UNIVERSIDAD DE GRANADA

TESIS DOCTORAL

PROGRAMA OFICIAL DE DOCTORADO EN TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES

# Supervised-learning methods for pattern recognition in fMRI data for the identification of informative brain regions in psychological contexts


DOCTORANDO

JUAN ELOY ARCO MARTÍN

DIRECTORES

MARÍA RUZ CÁMARA Y JAVIER RAMÍREZ PÉREZ DE INESTROSA

FEBRERO, 2019

Editor: Universidad de Granada. Tesis Doctorales
Autor: Juan Eloy Arco Martín
ISBN: 978-84-1306-172-6
URI: http://hdl.handle.net/10481/55521

# DECLARACIÓN

El doctorando Juan Eloy Arco Martín y los directores de tesis María Ruz Cámara y Javier Ramírez Pérez de Inestrosa garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de los directores de tesis. Hasta donde nuestro conocimiento alcanza, en la realización del trabajo se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

*Granada, Febrero de 2019*

Directores de la Tesis:

María Ruz Cámara

Javier Ramírez Pérez de Inestrosa

Doctorando:

Juan Eloy Arco Martín

# ABSTRACT

In the last years, there has been an exponential increase in the use of multivariate analysis in neuroimaging data. This has led to a new perspective in the study of brain function, which lets identify the brain areas involved in a cognitive function and the characterization of the patterns of information associated with them. This kind of analyses are based on a classification framework that increases the sensitivity of classic univariate approaches and detects subtle changes in neural activity of different experimental conditions.

This thesis focuses on three main aspects of the classification pipeline. First, an optimal estimation of the activation patterns to isolate the contribution of each experimental condition to the hemodynamic response. We have employed a Least-Squares Separate method, an approach that iteratively fits a new model for each trial in the experiment. Each model has two regressors: one for the target trial and another one for the rest. We show for the first time that this method can effectively identify the activity associated with different events even though they belong to different cognitive processes with a substantial difference in their duration. Second, a classification algorithm based on Multiple Kernel Learning (MKL). This approach combines different brain regions from an atlas to build the classification function and identifies the relevance of each region in a specific psychological context. We propose a modification of this algorithm and employ an L2-regularization to avoid the sparsity that L1-regularization entails. This approach yields simultaneously the high sensitivity of multivariate methods and the directionality of univariate approaches. Third, we employed a non-parametric approach based on permutation testing that computes more accurately the significance thresholds for each voxel in the brain. Hence, it takes into account the differences in maximum decoding accuracy that different brain regions have, enhancing sensitivity and detecting true informative that otherwise would not be marked as significant. Our results show that differences between parametric and non-parametric methods can be much larger when trying to detect subtle changes in neural activity.

We have shown that information derived from just above-chance accuracies should

not be underestimated if these accuracies are significant. We should expect high values of accuracy when contrasting stimuli with large perceptual differences. If these differences are minimal, the accuracy will be small. Future research should continue the development of machine learning methods especially optimized for Cognitive Neuroscience, where obtaining large accuracies is not of first interest but to provide a clear interpretation of the mathematics behind these algorithms.

# Agradecimientos

En primer lugar me gustaría dar las gracias a mis dos directores de tesis, María y Javier. Gracias, María, por haberme dado la oportunidad de realizar esta tesis y por todo el tiempo dedicado a corregir mis muchos errores. Gracias, Javier, por haber seguido guiándome en este camino que empezamos hace ya 6 años. Sin vuestro apoyo, este trabajo no habría llegado a buen puerto.

Tampoco puedo olvidarme de Carlos García Puntonet, la persona que me inició en este mundo de la investigación y del que siempre obtengo buenos consejos.

Gracias a todo el grupo de Neurociencia Cognitiva, por acogerme con tanto cariño y hacer sentir a un ingeniero un poco más psicólogo cada día.

Thanks to Janaina for letting me be part of the Machine Learning and Neuroimaging Lab at UCL. Thanks to João, Anil, Maria and Michele for being my partners during my stay in London.

Gracias a todos los amigos que me han apoyado durante este periodo tan complicado. Las tardes de Fifa, las partidas de bolos o las noches de risas y charlas también han sido claves para que esta tesis pudiera llevarse a cabo.

A mis compaños del CIMCYC, por tantas horas juntos tanto de trabajo como de diversión. Gracias a Javi por esas "Tardes de socorro y desahogo" y esas reuniones donde se mezclaban ciencia y cualquier otra cosa. Gracias a Fernando por los buenos ratos y por sus magníficas habilidades culinarias. A Maïka, por esas conversaciones donde el misterio y la imaginación siempre tenían cabida. A Enzo, por los viernes por la tarde y esas rutas por todos los rincones de la ciudad. Gracias a todas las personas que han hecho que esos días de trabajo fueran mucho mejores: Luis, Vicente, Itsaso, Daniel, sin olvidar a Juan Carlos y sus arreglos informáticos sin los que esta tesis nunca podría haberse acabado.

Gracias a todos los componentes del despacho 345, porque han sido como una familia para mi. A Nuria, porque aunque su estancia fue breve, siempre he podido contar con ella. A Sonia, por todos esos consejos y ánimos en los momentos complicados. A Carlos, por haber sido siempre un referente para mi, y conseguir hacer fácil lo difícil. Gracias

a Alberto por hacernos siempre sonreir, aunque sea siendo el protagonista de historias reales o ficticias. Gracias a David por equilibrar la ratio ingenieros/psicólogas, además de fomentar el partido de pádel de después el trabajo. A Chema, por su entusiasmo y dedicación que seguro tendrán su recompensa.

También quiero dar las gracias a PAP, las personas con las que he compartido más tiempo a lo largo de este trayecto. A Ana Paqui (faltaría más), porque tanto en los buenos como en los malos momentos siempre he sentido su apoyo, y ha sido una verdadera amiga a la que acudir. A Paloma, por idénticas razones, por esa disposición que siempre tiene a ayudar a los demás y resolver problemas que a los demás nos parecen imposibles.

Muchas gracias a Paco y Mari Carmen por todo ese ánimo, cariño y apoyo incondicional que me habéis demostrado en todo este tiempo.

A Virginia, mi compañera de viaje, la persona con más paciencia para aguantarme. Gracias por estar siempre conmigo, ayudándome a levantarme con cada golpe y haciéndome ver que juntos las cosas son mucho más sencillas.

Y por último, y más importante, a mis padres, las personas que más admiro, por haberme enseñado desde siempre que con trabajo y esfuerzo hasta las cosas más difíciles pueden realizarse. Gracias por haber hecho todo lo que estaba en vuestra mano porque yo tuviera la mejor educación posible. Gracias, porque sin vosotros no sería la persona que soy hoy.

# CONTENTS

# List of Figures

# List of Tables

# Acronyms

**ABLA** Atlas-Based Local Averaging

**AC** Anterior Commissure

**ACC** Anterior Cingulate Cortex

**ADHD** Attention Deficit Hyperactivity Disorder

**BASC** Bootstrap Analysis of Stable Clusters

**BCI** Brain Computer Interfaces

**CAD** Computer-Aided Diagnosis

**CBF** Cerebral Blood Flow

**CBV** Cerebral Blood Volume

**$CMRO_2$** Cerebral Metabolic Rate of Oxygen

**CT** Computerized Tomography

**CV** Cross Validation

**DCT** Discrete Cosine Transform

**dMRI** Difussion Magnetic Resonance Imaging

**EC** Euler Characteristic

**EEG** Electroencephalography

**EET** Epoxyeicosatrienoic

**ERP** Event-related potentials

**FaOR** Faces and Objects Representations dataset

**FDR** False Discovery Rate

**FI** Faces Identificaton dataset

**FIR** Finite Impulse Response

**fMRI** Functional Magnetic Resonance Imaging

**FN** False Negative

**FOV** Field Of View

**FP** False Positive

**FSL** FMRIB Software Library

**FWE** Family-wise Error

**FWHM** Full Width at Half Maximum

**GLM** General Linear Model

**HDR** Hemodynamic Response

**HRE** Hemodynamic Response Efficiency

**HRF** Hemodynamic Response Function

**ICA** Independent Component Analysis

**IFG** Inferior Frontal Gyrus

**ISI** Interstimulus Interval

**LFP** Local Field Potential

**LORO** Leave-One-Run-Out

**LSA** Least-Squares All

**LSS** Least-Squares Separate

**LSU** Least-Squares Unitary

**LTI** Linear Time Invariant

**MEG** Magnetoencephalography

**MFC** Medial Frontal Cortex

**MFG** Medial Frontal Gyrus

**MKL** Multiple Kernel Learning

**ML** Machine Learning

**MNI** Montreal Neurological Institute

**MR** Magnetic Resonance

**MTG** Middle Temporal Gyrus

**MUA** Multiple Unit Activity

**MVPA** Multi-Variate Pattern Analysis

**NMF** Nonnegative Matrices Factorization

**PC** Posterior Commissure

**PCA** Principal Component Analysis

**PET** Positron Emission Tomography

**PFC** Prefrontal Cortex

**PLS** Partial Least Squares

**RBF** Radial Basis Function

**RFT** Random Field Theory

**RMF** Resonancia Magnética Funcional

**ROI** Region of Interest

**SFG** Superior Frontal Gyrus

**SL** Searchlight

**SMA** Supplementary Motor Area

**sMRI** Structural Magnetic Resonance Imaging

**SPECT** Single Positron Emission Computed Tomography

**SPGR** Spoiled Gradient Recall

**SPM** Statistical Parametric Mapping

**SVM** Support Vector Machines

**TDT** The Decoding Toolbox

**TE** Echo Time

**TFCE** Threshold-Free Cluster Enhancement

**TN** True Negative

**TR** Repetition Time

**TP** True Positive

**UG** Ultimatum Game dataset

**vmPFC** Ventromedial Prefontral Cortex

# Part I

# Introduction

CHAPTER 1

# INTRODUCTION

## 1.1 Motivation

In recent years, the use of functional magnetic resonance imaging (fMRI) has increased exponentially. This technique has become an essential tool for a better understanding of the human brain due to its spatial precision. This method provides an indirect measure of the changes produced in neural activity when participants are performing tasks of different nature or while they are at rest. Most fMRI studies measure Blood-Oxygen-Level-Dependent (BOLD) contrast, which reflects regional changes in cerebral blood flow (CBF), cerebral blood volume (CBV) and blood oxygenation. These three vascular responses reflect local increases in neural activity (Logothetis, 2003, 2014), which study the role of different brain regions in specific functions.

Classic univariate analysis focuses on the fMRI signal in individual voxels or the average signal in a brain region, evaluating if there are differences between two experimental conditions (Henson and Friston, 2007; Worsley and Friston, 1995). Thus, these analyses can identify the regions that are involved in a certain cognitive function. However, there are situations in which the average activity is the same for different experimental conditions, but they differ in how the activity pattern is distributed (Mur et al., 2009). Multivariate pattern analysis (MVPA, Haxby et al., 2001; Mahmoudi et al., 2012) evaluates the distribution of BOLD activation across multiple voxels rather than the average level of activity. This increases the sensitivity compared to univariate approaches, identifying minimum changes that otherwise would not be identified (Coutanche et al., 2011;

Haynes and Rees, 2005; Kamitami and Tong, 2005; Norman et al., 2006).

MVPA combines multidimensional analysis with machine learning (ML), where a classifier tries to learn the relationship between spatial activity patterns and experimental conditions. This kind of analysis has been previously used in neuroimaging studies, as an automatic tool for diagnosis of different neurological and psychiatric disorders (Adeli et al., 2017; Arco et al., 2015; Choi et al., 2017; Del Gaizo et al., 2017; Khedher et al., 2017; Plant et al., 2010; Salvatore et al., 2014). In this clinical context, the aim of the classification framework is to obtain a model that predicts with the largest accuracy if a person suffers from a certain disease or not. However, in Cognitive Neuroscience the goal is to study different functions of the human brain. Thus, obtaining the largest accuracy is not of first interest, but to identify the regions involved in a certain cognitive process. This task is considerably hard since neural differences between experimental conditions are often subtle.

One of the main problems to overcome in fMRI studies is related to the sluggishness of the BOLD signal. This signal peaks on average at 6-8 seconds after the beginning of the neural activity, and lasts for as long 16 seconds before returning to baseline (Logothetis, 2003, 2004; Zaidi et al., 2018). However, the time between different stimuli in psychological experiments, known as inter-stimulus-interval (ISI), is usually much shorter (González-García et al., 2017; Palenciano et al., 2018; Visconti di Oleggio Castello et al., 2017). This means that the signal acquired by the MR scanner at each specific moment is not due to the neural activity produced by a stimulus, but to a combination of previous ones. The shorter the interval between adjacent stimuli, the more difficult to estimate their contribution to the hemodynamic signal (Abdulrahman and Henson, 2016; Mumford et al., 2014; Turner et al., 2012) and the subsequent classification.

The large difficulty that fMRI classification entails is related to how information is distributed in the underlying neural populations. Despite the high spatial resolution that MR scanners offer, a voxel contains heterogeneous neuronal populations. Nonetheless, the classifier is able to identify these neuronal subpopulations and decode information from its distribution. Haynes and Rees (2005) demonstrated that voxels of the visual cortex contain information regarding the orientation of different stimuli. However, the distribution is different in each voxel, so that the classifier was able to exploit these differences to decode orientation from multi-voxel patterns. The extreme difficulty that this kind of classifications entails results in low classification accuracies, specially when evaluating a brain region that is involved in different cognitive processes. Previous studies have shown the important role of the prefrontal cortex (PFC) in cognitive control

(Botvinick et al., 2015), learning (Botvinick et al., 2015), multitasking (Badre and Frank, 2011) and decision making (Frank and Badre, 2011). Bhandari et al. (2018) carried out a systematic meta-analysis of fMRI studies that tried to decode information from the PFC. They showed that the median classification accuracy for these studies was 55.4%, which is much lower than in other brain regions such as the visual cortex (66.6%) or temporal cortices (71%). Thus, low accuracies are not uncommon in Cognitive Neuroscience.

MVPA can be used to detect small spatial differences. The activation patterns and how information is organized can highly vary between different people. In most experiments, it is not possible to successfully train the classifier with data from a person and test its performance with data from another one (Bode and Haynes, 2009; Coutanche and Thompson-Schill, 2012; De Martino et al., 2008; Kamitami and Tong, 2005). For this reason, classification in Cognitive Neuroscience is usually performed at the subject level.

