore, the temporal and dispersion derivatives of the canonical HRF (depicted by the dashed red lines in figure 2.4) were also designed by the same SPM function. The integration of these is a popular alternative to increase the flexibility of the canonical HRF because they can model small differences in the latency and the duration of the peak response, respectively. They were calculated based on the finite difference, $\frac{d(HRF)}{dp} \sim \frac{HRF(p+a)-HRF(p)}{a}$, with a temporal step of 1 sec, $p = p_6$ (the onset in sec) and $a = 1$, and a dispersion step of 0.01, $p = p_3$ and $a = 0.01$.

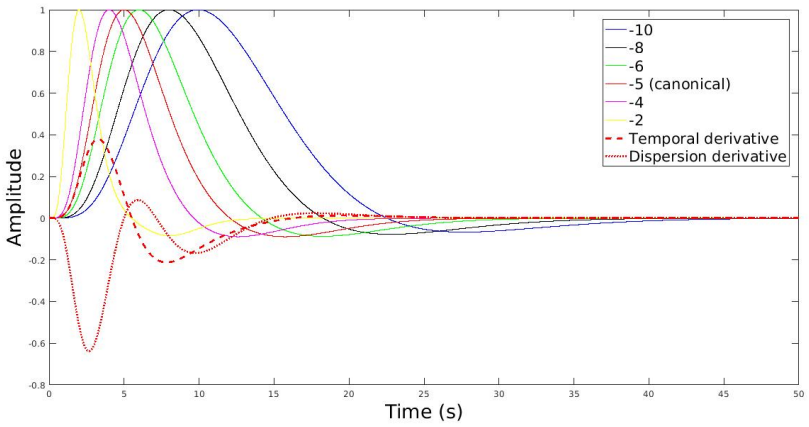

Figure 2.4: The different HRFs to be convolved with each regressor. The canonical HRF with the peak at 5 sec, represented by a red line, and the remaining five HRFs peaking at 10,8,6,4 and 2 sec. The temporal and dispersion derivatives of the canonical HRF are also represented by the dashed red lines.

The feature dimension of the LC metric can be decomposed into channels (CH), the 5 frequency bands (FQ), different delays of the HRF ($DELAY$) and time points ($TIME$), as represented:

$$[CH] \times [FQ] \times [DELAY] \times [TIME] \tag{2.2}$$

Regarding the LC metric, if each regressor was convolved with the six HRFs, $5[FQ] \times 6[DELAY] = 30$

predictors were constructed for each channel. Therefore if the analysis includes the information from the 31 channels, the number of predictors increases to $30 \times 31 = 930$ in total.

In relation to RMSF and TP, $1[TP/RMSF] \times 6[DELAY] = 6$ (example in figure 2.5) predictors per channel. However, when the 31 channels are considered, the number of predictors increases to $6 \times 31 = 186$. When the problem was reduced to the canonical HRF $1[DELAY]$, $1/6$ of the total number of predictors are integrated through linear regression.

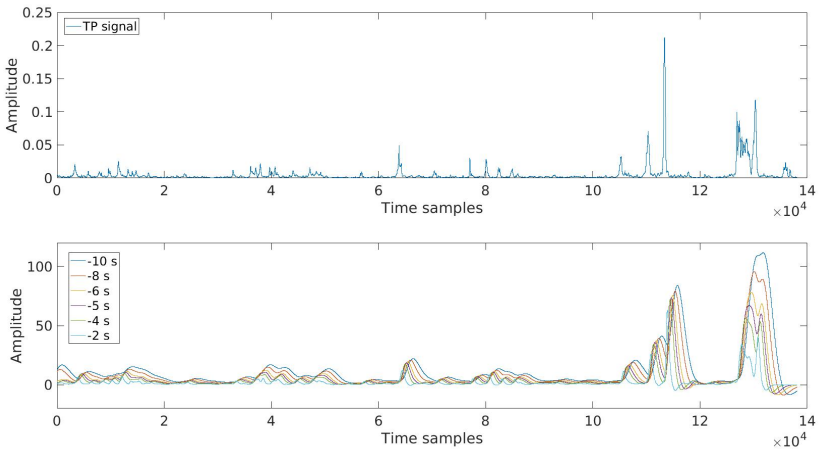

Figure 2.5: The effect of the convolution on the power mean signal from channel 3 of Patient 1. The upper figure is the power mean signal before convolution and the lower figure is the result of convolving the previous signal with the six HRFs in figure 2.4, with 6 different delays (indicated in the legend, in sec).

The convolved EEG signals were downsampled to the intermediate frequency of 4 Hz and normalized (zero mean and one standard deviation). This step is also important to facilitate the interpretability of the prediction weights. For a simpler notation, the post-processed BOLD signal was denominated as Y signal and the matrix of regressors whose rows and columns are the observations and features, respectively, as the X matrix.

The metrics were divided into six main models summarized in table 2.2, depending on whether each signal (LC, TP or RMSF signals) was convolved with the six HRFs or only with the canonical HRF. Not all models were used in all the steps of the work. Therefore, at each part, the models that were used are stated.

2.2.2 Models Evaluation

In this section different regression methods were applied for the estimation of the EFPs and determination of the prediction performance with the principal aim of finding which method and EEG model best predicts the BOLD signal, according to some selection criteria. Although it was not constantly mentioned throughout the work, each of the steps described in this chapter was always applied to each of the five patients separately.

This section was divided into two main parts, where in the first part the regressors from all channels were introduced in the regression method. However, the results from the first part motivated the formulation of a second part where a separate model was estimated for each channel. Still in the first part, a

Table 2.2: The six models of the EEG metrics used throughout the work. The three metrics were divided into six models, depending on whether each metric (LC, TP or RMSF signals) was convolved with the six HRFs or only with the canonical HRF.

Models	**HRFs**	
	Six delays (10,8,6,5,4 and 2 sec)	Canonical HRF (delay 5 sec)
LC metric	LC_{HRFs}	LC_{canHRF}
TP metric	TP_{HRFs}	TP_{canHRF}
RMSF metric	$RMSF_{HRFs}$	$RMSF_{canHRF}$

nested model approach was developed in order to determine a reduced subset of regressors for each dimension of the features' space.

Every time a specific model was used in a specific regression method, two main steps were performed: a CV procedure to optimize the method's parameters (when applicable) and estimate the prediction error and then model fitting using the original data to obtain the final estimated EFP. For each of these steps, the following performance measures were determined: the fitting error using NMSE (equation 1.31), the BIC (equation 1.33), the adjusted coefficient of determination (R^2_{adj}, in equation 1.37) and the DOF. However, in the case of the CV, all measures were averaged across the test sets and the fitting error was replaced by the average prediction error. Although the AIC had been calculated in the beginning of the work (by formula in 1.35), it was left behind because it penalizes less the number of predictors than BIC. Since the large number of predictors was a concerning issue, more attention was given to BIC, aiming to prevent the overfitting problem.

2.2.2.1 Models with all channels

In this section, the models LC_{HRFs}, $RMSF_{HRFs}$ and TP_{HRFs} were the first being studied by the methods subsequently introduced. Afterwards, the methods were compared and SR came out as the one that retrieved the best prediction results, in most of the patients. Therefore, SR was selected for conducting further tests, as studying the effect of only convolving the signals with the canonical HRF, models TP_{canHRF} and $RMSF_{HRFs}$, among others. Since LS is the less time-consuming and the simplest method (namely regarding the estimation of the DOF), it was always used for comparison purposes. The LC_{canHRF} was only used in the next section.

2.2.2.1.1 Ridge and Lasso

The Elastic Net Regularization formulated in equation 1.27 has two different parameters, α and λ, that control the type and the strength of regularization, respectively. In order to determine the EFP of the model, the two parameters must be first determined. Trying to use a large family of parameters becomes unfeasible, therefore the number of combinations of the two parameters had to be shrunk and the optimal solution among the options was selected based on selection criteria.

The Elastic Net Regularization procedure was mainly divided into two parts: the hybrid and the non-hybrid approaches. In the first part, one combination of parameters was selected among all subset and in the second part, each α was studied separately.

Considering the α parameter, five values were evaluated $\alpha = [0, 0.25, 0.5, 0.75, 1]$. The three values

in the middle correspond to the combination of the two types of regularization (Ridge and Lasso) and the $\alpha = 1$ to the pure Lasso regularization. The *lasso* function from Matlab was used for all values of α except when $\alpha = 0$. Since it was not possible to use a pure Ridge regularization ($\alpha = 0$) with the *lasso* function in Matlab, the *ridge* function was used instead every time $\alpha = 0$ was selected in the implementation.

The data inputs used for determining the range of λ values were the X matrix and Y target signal but with a shuffle on the observation's dimension. Whenever the CV procedure was performed, a shuffle of the observations must be always done before attribution to the different sets, by *randompartition* function from Matlab.

Afterwards, the λ range was differently set for each value of α. The *lasso* function calculates the largest value of λ that gives a non-null model (at least one DOF). Then, the lower limit was determined by establishing a ratio value between the smallest and the largest value of the sequence. The default value of this parameter was not changed, $1e^4$.

The range of Ridge regularization parameters was determined using the singular values, obtained from the SVD of the X matrix. Its range was bounded between the minimal and the squared maximal singular values. By examining equation 1.23, when $\lambda = d_j$, with j corresponding to the minimal value, the DOF will tend to $df(\lambda) = p$ and when $\lambda = d_j^2$, with j corresponding to the maximum value, the DOF will tend to $df(\lambda) = 1$ (no regularization). For computational efficiency, 20 logarithmically spaced λ values between the upper and lower values, for each α, were generated. To conclude, $5\alpha \times 20\lambda = 100$ different combinations were possible.

- **Hybrid approach**

In order to select the best combination, a nested M-K-fold CV was applied as it was described in figure 1.11, with M=K=10 folds. For each fixed test and validation sets, the model was estimated for all the 100 combinations of parameters in the training set by applying the *lasso* and *ridge* (for $\alpha = 0$) functions from Matlab. Then the chosen combination was the one that yielded the lowest average BIC value on the validation set. Hereinafter, the optimal combination for a specific test set was selected based on the number of occurrences: the combination that showed up more across the M times, for a given k. Depending on the determined parameters, the model was fitted in the learning set using again *lasso* or *ridge* functions and the prediction error (using the NMSE) was determined in the test set. The average prediction error was the result of the average of the values across the K times.

- **Non-hybrid approach**

The only difference comparing with the hybrid approach was that the average prediction error was determined for a fixed α. Therefore, only $1\alpha \times 20\lambda = 20$ combinations was evaluated for a fixed test and validation sets. However, this analysis was performed for each α.

The way that the number of DOF was calculated depended on the type of regularization in question: by the number of non-zero parameters in the case of Lasso regularization or by equation 1.22 in the Ridge and Elastic Net Regularization cases. More details about this is explained in Ridge and Elastic

Net regularization in the Introduction (1.3.1.3.1 and 1.3.1.3.3). The nested CV scheme is only necessary when parameter/s need to be optimized.

Subsequently, a specific type of regularization must be chosen in order to apply it later on each patient's data. The selected α was the most frequent in the hybrid approach and the selected λ was the most frequent in the non-hybrid approach for that specific α.

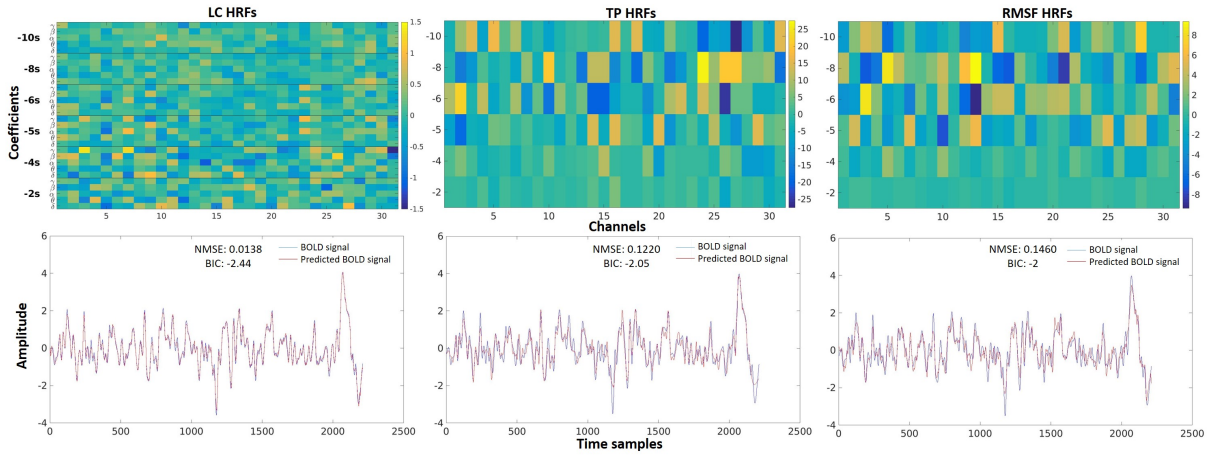

Figure 2.6: The figures above show the estimated model coefficients that best predict the BOLD activity (EFP) using Ridge. The predicted signal is shown (red color in the figures below). The original BOLD signal is pictured in blue. The first, second and third columns of figures show the result of using the LC_{HRFs}, TP_{HRFs} and RMSF_{HRFs} predictors respectively, derived from Patient 1. The values of BIC ($\times 10^4$) and error resulted from the fitting are also represented.

The EFP represented in figure 2.6, β_j, with $j = 1, \ldots, p$ indicate relevant frequencies and time delays of the EEG activity. In the exemplified models, the first column of the figure, p is equal to 930 and in the second and third columns, p is equal to 186 (shown by the equation 2.2).

2.2.2.1.2 Least Angle Regression (LAR)

The LAR was performed by the *lar* function of the *SpaSM tootbox* from Matlab, which can be found in `https://www.jstatsoft.org/article/view/v084i10` (Sjöstrand et al., 2018). Its implementation is described in Efron et al. (2004). This function returns a structure that contains various useful measures, namely the goodness-of-fit estimates BIC across the entire solutions' path, starting from zero active variables to the ordinary least squares solution. It also includes the shrinkage value, s, defined as the L_1 size of the coefficients, β, at each step in the range $[0, 1]$.

Since this implementation allows to find the optimal solution (the step that offers the minimal value of BIC) for the input variables, it was just necessary to design an external K-fold CV, with $K = 10$, in order to assess the prediction error. This means that for each test set, the *lar* was computed on the learning set. As done in the Elastic Net Regularization method, besides the prediction error, the other measures were also determined in the test sets and subsequently averaged. As before, the number of DOF at each step of the LAR algorithm was well approximated by the number of non-zero elements of β. The estimated coefficients and the predicted signals by using the three models from Patient 1, are represented in figure 2.7.

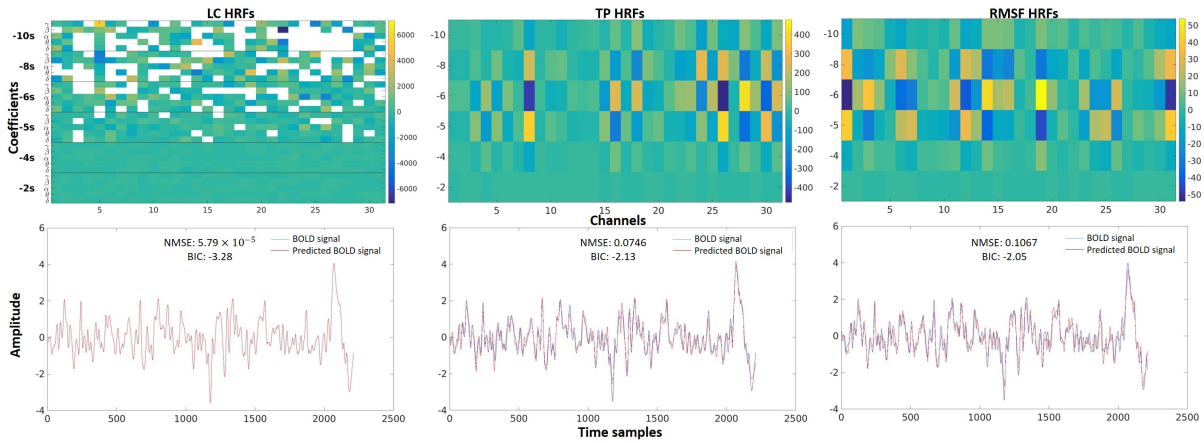

Figure 2.7: The figures above show estimated model coefficients that best predict the BOLDactivity (EFP) derived from applying the LAR. The figures below show the BOLD signal and the respective predictions by using the LC$_{HRFs}$ (first column), TP$_{HRFs}$ (second column) and RMSF$_{HRFs}$ (third column) predictors, derived from Patient 1. The values of BIC ($\times 10^4$) and error resulted from the fitting are also represented. The white squares represent the coefficients whose regressors were not included in the optimal solution determined by BIC criterion.

2.2.2.1.3 Least-squares Regression (LS)

The LS is the simplest regression method and due to this reason, it was used for comparison purposes.

The *glmfit* function from Matlab was used for model estimation (solution of equation 1.15) using the responses in Y and the predictors in X. The distribution of the data was assumed to be normal, therefore the required *distr* parameter on the function was set to *normal*.

Since no parameters had to optimized, a K-fold cross-validation with $K = 10$ was the solution for obtaining the average of the mentioned measures. The number of the DOF only depends on the number of columns of the X matrix, p, and not on the fold where the regression is applied. The same was performed on the patients' data as figured in 2.6 and 2.7.

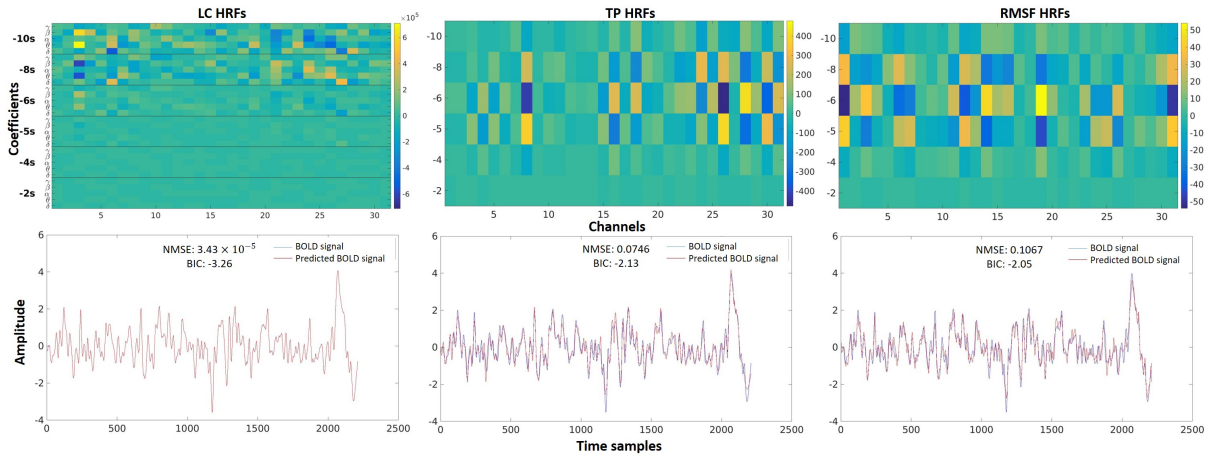

Figure 2.8: The figures above show the estimated model coefficients that best predict the BOLD activity (EFP) using LS. The figures below show the BOLD signal and the respective predictions by using the LC$_{HRFs}$ (first column), TP$_{HRFs}$ (second column) and RMSF$_{HRFs}$ (third column) predictors, derived from Patient 1. The values of BIC ($\times 10^4$) and error resulted from the fitting are also represented.

Nested model approach

The main purpose of this approach was to assess a minimum number of regressors for each dimen-

sion of the feature space (combination of delay/frequency band, included) required to explain the data. This study was performed on the LC$_{HRFs}$, TP$_{HRFs}$ and RMSF$_{HRFs}$ models. Considering as example the LC metric and the channels' dimension and assuming $c = 1, \ldots, C$, with $C = 31 = \#channels$, the path was constructed as follows:

1. At the first step, $c = 1$, the regressors from each channel (frequency bands and delays), were used by LS to fit the model to the response Y. For each of the C times, the NMSE was determined. At the final, the channel that yielded the lowest NMSE was selected and incorporated into the nested model.
2. Subsequently, $c = 2$, the best combination of two channels was determined by using the same method and selection criteria. By keeping fixed the channel selected in the first step and varying the remaining 30 channels. In other words, in the step c, $C - c + 1$ channels/combinations were tested.
3. The search continues until reaches $c = C - 1$, which means that just one channel is still outside the model.
4. After the integration of the last channel, a LS fitting is applied to the full model (all channels).

Every time a channel is added to the model some measures were computed, namely: the NMSE, the BIC and the R^2_{adj}. Therefore, by subtracting the R^2_{adj} values from two consecutive steps, at the end $C-1$ VE values were determined.

After applying this approach to the original patient's data, two different procedures were used. In order to perform statistical tests on the VE values, CV method with K=10 and later bootstrap method with K=100 were designed to increase its sampling distribution. First, the original data was divided into ten folds roughly equal sized, and in each fold the described nested model was again applied. However, the VE matrix, with a dimension of $[C - 1, 10]$, was not enough to approximate a normal distribution. Then, 100 bootstrap samples were randomly derived from the original data and the VE matrix, with a dimension of $[C - 1, 100]$ was obtained. In this approach, these two procedures were not used to determine the prediction error as done before, but instead to increase the sample data under study.

Then, a one-sided *t*-test, $t = \frac{\bar{x}-\mu}{s-\sqrt{n}}$, was applied for each row of the VE matrix, with $\bar{x}$ the sample mean, μ the hypothesized population mean, s the sample standard deviation and n the sample size. The *ttest* function from Matlab was used for this task. It returns a test decision (and the respective p-values) for the null hypothesis that the data in VE comes from a normal distribution with mean equal to zero and unknown variance, against the alternative that the mean is greater, at a 5% significance level.

The figures 2.9 and 2.10 represent the result of first step of the nested model. The left figure in 2.9 denotes the application of the nested approach on the channels' dimension. On the other hand, in the right figure, a specific combination of delay/frequency band across all channels was used to fit the model at each time.

2.2.2.1.4 Stepwise Regression (SR)

The *stepwisefit* function from Matlab was used to carry out the SR algorithm. This implementation is a

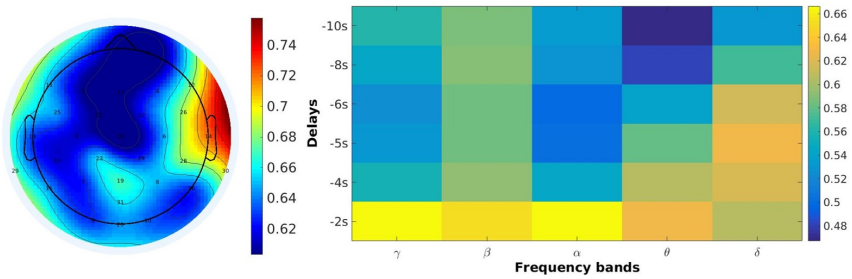

Figure 2.9: Result of the first step of the nested model approach on the LC$_{HRFs}$ model from Patient 1. The color scale represent the training error. In the left figure, the predictors from each channel were used to fit the target BOLD signal, whereas on the right figure, the predictors from each combination of delay/frequency band were used on fitting. The channel 18 and delay/frequency band, $-10s/\theta$, were selected to proceed to the second step of the approach.