The classification is performed for each subject, but the subsequent results are then evaluated to confirm if they are consistent across subjects. To do so, it is crucial to assess if results are statistically significant. Parametric approaches have been used in hundreds of fMRI studies (Forman et al., 1995; Friston et al., 1994; Hayasaka and Nichols, 2003; Misaki et al., 2010; Woo et al., 2014). They assume that accuracies follow a Gaussian distribution, something that is not always met. Besides, significance thresholds for cluster-correction are usually computed assuming that MR images have a certain amount of smoothness (Stelzer et al., 2013; Woo et al., 2014). For this reason, recent studies have claimed that their use for testing accuracies significance can be inappropriate (Eklund et al., 2016; Stelzer et al., 2013). As an alternative, non-parametric approaches based on permutation testing partially address these issues. They compute empirically the distribution rather than assuming that it follows a specific shape. Different methods have been proposed, and they mainly differ in how permutations are computed and the way results are combined at the group level (Smith and Nichols, 2009; Stelzer et al., 2013).

Since spatial information is of great relevance in Cognitive Neuroscience, the classification framework has to adapt to this identification goal. A classifier that yields a large accuracy is not necessarily useful from the neuroscience perspective if the discriminative regions can not be identified. Thus, all methods in the classification framework must preserve spatial information, which considerably limits the approaches that can be used. In ML, it is common to employ feature extraction methods to reduce the dimensionality of the input data. These approaches apply a geometric transformation to project data from the original space to a new one. Despite this operation can enhance the classifier

performance, spatial information is missed, which invalidates its use in psychological contexts unless it is applied in combination with other operations that preserve spatial information. Hence, all the intermediate stages in the classification framework must be chosen according to the main goal of this process: to identify the voxels that lead to these classification results.

## 1.2 Aims and objectives

This thesis aims to compare different methods for estimation of the activity patterns, in addition to different classification approaches for an optimal identification of informative brain regions. We carry out this comparison in multiple contexts: complex scenarios where each trial contains several events of different duration, event-related designs with an intermediate ISI, block-designs, etc. We will take two main approaches: isolating the contribution to the BOLD signal of each event in all these scenarios and proposing new classification alternatives that provide additional information about the brain regions involved in a psychological process. Therefore, we can define the following objectives:

1. Compare performance of different classifiers and find the one that identifies the maximum number of informative brain regions when the overlapping between adjacent events is high.

2. Evaluate different approaches and identify the best choice to optimize the estimation of the activity patterns in the aforementioned scenarios.

3. Evaluate different methods for assessing statistical significance in these contexts, controlling for false positives while offering the largest sensitivity.

4. Develop new strategies to maximize the detection of informative brain regions, providing new details about how information is distributed in these regions.

We have work in different studies to achieve the goals previously mentioned :

1. A study with the algorithm most commonly employed in current fMRI experiments: Searchlight. We evaluate the variability of the informative regions depending on factors such as the dimensionality of the input data, the classifier used and its hyperparameters.

2. A study for an optimum estimation of activation patterns. Specifically, we employ three alternatives. In the first one, all trials of the same type within each run are collapsed into one single regressor. In the second, a different regressor is used for each trial. The last one relies on an iterative process in which the activity due to each trial is computed in a separate model. Hence, each model has two regressors: one for the target trial and another one for the rest.

3. The comparison of three statistical methods (a parametric one and two non-parametric) for assessing statistical significance of the resulting accuracies of the classification process.

4. The development of different classification methods based on atlas that provide an alternative measure of the classifier performance. This measure is based on the weights of a linear classifier, and provides new insights into how information is distributed.

## 1.3 Organization of the thesis

This thesis is organized in three different parts, each of which subdivided in several chapters. In Part I, we introduce the motivation and main aims of this work (Chapter 1). Then, we provide an overview of the physical basis of fMRI (Chapter 2). Chapter 3 summarizes the most common preprocessing steps used in fMRI studies. Then, we examine the state of the art in multivariate analyses in neuroscience (Chapter 4). The last two chapters of this part focus on the different classification methods employed in neuroscience (Chapter 5) and the approaches for assessing statistical significance in fMRI studies (Chapter 6).

Part II refers to each of the solutions provided in this thesis. Chapter 7 evaluates the influence of different parameters in Searchlight performance. Moreover, we test the effect of different algorithms in the classification results. In Chapter 8, we provide the optimum method for estimating the activation patterns previous to the classification step in a context with high collinearity and overlap. Besides, we test parametric and non-parametric methods for assessing significance of classifier accuracy. Chapter 9 offers different atlas-based methods as an alternative to Searchlight for the identification of informative brain regions.

Finally, in Part III we provide a general discussion of the results obtained in this thesis and conclusions about the implications that these results have in the neuroscience

context.

CHAPTER 2

# FUNCTIONAL MAGNETIC RESONANCE IMAGING

This chapter provides an overview of fMRI, a technique that measures changes in neural activity in the brain and computes images according to these variations. Specifically, most fMRI experiments are based on the blood oxygen-level dependent (BOLD) signal, a measure of the ratio of oxygenated to deoxygenated hemoglobin which is associated with the activity/inactivity of a brain region during a specific cognitive function. This technique has been used in thousands of studies mainly due to its large spatial resolution, a crucial feature when the main goal is to draw inferences about cognitive brain states. The different sections of this chapter summarize the physical basics of fMRI, the mathematical framework behind image reconstruction, and the vascular and metabolic mechanisms underlying the fMRI hemodynamic response.

## 2.1 Introduction

Scientists have always aimed at enhancing their knowledge about the human brain. In the Renaissance, the philosopher Rene Descartes speculated about the role of the pineal gland and its involvement in imagination and memory tasks (Stanford Encyclopedia of Philosophy, 2013). He assumed that this structure must be of great importance since it was the only part of the nervous system not divided into two bilateral and symmetrical regions. Besides, its proximity to the ventricles would imply that this gland controlled the flow of cerebrospinal fluid, discarding that the cortex had any relevant function (Kolb and Whishaw, 2003). In the nineteenth century, Franz Josef Gall and Johann

Casper Suprzheim discovered that the cortex is a functioning part of the brain and not just a covering for the pineal body (Kolb and Whishaw, 2003). Following this finding, they established that different brain regions would be related to different aspects of the human mind. Further, they assumed that the amount of brain tissue and, subsequently, the shape of the skull would provide information about the function of the underlying brain region (Huettel et al., 2009), which was called phrenology.

In the 1880s, experiments conducted by Angelo Mosso suggested that there is a relationship between blood flow and brain function (Mosso, 1881). However, this was not characterized until a few years later, when Roy and Sherrington (1890) concluded that there are local variations in the blood flow related to the neural activity of a specific region. This finding supposed a large breakthrough for brain researchers and paved the way for the studies conducted in the next decades. However, brain-imaging did not begin until the early 20th-century, with the appearance of the pneumoencephalography (Dobbs, 2005). This method consisted in draining the cerebrospinal fluid from the brain and replacing it with air to enhance the quality of an X-ray image (Raichle, 2009). This technique was improved by the angiography, which consists in injecting a radio-opaque substance in the bloodstream that absorbs X-rays, producing a high-contrasted image of blood vessels.

The modern era of the brain imaging started in the early 1970s with the development of the computerized tomography (CT). This approach, considerably less invasive than previous techniques, is based on the fact that tissues absorb more or less amount of radiation depending on their density. These differences are then employed to build a contrasted image of the brain (Kolb and Whishaw, 2003). CT is a very useful anatomical tool for diagnosis of neurological disorders, but unable to provide any information about the functionality of the brain. The first functional imaging technique was the Positron Emission Tomography (PET). In this technique, a small amount of radioactive elements is injected to the bloodstream. These elements can trace the path followed by glucose molecules to the brain. Mental processes consume glucose, so that areas with a larger activity use more blood than those with a low activity, and thus they have more radioactive molecules. This substance releases positrons which collide with electrons in the brain, emitting photons which are detected by a camera. Finally, an MR image is reconstructed from variations of photons of different brain areas (Raichle, 1983). In 1990s, fMRI revolutionized the cognitive neurosciences because it is a non-invasive technique (no injection of radioactive substances is required) which offers a large spatial resolution compared to PET. The physical basis of MR is detailed in next section.

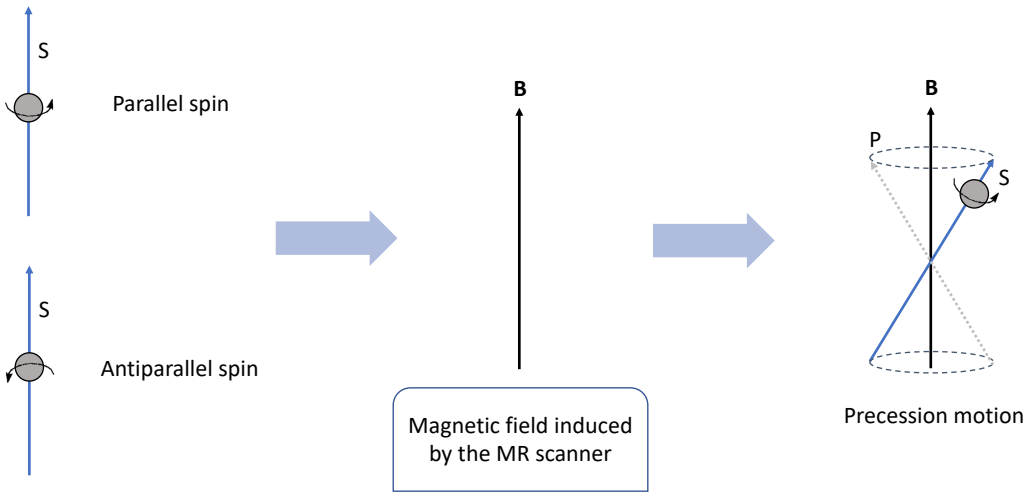


**Figure 2.1:** Scheme of the magnetization of a spin system.

## 2.2 Basic principles of Magnetic Resonance

Magnetic resonance (MR) is based on a property of the nuclei of some elements which depends on the number of protons that they have. This property, known as spin, consists in that the proton spins on its own axis, generating an electrical current on its surface which induces a magnetic moment. This produces a local magnetic field with their corresponding north and south poles. One of the elements that meets this physical principle is hydrogen, which is present in most atoms of the human body. For this reason, MR is focused on inducing changes in the magnetic field of the hydrogen atoms. Depending on the sense of rotation around its axis, the proton is in parallel or antiparallel state. The first one is the most common and corresponds to a low-energy state, whereas the antiparallel sense corresponds to a high-energy state. Each proton is rotating on its axis without a fixed direction, but applying an external magnetic field makes them to align with this field. However, this alignment is not abruptly done but it requires a certain time where protons start a gyroscopic motion known as precession (see Figure 2.1). This consists in a combination of the original spin rotation around its axis and another around the direction of the axis of the magnetic field. The angular frequency (also known as Larmor frequency) around the field direction is given by:

$$\omega_0 = \gamma B_0, \tag{2.1}$$

where $\gamma$ is the gyromagnetic ratio of hydrogen and $B_0$ is the magnetic field. At this point, all the protons would be aligned to the magnetic field, but not all of them would be in parallel state (high-energy level). It is possible to transfer energy to those protons to

make them change to a high-level state. To do so, a brief radio pulse is applied at the Larmor frequency, generating a second magnetic field. Once this field is turned off, the protons tend to return to their natural state, namely, the initial low-energy state in which they previously were (Grover et al., 2015). Again, this is not an abrupt procedure: protons that spin synchronized start to relax progressively in an exponential decay described by two constants: $T_1$ (longitudinal relaxation) and $T_2$ (transversal relaxation). $T_1$ is the time required for recovering 63% of longitudinal magnetization after the RF pulse is turned off. $T_2$ is the time required for losing 63% of transversal magnetization. The amount of longitudinal magnetization varies along time as follows:

$$M_z(t) = M_0\,(1 - e^{-\frac{t}{T_1}}) \tag{2.2}$$

where $M_0$ is the original longitudinal magnetization. Thus, longitudinal magnetization increases over time until reaching its original value. However, the total magnetization is constant, so that a rise in the longitudinal component results in a decrease in transverse magnetization, as follows:

$$M_{xy}(t) = M_0\, e^{-\frac{t}{T_2}} \tag{2.3}$$

Figure 2.2 illustrates how longitudinal and transverse magnetization vary along time. In this example, $T_1 = 5s$ and $T_2 = 3.5s$. The value of these times have a large importance because image acquisition highly depends on the moment of the relaxation process measured. Besides, different brain tissues have different T1 and T2 times, so that it is crucial to select the proper moment to acquire the image and obtain an optimal intensity in the desired brain tissue (white/gray matter or cerebrospinal fluid).

## 2.3 Image acquisition

We have described in previous sections the need of emitting an RF pulse to modify the state of hydrogen protons. It is also necessary to detect these modifications, especially when that pulse finishes. To do so, MR scanners have RF receptors in addition to RF transmitters. Receptors can measure the magnetic field in each position of the brain by identifying the corresponding precession frequency. Thus, it is possible to select a specific slice by exciting only the protons associated with that position. However, all the protons of the brain spin at the same frequency since an only RF pulse is emitted. MR scanners use an additional spatial magnetic field gradient to make the spins of

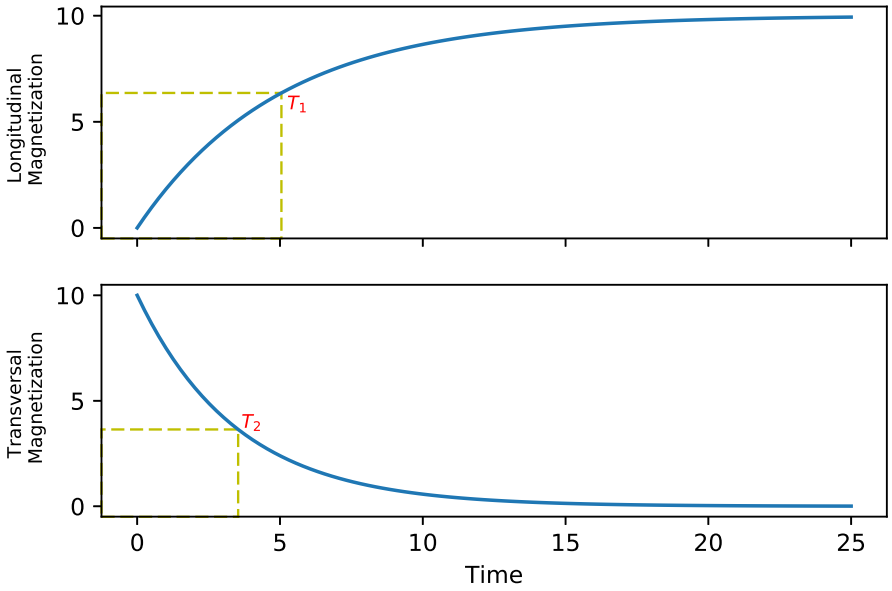


**Figure 2.2:** $T_1$ represents the time needed to recover 63% of the original longitudinal magnetization (top). $T_2$ represents the time required to lose 63% of the original transversal magnetization (bottom).

different positions precess at different speeds, which enables the selection of a particular slice by applying an RF pulse that contains the band of frequencies of the spins within the desired slice. Besides, it is possible to select different values of slice thickness by modifying the frequency range of the RF pulse.

Similarly to the slice selection, all the protons in a slice precess at the same frequency. However, different locations should have different frequencies to be able to identify each one of them and form the MR image. The solution is exactly the same as in slice selection: a gradient is applied so that precession frequencies within a slice change over space, in a procedure called frequency encoding (Huettel et al., 2009). This would be enough to reconstruct the information associated with two dimensions, but with reference to three-dimensional spaces, the frequency would be the same over the third dimension, difficulting again the signal acquisition. To enable a correct image formation, MR scanners apply an additional gradient for a short amount of time that modifies the phase of the spins (the angle of precession). This operation is known as phase encoding. Hence, employing frequency and phase encoding leads to an unique combination of frequency and phase for each voxel of the brain.

Mathematically, the magnetization for a given location $(x, y)$ in a slice with a thickness $\Delta z$ is given by:

$$\mathbf{M}(x,y) = \int_{z_0-\frac{\Delta z}{2}}^{z_0+\frac{\Delta z}{2}} \mathbf{M_{xy0}}(x,y,z)dz \tag{2.4}$$

Once the slice is selected, the magnetization only depends on $z$ direction, so that Equation 2.4 can be expressed as follows:

$$S(t) = \int_x \int_y \mathbf{M}(x,y) e^{-i\gamma \int_0^t (G_x(\tau)x + G_y(\tau)y)d\tau} dxdy \tag{2.5}$$

where $\mathbf{M}$ is the magnetic field along the $z$-direction for a thickness $\Delta z$, $G_x$ and $G_y$ are the strength of the gradient field in each specific direction. It is not straightforward to interpret this equation, and it would be extremely difficult to obtain an MR image from this information without an additional processing. Software from MR scanners relies on different mathematical tools to transform the initial data acquired (frequency and phase of each specific voxel along time) into the resulting MR image. First, there is a change in the coordinates system to a new one known as $k$-space, which collects all the spatial frequencies needed for later generating the image. Second, an inverse Fourier transform is applied. This method can reconstruct a signal from a series of simpler functions in a spatial-frequency domain, transforming the $k$-space to image-space data. The $k$-space trajectories can be defined as:

$$k_x(t) = \frac{\gamma}{2\pi} \int_0^t G_x(\tau)d\tau \tag{2.6a}$$

$$k_y(t) = \frac{\gamma}{2\pi} \int_0^t G_y(\tau)d\tau \tag{2.6b}$$

where $G_x$ and $G_y$ are the gradients in direction $x$ and $y$, respectively (see Huettel et al., 2009 for a detailed explanation of the mathematics behind these operations). We can rewrite Equation 2.5 taking into account the transformation to the $k-$space as follows:

$$S(t) = \int_x \int_y M(x,y) e^{-i2\pi k_x(t)x} e^{-i2\pi k_y(t)y} dxdy \tag{2.7}$$

Once gradients are applied for phase and frequency encoding, the signal is in $k-$space format. This change in the coordinates system implies that the new space has units in spatial frequency (1/distance) instead of distance. Then, a large number of $k$ samples is collected to obtain a proper reconstruction of the image. This process is known as filling the $k$-space. Interestingly, manipulating the strength of the gradients over time changes the path followed by the spatial acquisition, which results in different sequences. For example, the gradient-echo sequence (widely used in anatomical imaging sequences)

employs the Cartesian method, in which the image is acquired line by line, from left to right and top to bottom. There are other alternatives with different patterns of acquisition, such as radial (Pruessmann et al., 2001), spiral (Tan and Meyer, 2009) or zig-zag (Breuer et al., 2008), but the Cartesian method is traditionally the approach most used in brain imaging.

### 2.3.1 The EPI sequence

We have previously explained that MR builds an image from a series of discrete signal samples. The time between consecutive excitation pulses, known as repetition time (TR), depends on the magnetization parameter of the target tissues. For example, to form a T1-weighted image, the TR is required to be two or three times longer than the T1 relaxation of the specific tissue. The average T1 value is typically on the order of a second (1.331 according to Wansapura et al., 1999), so that the TR should be 4 seconds. For each TR, one line of the $k$-space is collected. For filling the $k$-space, it is necessary to extract data from all lines of the $k$-space. To do so, the imaging time required is equal to the product of the TR and the number of phase-encoding steps. If the number of phase-encoding steps is 128 and the $TR = 4s$, the imaging time is about 512 seconds, or more than 8.5 minutes. Due to its sluggishness in the signal acquisition, the use of this sequence for functional imaging is automatically discarded.

Echo-planar imaging (EPI) is capable of significantly decreasing the MR imaging time since multiple lines of the $k$-space are acquired after a single RF pulse. For this reason, this procedure is known as single-shot echo-planar imaging. Like in the standard spin-echo sequence explained in section 2.3, the acquisition process begins with an RF pulse. Then, a particular slice is selected. In this case, the gradient applied for frequency encoding oscillates rapidly from a positive to negative amplitude, resulting in a train of gradient echoes (Poustchi-Amin et al., 2001). Each oscillation of the frequency-encoding corresponds to one line of the $k$-space. Specifically, data from all the $k$-space of an MR slice is collected after a single RF excitation pulse. This means that individual MR slices can be collected in a time ranging from 50 to 100 ms, providing a fast acquisition that has been employed in a vast number of fMRI studies (Coutanche and Thompson-Schill, 2012; LaConte et al., 2005; Misaki et al., 2010; Mur et al., 2010; Visconti di Oleggio Castello et al., 2017).

In early EPI acquisition, the filling of the $k$-space follows a zig-zag pattern: the frequency-encoding is performed along the horizontal axis, whereas phase-encoding is performed along the vertical one (see bottom of Figure 2.3). This different trajectory can

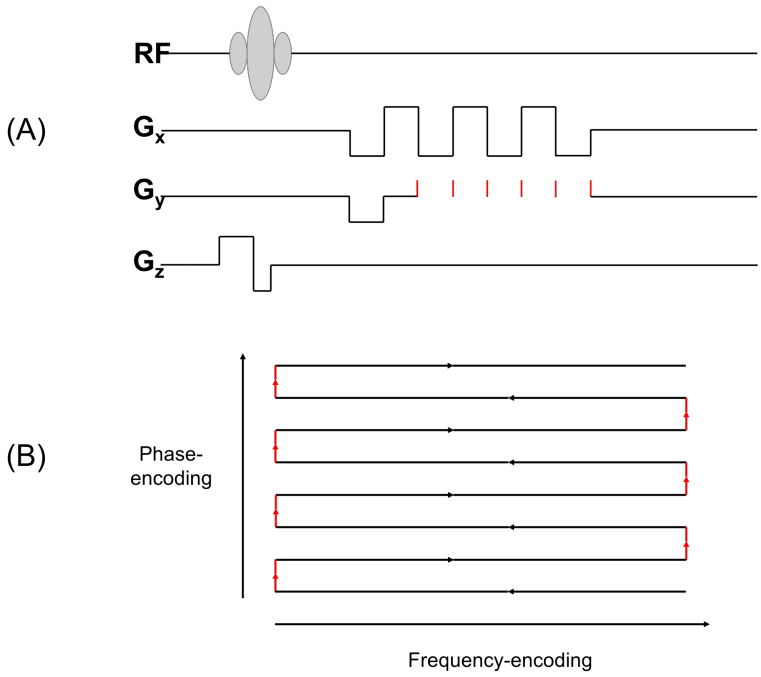


**Figure 2.3:** (A) Echo-planar imaging. Data from multiple lines are collected within each TR period. (B) In this case, the $k$-space is filled following a zig-zag trajectory. Each axis corresponds to a different way of encoding: frequency-encoding in the horizontal one and phase-encoding in the vertical. Red lines in (A) and (B) correspond to phase encoding.

lead to some artifacts when applying Fourier transformation. The most common results from imperfections in the magnetic field used to collect the images, which can lead to signal loss and geometric distortions (Huettel et al., 2009). To address these issues, multi-shot echo-planar imaging can be used (McKinnon, 1993). As its name suggests, the $k$-space is now collected by multiple shots. Specifically, the $k$-space is divided into several portions and each one of them is acquired by one shot. Each portion is termed segment, and the shots are repeated until the information contained in all segments is collected (Cohen, 2000). Multi-shot EPI provides a high improvement in the signal to noise ratio compared to single-shot. According to Cohen (2000), the SNR when using two shots is about 40% better than when the single shot is used (see Rzedzian (1987) for a deeper understanding of the different EPI sequences and their main features).

## 2.4 The BOLD response

The appearance of fMRI has revolutionized the study of the human brain since it provides a powerful tool to explore the neural basis of human behavior. This approach relies on changes in oxygen concentration (known as BOLD signal) and the subsequent

modification of the magnetic properties that these changes entail. Measuring these magnetic alterations is the basis of fMRI. This section provides a detailed description of the BOLD signal and the underlying biological mechanisms that simultaneously concurr.

### 2.4.1 Vascular and metabolic factors

BOLD contrast was first described by Ogawa and Lee (1990). This signal is related to hemoglobin, a protein that transports oxygen in the red blood cells. Depending on the amount of oxygen, hemoglobin has different magnetic properties and different names. When cerebral tissues are oxygenated, oxyhemoglobin is diamagnetic. In contrast, oxygen consumption changes deoxyhemoglobin properties to paramagnetic, which produces alterations in the magnetic fields that MR detects. Oxygen consumption increases after a rise in the neural activity so that BOLD contrast provides an indirect measure of neural activity. However, there are multiple parameters influenced by the increase of the neural activity, and they belong to a process known as neurovascular coupling (D'Esposito et al., 2003). In summary, neural activity changes three main aspects: the cerebral blood flow (CBF), the cerebral blood volume (CBV) and the cerebral metabolic rate of oxygen ($CMRO_2$). Specifically, the concentration of deoxyhemoglobin decreases when the CBF increases, but increases after a rise in CBV or $CMRO_2$ (Simon and Buxton, 2015).

It is widely accepted that changes in the BOLD follow a specific temporal course termed hemodynamic response (HDR, Lindquist et al., 2009). Figure 2.4 shows the variation of this signal along time, from which we can differentiate between three main phases. After the initial stimulation, neural activity increases the deoxygenated hemoglobin in the capillaries of the regions involved, which makes BOLD to decrease below baseline. This initial phase is referred to as the initial dip or the early response and there has been a large controversy surrounding it. Hu and Yacoub (2012) carried out a wide revision of the different historical findings of the initial dip. Some studies suggested that the dip is a consequence of an increase in oxygen consumption (Röther et al., 2002). Other studies reported that this phase was originated from an increase in deoxygenated hemoglobin caused by changes in neuronal activity (Hu and Yacoub, 2012). It seems that this initial decrease is associated with changes in the capillaries close to the neurons and it has been shown more spatially specific than the other phases of the HDR, so that it might represent focal neuronal activity (Duong et al., 2001; Zaidi et al., 2018).

Two or three seconds after the beginning of the neural activity, the signal starts increasing. During this time, blood travels from arteries to capillaries and drain veins.

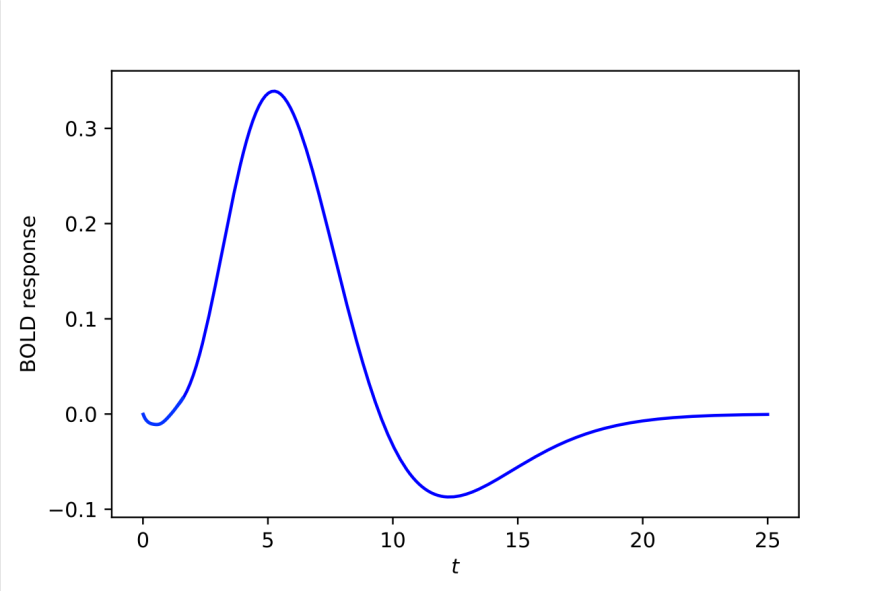


**Figure 2.4:** Illustration of the canonical hemodynamic response function (HRF).

Thus, the hemodynamic response is always shifted with respect to neural activity because of the different rhythms of nervous and circulatory systems. The signal rises to a plateau 6-8 seconds after stimulation, and then returns to baseline. After that, it is often observed a post-stimulus undershoot (Logothetis, 2004). There are different theories about the appearance of this phenomenon. A possible interpretation is related to the different speed in the recovery of blood volume and blood flow. Once the stimulation finishes, the CBV falls more slowly than the CBF, so that blood vessels keep dilated a long interval which results in a larger amount of deoxygenated hemoglobin in the regions where the BOLD decreases (Buxton et al., 1998). A different hypothesis proposes that the decrease of the BOLD would be caused by a sustained oxygen consumption after the end of the stimulus (van Zijl et al., 2012). Donahue et al. (2009) found that the post-stimulus undershoot highly varies depending on the brain region evaluated and they related these differences to oxygen metabolism.