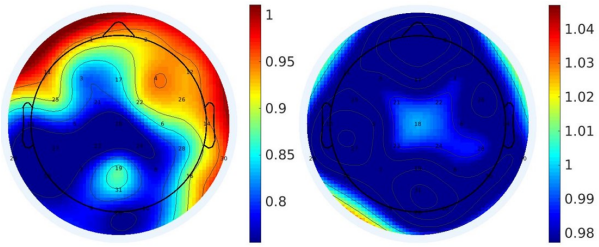

Figure 2.10: Result of the first step of the nested model approach on the TP$_{HRFs}$ (left figure) and RMSF$_{HRFs}$ (right figure) models from Patient 1. The channel 18 and the channel 2 gave the best results by using the TP and RMSF metrics, respectively.

hybrid stepwise-selection which considers both forward and backward moves at each time and selects the "best" of the two (more details in 1.3.1.2). Among all the parameters that this function returns, more attention was given to: the DOF, the predictors and the correspondent coefficients at each step and in the final model. The default entrance/exit p-values were left equal 0.05/0.1. Moreover, the fact that the exit p-value is higher is also suggested by Draper and Smith (1998) in order to provide some "protection" for predictors already admitted to the equation.

The same measures calculated in the previous methods, were also determined by a K-fold cross-validation, with $K = 10$, and then on the patient's whole data.

Besides the common procedures, a further analysis was performed in the case of the SR for all patients and models mentioned before, in order to better characterize the obtained EFP. For each dimension of the feature space was pointed out the channel/s, frequency band/s, HRF-delay/s and combination/s of delay/frequency band which had the largest coefficient's mean. Moreover, along each dimension, the percentage of non-zero coefficients was also assessed. In the case of the TP$_{HRFs}$ and RMSF$_{HRFs}$, this task was simplified since just two dimensions (HRF-delay and channel) were taken into account. This attempt originated a variety of results for each of the features across patients and models, leading to difficulties of interpretation and no direct link to the patients' condition, described in table 2.1, could be established . For this reason, the results are not presented in the Results chapter.

Afterwards, the predictive performance of all fitting methods were analyzed and compared across the five patients (3.1 to 3.5) and their differences were computed by statistical tests. As already mentioned, the associated results motivated the use of the LS and SR approaches on further tests, described below.

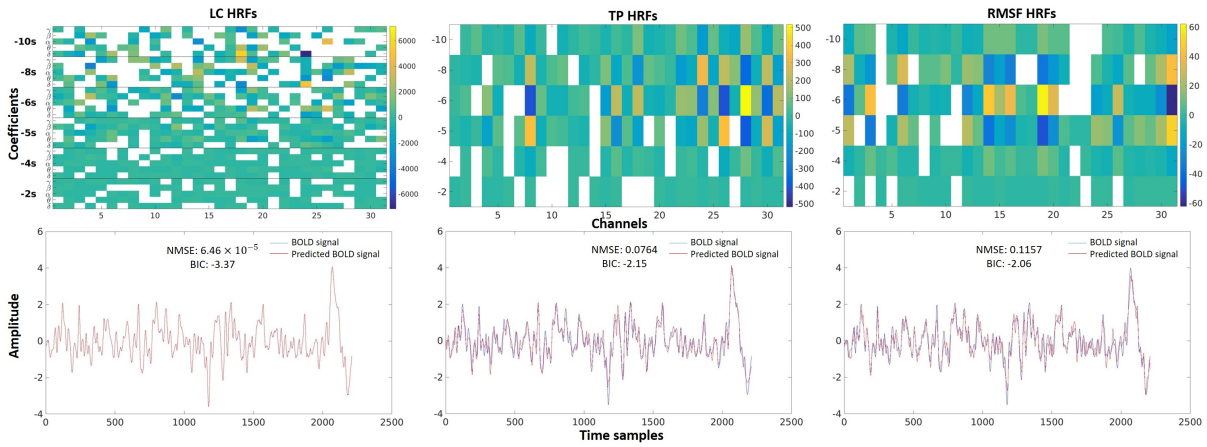

Figure 2.11: The figures above show the estimated model coefficients that best predict the BOLD activity (EFP) using SR. The figures below show the BOLD signal and the respective predictions by using the LC$_{HRFs}$ (first column), TP$_{HRFs}$ (second column) and RMSF$_{HRFs}$ (third column) predictors, derived from Patient 1. The values of BIC ($\times 10^4$) and error resulted from the fitting are also represented. The white squares represent the coefficients whose regressors were not included in the optimal solution determined by BIC criterion.

2.2.2.1.5 Further tests using LS and SR

The models TP$_{canHRF}$ and RMSF$_{canHRF}$ were used to fit the BOLD signal by LS and SR. Each of the models include 31 regressors (one for each channel), thus the 31 estimated coefficients can be observed by means of a topography (example in figure 2.12). Like has been done before, the average measures were estimated for the two models and methods.

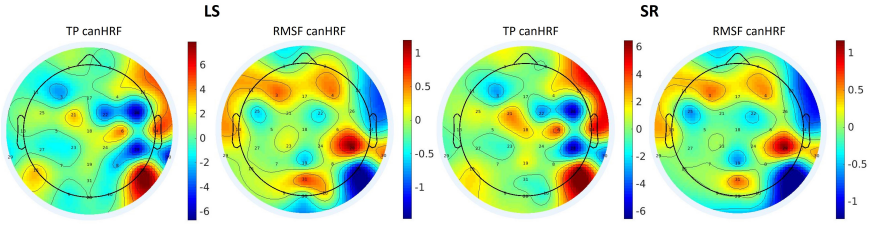

Figure 2.12: Topography of the 31 coefficients estimated by LS (the two first images) SR (the two last images) for Patient 1. For each method, the model TP$_{canHRF}$ and RMSF$_{canHRF}$ are represented in the left and right figures, respectively. In the SR scope, it performs an interpolation of the values across channels: the TP$_{canHRF}$ model has 24 non-zero coefficients and the RMSF$_{canHRF}$ model has 21.

Next, the ROR formula (equation 1.39) was applied to three models: TP$_{canHRF}$, TP$_{HRFs}$ and LC$_{HRFs}$ and emerged as other test to reinforce the results from BIC and R^2_{adj}. The TP was chosen instead of the RMSF due to the better predictive performance. From the first to the second models, the variability of the HRF was introduced into the model and from the second to the third, the power was discriminated across different frequency bands. This measure allows to estimate how the addition of the predictors influenced the overfitting problem.

Afterwards, the TP$_{canHRF}$ and RMSF$_{canHRF}$ were the models selected to perform some control tests because, by having one predictor per channel, the interpretability of the results becomes easier. The regressor that exhibited the highest correlation with the BOLD signal was kept and the remaining 30 regressors were generated by performing a phase randomization of the first. The Pearson correlation was calculated by using the *corr* function from Matlab. This procedure creates a surrogate data with the

same correlation properties as the original signal. One performs a Fourier transform on the original time series, preserving the Fourier amplitudes but randomizing the Fourier phases. Finally, one performs an inverse Fourier transform to create surrogate data. A table with the measures obtained from CV procedure is exhibited in Appendix C. Since it is a random process, this was repeated ten times. Therefore, the final values were the average resulted from CV over ten repetitions.

Subsequently, the MPs were regressed out from the models TP_{canHRF} and RMSF_{canHRF} and the results with this additional clean-up were compared with the ones previously obtained.

2.2.2.2 Models with each channel individually

The difficulty of finding an EFP's pattern along a specific dimension of the feature space for each patient, as can be noticed by the figures 2.6 and 2.7, called for a simplification of the problem. Moreover, by fitting a model with the predictor/s from all channels at the same time, the correspondent channel's coefficient/s could not be directly associated to a better or a worse prediction power of the BOLD signal. The resulted pattern just allows to interpret in the light of its contribution in the context of the combination of all predictors. This motivated the following study, which uses each channel individually to predict the BOLD signal and enabling to reduce one feature dimension to the EFP. Thus, the channels can be ordered according to their prediction performance.

By just considering each channel individually, the maximum number of associated DOF across methods and models was considerably reduced. For example, the LC_{HRFs} model, which is the one with the highest number of predictors, would have a maximum value of $p = 30$.

In this section, SR and LS were the chosen methods because SR was associated to the best results in the previous section and LS for being the traditional linear regression method.

In contrast with the 2.2.2.1 section, the six models studied hereafter, were first submitted to the MPs additional pre-processing step.

Furthermore, besides model LC_{HRFs} and LC_{canHRF}, a third model was derived from the LC metric: the signals were convolved with the canonical HRF plus its temporal and dispersion derivatives. It was just performed on the LC metric due to its better prediction results. Since both derivatives just make sense when tested together with the canonical HRF, no sparse methods were used (only LS method).

Lastly, the Ridge was used again to verify if the decrease on the DOF would influence the trend of the prediction results and the interpretability of the EFP, due to a much restricted range of values.

2.2.2.2.1 Stepwise Regression (SR)

For each of the six models, the predictors were separated by the correspondent channels and used to fit the target response by the SR. At each channel, the average NMSE and the average BIC value were computed. Afterwards, the two channels that yielded the lowest error and BIC in each model were selected and compared.

Then the predictors from the best channel (among all models) were used to fit the target BOLD signal (figure 2.13). The same was performed using the BIC criterion. Although the selection criterion was different, most of the times the two best channels were the same.

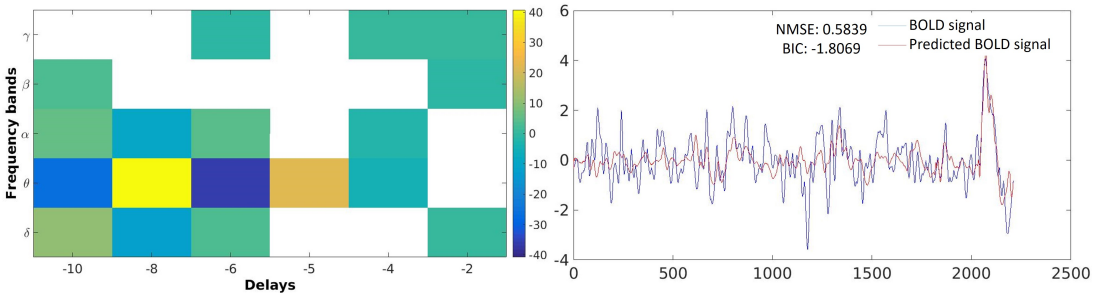

Figure 2.13: Example of the estimated model coefficients (left figure) that best predict the BOLD activity (EFP) using SR on the LC$_{HRFs}$ predictors from the best channel which was selected by the prediction error criterion. The right figure shows the BOLD signal and the respective prediction, derived from Patient 1. The values of BIC ($\times 10^4$) and error resulted from the fitting are also represented.

The results from each model and channel were subsequently compared through statistical analysis.

2.2.2.2.2 Least-squares Regression (LS)

As explained for the SR, the same procedures were conducted to achieve an average of the NMSE and of the BIC values at each channel. The two channels that yielded the lowest NMSE (and BIC) for each model were analyzed as well. Then the predictors derived from best channel among all models were selected for posterior BOLD fitting (figure 2.14).

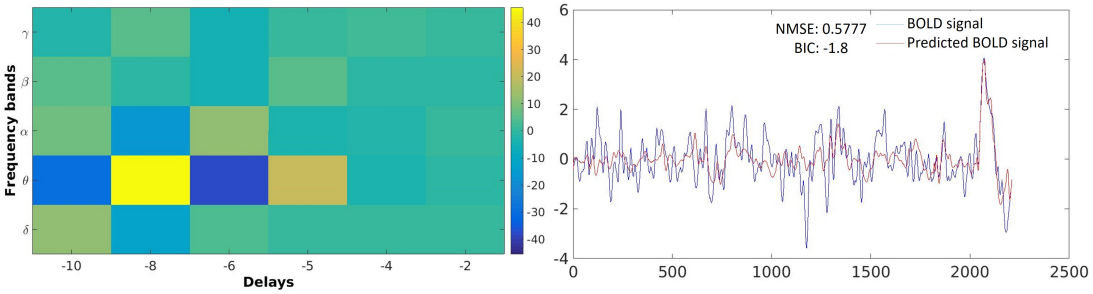

Figure 2.14: Example of the estimated model coefficients (left figure) that best predict the BOLD activity (EFP) using LS on the LC$_{HRFs}$ predictors from the best channel which was selected by the prediction error criterion. The right figure shows the BOLD signal and the respective prediction, derived from Patient 1. The values of BIC ($\times 10^4$) and error resulted from the fitting are also represented.

Just as carried out in the stepwise regression, the LS results were also submitted to the same statistical analysis.

As mentioned in the beginning of the section, for the LC metric, a new model was proposed, LC$_{canHRF+T+D}$. Mainly the BIC results from LC$_{canHRF+T+D}$ and LC$_{HRFs}$ demonstrated to be particularly similar across patients. For this reason, the results from each patient individually were submitted to statistical analysis.

Moreover, the LC$_{canHRF+T+D}$ predictors from the best channel were used to fit the BOLD signal and the EFP of the estimated coefficients were compared to the most complex LC model (figure 2.15).

2.2.2.2.3 Ridge Regression

This method was just used on the model LC$_{HRFs}$ because after model comparison it was the one that exhibited the best prediction performance of BOLD signal, based on the average NMSE and BIC.

In this part, the non-hybrid approach was applied but just for the Ridge regularization ("$\alpha = 0$"). The

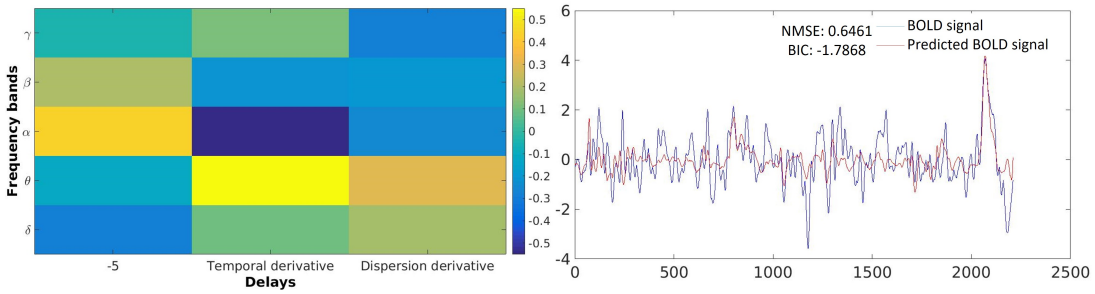

Figure 2.15: Example of the estimated model coefficients (left figure) that best predict the BOLD activity (EFP) using LS on the $LC_{canHRF+T+D}$ predictors from the best channel which was selected by the prediction error criterion. The right figure shows the BOLD signal and the respective prediction, derived from Patient 1. The values of BIC ($\times 10^4$) and error resulted from the fitting are also represented.

main difference is that, instead of applying it to a model containing the regressors from each of the 31 channels, the regularization parameter (λ) had to be optimized for each channel. The range of λ values determined for each patient in the previous section, was used again in this study.

The averaged measures at each channel (error and BIC), particularly from the one that yielded the lowest measures, were compared with the results obtained with LS and SR. Then, the predictors from the best channel of the best model were used to fit the BOLD signal by applying the Ridge with determined λ for that channel, as can be seen in the following figure (2.16). The Ridge EFPs of the remaining patients are presented in Appendix G.

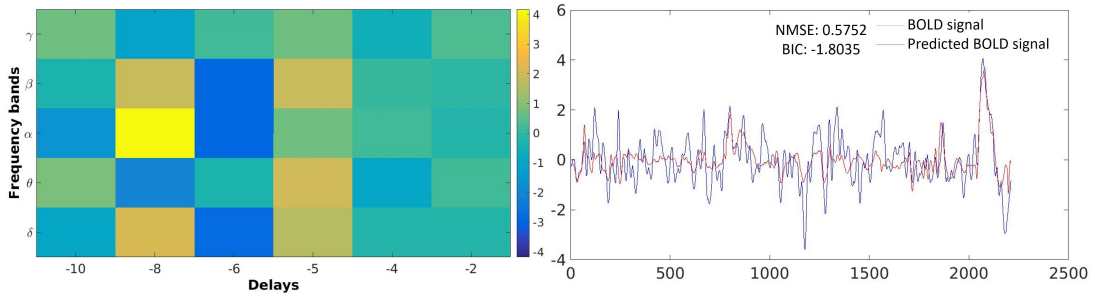

Figure 2.16: Example of the estimated model coefficients (left figure) that best predict the BOLD activity (EFP) using Ridge on the LC_{HRFs} predictors from the best channel which was selected by the prediction error criterion. The right figure shows the BOLD signal and the respective prediction, derived from Patient 1. The values of BIC ($\times 10^4$) and error resulted from the fitting are also represented.

Lastly, the results provided by the best model and method with and without the additional clean-up, based on the ICA, on the EEG data, were compared.

2.2.2.3 Statistical analysis

The main results were evaluated by means of repeated measures Analysis of Variance (ANOVA) in order to investigate whether significant differences were found between the investigated models and methods, considering a $p < 0.05$ level of significance. The NMSE and BIC results were tested separately. The different tests are chronologically displayed below. The 1. was conducted in the scope of all channels and 2. - 6. for each channel individually. The five subjects and the 10 folds were assigned to random factors and repeated measures, respectively. Regardless of the statistical test, a significant effect was

found for the factor subjects.

1. For the study considering all channels, after evaluating the prediction performance of the three models (LC_{HRFs}, TP_{HRFs} and $RMSF_{HRFs}$), a 2-way repeated measures ANOVA was performed with factors Model (LC, TP and RMSF) and Methods (LS, LAR, SR and the five variants of the Elastic Net Regularization, including Ridge and Lasso). After this step, a method was selected to perform some further tests as previously explained.
2. In the "Models with each channel individually" section (2.2.2.2) a set of tests were also accomplished. A 2-way ANOVA was performed with factors Channels (the 31) and Models (the 6). This test was done for both LS and SR methods, separately. Afterwards, this test was conducted by only inserting information regarding the best channel from each of the 6 models.
3. Afterwards, a 1-way ANOVA was performed to compare the LS and SR methods, by just using the results from the best channel across the six models.
4. Moreover, the results from applying the LS on the LC_{HRFs}, LC_{canHRF} and $LC_{canHRF+T+D}$ models were submitted to a new test to infer about their differences, for each patient individually. For each patient, the three models, LC_{HRFs}, $LC_{canHRF+T+D}$ and LC_{canHRF} and the 31 channels were assigned to fixed factors.
5. Within the Ridge results, an ANOVA test was executed to study if there were significant differences between the 31 channels. The differences between the results provided by the best channels from the three methods (SR, LS and Ridge) for the LC_{HRFs} model were also investigated.
6. Lastly, the results with and without ICA for the best model and method were compared by an ANOVA test. The fixed factors were the channels and with/without ICA.

These tests were retrieved by the *anova* function from Matlab. It also computes the statistics required for multiple comparison tests which were always performed next. The function *multcompare* performs pairwise comparison of the introduced factors. Just the most relevant results from multiple comparison are presented hereafter.

3. Results

This chapter shows the results for each division of the Methods chapter. Hereafter it will be presented the results from model estimation and evaluation, for each of the two types of models tested: including all channels and for each channel individually.

3.1 Models with all channels

This section is divided into 6 main parts. In the first, it is presented the average performance measures (DOF, BIC, R^2_{adj} and NMSE) from CV using the five values of α in the scope of the Elastic Net Regularization on LC$_{HRFs}$, RMSF$_{HRFs}$ and TP$_{HRFs}$ models. In the second, the same is presented regarding the SR, LAR and LS methods. The performance measures obtained by applying a specific regression method to estimate each EFP from the different EEG model, are shown in Appendix A. In the following part, there are presented the results from the nested model approach using LS. Then the three EEG models and the different methods were compared. Finally, results from further tests using LS and SR are shown: Appendices B and C present the CV results from using the LS and SR on the LC$_{canHRF}$ and RMSF$_{canHRF}$ models before MPs correction and from the control test, respectively.

3.1.1 Ridge and Lasso

Table 3.1: CV measures by using an Elastic Net Regularization methodology on the LC$_{HRFs}$ model obtained from each of the 5 patients. For both approaches, the results correspond to the average of the following measures across the ten test sets: BIC, R^2_{adj}, DOF and error. The grey cells highlight the methods that yielded the lowest BIC.

	LC$_{HRFs}$		Best α	Best λ	Average DOF	Average BIC values ($\times 10^4$)	Average R^2_{adj}	Average error (NMSE)
Patient 1	Hybrid		0	0.2965	267.0	-2.04	0.9701	0.0263
	Non-hybrid	Ridge	-	0.2965	267.0	-2.04	0.9701	0.0263
		Lasso 0.25	-	5.32×10^{-4}	562.1	-1.68	0.9315	0.0511
		Lasso 0.5	-	3.50×10^{-4}	253.3	-1.89	0.9369	0.0559
		Lasso 0.75	-	4.67×10^{-4}	290.2	-1.84	0.9275	0.0630
		Lasso	-	3.50×10^{-4}	253.3	-1.89	0.9369	0.0559
Patient 2	Hybrid		0	0.1390	517.1	-6.41	0.9055	0.0873
	Non-hybrid	Ridge	-	0.1390	517.1	-6.41	0.9055	0.0873
		Lasso 0.25	-	4.54×10^{-4}	662.4	-5.84	0.8032	0.1777
		Lasso 0.5	-	3.68×10^{-4}	556.7	-5.96	0.8136	0.1712
		Lasso 0.75	-	2.46×10^{-4}	490.8	-6.03	0.8211	0.1660
		Lasso	-	1.84×10^{-4}	483.6	-6.08	0.8346	0.1537
Patient 3	Hybrid		0	0.0212	558.6	-4.50	0.9825	0.0153
	Non-hybrid	Ridge	-	0.0212	558.6	-4.50	0.9825	0.0153
		Lasso 0.25	-	3.64×10^{-4}	684.2	-3.89	0.9358	0.0543
		Lasso 0.5	-	1.82×10^{-4}	577.9	-3.98	0.9372	0.0546
		Lasso 0.75	-	1.97×10^{-4}	519.0	-4.04	0.9388	0.0540
		Lasso	-	1.48×10^{-4}	507.3	-4.07	0.9428	0.0506
Patient 4	Hybrid		0	0.0457	134.7	-2.37	0.7194	0.2679
	Non-hybrid	Ridge	-	0.0457	134.7	-2.37	0.7194	0.2679
		Lasso 0.25	-	0.028	49.2	-2.16	0.2472	0.7404
		Lasso 0.5	-	0.0052	53.9	-2.18	0.3050	0.6824
		Lasso 0.75	-	5.02×10^{-4}	153.7	-2.22	0.5451	0.4314
		Lasso	-	3.76×10^{-4}	114.2	-2.30	0.6138	0.3714
Patient 5	Hybrid		0	0.0015	713.9	-6.49	0.9342	0.0589
	Non-hybrid	Ridge	-	0.0015	713.9	-6.49	0.9342	0.0589
		Lasso 0.25	-	1.52×10^{-4}	830.0	-5.78	0.8236	0.1549
		Lasso 0.5	-	3.26×10^{-4}	599.7	-5.85	0.7907	0.1909
		Lasso 0.75	-	2.17×10^{-4}	571.2	-5.91	0.8049	0.1788
		Lasso	-	1.63×10^{-4}	511.8	-5.97	0.8074	0.1781

Table 3.2: CV measures by using an Elastic Net Regularization methodology on the RMSF$_{HRFs}$ model obtained from each of the 5 patients. For both approaches, the results correspond to the average of the following measures across the ten test sets: BIC, R^2_{adj}, DOF and error. The grey cells highlight the methods that yielded the lowest BIC.