Thousands of studies have used the BOLD response to infer mental states in different cognitive processes, but there is a lot of information about this signal that remains unclear (Logothetis, 2014). Some authors have remarked that changes in the hemodynamic response vary across different regions of the brain (Handwerker et al., 2004). In fact, the coupling between neural activity and vascular response depends on the vascularization of the region, which determines the amplitude and resolution of the BOLD signal. The hemodynamic response efficiency (HRE; Logothetis, 2004) describes the efficiency of this coupling. Other studies have evaluated the variations in the BOLD signal induced by

ageing, concluding that elderly people present a reduced signal-to-noise ratio (D'Esposito et al., 2003). BOLD signal can also differ because of the vascular system of each person (Hillman, 2014) and some neurological disorders like epilepsy (Masteron et al., 2010).

### 2.4.2 The mathematics behind BOLD response

Understanding the relationship between BOLD signal and neural activity is crucial to make accurate inferences. From a mathematical perspective, the aim is to find a system that describes the relationship between an input and an output (Logothetis, 2004). The input would be the neural activity elicited by a stimulus, $A$, and the output is the resulting pattern of BOLD response after a certain stimulus, $B$. However, the input is not the neural activity associated only with that stimulus, but a joint of that activity ($A$) and unidentified neural responses ($N_N$). The activity is usually measured in a specific location of the brain, so that the variables previously defined ($B(x)$, $A(x)$ and $N_N(x)$) should show this dependence. Besides, each brain region has a different HRE, so that BOLD output at a specific location is determined by the equation:

$$B(x) = H(x)(A(x) + N_N(x)) \tag{2.8}$$

However, the signal measured at a certain location is not only due to the activity ocurred in this location, but also to the activity of the surrounding neighbors in the cortical surface. For a neighborhood $n(x)$, the BOLD signal can be measured as:

$$B(x) = \int_{n(x)} H(u)(A(u) + N_N(u))P(x-u)du \tag{2.9}$$

where $P(x)$ is a function that models the contribution of the adjacent positions to the signal measured at a specific location and the integral over $n(x)$ is the sum of all the contributions within a neighborhood. At this point, we have considered the noise attributed to the activity associated with unidentified sources that affect the BOLD signal. However, there are other sources of noise (Liu, 2016) caused by experimental factors (motion, cardiac, respiratory) and by the MR scanner (thermal, fluctuations in the magnetic field). Thus, the final expression for the estimation of the BOLD response can be computed by adding an additional noise term, $N_M(x)$, as follows:

$$B(x) = \int_{n(x)} H(u)(A(u) + N_N(u))P(x-u)du + N_M(x) \tag{2.10}$$

### 2.4.3 Neural correlates

Previous sections have described the neurovascular coupling, the relationship between neural activity and the subsequent changes in the cerebral blood flow. However, it is necessary to improve our knowledge about the type of neural activity underlying these changes to obtain a better interpretation of fMRI results. There are two types of signals according to the neural activity that they elicit: Local Field Potentials (LFP) and Multiple Unit Activity (MUA). LFP is a measure of brain activity that reflects the flow of information across neural networks related to the perisynaptic processes on the dendrites (Berens et al., 2010; Herreras, 2016), so that they integrate the different information inputs that a neuron receives. This activity is also known as low-frequency fluctuations ($f < 500$ Hz), and it is generated by membrane currents of the neurons (Burns et al., 2010). Moreover, MUA comprises the high-frequency fluctuations ($f > 1000$ Hz) associated with transmission that reflect the action potentials in populations of neurons (Mattia et al., 2010). Thus, both signals control different mechanisms, so identifying their implication in the hemodynamic changes would lead to a major understanding of the BOLD contrast.

Different studies have evaluated the implication of these signals in the hemodynamic changes. Logothetis et al. (2001) compared electrophysiological activity associated with LFP and MUA with BOLD fMRI responses from the visual cortex of monkeys. They first isolated spiking and synaptic activity and observed that both types of activity correlated with the BOLD response. However, while the increase in LFP was maintained along the duration of the stimulus, MUA adapted at the beginning to the BOLD signal but decayed until baseline after 2.5 seconds. Therefore, both signals allow a proper estimation of the BOLD response, but LFP predicted considerably better the hemodynamic response. Experiments conducted by Lauritzen et al. (2001) focused on dissociating the postsynaptic activity and action potentials. Results showed that CBF is independent from the neurons firing rate (number of action potentials per unit time), but not from their action potentials. This evidences that the hemodynamic response does not always reflect the firing rate, so that the presence of action potentials is not a necessary (nor enough) condition to induce a change in the BOLD signal (Mathiesen et al., 1998; Rauch et al., 2008).

In fMRI studies, measuring the changes that underlie the BOLD signal is an intermediate step for understanding the different neural mechanisms in a specific context. To do so, it is highly relevant to find how neural activity produces changes in the vascular system. Similarly to most concepts associated with fMRI, there is not a consensus yet about what exactly guides neurovascular coupling. Some authors suggest that the increase in oxygenated hemoglobin is driven by a metabolism demand of a certain tissue (Hoge

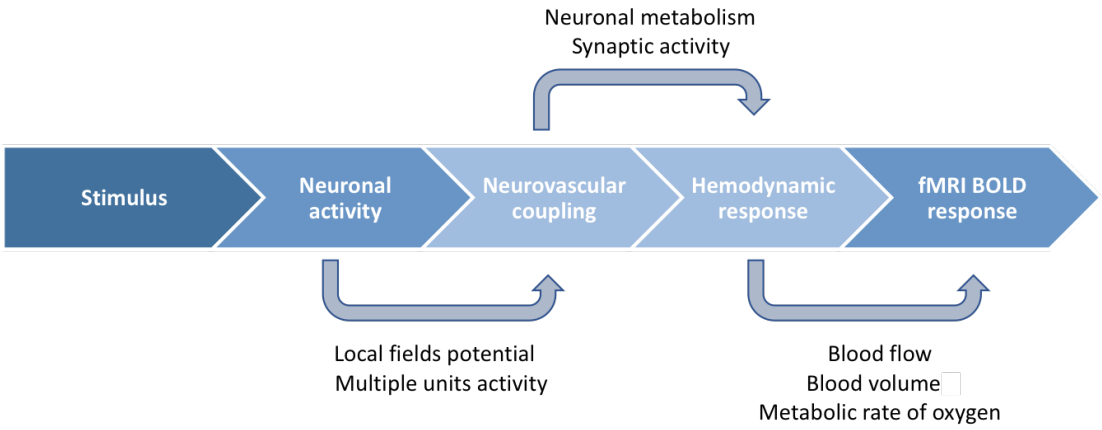


**Figure 2.5:** Diagram of the different steps behind the BOLD hemodynamic response.

et al., 1999; Shin, 2000). Specifically, the neural activity associated with a rise in energy consumption is due to postsynaptic currents and action potentials (Attwell and Iadecola, 2002). Filosa et al. (2006) proposed an alternative in which the potassium $K^+$ plays an important role in the neurovascular response, whereas results obtained by Attwell et al. (2010) showed the large contribution of epoxyeicosatrienoic (EET) acid to that response.

Other studies proposed a more direct neural control of the hemodynamic response. In this theory, changes associated with synaptic transmission would cause the increase in the CBF. Attwell and Iadecola (2002) concluded that the start of the hemodynamic response is not caused by signals arising from the energy deficit of a tissue, but it is driven by the glutamate signalling process. This procedure facilitates the release of other fast neurotransmitters and molecules that induce the vasodilation and the subsequent rise in the CBF. A recent study provided a state-of-the art of the different models associated with the neurovascular coupling and simulated the BOLD responses that different sources elicit (Mathias et al., 2018). Results suggest that potassium ions released during neural activity could be the most important agent in neurovascular coupling. Nonetheless, they suggested that other mechanisms can act together and increase the CBF. Besides, different experiments indicate that the vascular response is possibly influenced by multiple neural pathways. Thus, it is a very complex and integrated model the one that provides the required mechanisms to connect the neurovascular coupling and the resulting BOLD signal. Figure 2.5 shows a diagram of the different stages behind the BOLD signal.

### 2.4.4 Convolutional models

The information provided by the BOLD signal is an indirect measure: fMRI scanners do not detect neural activity but the changes in the magnetic properties of blood flow as a consequence of oxygen consumption, which is associated with the neural activity in a certain brain region. A crucial aspect of fMRI analyses is to find the exact relationship between the neural response and the BOLD signal. We mentioned in previous sections that this relationship depends on a large number of variables, so that it is not straightforward to estimate it properly. However, different studies have demonstrated that this relationship shows linear time invariant (LTI) properties (Henson and Friston, 2007). Linearity means that if an input signal is multiplied by a scalar, the output signal is also multiplied by the same scalar. Time invariance implies that if an input signal is shifted by $t$ seconds, the output signal would also be shifted the same amount of time. Following this description, there is a mathematic operation that establishes the defined relationship between neural activity and BOLD signal, termed convolution, which is defined as:

$$(h * f)(t) = \int h(\tau) f(t - \tau) d\tau \tag{2.11}$$

where $h$ is the hemodynamic response function (HRF), $f$ is the stimulus onset time and $\tau$ refers to the peristimulus time. It seems clear that obtaining a good fit of the hemodynamic response depends on choosing an appropriate HRF (Poldrack et al., 2011). The simplest option is to use only one basis function, the single canonical HRF. In a two basis function scheme, the single canonical HRF in addition to its partial derivatives with respect to its delay and dispersion are used. As explained in Poldrack et al. (2011), the expected output signal must be a delayed version of the input signal, as follows:

$$Y(t) = \beta X(t + \delta) \tag{2.12}$$

By the first-order Taylor expansion:

$$Y(t) = \beta(X(t) + \delta X'(t) + ...) \tag{2.13}$$

which can be approximated by:

$$Y(t) \approx \beta_1 X(t) + \beta_2 X'(t) \tag{2.14}$$

where $\beta_2$ can be interpreted by a linear combination of the HRF and its derivative.

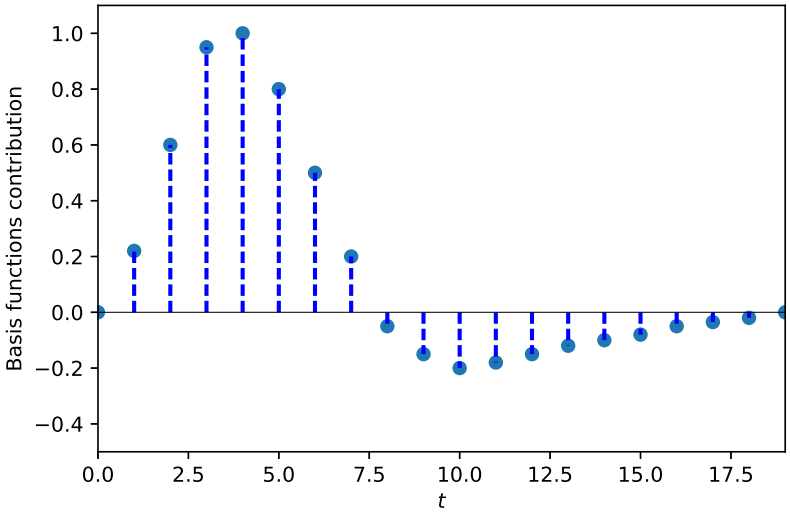


**Figure 2.6:** Illustration of the finite impulse model (FIR) model.

Another possibility is to employ finite impulse response (FIR) functions and Fourier basis sets, the most flexible option since it makes minimum assumptions about the shape of the response. The FIR consists in defining a window around the peristimulus interval. In this window, a series of impulse functions is defined and the HRF is modeled as a weighted combination of these functions. The weights correspond to the value of the HRF at each specific point (see Figure 2.6). This leads to a large flexibility because changing these values would lead to different HRF shapes that can adapt to the hemodynamic idiosyncrasies of each brain region. However, this flexibility also entails an increase in the variability of the estimates since a lower number of data points are employed to reproduce the hemodynamic function. For this reason, this approach is most recommended in studies where the aim is to model the hemodynamic function. Several studies have used it in experiments focused on activity detection (Bode and Haynes, 2009; Palenciano et al., 2018; Soon et al., 2008), but it does not always offer the best performance (Poldrack et al., 2011).

## 2.5 Experimental designs

Section 2.4 describes the metabolic factors behind the BOLD signal. However, the signal acquired by MR scanners does not provide a separate activity estimation for each stimulus, but a combination of them. For this reason, it is of great importance to choose an experimental design that provides the desired information. The easiest way

to evaluate the relationship between the independent variable and the dependent one is to compare an experimental condition with a control condition. In the experimental condition, subjects do a certain task that activates the neural mechanisms related to a desired cognitive process. In the control condition, the independent variable is absent or at a lower level.

The trials associated with experimental and control conditions can be organized in different ways. One possibility is to group together in time several trials of the experimental condition and do the same with trials of the control condition, leading to a block for each of the two conditions. As its name suggests, this is known as block-design. There are some experiments where the process of interest cannot be modulated over short intervals, such as vigilance or sustained attention tasks, so that a block design could be the right option (Huettel et al., 2009). Block designs are shown to be robust (Brockway, 2000; Tie et al., 2009), with large BOLD signal changes (Buxton et al., 1998) and statistical power (Friston et al., 1999).

An alternative to block designs are event-related designs, which assume that the activity of interest occurs for discrete and short intervals (Huettel et al., 2009). The stimuli that generate these bursts of neural activity are termed events, and each trial usually comprises several events. Unlike blocked designs, trials are presented in a random order rather than an alternating pattern, which allows to estimate the hemodynamic response function associated with each single event. This emphasizes that for each experimental condition, one stimulus is presented at a time, whereas in block designs stimuli of the same condition are presented consecutively. The time that separates two events, known as ISI, has a large importance in the design of an event-related experiment. An ISI longer to the duration of the hemodynamic response (10-12 s) leads to a slow event-related design (Buckner et al., 1996) whereas a rapid event-related design is obtained where the ISI is shorter than the hemodynamic response (Buckner et al., 1998). It has been demonstrated that rapid designs are quite effective at estimating the HRF response despite its short ISI (Lindquist et al., 2009). Rapid designs also provide a boost in the statistical power since the number of trials per each run increases. Moreover, the short ISI leads to a larger overlap of the hemodynamic responses, which results in a more difficult separation of the contribution of different events. See Figure 2.7 for an illustration of this issue and for a schematic of the two different experimental designs.

It seems clear that acquisition of fMRI signal is a complex process, and understanding the neural changes that lead to variations in the signal is not straightforward. A rise in activity in a specific brain region can be due to multiple factors since different neural

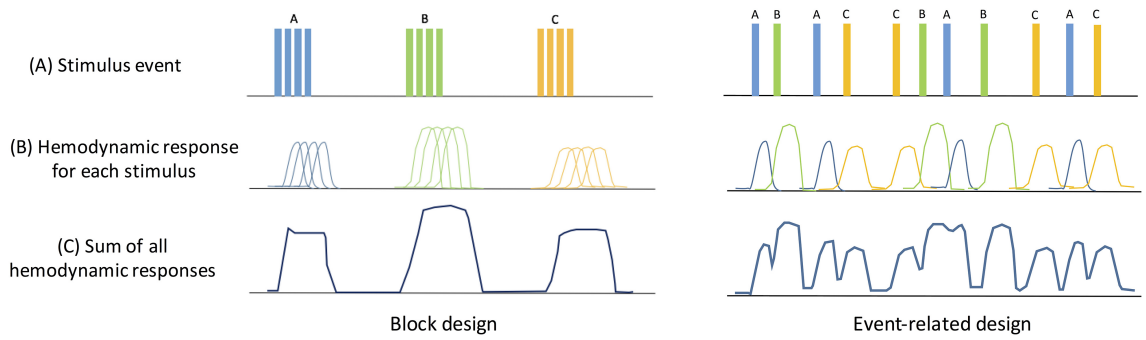


**Figure 2.7:** Schematics of two different fMRI designs: block and event-related. The first row corresponds to the timing of event onsets. In block designs, several stimuli of the same condition are presented consecutively, in what is known as epoch or block, and different conditions usually alternate in time, so relatively large signal changes are measured. In event-related designs, interleaved short-duration stimuli are employed. Given the delayed nature of the BOLD signal, the data produced by different stimuli overlaps, and thus extracting the signal caused by each one of them becomes more difficult.

processes occur at the same time. Convolutional models estimate the relationship between the neural activity and the BOLD signal, providing a powerful tool to disentangle the different sources of hemodynamic changes.

CHAPTER 3

# PREPROCESSING IN FMRI

We have previously explained the physical basis of fMRI in addition to the phenomenon of neurovascular coupling. Once the images are acquired, a series of preprocessing steps is applied to remove artifacts and prepare the data for the subsequent statistical and classification analyses. This chapter focuses on the mathematic mechanisms behind these procedures and defining statistical parametric mapping, the standard approach for evaluating differences in brain activity during fMRI experiments.

## 3.1 Preprocessing

When acquisition begins, the scanner usually requires some time to stabilize its gradients, whereas the brain tissues need take some seconds to reach the excitation. For this reason, it is common to discard the first 3-4 volumes (depending on the TR) to allow the saturation of the signal (Soares et al., 2016). This preliminary step is applied before any of the subsequent preprocessing stages.

### 3.1.1 Slice timing correction

FMRI scanners usually acquire each slice of the brain volume once at a time, so they need a specific time to acquire the whole volume. This time is TR (see section 2.3.1) and values of a few seconds are typical, although an optimum value for this parameter

has been broadly discussed (Constable and Spencer, 2001; Georgiopoulos et al., 2018). Due to the lack of synchronization between the acquisition of different slices, it is not correct to assume that all of them have been acquired at the same time. Since subsequent statistical analyses typically make this assumption, a slice-timing correction is applied to account for these differences. Specifically, this operation adjusts the voxel time series to a common reference for all voxels. For all the slices forming a volume, one of them is selected to be the reference. The election depends on the way fMRI data has been acquired. One approach is to use an ascending/descending slice acquisition, in which the process starts in one slice and collects the others consecutively. The reference slice would be the first/last for an ascending/descending acquisition. Nonetheless, most fMRI studies use an interleaved slice acquisition in which first odd slices are collected and then even slices. In this case, the reference slice would be the central one (e.g. the 15th in a 30-slices acquisition).

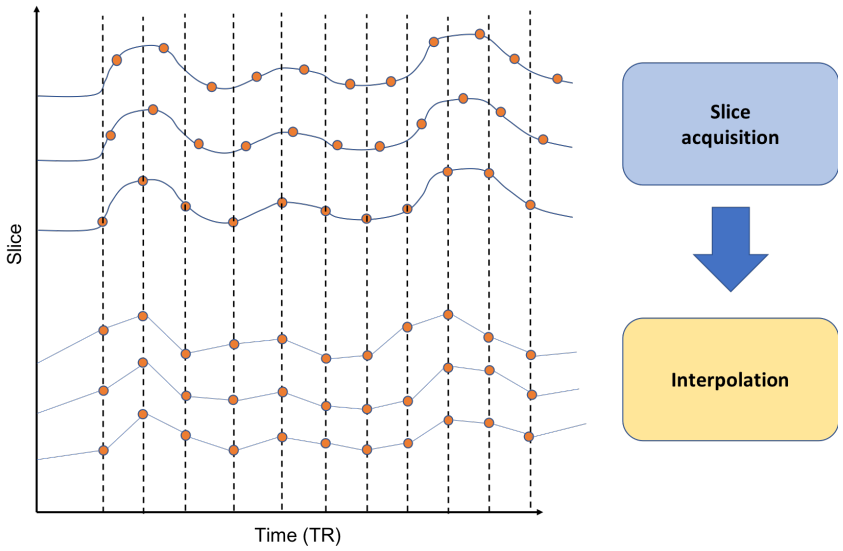


**Figure 3.1:** Representation of the slice-timing problem and the way interpolation adjusts the voxel time series to a common reference.

Correction of slice-timing discrepancies is possible by employing temporal interpolation. This method uses data from nearby time points to estimate the amplitude of the MR signal at a point that was not originally collected (see Figure 3.1 for an illustration of the problem and the way interpolation works). Linear interpolation is the simplest approach and can be expressed mathematically as follows:

$$y_n^{(r)} = \frac{(t(r) - t(n-1))\, y_n + (t(n) - t(r))\, y_{n-1}}{t(n) - t(n-1)} \tag{3.1}$$

where $y$ is the time series of slice number $n$ at time point $t(n)$ and $t(r)$ is the time point of the referent slice. To minimize the possible errors within each volume, each interpolated point is usually collected at TR/2 time. This kind of correction can introduce undesired temporal smoothing (Sladky et al., 2011) so that a *sinc* linear interpolation is usually preferred. This approach uses a fast Fourier transformation that leads to a phase shift in the frequency domain (Calhoun et al., 2000). This type of correction is given by:

$$y_n^{(r)} = \sum_{i=-\infty}^{\infty} x_i sinc\left(\frac{\pi}{TR}(r - iTR)\right) \tag{3.2}$$

There are possibly some situations in which slice-timing correction could be avoided, such as in block-designs. However, this has not been clearly demonstrated in the literature. In fact, there are several studies that show a clear improvement in the parameter estimation on each subject and a subsequent significant boost in sensitivity at the group level (Sladky et al., 2011). Moreover, Parker et al. (2017) demonstrated the effectiveness of this operation in a wide range of fMRI scenarios. For this reason, this preprocessing step is currently included in most fMRI software packages such as SPM (Wellcome Centre for Human Neuroimaging, 2018), FSL (Jenkinson et al., 2012) or BrainVoyager (Goebel et al., 2006).

### 3.1.2 Motion correction

Head motion is one of the major problems during fMRI acquisitions. Movements in the range of 1-2 mm have a dramatic effect in data quality, leading to displacements of the anatomic brain features and temporal variation of the voxel time course (Haller and Bartsch, 2009; Van Dijk et al., 2012). Previous research has shown that small head motion can lead to artifacts in activation maps, specially when there is a correlation between the motion and the neural mechanisms associated with the paradigm studied (Field et al., 2000; Johnstone et al., 2006). Zaitsev et al. (2017) listed the most common effects than motion has in fMRI data in addition to the severity that they usually show. Head motion results in inhomogeneities in the magnetic field that can produce differences in contrast and BOLD sensitivity (Deichmann et al., 2002; Yarach et al., 2016). The rotation of the head about one of the axes non-parallel to the main magnetic field leads to deviations in different brain tissues (Maclaren et al., 2013).

### Principles of motion correction

Although using head restraints can reduce the effect of motion during an fMRI session, it does not fully eliminate it. For this reason, there are a large number of motion correction algorithms that attempt to minimize this effect over fMRI signal. Most of these algorithms employ coregistration, a procedure which uses a cost function that relies on a voxel similarity measure to evaluate the alignment of images. In this approach, a reference image is selected and the others are aligned with this one. Motion usually occurs in the middle of the scanning session, so that one option is to select a single image or an average of images from the first part of the session (Orchard et al., 2003). Then, this method employs a model based on 6 parameters (translations and rotations about the three axes $x, y, z$) and applies the spatial transformation that provides the best match to the reference (Power et al., 2012). Each transformation is computed by minimizing the registration error, which is given by:

$$\epsilon_i = \left\langle \left( sI_i(T(x) - I_0(x)) \right)^2 \right\rangle \tag{3.3}$$

where $I_i$ is the image intensity at volume $i$, $T$ corresponds to the transformations applied, the angle brackets refer to the spatial average over the brain, $I_0$ denotes the reference image and $s$ is a scalar factor that accounts for small changes in signal intensity. The final transformation is obtained from a combination of rotation and displacement components, as follows:

$$T_i = \begin{bmatrix} R_i & d_i \\ 0 & 1 \end{bmatrix} \tag{3.4}$$

where $R_i$ is a 3x3 rotation matrix and $d_i$ is a 3x1 column vector of displacements. The complete mathematical calculations are detailed in Power et al. (2012).

All the algorithms available in the different packages for fMRI preprocessing are based on the aforementioned procedure and they only differ in the interpolation algorithms, the cost functions and the optimization methods of these cost functions. Despite this mathematical framework, current approaches are not perfect mainly due to the extreme difficulty in eliminating from brain images the nonlinear distortion that motion leads to. Besides, they sometimes make assumptions that are not always met. One example is related to the time point when motion appears. They usually assume that this occurs among different volumes, ignoring that this misalignment can also appear between the slices of a volume. Different alternatives have been proposed to account for this intra-volume variability by performing a slice-by-slice registration (Beall and Lowe,

2014; Bhagalla and Kim, 2008). These methods provide a fine-grained correction since they take into account the brain tissue that each voxel belongs to (white/gray matter or cerebrospinal fluid) to estimate the required transformations.

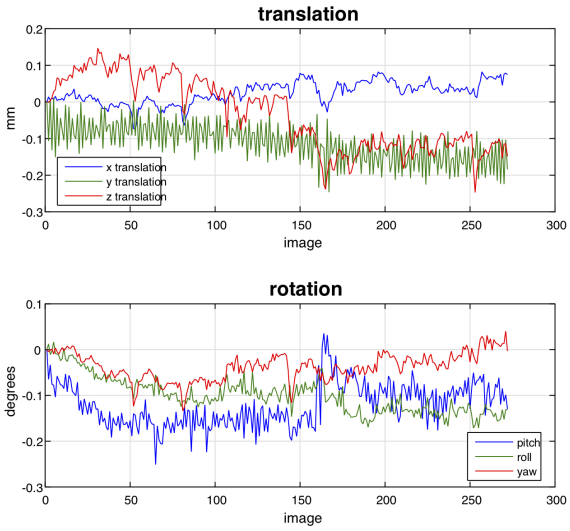


**Figure 3.2:** Parameters associated with the spatial transformation applied about the three axes. Pitch, yaw and roll refer to modifications in axes $x$, $y$ and $z$, respectively. Images generated from SPM package (Wellcome Centre for Human Neuroimaging, 2018).

Another possibility to minimize the effects of head motion is to include the motion parameters (see Figure 3.2) as nuisance regressors (covariates or no interest) in the subsequent general linear model analysis. Experiments conducted by Johnstone et al. (2006) showed that the inclusion of these parameters in the model is beneficial since it usually increases the sensitivity for detecting the desired effect and decreases the absolute error in the model estimation. However, if there is a correlation between motion and task design, modeling motion parameters could also remove activity associated with task. Furthermore, some artifacts can remain in the fMRI data even when these parameters are regressed out from the model, which explains why head motion is probably the major problem in fMRI acquisition.

### 3.1.3 Spatial normalization

Most fMRI experiments aim to make some findings about brain function that can be generalized across individuals. However, it is difficult to compare the activity maps derived from each subject because of the large variability in the size and shape of each brain. For this reason, an intermediate step is required to transform and align all

brains into a common space in a process known as spatial normalization. This procedure reduces the variability between subjects and determines the regional correspondence for all brains necessary for the subsequent group analysis.

### Defining a template

The idea of using a template in a common coordinate space was proposed by Talairach and Szikla (1980). They defined a three-dimensional Cartesian space based on a set of anatomical landmarks: the anterior commissure (AC), the posterior commissure (PC), the midline sagittal plane and the exterior boundaries of the brain. They later created an atlas from Brodmann's areas and anatomical structures to facilitate the location in the new coordinate space (Talairach and Tournoux, 1988). However, this atlas has severe limitations since it does not provide an MR scan of the person from the atlas was computed, which restricts the maximum spatial accuracy that it can be obtained. Besides, constructing a template from data of one single individual does not guarantee a good representation of the population. For this reason, a new template was later computed to match the Talairach atlas but considering MR scans of hundreds of subjects. First, 241 scans were used for identifying the anatomical landmarks proposed by Talairach and Szikla (1980). Then, a nine-parameter affine linear transformation was applied to 305 MR scans to match the average brain computed from 241 scans (Evans et al., 1993). The average of the 305 transformed images is known as MNI template (Brett et al., 2001), and has been commonly used in most fMRI studies.