	RMSF$_{HRFs}$		Best α	Best λ	Average DOF	Average BIC values ($\times 10^4$)	Average R^2_{adj}	Average error (NMSE)
Patient 1	Hybrid		0	0.0239	159.4	-1.74	0.8102	0.1761
	Non-hybrid	Ridge	-	0.0239	159.4	-1.74	0.8102	0.1761
		Lasso 0.25	-	9.38×10^{-5}	182.0	1.68	0.7608	0.2195
		Lasso 0.5	-	4.69×10^{-5}	182.2	1.69	0.7764	0.2052
		Lasso 0.75	-	3.13×10^{-5}	182.5	-1.70	0.7853	0.1970
		Lasso	-	2.34×10^{-5}	182.6	-1.70	0.1945	0.1945
Patient 2	Hybrid		1	5.48e-04	110.4	-5.61	0.4280	0.5627
	Non-hybrid	Ridge	-	0.0726	160.0	-5.61	0.4607	0.5266
		Lasso 0.25	-	0.0022	130.0	-5.58	0.4154	0.5735
		Lasso 0.5	-	0.0011	116.0	-5.60	0.4219	0.5683
		Lasso 0.75	-	7.3×10^{-4}	110.8	-5.61	0.4254	0.5653
		Lasso	-	5.48×10^{-4}	105.5	-5.61	0.4254	0.5657
Patient 3	Hybrid		0	0.0288	162.3	-3.54	0.5918	0.3932
	Non-hybrid	Ridge	-	0.0028	162.3	-3.54	0.5918	0.3932
		Lasso 0.25	-	1.72×10^{-4}	166.9	-3.49	0.5488	0.4342
		Lasso 0.5	-	3.67×10^{-4}	132.2	-3.51	0.5352	0.4509
		Lasso 0.75	-	6.46×10^{-4}	124.0	-3.52	0.5400	0.4471
		Lasso	-	2.98×10^{-4}	126.0	-3.52	0.5026	0.4832
Patient 4	Hybrid		0	0.0100	155.7	-2.34	0.7074	0.2773
	Non-hybrid	Ridge	-	0.0100	155.7	-2.34	0.7074	0.2773
		Lasso 0.25	-	1.44×10^{-4}	174.3	-2.26	0.6284	0.3499
		Lasso 0.5	-	7.22×10^{-5}	170.8	-2.28	0.6472	0.3326
		Lasso 0.75	-	2.06×10^{-4}	153.5	-2.28	0.6365	0.3448
		Lasso	-	2.51×10^{-4}	122.7	-2.30	0.6281	0.3566
Patient 5	Hybrid		0	2.5479	124.2	-5.49	0.3170	0.6706
	Non-hybrid	Ridge	-	2.5479	131.1	-5.49	0.3210	0.6659
		Lasso 0.25	-	0.0012	152.7	-5.46	0.3099	0.6746
		Lasso 0.5	-	5.87×10^{-4}	140.6	-5.48	0.3121	0.6737
		Lasso 0.75	-	6.35×10^{-4}	128.5	-5.48	0.3090	0.6780
		Lasso	-	0.0013	116.9	-5.49	0.3031	0.6850

Table 3.3: CV measures by using an Elastic Net Regularization methodology on the TP$_{HRFs}$ model obtained from each of the 5 patients. For both approaches, the results correspond to the average of the following measures across the ten test sets: BIC, R^2_{adj}, DOF and error. The grey cells highlight the methods that yielded the lowest BIC.

	TP$_{HRFs}$		Best α	Best λ	Average DOF	Average BIC values ($\times 10^4$)	Average R^2_{adj}	Average error (NMSE)
Patient 1	Hybrid		0	0.0023	143.1	-1.77	0.8261	0.1626
	Non-hybrid	Ridge	-	0.0023	143.1	-1.77	0.8261	0.1626
		Lasso 0.25	-	2.03×10^{-4}	176.8	-1.65	0.7253	0.2527
		Lasso 0.5	-	1.01×10^{-4}	172.8	-1.68	0.7585	0.2226
		Lasso 0.75	-	1.1×10^{-4}	154.5	-1.69	0.7580	0.2251
		Lasso	-	1.34×10^{-4}	138.9	-1.71	0.7645	0.2207
Patient 2	Hybrid		0	0.333	116.3	-5.69	0.5010	0.4905
	Non-hybrid	Ridge	-	0.333	116.3	-5.69	0.5010	0.4905
		Lasso 0.25	-	9.44×10^{-5}	175.7	-5.64	0.4947	0.4923
		Lasso 0.5	-	2.02×10^{-4}	148.6	-5.65	0.4874	0.5014
		Lasso 0.75	-	1.35×10^{-4}	129.6	-5.66	0.4859	0.5043
		Lasso	-	1.64×10^{-4}	122.5	-5.67	0.4881	0.5027
Patient 3	Hybrid		0	0.1680	133.7	-3.56	0.5945	0.3932
	Non-hybrid	Ridge	-	0.1680	133.7	-3.56	0.5945	0.3932
		Lasso 0.25	-	2.04×10^{-4}	172.6	-3.51	0.5752	0.4082
		Lasso 0.5	-	1.66×10^{-4}	158.8	-3.53	0.5788	0.4061
		Lasso 0.75	-	7.69×10^{-4}	116.3	-3.54	0.5566	0.4317
		Lasso	-	2.19×10^{-4}	123.4	-3.55	0.5773	0.4109
Patient 4	Hybrid		0	0.003	63.5	-2.30	0.5664	0.4244
	Non-hybrid	Ridge	-	0.003	63.5	-2.30	0.5664	0.4244
		Lasso 0.25	-	0.0044	25.2	-2.16	0.1808	0.8123
		Lasso 0.5	-	0.002	62.2	-2.17	0.3040	0.6815
		Lasso 0.75	-	4.97×10^{-4}	61.9	-2.20	0.3801	0.6070
		Lasso	-	2.3×10^{-4}	71	-2.24	0.4586	0.5285
Patient 5	Hybrid		0	0.0075	159.6	-5.64	0.4820	0.5059
	Non-hybrid	Ridge	-	0.0075	159.9	-5.64	0.4820	0.5059
		Lasso 0.25	-	0.0028	132.7	-5.57	0.4048	0.5836
		Lasso 0.5	-	8.58×10^{-4}	109.5	-5.59	0.4066	0.5839
		Lasso 0.75	-	5.72×10^{-4}	101.2	-5.60	0.4125	0.5788
		Lasso	-	4.29×10^{-4}	91.5	-5.62	0.4156	0.5766

Regarding the "best α" and the associated "best λ", only the combination with the largest number of occurrences throughout the ten learning sets was presented in the tables above. The selection criteria of the parameters during the inner CV was BIC, therefore the average error of the hybrid approach may not be the lowest. As it was expected, the training error (results showed in Appendix A) resulted in lower NMSE than the prediction error exhibited throughout the work.

The average of the BIC values in the hybrid approach was always equal or lower than the average of the BIC values found for each of the α in the non-hybrid approach. However, if no mixture of α values had occurred in the first approach, its results will equal the best result of the second approach. Overall, a predominant α always stood up in the hybrid approach, leading to similar results for both implementations. Moreover, the LC metric, comparing with the other two, always exhibited the lowest BIC. Except for patient 4, TP provided better results in relation to RMSF.

3.1.2 SR, LAR and LS

In this section, the CV results from applying the SR, the LAR and the LS on the LC$_{HRFs}$, the TP$_{HRFs}$ and on the RMSF$_{HRFs}$ models are presented afterwards.

Table 3.4: CV measures by using SR, LAR and LS on the LC$_{HRFs}$ model obtained from each of the 5 patients. For the three different methods, the results correspond to the average of the following measures across the ten test sets: BIC, R^2_{adj}, DOF and error. The grey cells highlight the methods that yielded the lowest BIC.

LC$_{HRFs}$		Average DOF	Average BIC values ($\times 10^4$)	Average R^2_{adj}	Average error (NMSE)
Patient 1	**SR**	552.7	-2.79	0.9997	1.99×10^{-4}
	LAR	742.4	-2.7	0.9998	1.51×10^{-4}
	LS	930.0	-2.53	0.9997	1.91×10^{-4}
Patient 2	**SR**	825.1	-7.69	0.9922	0.0069
	LAR	929.8	-7.60	0.9920	0.0069
	LS	930.0	-7.60	0.9920	0.0069
Patient 3	**SR**	816.8	-5.72	0.9992	6.74×10^{-4}
	LAR	927.5	-5.62	0.9992	5.96×10^{-4}
	LS	930.0	-5.62	0.9992	6.12×10^{-4}
Patient 4	**SR**	610.3	-3.90	0.9997	2.77×10^{-4}
	LAR	700.4	-3.71	0.9995	3.86×10^{-4}
	LS	930.0	-3.74	0.9997	1.82×10^{-4}
Patient 5	**SR**	793.3	-7.36	0.9860	0.0124
	LAR	926.8	-7.25	0.9859	0.0122
	LS	930.0	-7.24	0.9858	0.0123

Table 3.5: CV measures by using SR, LAR and LS on the RMSF$_{HRFs}$ model obtained from each of the 5 patients. For the three different methods, the average results across the ten folds of the following measures are presented: BIC, R^2_{adj}, DOF and error. The grey cells highlight the methods that yielded the lowest BIC.

RMSF$_{HRFs}$		Average DOF	Average BIC values ($\times 10^4$)	Average R^2_{adj}	Average error (NMSE)
Patient 1	**SR**	159.6	-1.79	0.8538	0.1356
	LAR	185.0	-1.81	0.8581	0.1300
	LS	186.0	-1.78	0.8582	0.1299
Patient 2	**SR**	98.1	-5.64	0.4471	0.5449
	LAR	178.5	-5.61	0.4702	0.5159
	LS	186.0	-5.60	0.4708	0.5148
Patient 3	**SR**	121.9	-3.57	0.5957	0.3931
	LAR	176.8	-3.54	0.6032	0.3809
	LS	186.0	-3.53	0.6047	0.3787
Patient 4	**SR**	135	-2.38	0.7295	0.2582
	LAR	183.0	-2.35	0.7402	0.2438
	LS	186.0	-2.35	0.7408	0.2430
Patient 5	**SR**	93.1	-5.53	0.3329	0.6580
	LAR	154.3	-5.49	0.3405	0.6446
	LS	186.0	-5.47	0.3492	0.6330

Table 3.6: CV measures by using SR, LAR and LS on the TP$_{HRFs}$ model obtained from each of the 5 patients. For the three different methods, the results correspond to the average of the following measures across the ten test sets: BIC, R^2_{adj}, DOF and error. The grey cells highlight the methods that yielded the lowest BIC.

TP$_{HRFs}$		Average DOF	Average BIC values ($\times 10^4$)	Average R^2_{adj}	Average error (NMSE)
Patient 1	**SR**	156.2	-1.87	0.8992	0.0937
	LAR	185.8	-1.85	0.9015	0.0902
	LS	186.0	-1.85	0.9014	0.0903
Patient 2	**SR**	132.9	-5.73	0.5380	0.4530
	LAR	181.9	-5.70	0.5504	0.4276
	LS	186.0	-5.70	0.5515	0.4363
Patient 3	**SR**	107.6	-3.59	0.6067	0.3837
	LAR	157.3	-3.55	0.6064	0.3796
	LS	186.0	-3.54	0.6115	0.3721
Patient 4	**SR**	149.0	-2.46	0.8102	0.1803
	LAR	186.0	-2.44	0.8161	0.1724
	LS	186.0	-2.44	0.8161	0.1724
Patient 5	**SR**	125.1	-5.66	0.4820	0.5085
	LAR	181.0	-5.64	0.4969	0.4897
	LS	186.0	-5.63	0.4975	0.4888

The SR retrieved lower BIC results in relation to the hybrid approach in the Elastic Net Regularization. Except for the RMSF$_{HRFs}$ model in Patient 1, SR outperformed LAR and LS. Although these two last methods were associated to the lowest NMSE, the same did not occur in relation to BIC. Independently of the method considered, the predictive power of LC metric was always better than TP and RMSF.

3.1.3 Nested model approach

The nested model approach was implemented by using LS regression method to fit the BOLD signal. For each of the LC$_{HRFs}$, TP$_{HRFs}$ and RMSF$_{HRFs}$ models (with no prior division of data into folds by CV or

bootstrap), the outcomes from the first step of the nested model approach will be presented in the following tables. Considering the LC$_{HRFs}$ model, this was divided into the following features' dimension: channels, frequency bands, HRF-delays and exceptionally in combinations of HRF-delay/frequency band.

Table 3.7: Results from the first step of the nested model approach on the LC$_{HRFs}$ model. For each patient and feature dimension, it is presented the one that yielded the lowest training error and the correspondent value.

LC$_{HRFs}$ model	**Channel**		**HRF-delay**		**Frequency band**		**Combination HRF/ Frequency band**	
	Best	NMSE	Best	NMSE	Best	NMSE	Best	NMSE
Patient 1	18	0.5745	10	0.0751	δ	0.0761	10/θ	0.4671
Patient 2	15	0.5535	4	0.3762	δ	0.3170	4/θ	0.6066
Patient 3	12	0.5003	6	0.2319	γ	0.2399	5/γ	0.5339
Patient 4	8	0.6093	10	0.1532	δ	0.1603	6/δ	0.5577
Patient 5	30	0.6569	8	0.4367	δ	0.4590	5/γ	0.6966

The RMSF$_{HRFs}$ model were divided into the following features' dimension: channels and HRF-delays.

Table 3.8: Results from the first step of the nested model approach on the RMSF$_{HRFs}$ model. For each patient and feature dimension, it is presented the one that yielded the lowest training error and the correspondent value.

RMSF$_{HRFs}$ model	**Channel**		**HRF-delay**	
	Best	NMSE	Best	NMSE
Patient 1	2	0.9306	10	0.5635
Patient 2	28	0.9401	4	0.8135
Patient 3	29	0.7866	6	0.6644
Patient 4	24	0.8019	8	0.7187
Patient 5	30	0.9170	8	0.8083

The TP$_{HRFs}$ model were divided into the following features' dimension: channels and HRF-delays.

Table 3.9: Results from the first step of the nested model approach on the TP$_{HRFs}$ model. For each patient and feature dimension, it is presented the one that yielded the lowest training error and the correspondent value.

TP$_{HRFs}$ model	**Channel**		**HRF-delay**	
	Best	NMSE	Best	NMSE
Patient 1	18	0.7204	10	0.5482
Patient 2	15	0.7595	4	0.6861
Patient 3	1	0.6995	4	0.5930
Patient 4	4	0.8050	6	0.5566
Patient 5	5	0.7330	6	0.7211

Regardless of the resampling procedure that was used (CV or bootstrap), the performed t-test on the VE values along folds had always retrieved p-values lower than 0.05, indicating that for the three metrics and correspondent features, the full model was recommended. In other words, there was no subset of channels, frequency bands, HRF-delays or combinations of HRF-delay/frequency band (in the LC$_{HRFs}$ case) that performed better than the model containing all predictors. The same results were obtained for the RMSF$_{HRFs}$ and TP$_{HRFs}$. For this reason, the sequence that was found from the first step of the implementation until p predictors were introduced, for each features' dimension, was not presented in this section.

3.1.4 Comparison of methods and models

The results of the performance measures BIC and NMSE from the different methods and the three models (LC_{HRFs}, TP_{HRFs} and $RMSF_{HRFs}$), are presented in figures 3.1 to 3.5, for each of the 5 patients.

A general conclusion can be stated regardless of the patient: the differences of the results from the first three methods (LS, LAR and SR) and the remaining ones (Ridge, Lasso and the three Elastic Net Regularization methods) are larger when the LC model is used. Moreover, the errors from the three first methods are near zero making the differences between them almost imperceptible.

However, when the number of predictors was reduced as in case of the RMSF and TP models, the differences between methods are not so obvious, mainly in patients 2, 3 and 5, where Ridge, Lasso or some variants of the Elastic Net Regularization methods are not associated to the worst results.

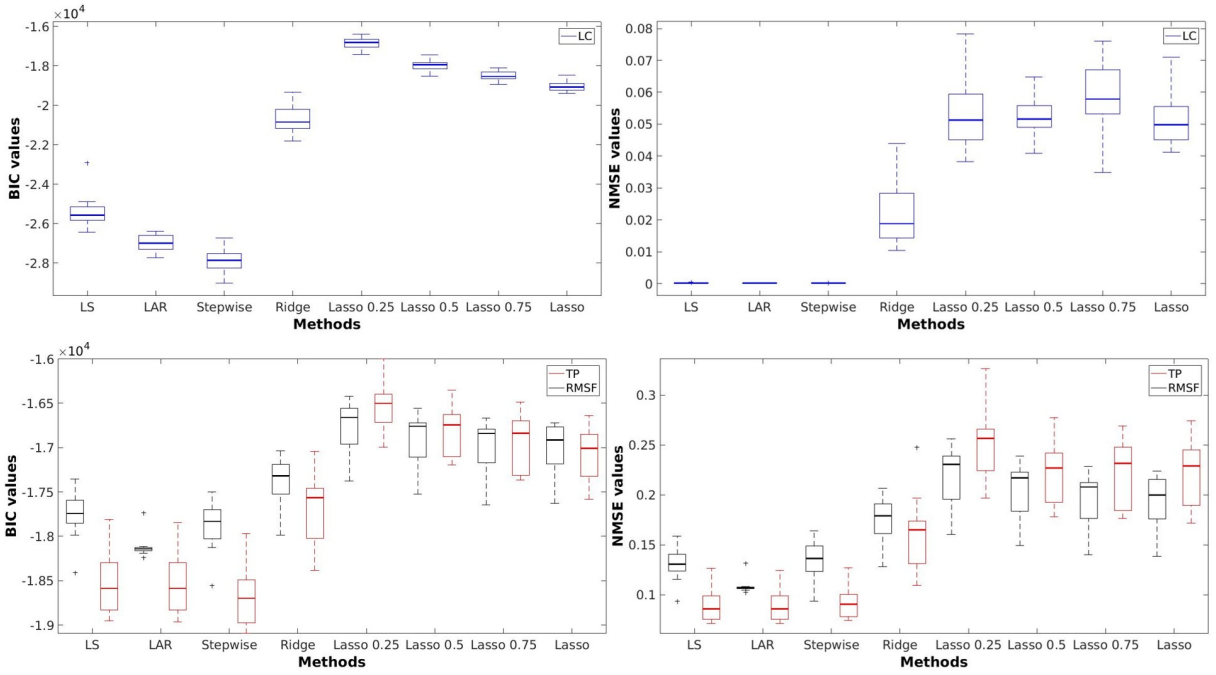

Figure 3.1: Boxplots of the prediction error (NMSE) and the BIC values obtained by LC_{HRFs} (top), $RMSF_{HRFs}$ and TP_{HRFs} (bottom) models across the 10 test folds for the different regression methods for Patient 1.

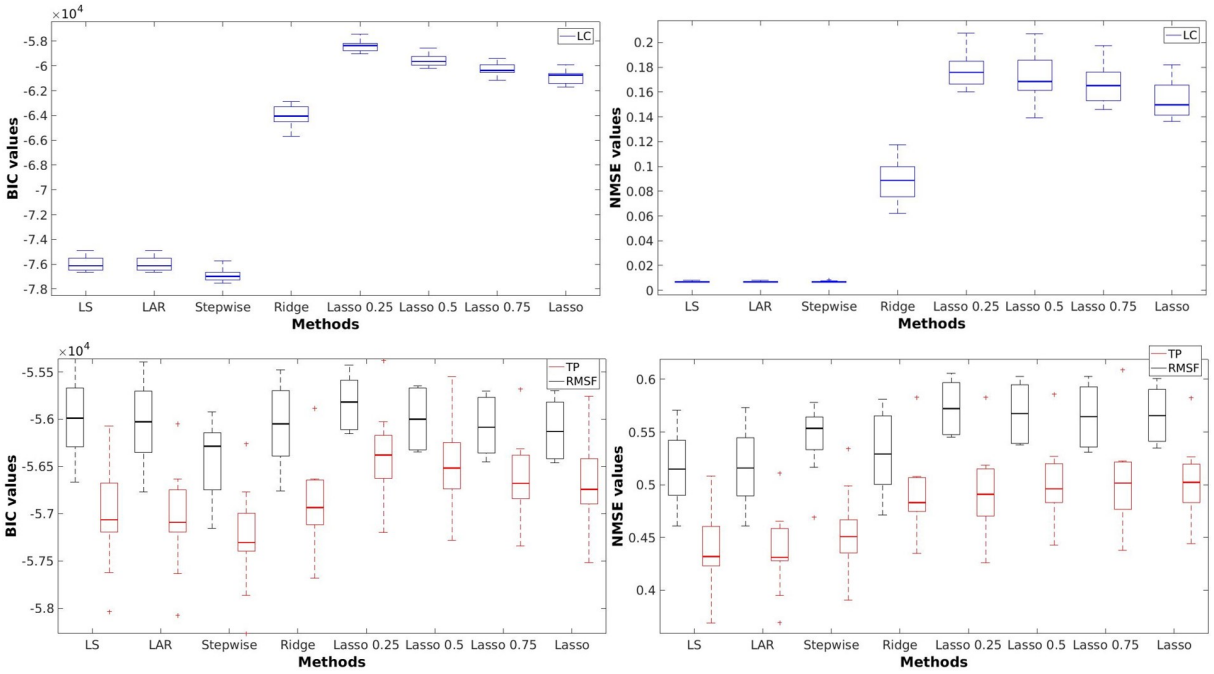

Figure 3.2: Boxplots of the prediction error (NMSE) and the BIC values obtained by LC_{HRFs} (top), $RMSF_{HRFs}$ and TP_{HRFs} (bottom) models across the 10 test folds for the different regression methods for Patient 2.