### Images normalization

Once the reference space is built, a transformation is applied to different brain images to register them to a common space. There are several algorithms for performing spatial normalization, but most of them rely on volumetric-based registration. These approaches focus on minimizing the sum of squares difference between the images to be normalized and the template image (Penny et al., 2006). First, they compute an affine transformation based on 12 parameters (e.g. rotation, translation, scale, squeeze, shear, etc). Mathematically,

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{3.5}$$

After the affine transformation, a series of nonlinear deformations is usually performed. These transformations are obtained from a linear combination of three discrete cosine transform (DCT) basis functions (Ashburner and Friston, 1999). Additionally, this framework employs diffeomorphisms (differentiable homeomorphism, Holden, 2008), complex mathematical processes that provide, for every voxel, a measure of the difference between the original and the transformed image (Ashburner, 2007). These transformations have a large number of parameters but they use regularization to find the optimum values that lead to smooth deformations (Poldrack et al., 2011).

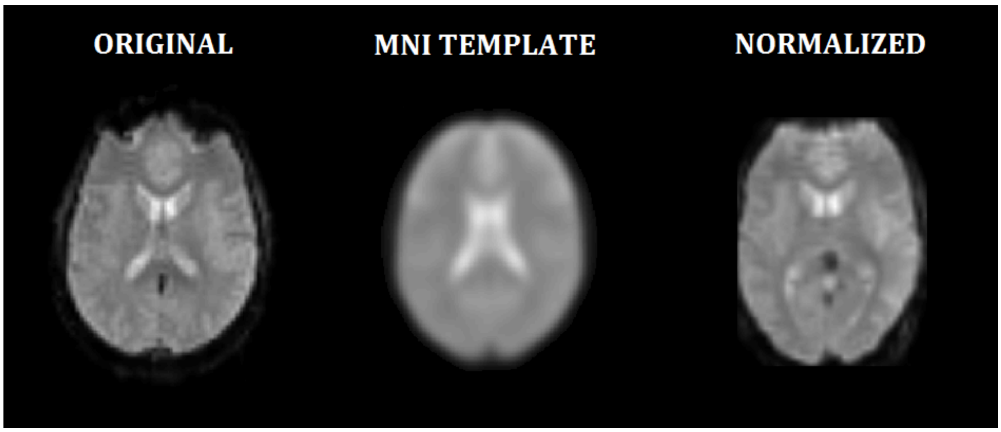


**Figure 3.3:** Different images involved in the normalization process: the original image, the MNI template that the original has to match and the resulting image after normalization.

### 3.1.4 Spatial smoothing

Spatial smoothing has been accepted as one of the main preprocessing steps in fMRI data. This operation involves convolving the image with an average filter: the value of a voxel in the resulting image is computed as a weighted sum of the value of the adjacent voxels in the original image. One of the reasons for using smoothing is to minimize the errors in spatial normalization that can affect the results derived from second-level analyses. Another important use of this operation is to modify MR images to satisfy the assumptions of Gaussian Random Field Theory (Worsley et al., 1992) for a subsequent multiple comparisons correction (Benjamini and Hochberg, 1995). Under some circumstances, smoothing can also reduce the thermal noise present during the images acquisition, leading to an increase in the signal-to-noise ratio (Molloy et al., 2014; Triantafyllou et al., 2006; Welvaert and Rosseel, 2013).

One of the main features of the spatial smoothing is the size of the kernel of the filter. This is defined by the full width at half maximum (FWHM), a measure that represents

(as the name suggests) the width of the kernel at half of the peak value. Voxels closer to the central voxel have a large contribution to the final value of the central one than those that are more distant. Besides, the size of the kernel determines the number of voxels used to compute the new value of the central voxel: the higher the size of the filter, the more smoothed (blurred) the resulting image. As explained in Lindquist and Wager (2008), a convolution between the image and the smooth filter is given by:

$$\hat{F}(\mathbf{x}) = \int_{\Omega} F(\mathbf{u})K(\mathbf{x}-\mathbf{u})d\mathbf{u} = \int_{\Omega^*} \hat{F}(\mathbf{k})\hat{K}(\mathbf{k})e^{i\mathbf{xk}}d\mathbf{k} \tag{3.6}$$

where $K(\mathbf{x})$ is the filter kernel, $\hat{K}(\mathbf{x})$ is its Fourier transform, $\hat{F}(\mathbf{x})$ is the $k$-space version of the image acquired and $F(\mathbf{x})$ is the MR image, given by the inverse Fourier transform of the $k$-space. Thus, $\hat{F}(\mathbf{x})$ provides the value of the smoothed image at the coordinate $\mathbf{x}$. Most fMRI studies use a Gaussian filter kernel, which can be expressed by:

$$K_{\Phi}(\mathbf{x}) \propto exp\left\{-\frac{\mathbf{x}^2}{2\sigma^2}\right\} \tag{3.7}$$

The conversion between the image space and its $k$-space requires applying a Fourier transform. This operation must also be applied to the kernel:

$$\hat{K}_{\Phi}(\mathbf{k}) \propto exp\left\{-\frac{1}{2}(2\pi\sigma)^2\mathbf{k}^2\right\} \tag{3.8}$$

We previously mentioned the role that the kernel size plays in the smoothing process. Figure 3.4 shows the effect of different sizes of the Gaussian Kernel in the resulting images. The relation between the FWHM and the standard deviation of the Gaussian kernel is given by:

$$\sigma = \frac{FWHM}{2\sqrt{2\ln 2}} \tag{3.9}$$

where $\sigma$ denotes the standard deviation associated with the Gaussian kernel and FWHM is usually specified in millimeters or voxels. The optimum value of FWHM should be at least twice the size of the voxel (Worsley and Friston, 1995).

In MVPA analyses, the order of the preprocessing pipeline partially changes. For example, spatial normalization is usually performed after classification, in a previous step to group-level statistics. Besides, smoothing can have a detrimental effect on decoding accuracies (Hendriks et al., 2017), so that its application in this kind of studies can be avoided .

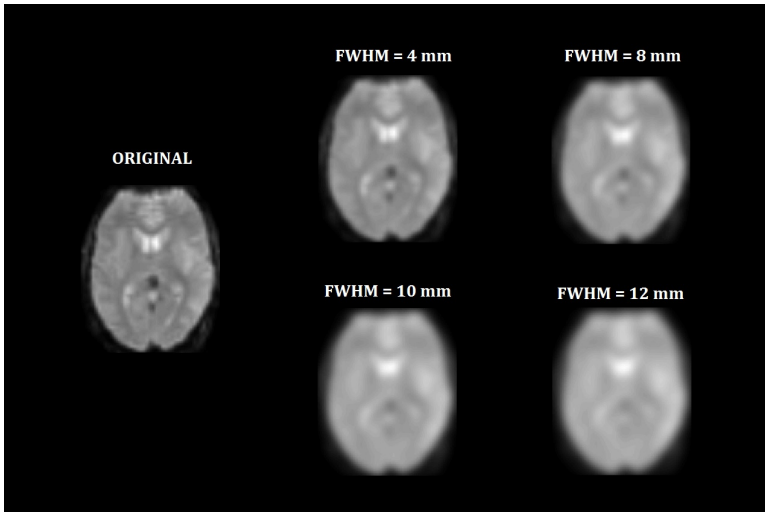


**Figure 3.4:** Effect of the size of the Gaussian kernel in the smooth operation.

## 3.2 The General Linear Model

Images preprocessing reduces noise and different artifacts caused during fMRI acquisition. However, it still necessary to find a mechanism that relates the measured fMRI signal to the different conditions in the experimental paradigm. To do so, the General Linear Model (GLM) has been employed from the early days of fMRI (Friston et al., 1995). This approach computes an estimation of the contribution of each predictor (experimental condition) to the variability observed in each voxel and temporal unit of the fMRI data. The method assumes that there is a linear relation between the independent (BOLD time courses) and dependent variables (experimental conditions), so that each voxel time course is modeled as a weighted sum of the experimental conditions. The intensity of the BOLD at each observation ($y$) is given by:

$$\begin{aligned} y_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \epsilon_1 \\ y_2 &= \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \epsilon_2 \\ &\vdots \\ y_n &= \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \epsilon_n \end{aligned} \tag{3.10}$$

where $x$ are the experimental conditions used as predictor variables, $\beta$ are the contribution of each variable to the observed signal and $\epsilon$ are the remaining unexplained data known as error term. The GLM assumes that errors are independent and identically

distributed following a Gaussian distribution with the same variance. Equation 3.10 can be rewritten by using a matrix notation:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_p \end{bmatrix} \tag{3.11}$$

or just by using the simplified notation:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.12}$$

where $\boldsymbol{Y}$ is a vector where $n$ provides the number of volumes acquired; $\mathbf{X}$ is usually known as design matrix where $p$ indicates the number of predictors; $\boldsymbol{\beta}$ contains the magnitude and direction of the relationship between the experimental conditions and fMRI data; and $\boldsymbol{\epsilon}$ indicates the errors associated with each observation. The main aim of the GLM is to estimate the value of the parameters $\beta$ (known as beta maps or beta images) to evaluate to what extent an experimental condition can explain the variance observed in the BOLD time course. The error term can be considered as the difference between the fMRI signal and its estimation:

$$\boldsymbol{\epsilon} = Y - \hat{Y} = Y - \mathbf{X}\boldsymbol{\beta} \tag{3.13}$$

It seems obvious that obtaining a good prediction requires that the error term is small. However, the methodology used in fMRI does not aim at minimizing the sum of the error values, but to find the optimal beta values that minimize the sum of squared errors. This is known as Least-Squares approach (Aitken, 1936), and there are a lot of variants applied to fMRI (see Waldorp, 2009 and Monti, 2011). The most simple one can be mathematically expressed by:

$$\boldsymbol{\sigma}_{\epsilon}^2 = (Y - \mathbf{X}\boldsymbol{\beta})^T (Y - \mathbf{X}\boldsymbol{\beta}) \tag{3.14}$$

According to descriptions in Poline and Brett (2012), beta maps can be computed as follows:

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y \tag{3.15}$$

and its variance is given by:

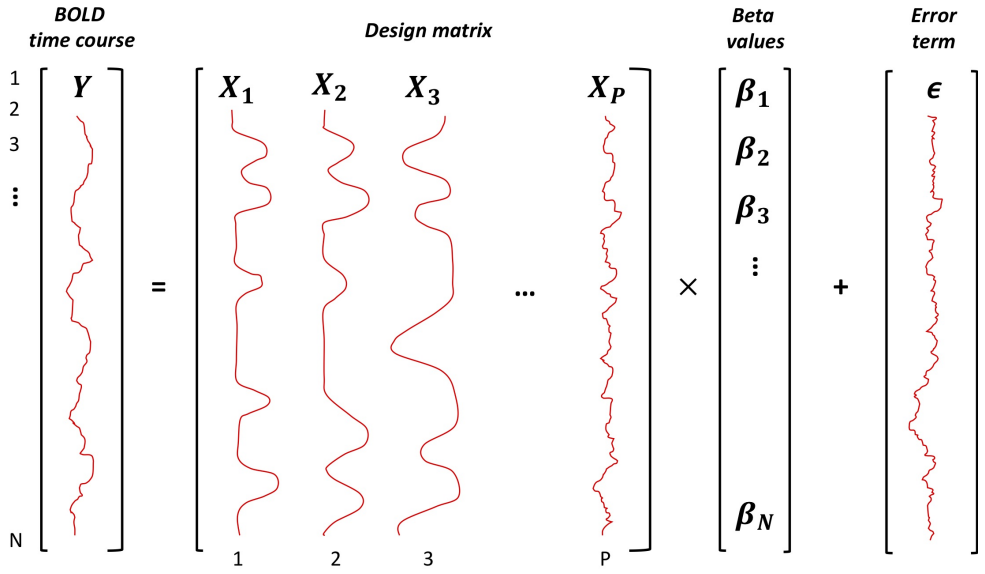


**Figure 3.5:** Diagram of the GLM model for a certain voxel.

$$var(\boldsymbol{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{3.16}$$

It is crucial to mitigate the effect of different artifacts appeared during the acquisition of fMRI signal. An important step is the proper estimation of activation patterns (Chapter 8 evaluates the performance of different alternatives). The way activation patterns are estimated affects the quality of the images fed into the classifier, so that the performance of the subsequent classification depends on the reliability of this estimation.

CHAPTER 4

# MACHINE LEARNING IN NEUROIMAGING

The use of machine learning in neuroimaging has represented a paradigm shift. This approach is a powerful tool to explore the human brain and to obtain a better interpretation of fMRI results. This chapter provides an overview of the standard approaches employed in each stage of the classification framework, from the different types of classifiers most commonly used in neuroimaging to the alternatives for dimensionality reduction.

## 4.1 Introduction

In recent years, the use of machine learning to analyze neuroimaging data has considerably increased. This mathematical framework extracts and learns the most relevant features from an input dataset and, based on this learning, it separates unseen data into different classes. When applying to brain imaging, these methods are known as Multi-variate pattern analysis (MVPA, Norman et al., 2006), and the features from which they extract information are the voxels contained in the images. The classifier evaluates if there is a pattern of information in a brain region that distinguishes between two (or more) experimental conditions. MVPA has been successfully applied to a wide range of imaging modalities, such as Single Positron Emission Computed Tomography (SPECT, Betancur et al., 2018; Górriz et al., 2017), Positron Emission Tomography (PET, Nakagawa et al., 2019; Segovia et al., 2017), structural magnetic resonance imaging (sMRI, Pagnozzi et al., 2018; Wang and Summers, 2012) or fMRI (Chen et al., 2019;

Mourão-Miranda et al., 2005). MVPA has also been used in the study of electrophysiological data such as Electroencephalography (EEG, Cichy and Pantazis, 2017; Zafar et al., 2018), Magnetoencephalography (MEG, Guggenmos et al., 2018; Hauswald et al., 2018) and Event-related potentials (ERPs, Bode and Stahl, 2014; Draschkow et al., 2018).

Classification methods are a fundamental part of Computer-Aided Diagnosis (CAD) systems. These approaches aim to assist clinicians in the diagnosis of diseases by providing an automatic tool that identifies patterns of information and makes accurate predictions about the development of these disorders. CAD systems have successfully diagnosed different neurological disorders such as Alzheimer's (Arco et al., 2015), Parkinson's (Choi et al., 2017) and epilepsy (Del Gaizo et al., 2017). Besides, this framework can play a crucial role in the therapeutic strategies for the recovery of several disorders. As an example, brain computer interfaces (BCI) record different neural signals and identify an information pattern for each one them. Then, these signals are converted to control assistive devices and computers (Blankertz et al., 2007; Mohanty et al., 2018; Nurse et al., 2015).

The main purpose in CAD and BCI systems is to obtain accurate predictions. However, there are other contexts in neuroscience where obtaining the largest accuracy is not of primary interest. MVPA can also be applied to study brain function, where the main goal is to detect the presence of a particular cognitive state. The adoption of MVPA has led to novel discoveries about the visual system (Cox and Savoy, 2003; Haxby et al., 2001), the auditory cortex (Formisano et al., 2008; Staeren et al., 2009), working memory (Harrison and Tong, 2009) or more abstract brain states such as intentions (Gilbert and Fung, 2018; Haynes et al., 2007). Hebart and Baker (2017) remarked the importance of differentiating multivariate analysis for prediction (CAD and BCI) and for interpretation/identification (brain states) as two independent frameworks. Both techniques share the main parts of a classification system (see Figure 4.1), but they also have some peculiarities. Throughout this chapter, we will tackle the state of the art for each stage of fMRI classification remarking the differences between prediction and identification contexts.

## 4.2 fMRI signal and classification

A crucial step in classification is to determine which images are fed into the classifier. In fMRI analyses, there are different alternatives that have a large influence in the subsequent results. One possibility is to employ RAW fMRI responses (top of Figure 4.2). The volumes acquired by the MR scanner are preprocessed and those corresponding to

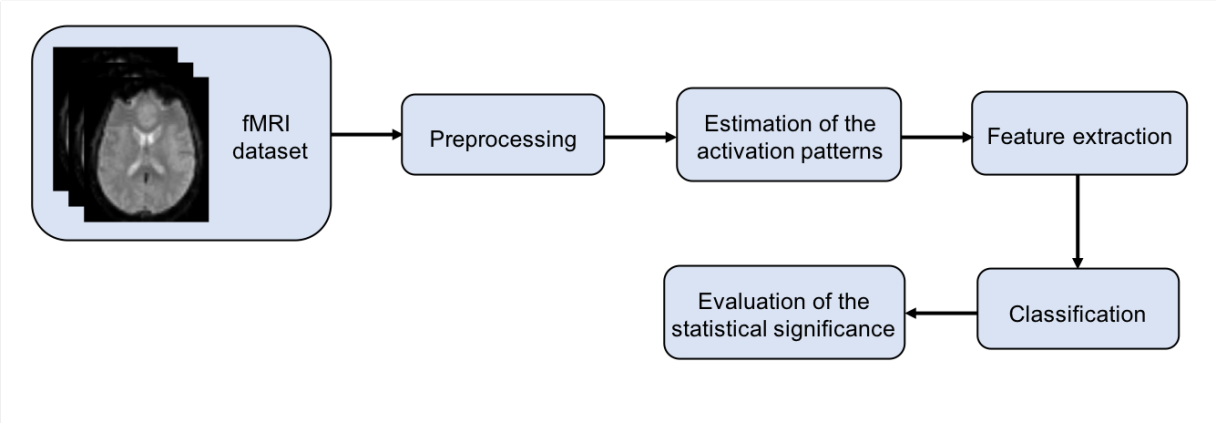


**Figure 4.1:** Illustration of the general framework in fMRI classification.


**Figure 4.2:** Different ways of computing the inputs of the classifier in fMRI studies.

the conditions of interest are used to build the classification model (LaConte et al., 2005; Mourão-Miranda et al., 2006). Another option is to average the volumes associated with each condition to obtain a more representative measure of fMRI response (middle of Figure 4.2, Cox and Savoy, 2003; Haynes and Rees, 2005; Kamitami and Tong, 2005; Mitchell et al., 2004; Mourão-Miranda et al., 2006). However, these approaches have two main drawbacks. First, they can only be applied when estimating the activity associated with a block, but not to a series of events of different conditions. Second, they do not take into account the sluggishness of the hemodynamic response. A specific fMRI volume does not represent the activity elicited by the event that starts/finishes at that point but a combination of the neural activity due to previous events. To address this issue, a GLM is usually applied (Friston et al., 1995). This method computes, for each time point, the contribution of each event defined in the experiment to the hemodynamic response (bottom of Figure 4.2). The resulting beta maps are obtained as the deconvolution of the hemodynamic response from fMRI signal, which are then used to input the classifier (Arco et al., 2018; De Martino et al., 2008; Haxby et al., 2001). Other studies employ $t$-values, which can be obtained by dividing the beta estimate for each voxel by its standard-error estimate (Kriegeskorte et al., 2008; Martínez-Ramón et al., 2006; Misaki et al., 2010).

## 4.3 Dimensionality reduction

One of the main challenging problems of classification in neuroimaging is related to the small sample size problem (Button et al., 2013). Most neuroscience datasets contain imaging from a few tens of subjects, but there are hundreds of thousands voxels in each image. This means that the number of parameters that the classification model has to deal with is much larger than the number of samples available. In this situation, classifiers are not able to identify the actual patterns of information but they adapt to the idiosyncrasies of these particular data, which leads to a poor generalization to unseen data. This issue is termed overfitting, and it is one of the main obstacles to achieve a good classification performance.

A solution to this issue is to reduce the number of features (i.e. voxels in neuroimaging) that the classifier receives as input. One possibility is to find an optimal subset of features, discarding irrelevant ones, that improves classification performance. This operation is known as feature selection and there are three main alternatives based on the criterion used: filtering, wrappers and embedded.

The simplest approach is filtering methods (Martinez-Murcia et al., 2017), and they evaluate the relevance of different features based only on their intrinsic properties. Specifically, they compute a score for each feature (voxel) and rank them according to this score. As feature selection is computed before classification, this process is not done in interaction with the classifier. Most of these scores are derived from statistical tests, such as the *t*-test (Bron et al., 2015). In this case, a *t*-test is performed for each voxel of the images, and only those that show significant differences between the two classes are then used as input of the classifier. The main drawback of filtering methods is that they do not guarantee a boost in performance since feature selection is not done according to the classifier performance, but they guarantee that the subsequent classification is based on statistically significant voxels. Another issue is related to its univariate nature: although differences in a single voxel are not significant, information distributed along a group of voxels can be significant.

Wrapper approaches (Chen and Chen, 2015) also compute a score for each feature but they perform the voxels selection in conjunction with the classifier performance. They eliminate a subset of voxels and use the remaining ones as input to the classifier, measuring the resulting classifier accuracy. This procedure is iteratively repeated, so that different subsets of voxels are eliminated in each iteration. The classifier finally selects the voxels that lead to a higher performance. Wrapper methods are simple because no previous assumptions are required. However, the iterative search that they rely on produces a large computational cost. Embedded approaches (Zhang et al., 2015) emerged as an optimization of wrapper ones. They also perform the aforementioned iterative process, but they also evaluate different hyperparameters associated with the classifier to select the one that leads to a maximum performance (Hanso and Halchenko, 2008; Hoeft et al., 2011).

There are other alternatives based on geometric transformations to reduce the dimensionality of neuroimaging data. Principal Component Analysis (PCA, Jolliffe, 2002) or Partial Least Squares (PLS, Segovia et al., 2013) project data to a lower dimension to facilitate the extraction of the information. Specifically, each image is decomposed as a linear combination of different components, so that only a few of them are used for the subsequent classification. Based on this concept, previous studies have evaluated different geometric transformations such as Nonnegative Matrices Factorization (NMF, Ghanbari et al., 2014) and Independent Component Analysis (ICA, Zhang et al., 2011).

All of these techniques overcome the dimensionality problem and increase the classifier performance. This can be quite beneficial in the classification context, where the

classifier performance is of primary interest. However, dimensionality reduction must be carefully done in the identification perspective. Overfitting is a problem to address also in this scenario, but solutions may differ compared to the classification context. As we mentioned before, most feature selection methods base the election of the proper features on the classifier performance. Voxels that are not included in the subset that leads to the maximum accuracy are automatically discarded. Nonetheless, accuracy and sensitivity are not always simultaneously obtained. A classifier that yields the maximum accuracy is not necessarily the one that identifies all the informative brain areas. Thus, feature selection in identification should not be derived from the classifier with the largest accuracy but from the one that marks as informative the highest number of voxels that truly contain information.

Another problematic question of dimensionality reduction in identification contexts is the loss of spatial information that these methods entail. As we mentioned before, PCA performs a projection from the original space (voxel-space) to a new one obtained from a linear combination of the original space (components-space). In the beginning, the information is contained in thousands of voxels, but once PCA is applied, only a few components remain. After classification, it would be possible to evaluate the importance of the components in the classification decision (i.e. how much each component has contributed to that decision). However, it would not be possible to associate each component to any of the initial voxels, missing the ability to identify the regions of the brain that contain information. For this reason, dimensionality reduction employs other alternatives for identification purposes.

The easiest approach for reducing the dimensionality is to limit the analysis to a specific brain region instead of using the whole-brain. This is known as region of interest (ROI) analysis, and it is particularly useful when there is an *a priori* hypothesis about the role of a region in a specific function (Poldrack, 2007). Kriegeskorte et al. (2006) proposed Searchlight, an alternative that does not require a previous definition of any region. This method builds a spherical region, so that classification is only performed from the voxels contained in it. The sphere sweeps all the positions in the brain, yielding an accuracy map where each position denotes the accuracy of the classification analysis when the voxel was the center of the sphere. Searchlight has become the most common approach for analysis of fMRI data and there is a large number of studies that have brought new insights into the localization of human brain function (Coutanche et al., 2011; González-García et al., 2017; Kahnt, 2018; Lee et al., 2017; Soon et al., 2008).

## 4.4 Classification

The classifier is the mathematical function that computes the decision boundary to separate the patterns associated with the different classes. There is a high number of classification algorithms, and they differ in the shape of the decision boundary. However, they can be divided in two main categories: linear classifiers and non-linear classifiers. Linear classifiers are simpler than non-linear ones, and they are commonly used in neuroimaging. As its name suggests, the decision function is derived from a linear combination of the inputs voxels. There are different linear classifiers depending on how the decision boundary is computed. Fisher's Linear Discriminant (LDA) selects the decision function that optimally discriminates the covariance matrices of the two classes while maximizing within-class variance (Bishop, 2006). Moreover, Support Vector Machines (SVM) take the decision function that maximizes the margin to the patterns of the two classes (Cristianini and Shawe-Taylor, 2000). In linear classifiers, each voxel has a certain weight that represents its contribution to the final classification decision. Moeover, in non-linear classifiers it is considerably harder to infer the informativeness of each voxel.

As in dimensionality reduction, the choice of the classifier depends on the context evaluated. In some situations, non-linear classifiers can adapt better to the idiosyncrasies of the data, performing consistently better than linear (LaConte et al., 2005; Mwangi et al., 2013; Rasmussen et al., 2011). Nevertheless, information about the specific localization of the sources of information that these algorithms provide is difficult to interpret. This can be problematic when applying to an identification context. Previous studies have overcome this issue by using non-linear classifiers in conjunction with ROI or Searchlight analyses (Misaki et al., 2010; Pereira et al., 2009). Thus, we can infer that the information that leads to a high accuracy relies on the region previously selected for the classification (ROI) or on the sphere centered in a specific voxel (Searchlight).

### 4.4.1 Within/Between-subjects

Another difference between classification and identification scenarios is how data from different subjects are analyzed. Classification in clinical studies usually employs data of different subjects of the dataset to train and test the classifier. For instance, images from all but one subject are used to train the classification model, whereas the images from the remaining one are employed to test its performance. This is known as between-subjects classification (Altaf et al., 2018; Arco et al., 2015; Martinez-Murcia et al., 2016). In most

psychological studies focused on identification, MVPA aims at detecting fine-grained differences from patterns of response associated with each individual. Since these differences are generally subtle, variability can emerge between different subjects regarding the specific distribution of the information pattern. For this reason, classification analyses are usually performed within individual subjects (Arco et al., 2018; Palenciano et al., 2018). Recent advances in data alignment have been proposed to overcome this problem. Hyperalignment (Haxby et al., 2011) is a promising framework that aligns brains based on their neural response and not on their anatomy as spatial normalization does. This method performs for each individual subject a transformation of the temporal trajectories of voxels to build a common model space. Recent studies have shown the promising results of this approach, suggesting a boost in classification performance (Guntupalli et al., 2016; Haxby et al., 2014).

Performing a within or a between-subjects classification has several implications regarding the independence of fMRI data. Data used for training the classifier need to be different from those employed for testing. Otherwise, the measure of the classifier performance would be biased. In between-subjects classification, data from training and test belong to different subjects, guaranteeing automatically their independence. However, in within-subject analysis, classification is performed separately for each subject, so that data both for training and testing belong to the same individual. The sluggishness of the BOLD signal makes close-in-time beta maps nonindependent (Lindquist et al., 2009), and the separation between images used for training/test must be done carefully. One solution is to take advantage of how fMRI studies are structured. Typical fMRI studies divide the total period of scanning into a number of smaller runs. This has more than an organizational impact since each run start reequalizes image intensity. MVPA usually employs runs as independent units for training and testing (Coutanche and Thompson-Schill, 2012): data belonging to all but one run are used for training and the remaining one for testing. Figures 4.3 and 4.4 show an schematic representation of between and within-subjects classification, respectively.