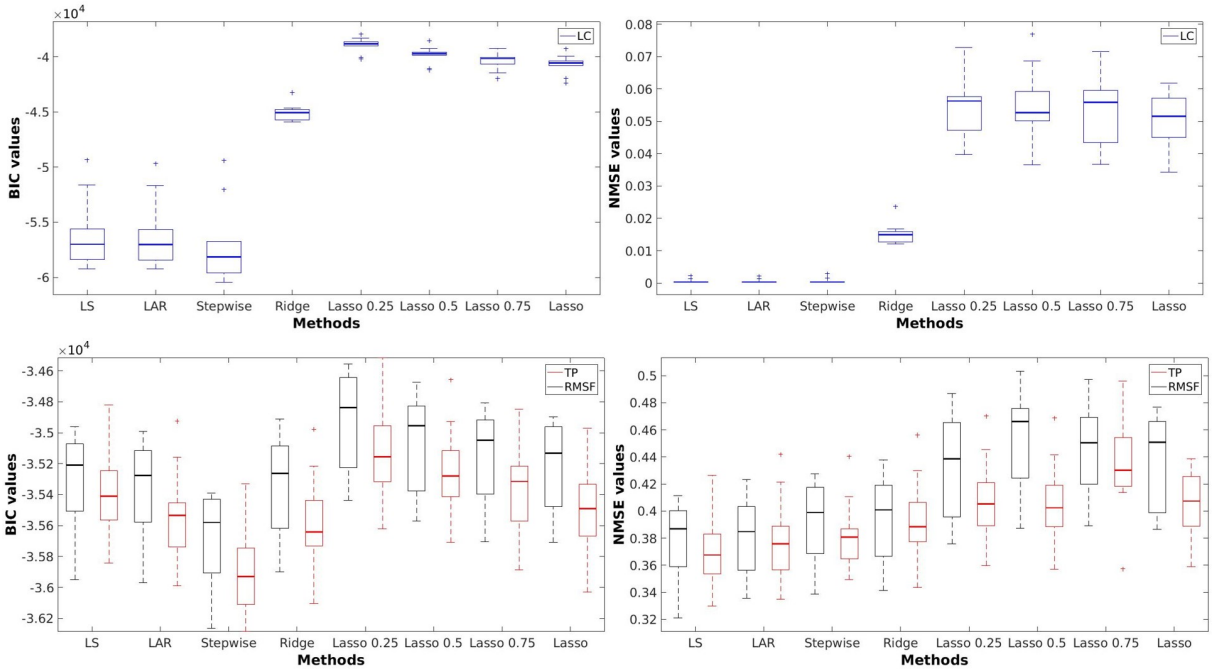

Figure 3.3: Boxplots of the prediction error (NMSE) and the BIC values obtained by LC_{HRFs} (top), $RMSF_{HRFs}$ and TP_{HRFs} (bottom) models across the 10 test folds for the different regression methods for Patient 3.

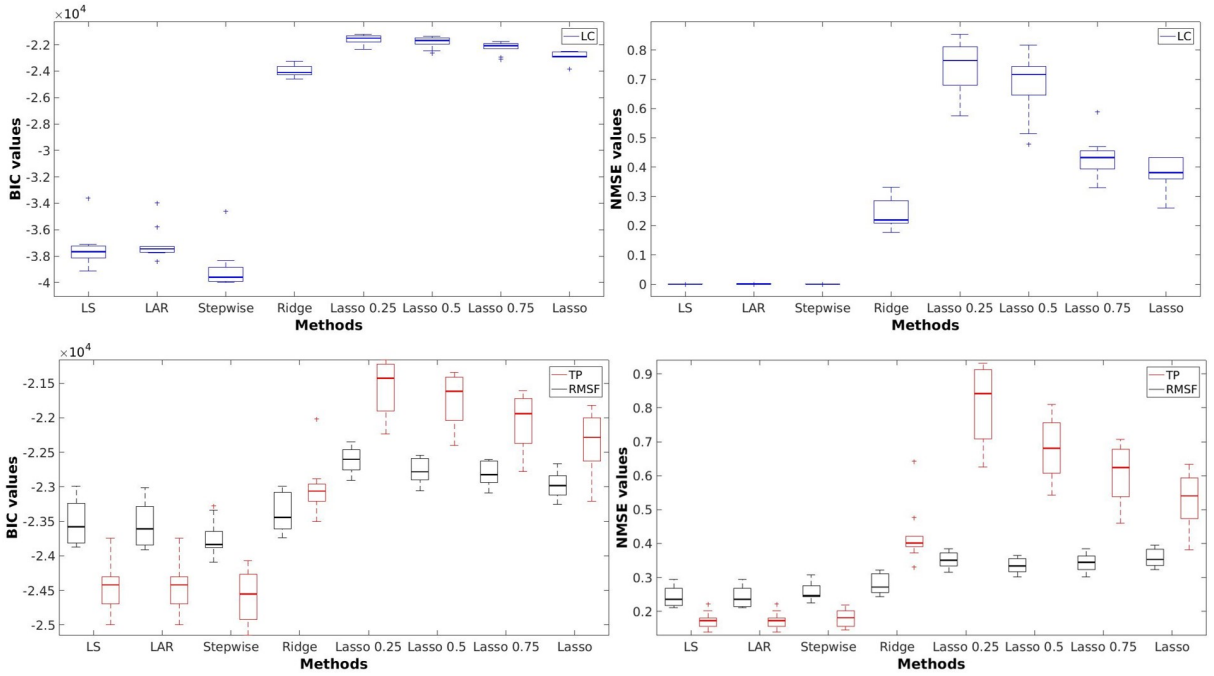

Figure 3.4: Boxplots of the prediction error (NMSE) and the BIC values obtained by LC_{HRFs} (top), $RMSF_{HRFs}$ and TP_{HRFs} (bottom) models across the 10 test folds for the different regression methods for Patient 4.

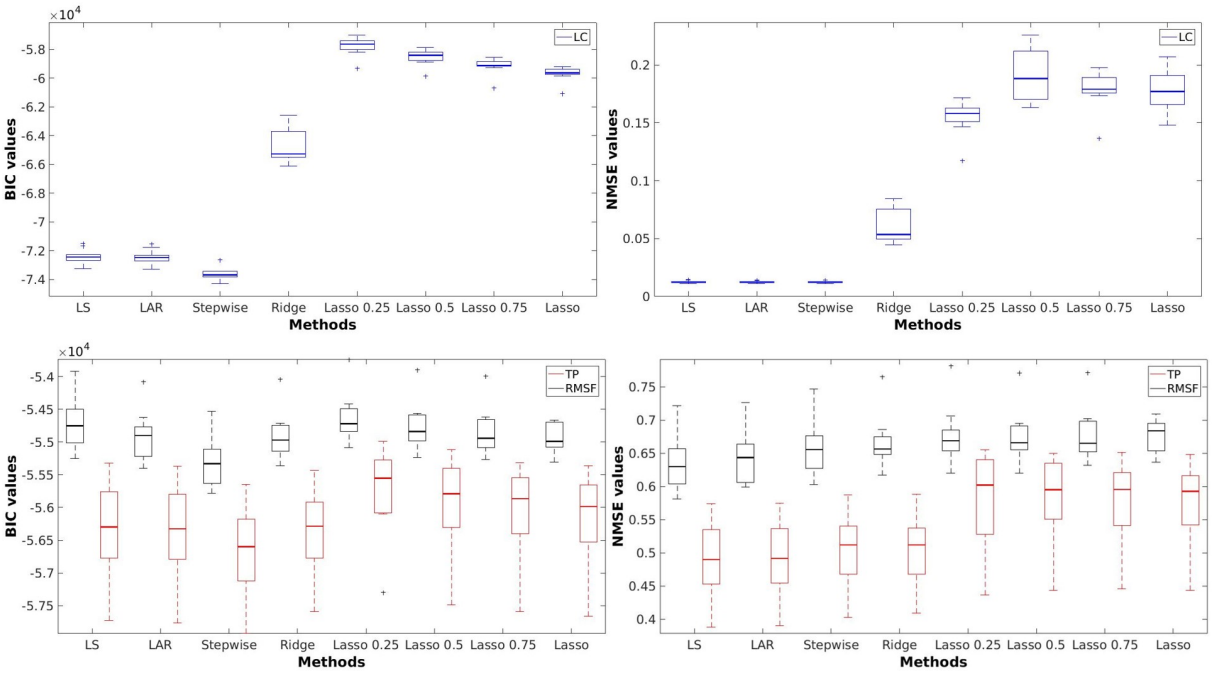

Figure 3.5: Boxplots of the prediction error (NMSE) represented on the right figures and the BIC values represented on the left figures obtained by LC_{HRFs}, $RMSF_{HRFs}$ and TP_{HRFs} models for the different regression methods for Patient 5. Both criteria were calculated on the 10 test sets. The two figures on the first row are the results from using the LC_{HRFs} and the two figures on the second row from using the TP_{HRFs} and $RMSF_{HRFs}$. The division into two rows of these three metrics is due to a more similar range of values of the TP and RMSF models. By giving more attention the BIC graphics, the first three regression methods (LS, LAR and SR) stand out.

In the results from the performed 2-way ANOVA, the three models (LC_{HRFs}, TP_{HRFs} and $RMSF_{HRFs}$) and eight methods (LS, LAR, SR, Ridge, Lasso 0.25, Lasso 0.5, Lasso 0.75 and Lasso) were analyzed. Regarding the BIC and NMSE, significant effect was found for the factors models and methods.

3.1.5 Further tests using LS and SR

The table 3.10 summarizes the average results originated by CV using the SR and LS methods and the RMSF$_{canHRF}$ and TP$_{canHRF}$ models, after MPs correction, to fit the BOLD signal. A similar table with the results obtained before MPs correction of the predictors, is exhibited in Appendix B. As can be observed by the highlighted results in the table, the SR method and the TP model retrieved the best results, independently of the patient. Comparing the results after and before MPs correction, for three of the five patients (patients 2, 3 and 4), the average BIC value decreased in relation to before correction.

Table 3.10: CV measures by using SR and LS on the RMSF$_{canHRF}$ and TP$_{canHRF}$ models obtained from each of the 5 patients, after MPs correction. For the two different methods, the average results across the ten folds of the following measures are presented: BIC values, R^2_{adj}, DOF and error. The grey cells highlight the methods that yielded the lowest BIC.

RMSF$_{canHRF}$ and TP$_{canHRF}$ models			Average DOF	Average BIC ($\times 10^4$)	Average R^2_{adj}	Average error (NMSE)
Patient 1	**SR**	TP	17.4	-1.5824	0.3360	0.6588
		RMSF	22.8	-1.5566	0.2594	0.7329
	LS	TP	31.0	-1.5764	0.3489	0.6420
		RMSF	31.0	-1.5539	0.2737	0.7161
Patient 2	**SR**	TP	21.5	-5.5308	0.2704	0.7273
		RMSF	19.5	-5.4640	0.1845	0.8132
	LS	TP	31.0	-5.5247	0.2718	0.7249
		RMSF	31.0	-5.4574	0.1873	0.8090
Patient 3	**SR**	TP	28.2	-3.4933	0.4193	0.5770
		RMSF	22.2	-3.4397	0.3288	0.6678
	LS	TP	31.0	-3.4929	0.4209	0.5750
		RMSF	31.0	-3.4346	0.3299	0.6654
Patient 4	**SR**	TP	24.4	-2.2385	0.3945	0.6005
		RMSF	21.8	-2.1566	0.1828	0.8112
	LS	TP	31.0	-2.2372	0.4059	0.5879
		RMSF	31.0	-2.1534	0.1935	0.7981
Patient 5	**SR**	TP	21.0	-5.5147	0.2495	0.7482
		RMSF	18.0	-5.4331	0.1403	0.8574
	LS	TP	31.0	-5.5055	0.2468	0.7498
		RMSF	31.0	-5.4230	0.1398	0.8563

Subsequently, the results of the ROR metric on the LC$_{HRFs}$, TP$_{HRFs}$ and TP$_{canHRF}$ models using the SR are shown in table 3.11. In general, no evidence of overfitting effect was found, given that all ROR values were near 0. Curiously, the LC$_{HRFs}$ showed the lowest value in relation to the other models. Although a higher number of regressors is considered in this model, according to this measure, the results were less affected by overfitting.

Table 3.11: For each patient, the ROR results using the SR on the three models: LC$_{HRFs}$, TP$_{HRFs}$ and TP$_{canHRF}$.

ROR	LC$_{HRFs}$	TP$_{HRFs}$	TP$_{canHRF}$
Patient 1	0.0078	0.0331	0.0586
Patient 2	0.0127	0.0417	0.0323
Patient 3	0.0041	0.0521	0.0291
Patient 4	0.0053	0.0348	0.0245
Patient 5	0.0191	0.0650	0.0320

As mentioned before, a control test was also performed. The first step was to find the regressor, for each of the TP$_{canHRFF}$ and RMSF$_{canHRF}$ models, whose correlation with the target signal was the largest. Regarding the first model, the channels/regressors expressing the highest correlation were: 18, 24, 1, 22 and 30, following the order of the patients (Patient 1 to Patient 5). For the second model were: 2, 16, 29, 24 and 30. Afterwards, as in table 3.10, the performance measures were calculated and represented in Appendix C. Except for Patient 1, the average values of BIC were higher in the control test in relation to the results presented in table 3.10.

It was noticed that the highest estimated coefficients in absolute value by SR or LS always corresponded to the regressor with the largest correlation with the BOLD signal. The TP$_{canHRF}$ for patients 1 and 3 using SR are given as examples, in figure 3.6, because they exhibited the most evident and focal topographies.

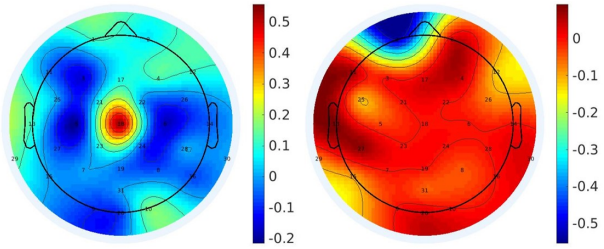

Figure 3.6: Example of the topography of the estimated TP$_{canHRF}$ model coefficients resulted by applying the SR on the largest correlation predictor and its correspondent phase randomization predictors, from patients 1 and 3 (left and right figures, respectively) in order to fit the BOLD response. The channel 18 (highlighted by the red color) and channel 1 (highlighted by the blue color) are associated to the predictors exhibiting the highest correlation with BOLD signal for patient 1 and 3, respectively.

3.2 Models with each channel individually

Hereafter are presented the results of the performance measures by using each of the six models of predictors from each channel separately. In this section, the SR (best predictive performance) and LS (comparison purposes) were the only considered methods. Then, the same procedure was only applied on the best model (LC$_{HRFs}$) but using Ridge. Finally, as in the previous section, the results from the three different methods and six models were compared.

3.2.1 Stepwise Regression (SR)

Figure 3.7 exhibit the average prediction error at each channel, for each of the six models and patients, by using SR. Some observations can be retrieved regarding the link between the spatial distribution of the errors across channels and the clinical condition of the patients (table 2.1). In summary, the TP metrics, specially in patients 1 and 2, displayed a pattern in which the lowest errors were concentrated in the left hemisphere (temporal lobe in case of patient 2). A similar relation can also be observed in case of the LC metric for patients 4 and 5.

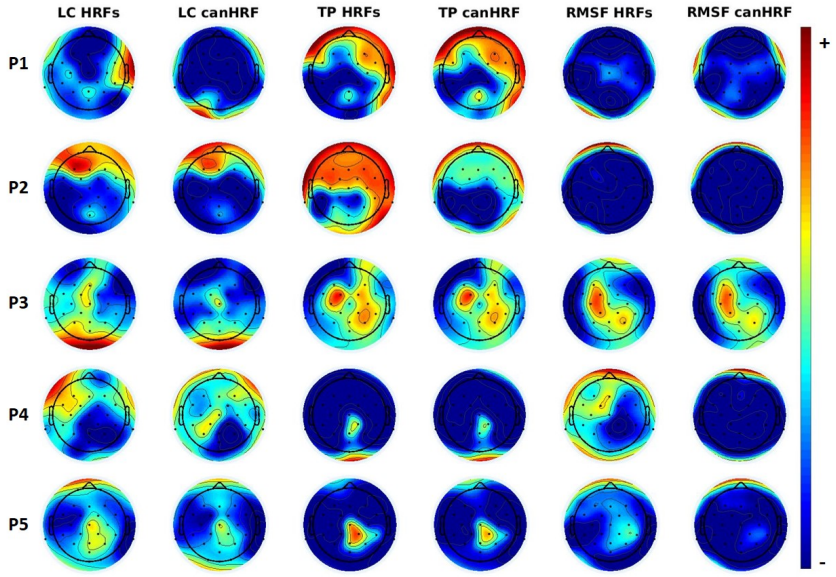

Figure 3.7: The prediction error at each channel by using the SR on each of the six models, for each patient. The different rows correspond to the different patients from patient 1 (P1) until patient 5 (P5). The different columns represent the six different models of predictors. In order to have a clear visualization of the errors distribution, for each model and patient, the color scale was bounded between the minimum and maximum for each representation. Therefore, no values were displayed in the limits of the colorbar.

Figures 3.8 to 3.12 show the results only from the two channels that yielded the lowest NMSE (also BIC) for each model and patient.

It can be concluded that the differences of results between models were shrunk in comparison to the range of values that was observed when the predictors from all channels were included into regression. This is mainly clear for LC_{HRFs}, RMSF_{HRFs} and TP_{HRFs} in relation to the outcomes in the Methods Comparison section.

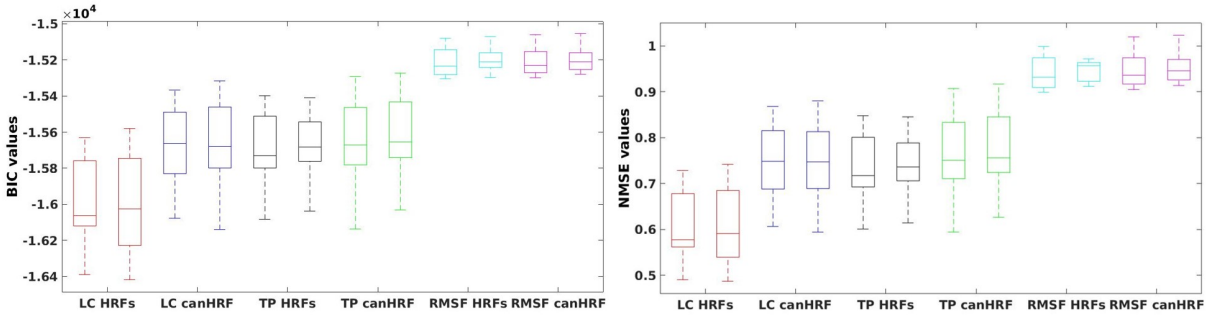

Figure 3.8: Boxplots of the BIC and NMSE across folds from the two channels that yielded the lowest prediction error (NMSE) and the lowest BIC values, for each of the six models, by using SR. This representation is derived from Patient 1.

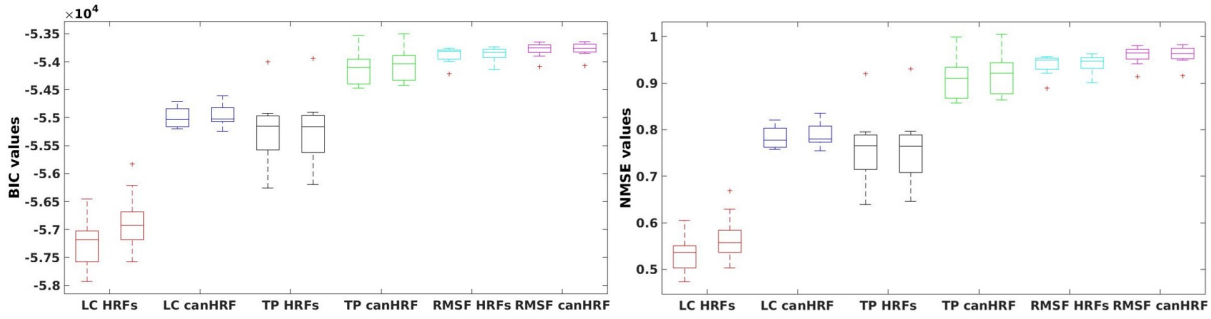

Figure 3.9: Boxplots of the BIC and NMSE across folds from the two channels that yielded the lowest prediction error (NMSE) and the lowest BIC values, for each of the six models, by using SR. This representation is derived from Patient 2.

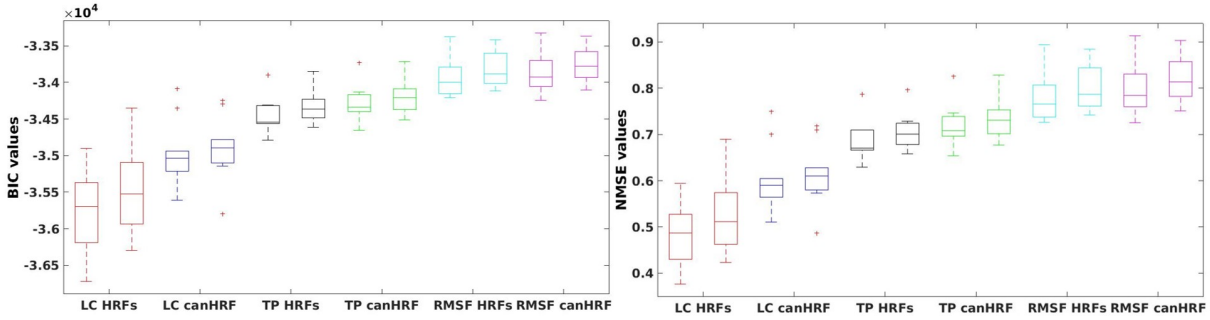

Figure 3.10: Boxplots of the BIC and NMSE across folds from the two channels that yielded the lowest prediction error (NMSE) and the lowest BIC values, for each of the six models, by using SR. This representation is derived from Patient 3.

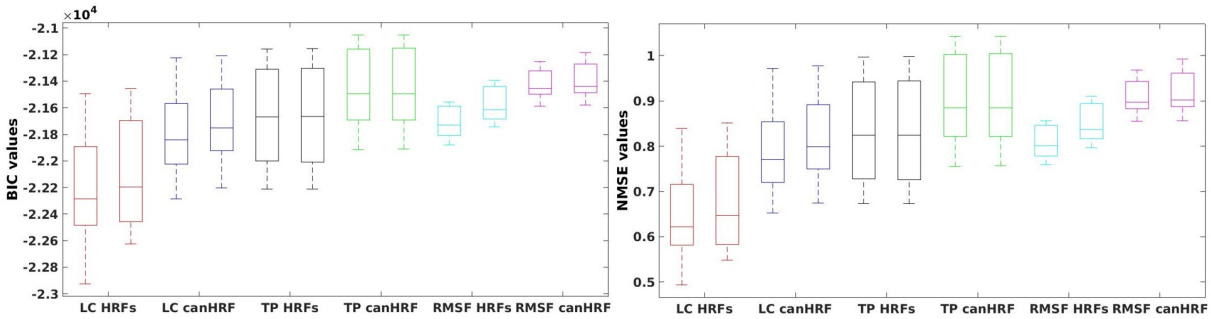

Figure 3.11: Boxplots of the BIC and NMSE across folds from the two channels that yielded the lowest prediction error (NMSE) and the lowest BIC values, for each of the six models, by using SR. This representation is derived from Patient 4.

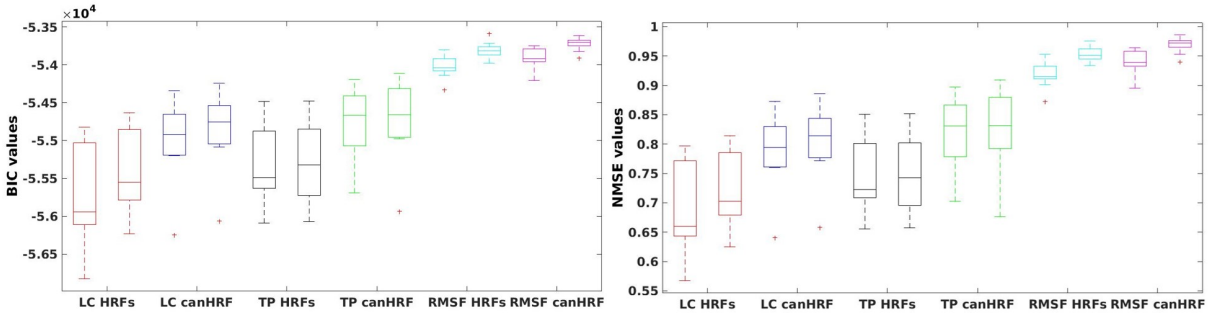

Figure 3.12: Boxplots of the prediction error (NMSE) represented on the right figure and the BIC values on the left figure derived from applying SR. This representation is derived from Patient 5. Both criteria were calculated on the 10 test sets. The six different colors represent the six different models under study: each model has two boxes presenting the results from the channel that yielded the lowest value (1^{st}) on the left side and the second lowest value (2^{nd}) on the right side, for each selection criteria.

Although the predictive performance values for each model considerably change across patients, it can be observed that the LC$_{HRFs}$ model outperforms the other five models, by always presenting the lowest NMSE and BIC median values across patients.

Then, the comparison of the results from the 31 channels across the six models was accomplished using the SR. For each selection criterion, only the factor models showed significant effect. Furthermore, by only using the results from the channels that yielded the lowest NMSE average and the BIC average, was found no significant differences between LC$_{canHRF}$/ TP$_{HRFs}$ and between RMSF$_{HRFs}$/ RMSF$_{canHRF}$, considering the BIC. By using NMSE, no significant differences were found just for the first pair of models.

3.2.2 Least-squares Regression (LS)

The topographies of the average prediction error at each channel, for each of the six models, by using the LS, exhibited similar results to the ones obtained by SR, apart of some subtle differences. For this reason, the resulted figure is presented in Appendix D. Also the boxplots that show the results from the two channels that yielded the lowest NMSE (also BIC) for each model and patient, demonstrated to be very similar to the ones presented by SR. Therefore these are also listed in Appendix E.

As it was concluded to SR, the model LC$_{HRFs}$ outperforms the other five models, by always presenting the lowest NMSE and BIC median values across patients. However, for Patient 5, it was not so evident due to the smaller differences between the LC$_{HRFs}$ and TP$_{HRFs}$ models.

As in SR, the comparison of the results from the 31 channels across the six models was conducted using the LS. For each selection criterion, only the factor models showed significant effect. Regarding the statistical test using the results from the best channels, the same conclusion was obtained as in SR.

The results from LS are not presented due to the similarity with the ones provided by SR, as it is subsequently shown by a statistical test in “Comparison of methods and models”.

Figures 3.13 to 3.17 present the results of using not only the canonical HRF but also its temporal and dispersion derivatives for the LC metric (LC$_{canHRF+T+D}$ model), represented in between the LC$_{HRFs}$ and the LC$_{canHRF}$ results, for each of the five patients.

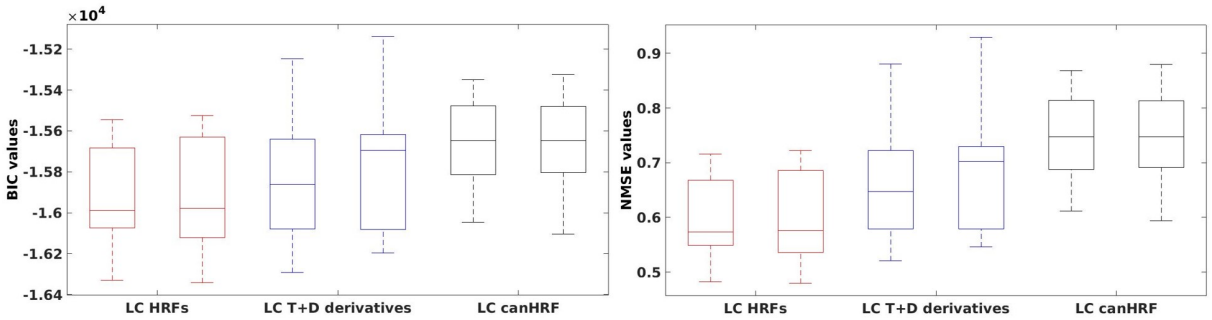

Figure 3.13: Boxplots of the prediction errors and the BIC values for Patient 1, using LS on the LC_{HRFs}, $\text{LC}_{canHRF+T+D}$ and LC_{canHRF} models.

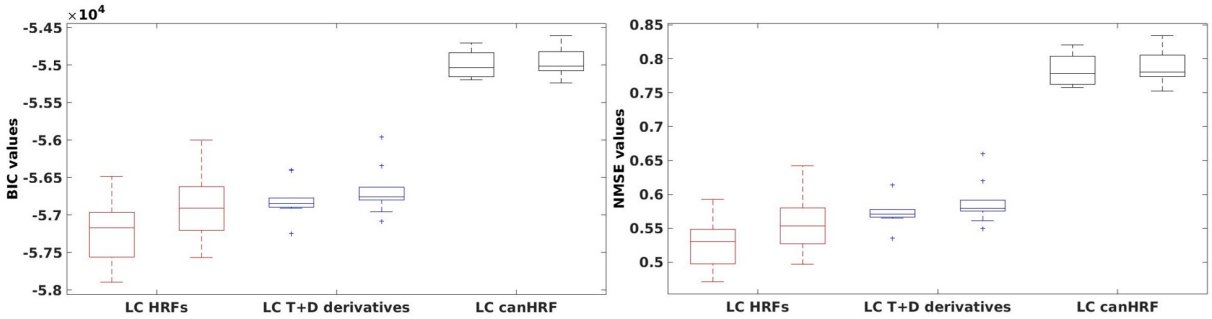

Figure 3.14: Boxplots of the prediction errors and the BIC values for Patient 2, using LS on the LC_{HRFs}, $\text{LC}_{canHRF+T+D}$ and LC_{canHRF} models.

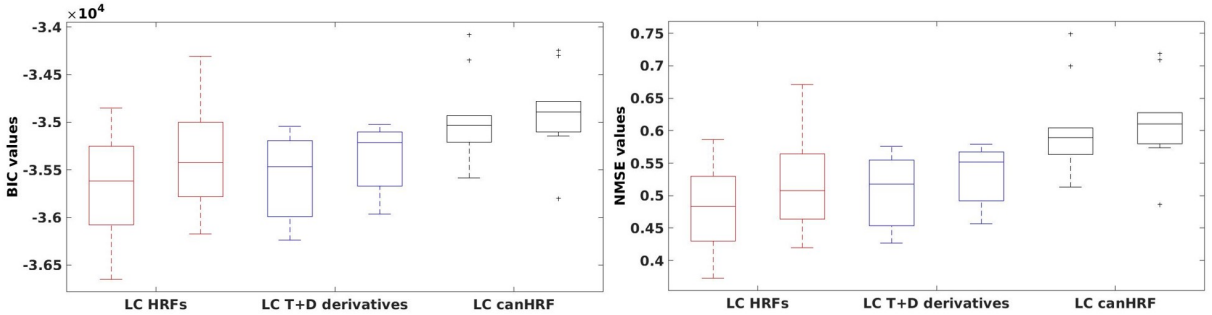

Figure 3.15: Boxplots of the prediction errors and the BIC values for Patient 3, using LS on the LC_{HRFs}, $\text{LC}_{canHRF+T+D}$ and LC_{canHRF} models.

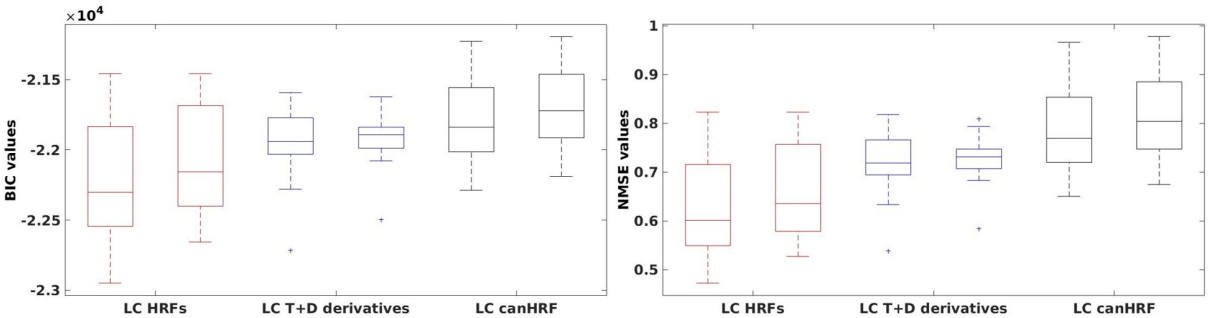

Figure 3.16: Boxplots of the prediction errors and the BIC values for Patient 4, using LS on the LC_{HRFs}, $\text{LC}_{canHRF+T+D}$ and LC_{canHRF} models.

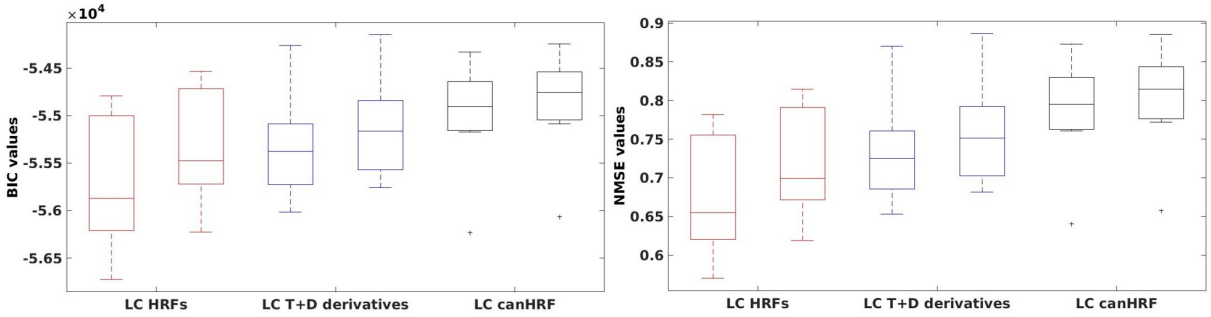

Figure 3.17: Boxplots of the prediction errors (in the right figure) and the BIC values (in the left figure), for the LC$_{HRFs}$ (red), LC$_{canHRF+T+D}$ (blue) and LC$_{canHRF}$ (black) models using LS, for Patient 5. For both selection criteria and each model, the results of the best channel (1^{st}) and the second best channel (2^{nd}) are represented on the left and right side, respectively.

Overall, a significant effect for factors models and channels was found, when the 31 channels were added. The same result was reached when only using the results from the best channels. However, after the posthoc test to compare models (multiple comparison step), only the results from Patient 3 stand out by achieving no significant differences between the BIC values of models LC$_{HRFs}$ and LC$_{canHRF+T+D}$. The EFPs derived from LC$_{HRFs}$ and LC$_{canHRF+T+D}$ for this specific patient, can be visualized in Appendix F.

An additional ANOVA test was also accomplished by introducing the results from all subjects together in the previous problem as random factor. For each criteria, significant effect were found for the factor models but not for the factor channels.

3.2.3 Ridge Regression

Concerning the results from applying Ridge on the model LC$_{HRFs}$, no significant effect were found for the factor channels, when the 31 were incorporated. However, the subjects exhibited a significant effect as before, regardless the chosen selection criterion. Afterwards, these results were compared with the ones obtained by SR and LS.

3.2.4 Comparison of methods and models

In relation to the results from the best channels across the 6 models, selected by BIC and NMSE, they were firstly compared across two methods: LS and SR. Regardless of the selection criteria that was used, SR and LS results showed no significant effect.

Then, LS, SR and Ridge were compared on the same model, LC$_{HRFs}$. The results retrieved by the three methods and the best channels showed no significant effect for the factor methods with values higher than 0.6, in case of using the BIC (similar results using NMSE). As usual, subjects presented significant differences between them. Since, within each patient, the differences between models were almost imperceptible, only one representation was acquired across all patients (BIC values were normalized within each patient before average), in figure 3.18.

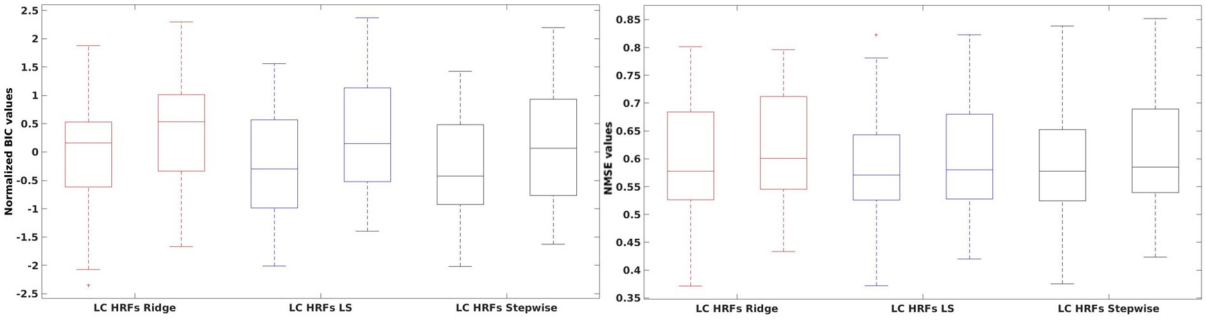

Figure 3.18: Boxplots of the prediction errors (in the right figure) and BIC values (in the left figure) by using Ridge (red), LS (blue) and SR (black) on the LC$_{HRFs}$ model, across the ten folds and the five patients. For both selection criteria and each model, the results of the best channel (1^{st}) and the second best channel (2^{nd}) are represented on the left and right side, respectively. Due to different ranges of BIC values across patients, the BIC values were normalized within each patient for visualization and comparison purposes.

The following table summarizes the two channels associated to the lowest average of BIC and NMSE values, originated by Ridge, LS and SR applied on the LC$_{HRFs}$ model from each patient.

Table 3.12: Summary table showing the number of the two channels that yielded the lowest average of the two selection criteria (NMSE and BIC) and the correspondent values acquired by using Ridge, LS and SR regression methods on LC$_{HRFs}$ model for each patient.

LC$_{HRFs}$ model		**Ridge**		**LS**		**SR**	
		Channel	**Value**	**Channel**	**Value**	**Channel**	**Value**
		BIC / NMSE	BIC ($\times 10^4$) / NMSE	BIC / NMSE	BIC ($\times 10^4$) / NMSE	BIC / NMSE	BIC ($\times 10^4$) / NMSE
Patient 1	**1^{st} channel**	2 / 2	-1.59 / 0.6006	18 / 18	-1.59 / 0.5936	18 / 18	-1.60 / 0.6029
	2^{nd} channel	18 / 18	-1.59 / 0.6153	2 / 2	-1.59 / 0.5960	2 / 2	-1.60 / 0.6025
Patient 2	**1^{st} channel**	15 /15	-5.71 / 0.5361	15 / 15	-5.71 / 0.5315	15 / 15	-5.72 / 0.5361
	2^{nd} channel	27 / 27	-5.68 / 0.5612	27 / 27	-5.68 / 0.5588	27 / 27	-5.68 / 0.5672
Patient 3	**1^{st} channel**	12 / 12	-3.56 / 0.4860	12 / 12	-3.56 / 0.4821	12 / 12	-3.57 / 0.4837
	2^{nd} channel	1 / 1	-3.54 / 0.5203	1 / 1	-3.53 / 0.5157	1 / 1	-3.54 / 0.5218
Patient 4	**1^{st} channel**	8 / 8	-2.20 / 0.6798	8 / 8	-2.21 / 0.6337	8 / 8	-2.22 / 0.6476
	2^{nd} channel	19 / 19	-2.19 / 0.7014	19 / 19	-2.20 / 0.6575	19 / 19	-2.21 / 0.6718
Patient 5	**1^{st} channel**	30 / 30	-5.56 / 0.6897	30 / 30	-5.56 / 0.6784	30 / 30	-5.57 / 0.6866
	2^{nd} channel	14 / 14	-5.53 / 0.7260	14 / 14	-5.53 / 0.7185	5 / 13	-5.53 / 0.7238

Except in Patient 1, the three regression methods selected the same best channel, across all patients and regardless of the selection criteria.

The results from applying SR or Ridge on LC$_{HRFs}$ model derived from the EEG with and without the ICA procedure, were compared. The results with and without ICA showed no significant effect, with a value higher than 0.7 for both selection criteria and methods. As consistently shown, the 31 channels showed no significant effect, contrary to subjects.

To conclude, the Ridge regularization method and the LC$_{HRFs}$ model (consistent with the previous section) were selected as the best method and model at predicting the BOLD changes, when the problem is restricted to each channel. The shrinkage property of Ridge allows a further analysis of the resulted EFPs in light of the clinical condition of each patient.

4. Discussion

The current study used a data-driven approach for the construction of an EEG prediction (EEG finger print; EFP) of BOLD-fMRI activity from patient-specific epileptic networks previously identified. Besides the EFP, the main focus was given on the determination of the regression method and model of EEG-derived regressors that best predicted the BOLD changes.

The followed framework used regression methods to estimate the EEG model, i.e., the linear combination of EEG features predictive of BOLD-fMRI: first from all electrodes and then from single electrodes. The “all electrodes approach” was motivated by recent evidence suggesting that epileptic networks are better reflected by the interaction patterns between networks of multiple electrodes (regions) (Stefan and da Silva, 2013). However, the increased dimension of the feature space motivated the use of a more parsimonious approach, which was further supported by a similar study (Meir-Hasson et al., 2014): single electrodes at each time.

The framework utilized the LC, TP and RMSF metrics, where each regressor was not coupled to an unique, pre-specified HRF-delay. In other words, a matrix of the regressors/delays representation was conceived to search for an optimal delay for each regressor, by introducing six HRFs with different delays into the model. Our results reinforce previous evidence of the HRF’s variability across subjects, electrodes or frequencies (in the LC case), by showing improved BOLD estimation when different HRFs were considered (de Munck et al., 2007). The same was concluded about the importance of including different EEG frequency bands on BOLD prediction, as related in Goense and Logothetis (2008).

The frequency range of neuronal activity and the slow hemodynamic changes is different (Wan et al., 2006), therefore their integration required a reduction of the sampling frequency of the EEG data to match the fMRI resolution. Thereby, in order to maintain the high frequency information of the EEG signals, their spectral power were submitted to the down-sampling step, instead of the raw data. In addition, the temporal resolution of the BOLD signal was increased in parallel to the same sampling frequency with the aim of reducing the loss of information in EEG.

Hereinafter, the discussion will be divided into the same structure as performed in Methods and Results.

4.1 Models with all channels

4.1.1 Ridge and Lasso

Some aspects must be pointed out in relation to the range of values of the regularization parameters. In case of Ridge regularization, the range was established in order to ensure that the DOF of the lowest and the largest λ value would correspond to the total number of predictors and to one DOF, respectively.

The drawback of Ridge is that its L_2 penalty does not allow to select predictors, including all in the final model even if with the respective shrunk coefficients near zero. Lasso appears as a combination of both shrinkage and subset selection methods. In cases with very large number of features, Lasso allows

to efficiently find the sparse model that involves a smaller subset of features. Having this in mind, it was expected that the nested CV would give preference to Lasso regularization, mainly in the LC metric which involves a higher number of highly correlated predictors. This strong correlation appeared after convolution with the HRFs. However, in the hybrid approach, the choice of Lasso did not occur (except for patient 2 in table 3.2), indicating that a model containing all predictors yields a better prediction than a sparse model. Moreover, the Ridge was selected in almost all learning sets for all patients, in the hybrid approach. This type of regularization was preferable given not only the lowest values of BIC but also prediction errors.

Other clear observation is that both DOF and the chosen regularization parameter from Ridge was overall larger than in Lasso's. This observation points to a larger amount of regularization, which is needed to deal with the fact that all predictors are different from zero.

In case of the Elastic Net Regularization cases (α =0.25, α =0.5, α =0.75) and pure Lasso, for each method, the smallest value of λ depended on the largest one (which is set to guarantee at least one non-zero parameter), because the implementation forces their ratio to meet a specific constant. It was noticed that as the parameter approached Lasso (as α increases), the magnitude of the coefficients decreases faster. Then, the largest λs decrease to achieve one non-zero parameter. Therefore, the smaller parameters also decreased. However, it was not sufficiently enough to achieve the same number of DOF offered by LS solution. Although this could have resulted in conflicts, the nested CV never picked the smallest values of the parameter as the "best" options by the BIC.

Regarding the LC$_{HRFs}$ model, the patient that yielded the largest average of the prediction errors was Patient 4. For this patient, the EEG predictors exhibited the largest absolute correlation between them (average Pearson correlation across all combinations of regressors $\approx$ 0.5665). This may be at least partially explained by the fact that the calculated DOF was the minimum, across all patients.

Concerning the RMSF$_{HRFs}$ model, an increase in prediction NMSE was observed, probably due to the reduction in the number of predictors relative to the LC$_{HRFs}$ model in all methods, across the five patients. The results from Patient 5 were the most hampered, rising from 0.0589 to 0.6706. Although the number of predictors were more limited, a larger BIC was also observed, comparing with the previous metric. Contrasting with the previous findings, Patient 4 suffered the lowest impact from this reduction, by changing its prediction error and BIC value from 0.2679 to 0.2773 and from -2.37 to -2.34 ($\times 10^4$), respectively. It can be concluded that the RMSF predictors from Patient 4 has a higher BOLD predictive power comparing with the remaining patients. For the first time, Lasso was chosen by the hybrid approach for Patient 2, by having more $\alpha = 1$ occurrences across the test folds comparing with the others, suggesting that correlated predictors were not contributing to the prediction.

The TP$_{HRFs}$ model was motivated by the better results from the LC (main difference is regarding the discretization of the power into frequency bands) and due to the fact that the first model has a reduced number of predictors (equalling the ones of the RMSF metric). This was seen as an advantage to limit the overfitting problem. Comparing the results from this metric with the two previously discussed, it was found that overall the results were in between the ones from the LC$_{HRFs}$ and RMSF$_{HRFs}$ models in terms of the BIC and error, except for Patient 4. In relation to Patient 4, the RMSF metric was favoured

by both selection criteria, exhibiting a larger average of DOF even with a weaker regularization (lower λ value).

Regardless the model considered until now, the R^2_{adj} results overall corroborated the observable BIC values, presenting a larger value (near 1) for lower BIC values. However in "SR, LAR and LS" section, the drastic reduction of the error values had a stronger impact than the fluctuations of DOF on the R^2_{adj} measures. This provided one more reason to use BIC as the selection criterion.

Overall, one can conclude that despite of the higher number of predictors offered by the LC$_{HRFs}$ model, it showed the best prediction even after taking into account the increased number of regressors, across all patients.

4.1.2 SR, LAR and LS

Comparing the average of the DOF across SR, LAR and LS methods, it was noticeable that the lowest value belonged to SR and the largest to LS, regardless of the model. Although the SR was not associated to the lowest values of error, it overall retrieved the lowest BIC values, across all methods (including the ones from the Elastic Net Regularization approach) and patients.

In the LC$_{HRFs}$ model, the value of the average prediction NMSE was drastically reduced in these three presented methods, which is possibly linked to the increase of the number of DOF. Although the resulted average of DOF was always larger in relation to Ridge, Lasso and the Elastic Net Regularization approach, the average BIC and error always presented lower values in these regression methods. Except for Patient 5, the average error from SR was at least one order of magnitude lower.

Regarding the RMSF$_{HRFs}$ model, the same was not observable because, except for Patient 1 which presented the same average of DOF as in Ridge, Lasso and the Elastic Net Regularization approach, the other patients had lower values. Although the average error remains lower, the values shared the same order of magnitude. In particular, Patient 3 presented a difference of only 0.001. For patient 1, as an exception, LAR was associated to the best prediction performance.

Concerning the TP$_{HRFs}$ model, for patients 3 and 5, the shrinkage methods presented a higher number of DOF and similar average error values. However, SR demonstrated to be once again the preferable method.

The LC$_{HRFs}$ emerged as the model associated to the best compromise between the prediction error and the number of DOF given by BIC, when predictors from all channels were added through linear regression.

4.1.3 Nested model approach

Independently of the feature dimension being considered, a great variability of the results originated by the first step of the nested model approach was found, across patients. Regardless of the model, LC$_{HRFs}$, RMSF$_{HRFs}$, or TP$_{HRFs}$, the channel associated to the lowest error was different across patients as expected, given the different spatial distribution of their epileptic networks.

Interestingly, by interpreting tables 3.7, 3.8 and 3.9, it can be stated that the "best" HRF-delay or

frequency band performed better than the “best” channel or combination HRF/frequency band. In other words, the information retrieved from all channels convolved with a specific HRF or belonging to the same frequency band, provided a lower error than the predictors estimated from an isolate channel can offer. For four of the five patients, the δ frequency was the choice when the LC_{HRFs} model was used, indicating that EEG fluctuations in this frequency band are more closely associated with BOLD in those patients. This is consistent with the low-frequency of the oscillations of the IEDs across patients, which varied between 3-10 Hz (Abreu et al., 2018b). The LC_{HRFs} and $RMSF_{HRFs}$ models shared the same results for the “best” HRF-delay except for patient 4 and regarding the “best” channel, also had the same result for Patient 5. The channel 30 for the LC_{HRFs} and $RMSF_{HRFs}$ models can also rise some questions regarding the link to the epileptogenic focus because the patient's clinical condition refers that IEDs propagated from the posterior right quadrant to frontal part.

The LC_{HRFs} and TP_{HRFs} models, had in common the “best” channel and HRF for patients 1 and 2. The fact that the channel 15 was associated to the lowest error in Patient 2 was particular interesting because it belongs to the active area (left temporal lobe) depicted by the epileptic networks determined by previous works (Abreu et al., 2018b) and described in 2.1.

Although the same was not so evident in the other patients, several questions remain on the best metrics to predict BOLD fluctuations across different patients, which may not be the same.

Lastly, the HRFs with a -5 and -6 sec delay, which are more approximated to the canonical HRF, dominated the results for the TP_{HRFs} model (except for Patient 1).

Concluding, the results from this approach motivated the use of a model containing all channels, frequencies and HRFs, even if it implies a much higher number of predictors. This conclusion is further supported by the values given by ROR, in which, even the most complex model, produced low ROR values, providing no evidence of overfitting.

4.1.4 Comparison of methods and models

The methods were also compared by means of boxplots in order to make a clearer idea not only about the mean of the measures across the ten folds, but also the median and variance.

The average prediction error and BIC values from different methods, LS, LAR, SR, Ridge, Lasso and the Elastic Net Regularization approach, described a common pattern across the five patients, although with underlying differences. Focusing on BIC results, the first three methods appeared to be the ones associated with the lowest values of this criterion at first sight, however the differences that set these apart from the other five methods depended on the considered method: LC_{HRFs}, $RMSF_{HRFs}$, or TP_{HRFs}. While in the LC metric, it is clearly visible that the first three methods outperformed the remaining five, the same was not so clear in relation to the metrics with lower number of predictors. Moreover, mainly in BIC, less variability was found across folds for a fixed method, by using the LC_{HRFs} model. The effect of the number of predictors (930 against 186) seemed to be the main difference behind these outcomes, indicating that the more the number of predictors given to the model, the higher the difference between the Elastic Net Regularization including Ridge and Lasso and the LS, SR and LAR methods, which can be explained by the greater degree of sparsity of the latter. Furthermore, mainly in

patients 2 and 5, some shrinkage methods seemed to perform better than LS or LAR.

4.1.5 Further tests using LS and SR

Some further studies were accomplished in order to investigate not only the effect of including other HRFs beyond the canonical but also the differences between the RMSF and TP metrics.

Although the variability of the HRF has been removed from the model, the TP results were consistently better than the RMSF, across all patients. This was interesting not only because it showed consistency across subjects but also because it went against the findings in epilepsy context presented by Leite et al. (2013), which showed a clear superiority of the RMSF in predicting the BOLD signal, when compared to other power-weighted metrics.

Even after the regression of the MPs from the RMSF$_{canHRF}$/ TP$_{canHRF}$ models, no significant changes were denoted: the TP metric remained the best. This step was motivated by the Abreu et al. (2018b) study, which found a higher VE by the MPs for the power-weighted metrics (comparing to synchronization metrics), mainly in TP, suggesting that even the epilepsy-related IC time-courses were still affected by motion.

The ROR approach (results in 3.11) was used to attribute a quantity to the level of overfitting presented on the prediction estimation. Although model selection based on the estimated BIC gave preference to the models with a higher number of predictors, the fact that the most complex model could better fit fluctuations that were just present in BOLD and not on the EEG predictors, raised questions regarding the extent of the overfitting problem. Concerning the two models, TP$_{HRFs}$ and TP$_{canHRF}$, the results from the first were higher than the second (except in Patient 1) which was expected given a larger number of predictors. However, the ROR results reinforced what has been observed, by obtaining lower values for the LC$_{HRFs}$ model across patients. Mainly in the “SR, LAR and LS”, the fitting adopted a strategy that consisted on constantly assigning large and symmetric coefficients’ values to two very correlated predictors in order to approximate the mentioned slight fluctuations, which was seen as overfitting. Moreover, since the attribution of the highest coefficients was not given to the predictors that separately predicted better the BOLD signal (when non-shrinkage methods were used), the control test was essential to make sure on the performance of the regression methods under study. As expected, the highest coefficient was allocated to the predictor exhibiting the highest correlation to the BOLD signal.

4.2 Models with each channel individually

4.2.1 SR and LS

Although in some models and patients, the spatial distribution of the prediction errors became difficult to interpret, in others a relation was found between the localization of the lowest errors and the focus and propagation of the epileptic networks, specially concerning the LC and TP metrics for the group of patients P4/P5 and P1/P2, respectively. For the first group of patients, the best predictive results were concentrated in the left temporal lobe: for patient 4 in the posterior occipital-temporal lobe and for patient

5 in the posterior right quadrant (although the left side was also associated to good results).

When exclusively analyzing the channel that yielded the lowest BIC value, no significant differences were detected between the LC$_{canHRF}$ and TP$_{HRFs}$, concluding that adding the mean power over the five frequency bands exclusively convolved with a canonical HRF or the total power over the spectogram but convolved with six different HRFs instead, led overall to the same results across patients. However, both models are linked to the same number of predictors. Furthermore, including the six delays or just the canonical HRF in the RMSF metric yielded no significant differences on the best channel's results. Given the lower number of predictors for the second model, this was preferable. The LC$_{HRFs}$ model exhibited the best results, as in the previous section. Exactly the same observations were registered for the LS case.

The boxplots clearly showed the results leading to the main conclusions already discussed. It can be firstly stated that the results achieved by each selection criterion led to almost identical findings. Although the results across models came closer to each other in some patients, these differences were very subtle. We can conclude that the weight that the BIC attributes to the additional DOF was not enough to produce considerable changes on the trend of the results. This can only be stated because, except for patient 5, both criteria picked the same best channels. The outcomes from the best channel for LC$_{canHRF}$ and TP$_{HRFs}$ models switched positions across patients. Regarding the RMSF$_{HRFs}$ and RMSF$_{canHRF}$ models, it was easier to infer about their similarities, as previously mentioned.

Contrary to what was observed during the integration of the predictors from all channels into the same model, when the fitting was performed at the level of the channel, no significant differences were found between LS and SR results from the best channel across models. This fact can be again interpreted in light of the number of predictors being added to the model to fit the BOLD signal. Although these two methods work in different ways, as explained in the Introduction chapter, the results came together when there was a limited number of regressors (maximum of 30 in case of model LC$_{HRFs}$) at disposal.

Moreover, as related in Rosa et al. (2010), the temporal and dispersion derivatives were able to model small differences in latency and duration of the peak in relation to the canonical HRF, in Patient 3. Since no significant differences were found between the LC$_{canHRF+T+D}$ and the LC$_{HRFs}$, which has more predictors than the first, the less complex model is favoured. As can be seen in Appendix F, the range of coefficients' values of the LC$_{canHRF+T+D}$, $[-0.6, 0.6]$, was reduced in relation to LC$_{HRFs}$, $[-4, 4]$). The increased values of the coefficients, once again is linked to a higher number of predictors. The representation of the EFPs can also be compared. The EFP derived from LC$_{HRFs}$ presented the highest coefficients in the -6 and -5 sec delays, evidencing the contributing of the delays that are the most similar to the canonical HRF. On the other hand, EFP derived from LC$_{canHRF+T+D}$ model demonstrated a clear importance of the dispersion derivative as a complementary predictor of the canonical one.

4.2.2 Ridge Regression

The Ridge regression was again computed to verify if the difference in the predictive power between this and the last two methods, would also change.

Since the prediction measures resulted from with and without the ICA additional clean-up did not show significant differences, it can be concluded that the choice of this method and LC$_{HRFs}$ model and its associated EFP were not significantly influenced by the existence of the selected artefactual ICs. However, the fact that the artefactual ICs were mainly chosen by visually inspection, arise some questions about the accuracy of the detection and removal.

This suggests again some discussion on the best method to be used: Ridge or SR. The Ridge offers some beneficial properties that were set apart due to its worse performance measured by BIC compared with the SR, when the predictors from all channels were integrated. The Ridge has two important advantages over the other (non-shrinkage) linear regression methods, whose the main one is that it penalizes the estimated coefficients in a non-arbitrarily way: by shrinking the independent variables that are highly correlated (data suffering of multicollinearity). This property is also linked to the reduction of overfitting. As has been noticed and discussed, the non-shrinkage techniques estimated coefficients whose values blew up when the features are highly correlated with each other. Still in this case, follows the second advantage which is that the penalty term guarantees a unique solution even when the rank of the predictors' matrix is lower than the number of predictors, p. One example of this can be observed in figures 2.13 and 2.16, which correspond to SR and Ridge, respectively. In the first, the values lie between $[-40, 40]$, and in the second between $[-4, 4]$.

Then, the observed EFP obtained by Ridge on the LC$_{HRFs}$ is more easily interpreted (Appendix G for Patient 2 until 5). The EFP suggested some observations on the EEG activity related to the fMRI target. Firstly, as it has been discussed throughout the work, the variability of the disease across patients was linked to the differences found between EFPs. For Patient 1, δ, θ, α (more strongly) and β bands contribute all together to the prediction of BOLD-fMRI signal modulation in the target region. These frequency bands were mainly distributed within the delays $-8, -6$ and -5 sec delays. For Patient 2, the main frequency bands were δ (more strongly), α, β and γ bands and the delays, -8 and -6 sec. For Patient 3 and 4, δ, θ, β and γ bands: however, the γ and β were the main frequency bands for Patient 3 and 4, respectively. Regarding delays, a wider range, in Patient 3, from -8 to -4 sec and $-10, -8$ sec delays and -5 sec delay for Patient 4. Lastly, for Patient 5, θ, β and γ bands (more strongly) and the delays, $-10, -8$ and -6 presented the highest coefficients. The delay -2 sec delay was overall the one that less contributed to the prediction, across all patients. This may suggest that the temporal gap between the EEG signals and the hemodynamic changes, for this specific patients, is higher than this period of time.

The fact that in Patient 2 the δ frequency was once again dominant, was in line with the clinical condition (Abreu et al., 2018b), which relates the presence of continuous spike wave discharges in slow wave sleep. The frequency of this wave is around 3 Hz which belongs to the frequency range of δ. Although Patient 5 was also characterized by this clinical condition, the correspondent EFP did not present dominant values on the δ coefficients. As in Meir-Hasson et al. (2014) study, the θ, α, β and γ seemed to be the frequency bands that all together contribute more to the prediction of BOLD-fMRI signal.

4.2.3 Comparison of methods and models

Therefore, Ridge regularization was seen as the best solution in the case of evaluating each channel individually, when no significant differences were found in relation to SR and LS. However, it is relevant to note that the other models were not tested with Ridge, therefore the same conclusion may not be applicable to the other cases.

The table 3.12 reports useful information, namely the fact that, with exception of Patient 1, the “best” channel is the same in all methods, for each patient. Although the three methods estimates the coefficients through different implementations, led to similar results.

Although the channel selected by NMSE in the LS method had demonstrated the lowest errors, when observing the BIC values, SR achieved better results.

In addition, the resulted average of the DOF for the “best” channel by Ridge and SR were also compared (although not reported in the table). It was noticed for the first method that the average value was consistently around 26, whereas for the second method, the values decreased lying within the following range, $[16.6, 20.70]$.

This discrepancy on the number of DOF in SR can be directly linked to some points of criticism that have been made to this technique. The truth is, an intensive search is developed in each step of the method, therefore it is important to consider more DOF than the number that is set in the final model. However, the number of DOF used to compute this measure were retrieved by the *DF* output from the *stepwise* Matlab function, which corresponded to the number of non-zeros in the final model (Draper and Smith, 1998).

5. Conclusion and future work

5.1 Conclusion

This work presented a new approach to EEG-fMRI data integration in the field of epilepsy. The adopted methodology allowed a deeper understanding on the applicability of some regression methods on identification of the EEG model predicting BOLD signal. Furthermore, the influence of the dimension of the feature space (i.e., number of regressors) on their performance was asserted. Besides different methods, models based on three different EEG metrics were investigated: LC, RMSF and TP.

Despite of the variability found among the five patients, when the predictors from all channels were incorporated, the SR applied on the LC$_{HRFs}$ model led, overall, to the lowest prediction error and BIC values. According to these results, the LC$_{HRFs}$ and the SR were entitled of the "best" model and method, respectively. Moreover, it was noticeable that the difference in the predictive power between the LC metric and the RMSF/TP was higher when all channels were included.

Moreover, even when the temporal variability of the predictors was reduced by exclusively using the canonical HRF in the TP and RMSF metrics, the first had demonstrated better results than the second, independently of the patient.

Some conclusions varied when the lower number of predictors were taken into account: each channel individually. The difference between the results from the six models were attenuated when the number of predictors decreased to the channel's dimension. Nevertheless, when investigating the differences of the results retrieved by the best channels of each of the six models, a significant effect was found for factor models, for each LS and SR. However, after the multiple comparison step, no significant differences were detected between LC$_{canHRF}$/TP$_{HRFs}$ and RMSF$_{HRFs}$/RMSF$_{canHRF}$, for both LS and SR. The LC$_{HRFs}$ model continued to exhibit the best prediction measures.

Focusing on the LC$_{HRFs}$ model, its variability was investigated by adding the temporal and dispersion derivatives to the canonical HRF. It was concluded that for Patient 3, no significant differences were found between this model and the LC$_{HRFs}$. Therefore, preference was given to the model associated to the lowest number of predictors in order to minimize the possible overfitting problem.

In a last attempt to mainly compare the SR, LS and the Ridge methods when dealing with a limited number of predictors, no evident differences were found between these techniques. The Ridge was proposed as the "best" method due to its preferential properties to prevent overfitting and interpretation of the derived EFP. Furthermore, except for Patient 1, the channels that yielded the lowest prediction measures were the same, regardless of the methods.

Lastly, despite slight changes across subjects, all frequency bands contributed to BOLD prediction, as seen by the EFP derived from the Ridge. Overall, the most prominent coefficients demonstrated to be localized within the following HRF-delays:-8, -6 and -5 (canonical HRF) sec.

The Meir-Hasson et al. (2014) was the only study that developed a similar approach but regarding the prediction of the amygdala's activity, by using EFP estimated by Ridge. After comparing its predictive performance with Lasso, they also concluded that the results from the first method were slight better. Our

results were consistent with their, by giving preference to Ridge. However, instead of HRF convolution, Meir-Hasson et al. (2014) performed a simple time-shift of the EEG features in relation to the average BOLD signal.

5.2 Limitations of this work

In the model identification problem addressed in this thesis, a number of sometimes highly collinear regressors was considered, which motivated the use of techniques that can deal with this problem, namely shrinkage methods of linear regression. Moreover, overfitting metrics were also computed. Although no evidence of overfitting was found, this should nevertheless be further investigated. Related to this issue is the fact that the CV procedure employed here might suffer from over-optimism due to the fact that data were previously submitted to temporal shuffling in order to eliminate auto-correlations. In the future, CV procedures for time series should be considered.

5.3 Future work

Learned self-regulation of localized brain activity in a epileptic network characteristic from a specific patient, in a long-term, could minimize symptoms of effective disturbances (Meir-Hasson et al., 2016).

The use of the fMRI technique to guide neurofeedback for self-regulation of the average BOLD signal across the epileptic network would be unpractical because it is an expensive and immobile procedure which does not allow to capture a more real and relaxed setting for the patient. On the other hand, the EEG comes up as an easier technique not only because has already been implemented in different environments but also it is relatively inexpensive. Nevertheless, as already discussed, the low spatial resolution of the EEG complicates its use on neurofeedback of targetting activity from distributed brain networks, potentially including deeper brain structures.

A great inhomogeneity of the EFP was found among subjects which was reflected by the EFPs derived by different methods and models to predict BOLD response from EEG data. Therefore, a common EFP, as performed in Meir-Hasson et al. (2016), should not be considered.

The main focus of the future work would be to demonstrate the ability to modulate brain-activity in a neurofeedback setting using the patient's EFP estimated off-line from a previous simultaneous EEG-fMRI study, as the one developed in this work. It would comprise the application of the same pre-processing steps to the on-line acquired EEG and then the generation of the next predicted BOLD signal time point by multiplying the EEG features with the EFP. Lastly, the result should be displayed as a NF signal back to the patient, whose main objective would be the amplitude attenuation of the BOLD-related epilepsy signal.

References

Abreu, R., Leal, A., and Figueiredo, P. (2018a). EEG-Informed fMRI: A Review of Data Analysis Methods. *Frontiers in Human Neuroscience*, 12:1–23.

Abreu, R., Leal, A., Lopes, F., and Figueiredo, P. (2018b). EEG synchronization measures predict epilepsy-related BOLD-fMRI fluctuations better than commonly used univariate metrics. *Clinical Neurophysiology*, 129:618–635.

Abreu, R., Leite, M., Jorge, J., Grouiller, F., Zwaag, W. V. D., Leal, A., and Figueiredo, P. (2016a). Ballistocardiogram artifact correction taking into account physiological signal preservation in simultaneous EEG-fMRI. *NeuroImage*, 135:45–63.

Abreu, R., Leite, M., Leal, A., and Figueiredo, P. (2016b). Objective selection of epilepsy-related independent components from EEG data. *Journal of Neuroscience Methods*, 258:67–78.

Aguirre, G., Zarahn, E., and D'Esposito, M. (1998). The Variability of Human, BOLD Hemodynamic Responses. *Neuroimage*, 8:360–369.

Aho, K., Derryberry, D., and Peterson, T. (2014). Model selection for ecologists: the worldview of AIC and BIC. *Ecology*, 95:631–636.

Akay, M. (2006). Wiley encyclopedia of biomedical engineering. *Editorial John Wiley & Sons, Estados Unidos da America*.

Allen, P. J., Josephs, O., and Turner, R. (2000). A Method for Removing Imaging Artifact from Continuous EEG Recorded during Functional MRI. *NeuroImage*, 12:230–239.

Arlot, S. and Celisse, A. (2009). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Bagheri, E., Dauwels, J., Dean, B. C., Waters, C. G., Westover, M. B., and Halford, J. J. (2017). Interictal epileptiform discharge characteristics underlying expert interrater agreement. *Clinical Neurophysiology*, 128(10):1994–2005.

Bagshaw, A. P., Hawco, C., Be, C.-g., Kobayashi, E., Aghakhani, Y., Pike, G. B., and Gotman, J. (2005). Analysis of the EEG – fMRI response to prolonged bursts of interictal epileptiform activity. *Neuroimage*, 24:1099–1112.

Bénar, C.-G., Gross, D. W., Wang, Y., Petre, V., and Pike, B. (2002). The BOLD Response to Interictal Epileptiform Discharges. *NeuroImage*, 1192:1182–1192.

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

Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics and Data Analysis*, 120:70–83.

Bianciardi, M., Fukunaga, M., van Gelderen, P., Horovitz, S. G., de Zwart, J. A., Shmueli, K., and Duyn, J. H. (2009). Sources of functional magnetic resonance imaging signal fluctuations in the human brain at rest: a 7 T study. *Magnetic Resonance Imaging*, 27(8):1019–1029.

Bradley, E. and Robert J., T. (1993). *An Introduction to the Bootstrap*. CRC press.

Britz, J., Ville, D. V. D., and Michel, C. M. (2010). BOLD correlates of EEG topography reveal rapid resting-state network dynamics. *NeuroImage*, 52(4):1162–1170.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33(2):261–304.

Buxton, R. B., Uludağ, K., Dubowitz, D. J., and Liu, T. T. (2004). Modeling the hemodynamic response to brain activation. *NeuroImage*, 23(1):220–233.

Carmichael, D. W., Hamandi, K., Laufs, H., Duncan, J. S., Thomas, D. L., and Lemieux, L. (2008). An investigation of the relationship between BOLD and perfusion signal changes during epileptic generalised spike wave activity. *Magnetic Resonance Imaging*, 26(7):870–873.

Chaumon, M., Bishop, D. V., and Busch, N. A. (2015). A practical guide to the selection of independent components of the electroencephalogram for artifact correction. *Journal of Neuroscience Methods*, 250:47–63.

de Munck, J. C., Gonçalves, S. I., Huijboom, L., Kuijer, J. P., Pouwels, P. J., Heethaar, R. M., and Lopes da Silva, F. H. (2007). The hemodynamic response of the alpha rhythm: An EEG/fMRI study. *NeuroImage*, 35(3):1142–1151.

de Munck, J. C., Gonçalves, S. I., Mammoliti, R., Heethaar, R. M., and Lopes da Silva, F. H. (2009). Interactions between different EEG frequency bands and their effect on alpha-fMRI correlations. *NeuroImage*, 47(1):69–76.

Debener, S., Ullsperger, M., Siegel, M., Fiehler, K., Cramon, D. Y. V., and Engel, A. K. (2005). Trial-by-Trial Coupling of Concurrent Electroencephalogram and Functional Magnetic Resonance Imaging Identifies the Dynamics of Performance Monitoring. *The journal of neuroscience*, 25(50):11730–11737.

Deichmann, R., Noth, U., and Weiskopf, N. (2010). The Basics of Functional Magnetic Resonance Imaging. In *EEG-fMRI Physiological Basis, Technique and Applications*, pages 39–62. Springer Science & Business Media.

Draper, N. and Smith, H. (1998). *Applied Regression Analysis*. John Wiley & Sons.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2):407–499.

Esposito, M. D., Zarahn, E., Aguirre, G. K., and Rypma, B. (1999). The Effect of Normal Aging on the Coupling of Neural Activity to the Bold Hemodynamic Response. *NeuroImage*, 14:6–14.

Fisher, R. S., Cross, J. H., French, J. A., Higurashi, N., Hirsch, E., Jansen, F. E., Lagae, L., Moshé, S. L., Peltola, J., Roulet Perez, E., Scheffer, I. E., and Zuberi, S. M. (2017). Operational classification of seizure types by the International League Against Epilepsy: Position Paper of the ILAE Commission for Classification and Terminology. *Epilepsia*, 58(4):522–530.

Fisher, R. S. and Scharfman, H. E. (2015). How Can We Identify Ictal and Interictal Abnormal Activity? *Adv Exp Med Biol*, 813:3–23.

Formaggio, E., Francesca, S., Bertoldo, A., Manganotti, P., Fiaschi, A., and Maria, G. (2011). Integrating EEG and fMRI in epilepsy. *NeuroImage*, 54(4):2719–2731.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1.

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M. D., and Turner, R. (1998). Event - Related fMRI : Characterizing Differential Responses. *NeuroImage*, 7(7):30–40.

Friston, K. J., Jezzard, P., and Turner, R. (1994). Analysis of Functional MRI Time-Series. *Human Brain Mapping*, 1:153–171.

Goense, J. B. M. and Logothetis, N. K. (2008). Neurophysiology of the BOLD fMRI Signal in Awake Monkeys. *Current Biology*, 18(9):631–640.

Goldman, R., Stern, J., Jr, J. E., and Cohen, M. (2002). Simultaneous EEG and fMRI of the alpha rhythm. *Neuroreport*, 13(18):2487–2492.

Gotman, J. (2008). Epileptic networks studied with EEG-fMRI. *Epilepsia*, 49(3):42–51.

Grouiller, F., Vercueil, L., Krainik, A., Segebarth, C., Kahane, P., and David, O. (2010). Characterization of the hemodynamic modes associated with interictal epileptic activity using a deformable model-based analysis of combined EEG and functional MRI recordings. *Human Brain Mapping*, 31(8):1157–1173.

Guyton, A. and Hall, J. (2006). *Textbook of Medical Physiology*. Elsevier Health Sciences.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*, volume 1. Springer series in statistics New York, NY, USA.

Herrmann, C. S. and Debener, S. (2008). Simultaneous recording of EEG and BOLD responses: A historical perspective. *International Journal of Psychophysiology*, 67(3):161–168.

Hoge, R., Atkinson, J., Gill, B., Crelier, G., and Marrett, S. (1999). Investigation of BOLD signal Dependence on Cerebral Blood Flow and Oxygen Consumption: The Deoxyhemoglobin Dilution Model. *Magnetic Resonance in Medicine*, 42:849–863.

Huettel, S. A., Singerman, J. D., and Mccarthy, G. (2001). The Effects of Aging upon the Hemodynamic Response Measured by Functional MRI. *NeuroImage*, 175:161–175.

Hutchison, D. and Mitchell, J. C. (2006). *Computational Science and its Applications - ICCSA 2006*. Springer.

Jacobs, J., Hawco, C., Kobayashi, E., Boor, R., Levan, P., Stephani, U., Siniatchkin, M., and Gotman, J. (2008). Variability of the hemodynamic response as a function of age and frequency of epileptic discharge in children with epilepsy. *Human Brain Mapping Journal*, 40(2):601–614.

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.

Kang, J. K., Be, C.-g., Al-asmi, A., Khani, Y. A., Pike, G. B., and Gotman, J. (2003). Using patient-specific hemodynamic response functions in combined EEG-fMRI studies in epilepsy. *NeuroImage*, 20:1162–1170.

Kilner, J. M., Mattout, J., Henson, R., and Friston, K. J. (2005). Hemodynamic correlates of EEG: A heuristic. *NeuroImage*, 28(1):280–286.

Kottlow, M., Jann, K., Dierks, T., and Koenig, T. (2012). Clinical Neurophysiology Increased phase synchronization during continuous face integration measured simultaneously with EEG and fMRI. *Clinical Neurophysiology*, 123(8):1536–1548.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods and Research*, 33(2):188–229.

Laufs, H., Holt, J. L., Elfont, R., Krams, M., Paul, J. S., Krakow, K., and Kleinschmidt, A. (2006). Where the BOLD signal goes when alpha EEG leaves. *NeuroImage*, 31(4):1408–1418.

Laufs, H., Kleinschmidt, A., Beyerle, A., Eger, E., Salek-Haddadi, A., Preibisch, C., and Krakow, K. (2003a). EEG-correlated fMRI of human alpha activity. *NeuroImage*, 19(4):1463–1476.

Laufs, H., Krakow, K., Sterzer, P., Eger, E., Beyerle, A., Salek-Haddadi, A., and Kleinschmidt, A. (2003b). Electroencephalographic signatures of attentional and cognitive default modes in spontaneous brain activity fluctuations at rest. *Proceedings of the National Academy of Sciences*, 100(19):11053–11058.

Leite, M., Leal, A., and Figueiredo, P. (2013). Transfer function between EEG and BOLD signals of epileptic activity. *Frontiers in Neurology*, 4(1):1–13.

Leite, Marco; Figueiredo, P. (2012). Estimation of the hemodynamic response to epileptic activity in EEG-fMRI data. In *IEEE 2nd Portuguese Meeting in Bioengineering (ENBENG)*.

Levan, P., Tyvaert, L., and Gotman, J. (2010). Modulation by EEG features of BOLD responses to interictal epileptiform discharges. *NeuroImage*, 50(1):15–26.

Lindquist, M. A., Meng Loh, J., Atlas, L. Y., and Wager, T. D. (2009). Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. *NeuroImage*, 45(1):187–198.

Liu, Z. and He, B. (2008). fMRI – EEG integrated cortical source imaging by use of time-variant spatial constraints. *NeuroImage*, 39:1198–1214.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, 453(12):869–878.

Logothetis, N. K., J, P., M, A., T, T., and A, O. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412:1–8.

Logothetis, N. K. and Wandell, B. A. (2004). Interpreting the BOLD Signal. *Annual Review of Physiology*, 66(1):735–769.

Lopes, F. (2010). EEG: origin and measurement. In Mulert, C. and Lemieux, L., editors, *EEG-fMRI Physiological Basis, Technique, and Applications*, pages 19–39. Heidelberg: Springer.

Lopes da Silva, F. (2013). EEG and MEG: Relevance to neuroscience. *Neuron*, 80(5):1112–1128.

Mantini, D., Perrucci, M. G., Del Gratta, C., Romani, G. L., and Corbetta, M. (2007). Electrophysiological signatures of resting state networks in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, 104(32):13170–5.

Meir-Hasson, Y., Keynan, J. N., Kinreich, S., Jackont, G., Cohen, A., Podlipsky-Klovatch, I., Hendler, T., and Intrator, N. (2016). One-class FMRI-inspired EEG model for self-regulation training. *PLoS ONE*, 11(5):1–17.

Meir-Hasson, Y., Kinreich, S., Podlipsky, I., Hendler, T., and Intrator, N. (2014). An EEG Finger-Print of fMRI deep regional activation. *NeuroImage*, 102(1):128–141.

Mizuhara, H., Wang, L.-q., Kobayashi, K., and Yamaguchi, Y. (2005). Long-range EEG phase synchronization during an arithmetic task indexes a coherent cortical network simultaneously measured by fMRI. *Neuroimage*, 27:553–563.

Montgomery, D., Peck, E., and Vining, G. (2006). *Introduction to Linear Regression Analysis*. John Wiley & Sons.

Moosmann, M., Ritter, P., Krastel, I., Brink, A., Thees, S., Blankenburg, F., Taskin, B., Obrig, H., and Villringer, A. (2003). Correlates of alpha rhythm in functional magnetic resonance imaging and near infrared spectroscopy. *NeuroImage*, 20:145–158.

Murphy, K. (2012). *Adaptive Computation and Machine-Learning*. The MIT ress, Massachusetts Institute of Technology 2012.

Murta, T., Chaudhary, U. J., Tierney, T. M., Dias, A., Leite, M., Carmichael, D. W., Figueiredo, P., and Lemieux, L. (2017). Phase–amplitude coupling and the BOLD signal: A simultaneous intracranial EEG (icEEG) - fMRI study in humans performing a finger-tapping task. *NeuroImage*, 146:438–451.

Murta, T., Hu, L., Tierney, T. M., Chaudhary, U. J., Walker, M. C., Carmichael, D. W., Figueiredo, P., and Lemieux, L. (2016a). A study of the electro-haemodynamic coupling using simultaneously acquired intracranial EEG and fMRI data in humans. *NeuroImage*, 142:371–380.

Murta, T., Hu, L., Tierney, T. M., Chaudhary, U. J., Walker, M. C., Carmichael, D. W., Figueiredo, P., and Lemieux, L. (2016b). A study of the electro-haemodynamic coupling using simultaneously acquired intracranial EEG and fMRI data in humans. *NeuroImage*, 142:371–380.

Murta, T., Leite, M., Carmichael, D. W., Figueiredo, P., and Lemieux, L. (2015). Electrophysiological correlates of the BOLD signal for EEG-informed fMRI. *Human Brain Mapping*, 36(1):391–414.

Niedermeyer, E., Schomer, D. L., and Lopes da Silva, F. (2011). *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Lippincott Williams & Wilkins.

Noachtar, S. and Rémi, J. (2009). The role of EEG in epilepsy : A critical review. *Epilepsy and Behavior*, 15(1):22–33.

Nunez, P. L. and Silberstein, R. B. (2000). On the Relationship of Synaptic Activity to Macroscopic Measurements: Does Co-Registration of EEG with fMRI Make Sense? *Brain Topographypography*, 13:79–96.

Ogawa, S., Lee, T.-M., Nayak, A. S., and Glynn, P. (1990). Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic Resonance in Medicine*, 14(1):68–78.

Olejniczak, P. (2006). Neurophysiologic Basis of EEG. *J Clin Neurophysiol*, 23(3):186–189.

Perronnet, L., Anatole, L., Mano, M., Clerc, M., Lotte, F., Barillot, C., and Umr, I. (2018). Learning 2-in-1 : towards integrated EEG-fMRI-neurofeedback. *bioRxiv*, pages 1–25.

Podlipsky, I., Ben-Simon, E., Hendler, T., and Intrator, N. (2012). Robust modeling based on optimized EEG bands for functional brain state inference. *Journal of Neuroscience Methods*, 203(2):377–385.

Rainer, G. and Fabrizio, E. (2010). The Added Value of EEG-fMRI in Imaging Neuroscience. In Mulert, C. and Lemieux, L., editors, *EEG fMRI: Physiological Basis, Technique, and Applications*, pages 97–112. Spinger Verlag.

Ramos, J., Chicherov, V., Herzog, M. H., and Figueiredo, P. (2018). An automatic pre-processing pipeline for EEG analysis (APP) based on robust statistics. *Clinical Neurophysiology*, 129(7):1427–1437.

Rosa, M. J., 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.

Scheeringa, R., Bastiaansen, M. C., Petersson, K. M., Oostenveld, R., Norris, D. G., and Hagoort, P. (2008). Frontal theta EEG activity correlates negatively with the default mode network in resting state. *International Journal of Psychophysiology*, 67(3):242–251.

Sitaram, R., Ros, T., Stoeckel, L., Haller, S., Scharnowski, F., Lewis-Peacock, J., Weiskopf, N., Blefari, M. L., Rana, M., Oblak, E., Birbaumer, N., and Sulzer, J. (2017). Closed-loop brain training: The science of neurofeedback. *Nature Reviews Neuroscience*, 18(2):86–100.

Sjöstrand, K., Clemmensen, L. H., Larsen, R., Einarsson, G., and Ersbøll, B. (2018). SpaSM: A MATLAB Toolbox for Sparse Statistical Modeling. *Journal of Statistical Software*, 84(10).

Stefan, H. and da Silva, F. H. L. (2013). Epileptic neuronal networks: Methods of identification and clinical relevance. *Frontiers in Neurology*, 4(8):1–15.

Tatum, W. O., Rubboli, G., Kaplan, P. W., Mirsatari, S. M., Radhakrishnan, K., Gloss, D., Caboclo, L. O., Drislane, F. W., Koutroumanidis, M., Schomer, D. L., Kastelijn-Nolst Trenite, D., Cook, M., and Beniczky, S. (2018). Clinical utility of EEG in diagnosing and monitoring epilepsy in adults. *Clinical Neurophysiology*, 129(5):1056–1082.

Thornton, R. C., Rodionov, R., Laufs, H., Vulliemoz, S., Vaudano, A., Carmichael, D., Cannadathu, S., Guye, M., Mcevoy, A., Lhatoo, S., Bartolomei, F., Chauvel, P., Diehl, B., Martino, F. D., Elwes, R. D. C., Walker, M. C., Duncan, J. S., and Lemieux, L. (2010). Imaging haemodynamic changes related to seizures : Comparison of EEG-based general linear model , independent component analysis of fMRI and intracranial EEG. *NeuroImage*, 53(1):196–205.

Tyvaert, L., LeVan, P., Grova, C., Dubeau, F., and Gotman, J. (2008). Effects of fluctuating physiological rhythms during prolonged EEG-fMRI studies. *Clinical Neurophysiology*, 119(12):2762–2774.

van Mierlo, P., Papadopoulou, M., Carrette, E., Boon, P., Vandenberghe, S., Vonck, K., and Marinazzo, D. (2014). Functional brain connectivity from EEG in epilepsy: Seizure prediction and epileptogenic focus localization. *Progress in Neurobiology*, 121:19–35.

Vulliemoz, S., Rodionov, R., Carmichael, D. W., Thornton, R., Guye, M., Lhatoo, S. D., Michel, C. M., Duncan, J. S., and Lemieux, L. (2010). Continuous EEG source imaging enhances analysis of EEG-fMRI in focal epilepsy. *NeuroImage*, 49(4):3219–3229.

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.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic-net. *Journal of the Royal Statistical Society*, 67(2):301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the ”degrees of freedom” of the lasso. *Annals of Statistics*, 35(5):2173–2192.

A. Fitting measures from original data

The following tables summarize the performance measures obtained by applying Ridge, Lasso, Elastic Net Regularization with three different values of $\alpha = [0.25, 0.5, 0.75]$, SR, LAR and LS on the LC_{HRFs}, $RMSF_{HRFs}$ and TP_{HRFs} models. This implies no division into different sets as was performed by CV.

Table A.1: Measures obtained by using an Elastic Net Regularization methodology on the LC_{HRFs} model obtained from each of the 5 patients. For the different methods, the considered measures are: BIC values, R^2_{adj}, DOF and error. The grey cells highlight the methods that yielded the lowest BIC.

LC_{HRFs}		DOF	BIC value ($\times 10^4$)	R^2_{adj}	Error (NMSE)
Patient 1	Ridge	266	-2.44	0.9843	0.0138
	Lasso 0.25	557	-1.98	0.9454	0.0408
	Lasso 0.5	398	-2.09	0.9478	0.0428
	Lasso 0.75	276	-2.15	0.9421	0.0507
	Lasso	253	-2.18	0.9466	0.0473
Patient 2	Ridge	505	-7.38	0.9233	0.0710
	Lasso 0.25	685	-6.65	0.8186	0.1632
	Lasso 0.5	519	-6.79	0.8223	0.1642
	Lasso 0.75	497	-6.87	0.8365	0.1516
	Lasso	447	-6.95	0.8472	0.1428
Patient 3	Ridge	562	-5.36	0.9905	0.0083
	Lasso 0.25	672	-4.61	0.9442	0.0473
	Lasso 0.5	613	-4.61	0.9526	0.0408
	Lasso 0.75	533	-4.61	0.9466	0.0470
	Lasso	510	-4.67	0.9509	0.0434
Patient 4	Ridge	161	-2.83	0.8443	0.1473
	Lasso 0.25	51	-2.44	0.2764	0.7112
	Lasso 0.5	68	-2.47	0.3488	0.6363
	Lasso 0.75	169	-2.54	0.5988	0.3784
	Lasso	109	-2.61	0.6410	0.3459
Patient 5	Ridge	749	-7.77	0.9670	0.0294
	Lasso 0.25	830	-6.67	0.8495	0.1322
	Lasso 0.5	614	-6.69	0.8132	0.1700
	Lasso 0.75	571	-6.78	0.8277	0.1579
	Lasso	527	-6.86	0.8405	0.1472

Table A.2: Measures obtained by using an Elastic Net Regularization methodology on the RMSF$_{HRFs}$ model obtained from each of the 5 patients. For the different methods, the considered measures are: BIC values, R^2_{adj}, DOF and error. The grey cells highlight the methods that yielded the lowest BIC.

RMSF$_{HRFs}$		DOF	BIC value ($\times 10^4$)	R^2_{adj}	Error (NMSE)
Patient 1	Ridge	160	-2.00	0.8426	0.1460
	Lasso 0.25	183	-1.93	0.7927	0.1901
	Lasso 0.5	184	-1.94	0.8080	0.1760
	Lasso 0.75	183	-1.95	0.8154	0.1693
	Lasso	182	-1.96	0.8241	0.1614
Patient 2	Ridge	160	-7.07	0.4895	0.4985
	Lasso 0.25	126	-6.31	0.4330	0.5565
	Lasso 0.5	113	-6.33	0.4407	0.5500
	Lasso 0.75	109	-6.34	0.4453	0.5458
	Lasso	112	-7.07	0.4478	0.5431
Patient 3	Ridge	163	-4.03	0.6268	0.3594
	Lasso 0.25	169	-3.98	0.5885	0.3958
	Lasso 0.5	139	-3.99	0.5757	0.4110
	Lasso 0.75	114	-3.99	0.5613	0.4274
	Lasso	117	-4.01	0.5787	0.4101
Patient 4	Ridge	156	-2.68	0.7427	0.2438
	Lasso 0.25	174	-2.58	0.6606	0.3196
	Lasso 0.5	177	-2.60	0.6826	0.2985
	Lasso 0.75	146	-2.60	0.6589	0.3244
	Lasso	119	-2.62	0.6592	0.3272
Patient 5	Ridge	131	-6.21	0.3528	0.6347
	Lasso 0.25	150	-6.18	0.3319	0.6534
	Lasso 0.5	142	-6.19	0.3417	0.6446
	Lasso 0.75	124	-6.21	0.3400	0.6480
	Lasso	98	-6.20	0.3171	0.6726

Table A.3: Measures obtained by using an Elastic Net Regularization methodology on the TP_{HRFs} model obtained from each of the 5 patients. For the different methods, the considered measures are: BIC values, R^2_{adj}, DOF and error. The grey cells highlight the methods that yielded the lowest BIC.

TP_{HRFs}		DOF	BIC value ($\times 10^4$)	R^2_{adj}	Error (NMSE)
Patient 1	Ridge	149	-2.05	0.8691	0.1220
	Lasso 0.25	175	-1.89	0.7495	0.2307
	Lasso 0.5	173	-1.93	0.7867	0.1966
	Lasso 0.75	156	-1.94	0.7840	0.2007
	Lasso	136	-1.95	0.7813	0.2052
Patient 2	Ridge	123	-6.43	0.5254	0.4660
	Lasso 0.25	178	-6.37	0.5124	0.4749
	Lasso 0.5	153	-6.38	0.5039	0.4850
	Lasso 0.75	149	-6.39	0.5077	0.4815
	Lasso	125	-6.41	0.5085	0.4825
Patient 3	Ridge	136	-4.04	0.6249	0.3635
	Lasso 0.25	173	-3.99	0.6017	0.3827
	Lasso 0.5	161	-4	0.6062	0.3795
	Lasso 0.75	107	-4.01	0.5720	0.4176
	Lasso	124	-4.03	0.6071	0.3819
Patient 4	Ridge	82	-2.67	0.6875	0.3038
	Lasso 0.25	24	-2.43	0.1945	0.7990
	Lasso 0.5	52	-2.45	0.3007	0.6871
	Lasso 0.75	61	-2.49	0.3961	0.5915
	Lasso	71	-2.52	0.4779	0.5097
Patient 5	Ridge	164	-6.38	0.5129	0.4753
	Lasso 0.25	108	-6.29	0.4080	0.5826
	Lasso 0.5	108	-6.32	0.4279	0.5630
	Lasso 0.75	104	-6.33	0.4335	0.5579
	Lasso	99	-6.34	0.4377	0.5541

Table A.4: Measures obtained by using SR, LAR and LS on the LC_{HRFs} model obtained from each of the 5 patients. The considered measures are: BIC values, R^2_{adj}, DOF and error. The grey cells highlight the methods that yielded the lowest BIC.

LC_{HRFs}		DOF	BIC value (10^4)	R^2_{adj}	Error (NMSE)
Patient 1	**SR**	597	-3.37	0.9999	6.46×10^{-5}
	LAR	751	-3.28	0.9999	5.79×10^{-5}
	LS	930	-3.26	0.9999	3.43×10^{-5}
Patient 2	**SR**	828	-8.91	0.9944	0.0049
	LAR	930	-8.84	0.9944	0.0048
	LS	930	-8.84	0.9944	0.0048
Patient 3	**SR**	820	-6.91	0.9998	1.52×10^{-4}
	LAR	930	-6.82	0.9998	1.49×10^{-4}
	LS	930	-6.82	0.9998	1.49×10^{-4}
Patient 4	**SR**	615	-4.68	0.9999	8.53×10^{-5}
	LAR	690	-4.41	0.9989	1.72×10^{-4}
	LS	930	-4.58	0.9999	5.27×10^{-5}
Patient 5	**SR**	786	-8.52	0.9895	0.0093
	LAR	921	-8.41	0.9895	0.0091
	LS	930	-8.41	0.9895	0.0089

Table A.5: Measures obtained by using SR, LAR and LS on the $RMSF_{HRFs}$ model obtained from each of the 5 patients. The considered measures are: BIC values, R^2_{adj}, DOF and error. The grey cells highlight the methods that yielded the lowest BIC.

$RMSF_{HRFs}$		DOF	BIC value (10^4)	R^2_{adj}	Error (NMSE)
Patient 1	**SR**	156	-2.06	0.8755	0.1157
	LAR	186	-2.05	0.8835	0.1067
	LS	186	-2.05	0.8835	0.1067
Patient 2	**SR**	80	-6.38	0.4594	0.5343
	LAR	177	-6.36	0.5008	0.4862
	LS	186	-6.35	0.5026	0.4838
Patient 3	**SR**	119	-4.05	0.6191	0.3706
	LAR	178	-4.03	0.6400	0.3455
	LS	186	-4.03	0.6410	0.3439
Patient 4	**SR**	138	-2.71	0.7579	0.2309
	LAR	184	-2.70	0.7766	0.2096
	LS	186	-2.70	0.7764	0.2096
Patient 5	**SR**	99	-6.26	0.3743	0.6166
	LAR	148	-6.22	0.3707	0.6156
	LS	186	-6.21	0.3893	0.5940

Table A.6: Measures obtained by using SR, LAR and LS on the TP_{HRFs} model obtained from each of the 5 patients. The considered measures are: BIC values, R^2_{adj}, DOF and error. The grey cells highlight the methods that yielded the lowest BIC.

TP_{HRFs}		DOF	BIC value (10^4)	R^2_{adj}	Error (NMSE)
Patient 1	**SR**	155	-2.15	0.9178	0.0764
	LAR	186	-2.13	0.9185	0.0746
	LS	186	-2.13	0.9185	0.0746
Patient 2	**SR**	142	-6.49	0.5721	0.4190
	LAR	183	-6.46	0.5748	0.4138
	LS	186	-6.46	0.5752	0.4132
Patient 3	**SR**	111	-4.08	0.6363	0.3546
	LAR	161	-4.04	0.6406	0.3463
	LS	186	-4.03	0.6465	0.3386
Patient 4	**SR**	149	-2.81	0.8308	0.1607
	LAR	186	-2.79	0.8355	0.1542
	LS	186	-2.79	0.8355	0.1542
Patient 5	**SR**	137	-6.42	0.5263	0.4641
	LAR	182	-6.40	0.5324	0.4552
	LS	186	-6.40	0.5340	0.4533

B. Measures from CV procedure: before MPs correction

Table B.1: CV measures by using SR and LS on the RMSF$_{canHRF}$ and TP$_{canHRF}$ models obtained from each of the 5 patients, before MPs correction. For the two different methods, the average results across the ten folds of the following measures are presented: BIC values, R^2_{adj}, DOF and error. The grey cells highlight the methods that yielded the lowest BIC.

RMSF$_{canHRF}$ and TP$_{canHRF}$ models			Average DOF	Average BIC ($\times 10^4$)	Average R^2_{adj}	Average error (NMSE)
Patient 1	**SR**	TP	18.2	-1.5826	0.3426	0.6519
		RMSF	21.6	-1.5612	0.2767	0.7162
	LS	TP	31.0	-1.578	0.3505	0.604
		RMSF	31.0	-1.55	0.2704	0.7193
Patient 2	**SR**	TP	20.7	-5.5023	0.2345	0.7631
		RMSF	24.4	-5.4406	0.1581	0.8388
	LS	TP	31.0	-5.4942	0.2347	0.7618
		RMSF	31.0	-5.44	0.1579	0.8383
Patient 3	**SR**	TP	28.2	-3.468	0.3809	0.6151
		RMSF	23.3	-3.4298	0.3131	0.6833
	LS	TP	31.0	-3.466	0.3820	0.6137
		RMSF	31.0	-3.42	0.3137	0.6815
Patient 4	**SR**	TP	24.4	-2.232	0.3848	0.6102
		RMSF	19.9	-2.1670	0.2109	0.7838
	LS	TP	31.0	-2.229	0.3872	0.6064
		RMSF	31.0	-2.16	0.2038	0.7879
Patient 5	**SR**	TP	19.8	-5.5181	0.2515	0.7464
		RMSF	16.4	-5.4307	0.1355	0.8624
	LS	TP	31.0	-5.5094	0.2527	0.7439
		RMSF	31.0	-5.4209	0.1372	0.8589

C. Results from Control test

Table C.1: CV measures by using SR and LS models obtained from phase randomization of the predictor with the highest correlation with BOLD signal from RMSF$_{canHRF}$ and TP$_{canHRF}$ models, for each of the 5 patients. For the two different methods, the average results across the ten folds of the following measures are presented: BIC values, R^2_{adj}, DOF and error. The grey cells highlight the methods that yielded the lowest BIC.

RMSF$_{canHRF}$ and TP$_{canHRF}$ models			Average DOF	Average BIC ($\times 10^4$)	Average R^2_{adj}	Average error (NMSE)
Patient 1	**SR**	TP	21.2	-1.6089	0.4302	0.5644
		RMSF	17.8	-1.5597	0.2617	0.7324
	LS	TP	31.0	-1.6083	0.4453	0.5459
		RMSF	31.0	-1.5570	0.2836	0.7064
Patient 2	**SR**	TP	22.5	-5.4543	0.1739	0.8231
		RMSF	20.6	-5.3984	0.0930	0.9040
	LS	TP	31.0	-5.4491	0.1754	0.8204
		RMSF	31.0	-5.3902	0.0934	0.9020
Patient 3	**SR**	TP	21.0	-3.4541	0.3496	0.6449
		RMSF	20.9	-3.4136	0.2803	0.7159
	LS	TP	31.0	-3.4429	0.3437	0.6512
		RMSF	31.0	-3.4132	0.2924	0.7020
Patient 4	**SR**	TP	19.9	-2.1816	0.2443	0.7501
		RMSF	21.5	-2.1739	0.2319	0.7619
	LS	TP	31.0	-2.1745	0.2482	0.7431
		RMSF	31.0	-2.1639	0.2225	0.7685
Patient 5	**SR**	TP	21.3	-5.5076	0.2373	0.7606
		RMSF	21.5	-5.4204	0.1269	0.8700
	LS	TP	31.0	-5.4963	0.2356	0.7605
		RMSF	31.0	-5.4110	0.1232	0.8723

D. Topographies of error using LS

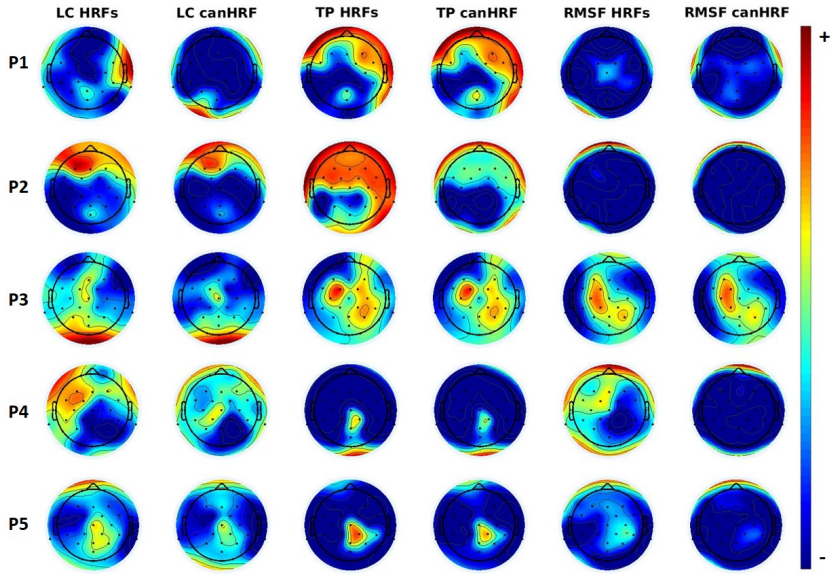

Figure D.1: The prediction error at each channel by using the LS on each of the six models, for each patient. The different rows correspond to the different patients from patient 1 (P1) until patient 5 (P5). The different columns represent the six different models of predictors. In order to have a clear visualization of the errors distribution for each model and patient, the color scale was bounded between the minimum and maximum for each representation. Therefore, no values were displayed in the limits of the colorbar.

E. Boxplots of NMSE and BIC using LS

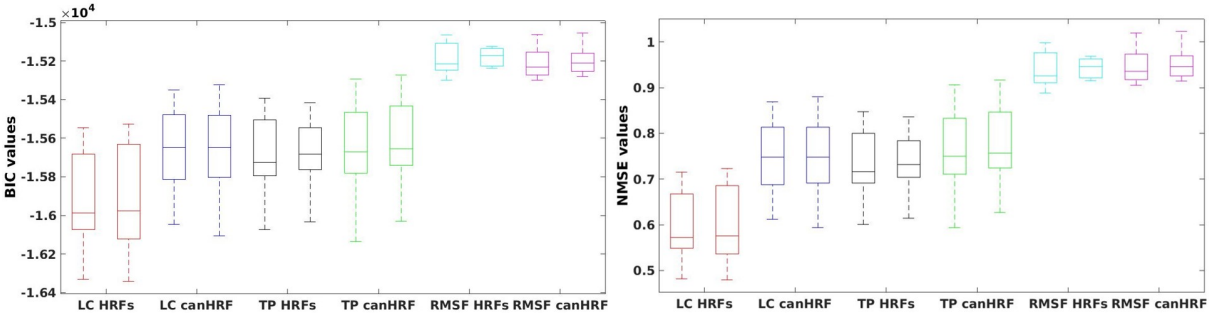

Figure E.1: Boxplots of the BIC and NMSE across folds from the two channels that yielded the lowest prediction error (NMSE) and the lowest BIC values, separately, for each of the six models, by using LS. This representation is derived from Patient 1.

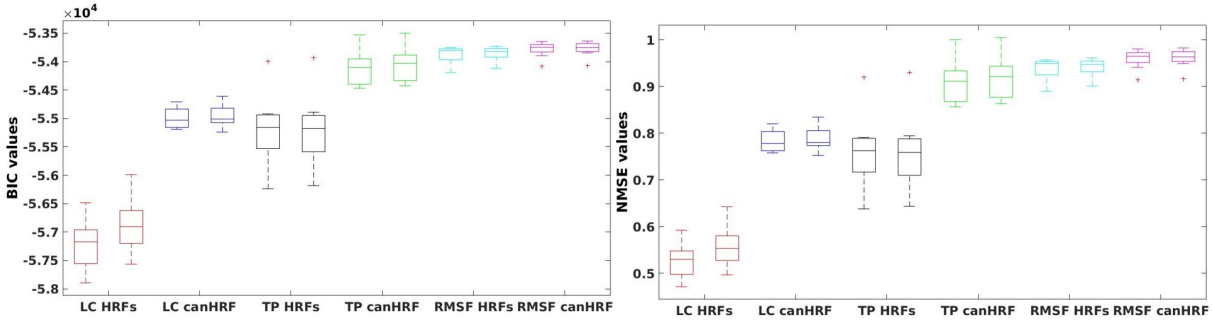

Figure E.2: Boxplots of the BIC and NMSE across folds from the two channels that yielded the lowest prediction error (NMSE) and the lowest BIC values, separately, for each of the six models, by using LS. This representation is derived from Patient 2.

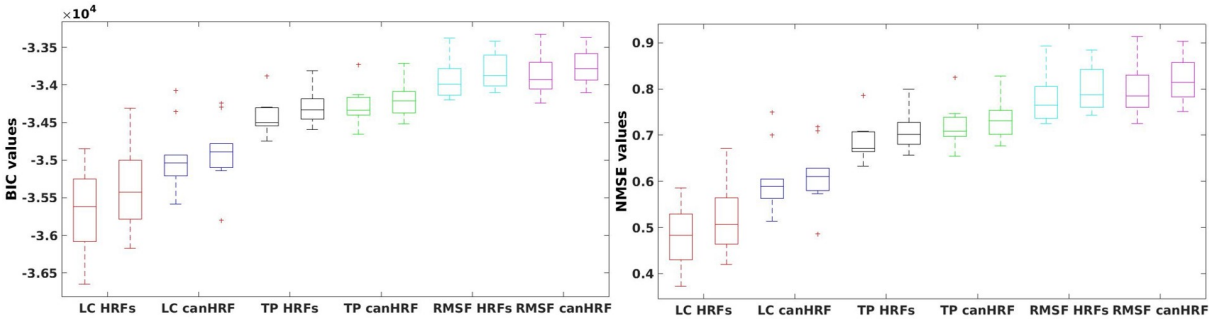

Figure E.3: Boxplots of the BIC and NMSE across folds from the two channels that yielded the lowest prediction error (NMSE) and the lowest BIC values, separately, for each of the six models, by using LS. This representation is derived from Patient 3.

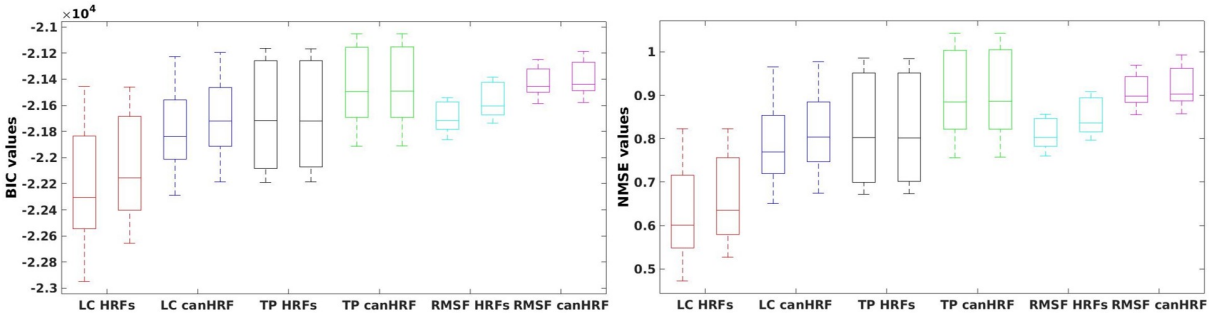

Figure E.4: Boxplots of the BIC and NMSE across folds from the two channels that yielded the lowest prediction error (NMSE) and the lowest BIC values, separately, for each of the six models, by using LS. This representation is derived from Patient 4.

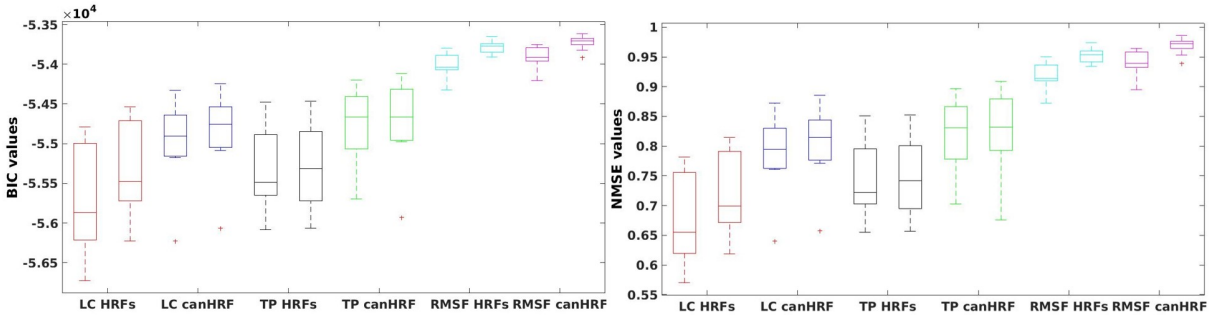

Figure E.5: Boxplots of the prediction error (NMSE) represented on the right figure and the BIC values on the left figure derived from applying LS. This representation is derived from Patient 5. Both criteria were calculated on the 10 test sets. The six different colors represent the six different models under study: each model has two boxes presenting the results from the channel that yielded the lowest value (1^{st}) on the left side and the second lowest value (2^{nd}) on the right side, for each selection criteria.

F. EFPs derived from LS for Patient 3: LC$_{HRFs}$ and LC$_{canHRF+T+D}$

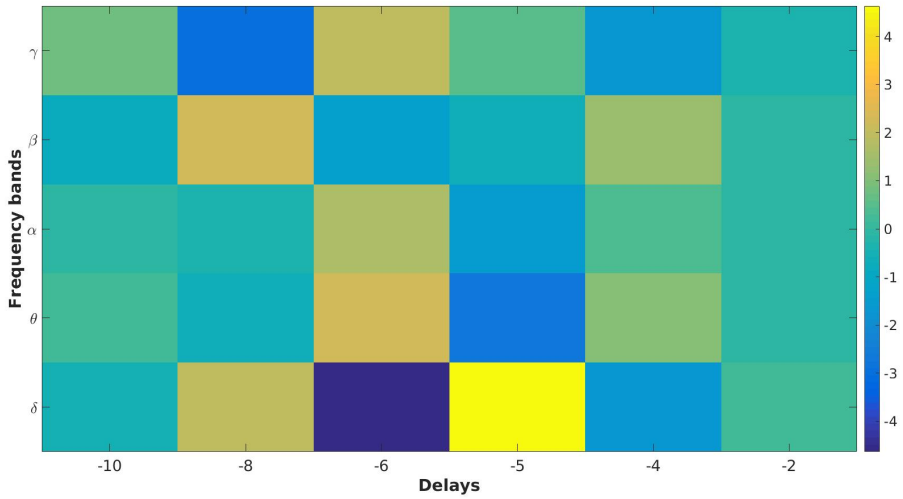

Figure F.1: The estimated model coefficients that best predict the BOLD activity (EFP) using LS on the LC$_{HRFs}$ model from the best channel which was selected by the prediction error criterion, derived from Patient 3.

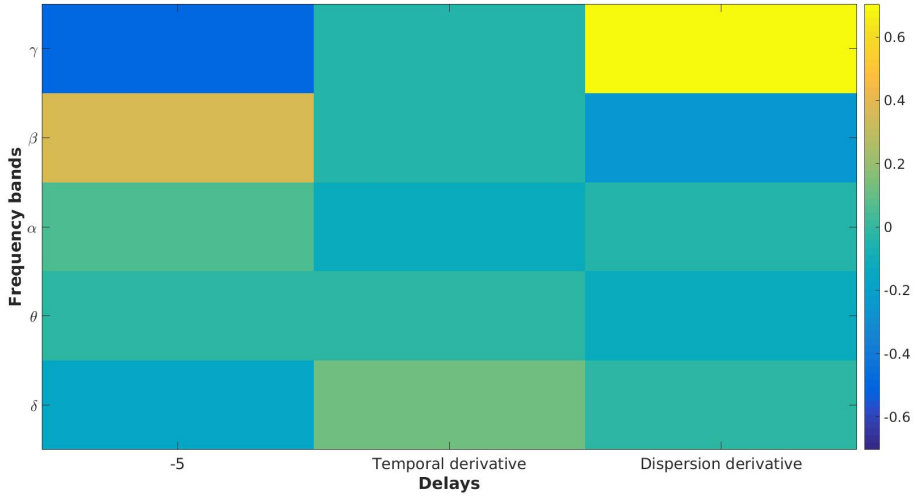

Figure F.2: The estimated model coefficients that best predict the BOLD activity (EFP) using LS on the LC$_{canHRF+T+D}$ model from the best channel which was selected by the prediction error criterion, derived from Patient 3.

G. EFPs derived from Ridge

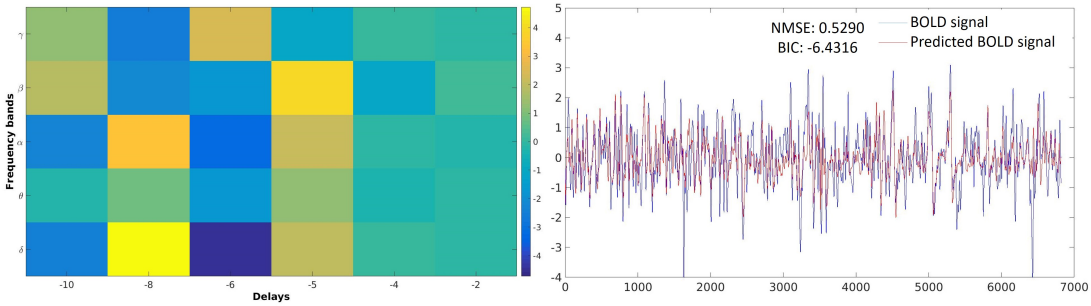

Figure G.1: The estimated model coefficients (left figure) that best predict the BOLD activity (EFP) using Ridge on the LC$_{HRFs}$ predictors from the best channel which was selected by the prediction error criterion, derived from Patient 2. The values of BIC ($\times 10^4$) and error resulted from the fitting are also represented.

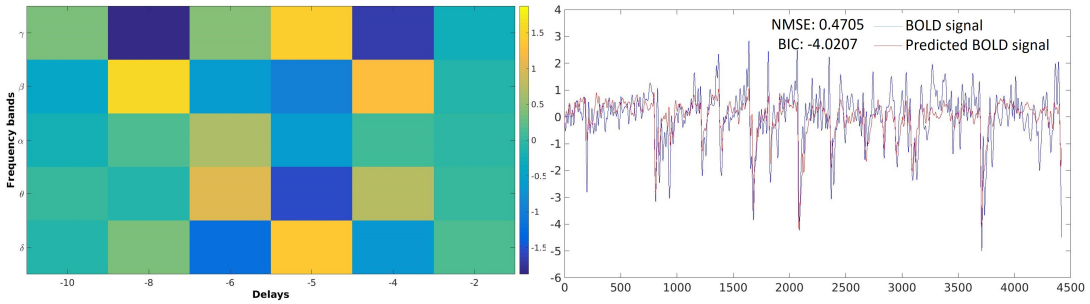

Figure G.2: The estimated model coefficients (left figure) that best predict the BOLD activity (EFP) using Ridge on the LC$_{HRFs}$ predictors from the best channel which was selected by the prediction error criterion, derived from Patient 3. The values of BIC ($\times 10^4$) and error resulted from the fitting are also represented.

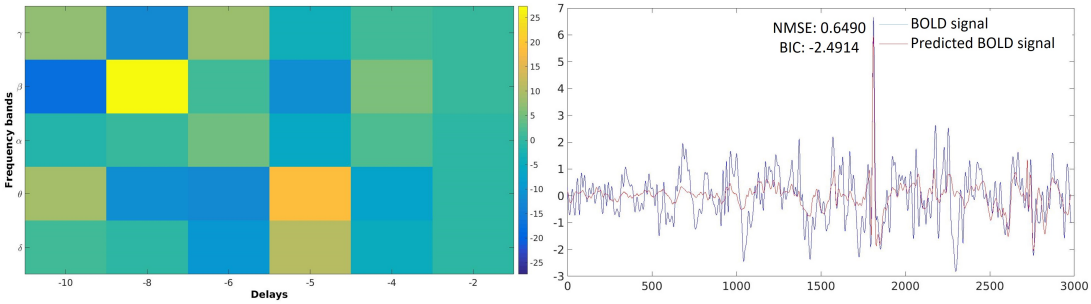

Figure G.3: The estimated model coefficients (left figure) that best predict the BOLD activity (EFP) using Ridge on the LC$_{HRFs}$ predictors from the best channel which was selected by the prediction error criterion, derived from Patient 4. The values of BIC ($\times 10^4$) and error resulted from the fitting are also represented.

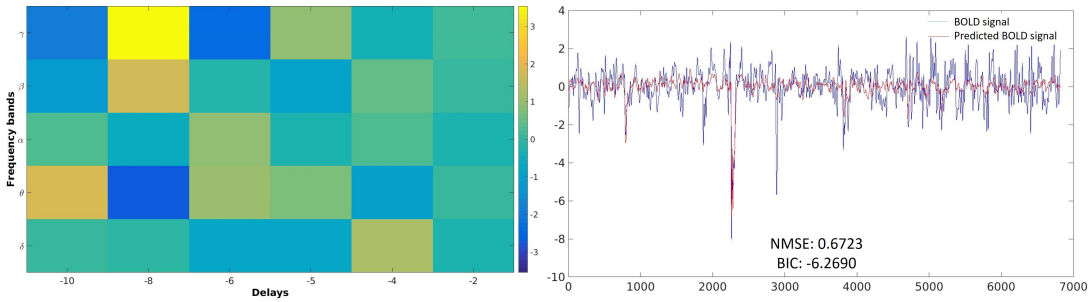

Figure G.4: The estimated model coefficients (left figure) that best predict the BOLD activity (EFP) using Ridge on the LC$_{HRFs}$ predictors from the best channel which was selected by the prediction error criterion, derived from Patient 5. The values of BIC ($\times 10^4$) and error resulted from the fitting are also represented.