## 4.5 Statistical significance

When applying classification to neuroscientific studies, measuring the performance of the classifier is important, but evaluating its significance is crucial. In this context, the main aim is to determine the probability of a classification result at the group level. However, as explained in previous section, classification is usually performed on each

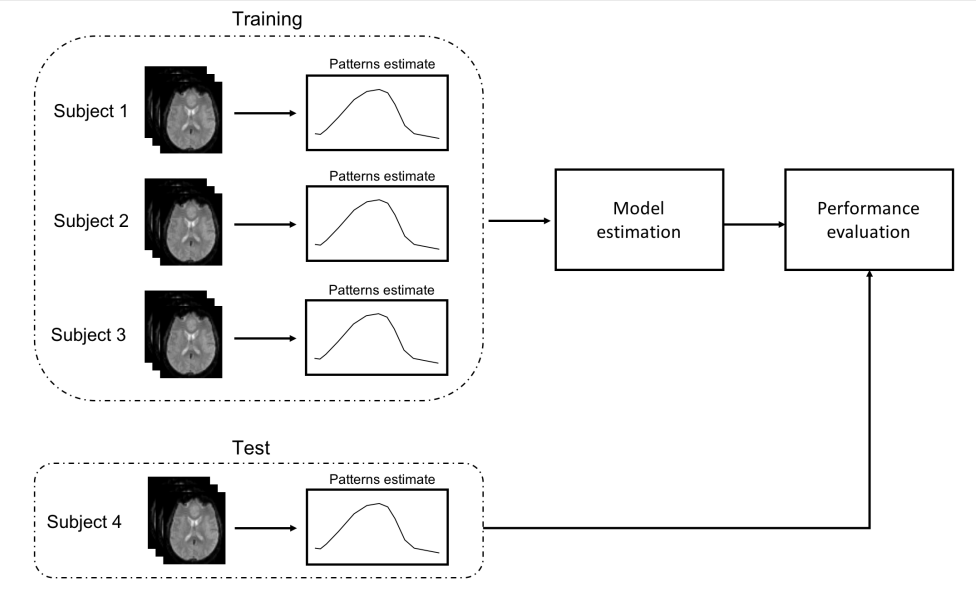


**Figure 4.3:** Schema of between-subjects classification

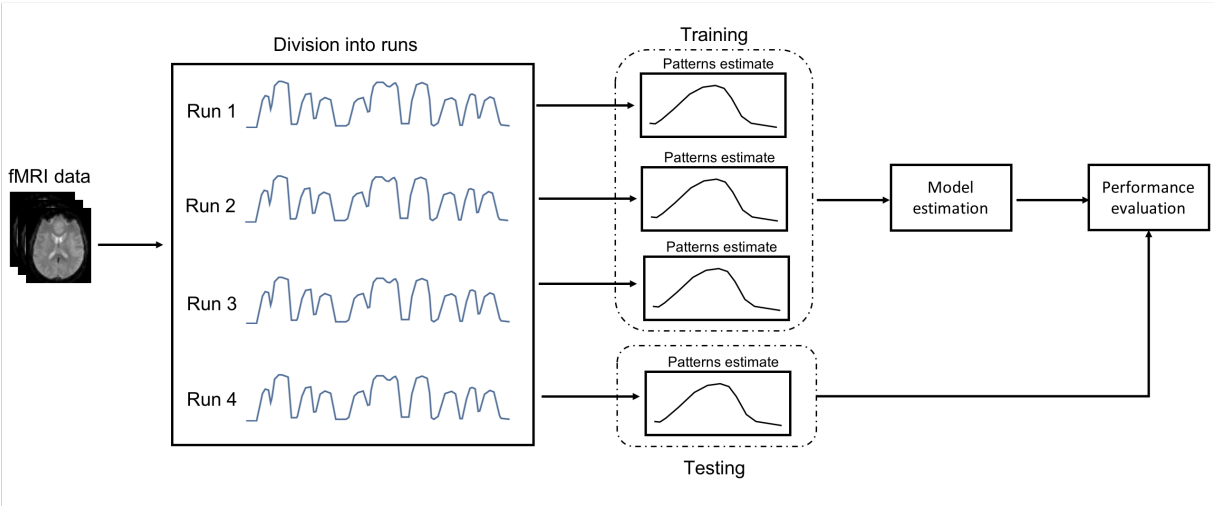


**Figure 4.4:** Schema of within-subjects classification

participant separately. A second-level analysis is required to group results for all subjects and evaluate if they are statistically relevant. The $t$-test is one of the simplest and most widely used statistical methods for assessing significance at the voxel level. Once the statistical group map is computed, we could infer a specific role of the voxels marked as significant in the cognitive function under study. The election of the significance level is not trivial in neuroimaging. The use of a significance level of $p < 0.05$ means that if one hundred statistical tests are performed, five would wrongly reject the null hypothesis. However, a MR image contains hundreds of thousands voxels, which means that thousands of voxels are potential false positives.

The study conducted by Bennett et al. (2009) highlighted the danger of an inadequate correction of statistical tests. They found significant voxels in an fMRI scanning session with a post-morten Atlantic Salmon as the subject. They argued that relying in standard statistical thresholds is ineffective for a proper control of false positives, even with restrictive significance levels ($p$<0.001). This is known as multiple comparisons problem, and it appears when a number of statistical tests are simultaneously performed. Usually, neuroscience studies employ two options for controlling the amount of false positives: Family Wise Error (FWE, Nichols and Hayasaka, 2003) and False Discovery Rate (FDR, Benjamini and Hochberg, 1995).

Despite it is not possible to fully eliminate Type 1 errors, these methods reduce considerably the number of false positives (Eklund et al., 2016). Nonetheless, recent claims have not recommended the use of parametric methods when assumptions about the Gaussianity of data are not met. Specifically, results obtained in Eklund et al. (2016) suggest that parametric statistical methods are conservative for voxel-wise inference and that they lead to inflated results for cluster-wise inference. A solution is to employ non-parametric methods based on permutation testing. These approaches rely on minimal assumptions, and do not assume a specific shape of data, but build an empirical distribution from a number of permutations.

All the stages in a classification framework are important to yield an optimal solution. When applied to fMRI data in a Cognitive Neuroscience context, the methods employed in each step of classification must preserve the spatial information. In this scenario, the optimal choice of the different approaches remains unclear. Next chapters focus on explaining the most common alternatives in Cognitive Neuroscience, in addition to provide different comparisons aiming to obtain the methods that lead to the best performance.

CHAPTER 5

# CLASSIFICATION APPROACHES IN FMRI

Mapping the human brain has been one of the main goals of neuroscientists in the last decades. The development of fMRI has made possible to study brain activity associated with each location in the brain while participants perform different tasks. This chapter summarizes the different alternatives used for fMRI classification where localizing informative regions is the main goal. Moreover, this chapter provides an overview of the most common classifier, SVM, in addition to the different measures employed in this context to evaluate the performance of the classification.

## 5.1 Introduction

In Chapter 4, we discussed the problem of the dimensionality that most fMRI experiments have. The number of features that the classifier has to learn from is much larger than the number of samples available. Most feature selection methods cannot be used in an identification context because they lead to solutions that lose the spatial information. These methods yield a classification solution, but they are not able to report the brain regions used to reach that classification. This considerably limits the approaches for classification that can be used in this kind of scenarios. Throughout this chapter, we explain the three main classification analyses usually employed in Cognitive Neuroscience.

## 5.2 Classification methods

### 5.2.1 ROI analysis

The simplest approach for identification contexts is to define an ROI based on *a priori* knowledge. In this case, the classifier uses only the voxels contained in the region, and the resulting performance highly depends on how well the *a priori* hypothesis fits the observed data. Results can be biased if ROIs are selected from previous analysis using the same data (Kriegeskorte et al., 2009). For this reason, the reason that motivates the ROI definition (either anatomical or functional) has to be derived from experimental hypothesis during the design stage. The best practice to define anatomical ROIs is to employ data from each individual subject due to the existent variability between subjects (Etzel et al., 2009). It is also possible to use the regions defined in different atlases, but there can be a lack of overlap between these atlases and the brains of different subjects (Nieto-Castanon et al., 2003). Functional ROIs can be generated from data of an individual subject, using a localizer that identifies the voxels in a brain region that show a specific response. Another alternative is to take the coordinates of the significant regions from a previous study and evaluate this ROI over the data of the current experiment. It is also possible to employ an ROI obtained in a meta-analysis of the domain of interest of the current study (Turkeltaub et al., 2002; Wager and Smith, 2003), which generates ROIs less sensitive to noise than those based on the results of a single study (Poldrack, 2007).

One of the main benefits of ROI analysis is that it is possible to draw conclusions about the region as a whole, which is particularly interesting from the psychological standpoint. Besides, ROI-analysis reduces Type 1 errors since only one statistical test is applied to the whole region (Poldrack, 2007). There is a large number of studies that use this kind of analyses. Haxby et al. (2001) employed an ROI in the ventral temporal cortex to study whether faces and objects are differentially represented. Results obtained by Haynes and Rees (2005) demonstrated the existence of an orientation-selective processing in the primary visual cortex (V1). Figure 5.1 illustrates how classification is performed in ROI analysis.

### 5.2.2 Whole-brain analysis

There are situations in which there is not a strong hypothesis about the regions involved in a certain function. In these cases, classification is not limited to a specific region, but

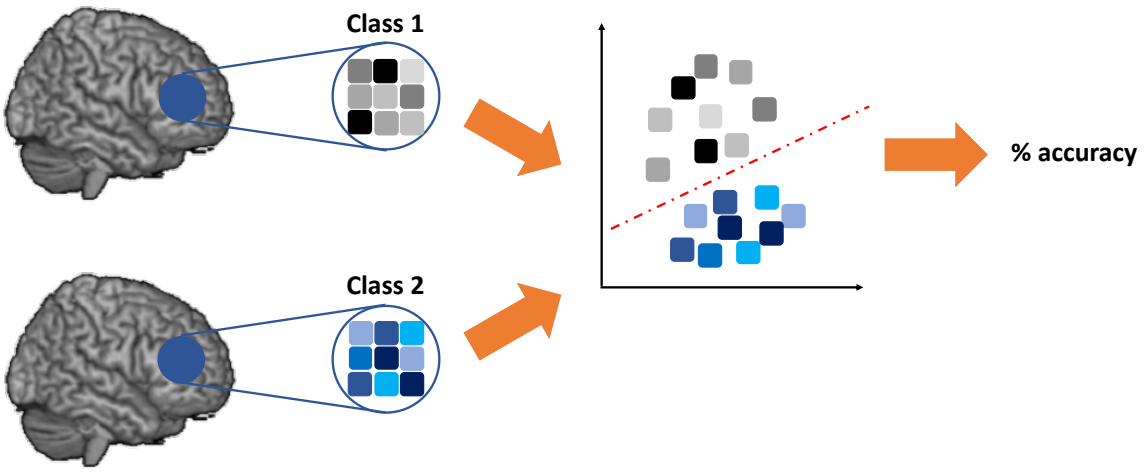


**Figure 5.1:** Schema of ROI classification analysis. Only voxels contained in the region of interest are used for the subsequent classification.

the whole brain is explored. This approach is termed whole-brain analysis, and like its name suggests, the classifier receives as input all voxels in the brain. The classification algorithm learns the information from tens of thousands voxels from each of the tens or hundreds fMRI volumes. As we mentioned before, the large difference between features (voxels) and samples (MR images), which is known as the curse of dimensionality, is an issue to overcome in this kind of analyses. In this case, finding the optimum hyperplane to properly separate the different classes is difficult since there is a lot of information to extract and a few examples to find a generalizable solution (Fort and Lambert-Lacroix, 2005). The use of a *t*-test for feature selection has been successfully applied in previous analyses (Balci et al., 2008; De Martino et al., 2008; Mwangi et al., 2014). However, this can be problematic because of the within-subject classification usually performed in Cognitive Neuroscience studies. Performing a significant test at the subject level and then evaluate again the statistical significance at the group-level could lead to a double-dipping, i.e. the use of the same data for selection and selective analyses (Kriegeskorte et al., 2009). For this reason, feature selection methods are not commonly used in identification scenarios.

When a classification is performed, the value of each weight is estimated in an optimization process according to some criterion. For instance, in an SVM classifier the resulting weights will be those that lead to a maximum margin between the support vectors. The resulting weight of each voxel represents its contribution to the classification function, whereas its sign is associated with the direction of this contribution. A large weight (either positive or negative) represents that the voxel is relevant according to the

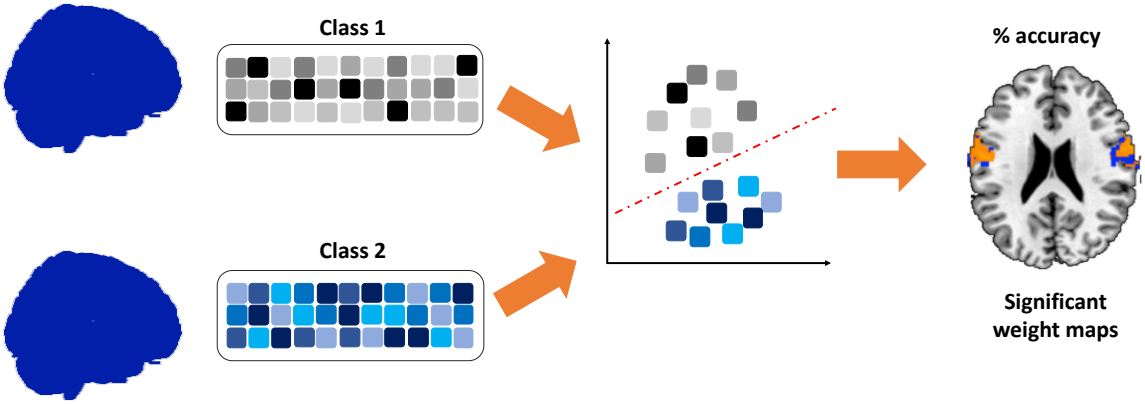


**Figure 5.2:** Schema of whole-brain classification analysis. All voxels in the brain are employed by the classifier, resulting in a weight map which informs about the contribution of each voxel to the decision function.

classification algorithm, whereas a positive/negative weight indicates that the voxel is associated with the first/second class in a binary classification. Although the classifier employs all voxels to build the hyperplane, weights inform about the importance of each separate voxel in the classification. In whole-brain analysis, spatial information is usually provided by the resulting weights. However, obtaining a map of tens of thousands weights (one for each voxel) is not informative enough to draw an inference from a specific brain region. One solution is to perform a statistical test in the resulting weight map to evaluate the significance of each voxel. Mourão-Miranda et al. (2005) performed permutation tests to generate a map of $p$-values for each voxel. Only those that surpassed the statistical threshold were marked as significant. Other studies have used a similar methodology since this is the simplest way to localize information when a large number of features are used for classification (Klöppel et al., 2012; Marquand et al., 2014). Figure 5.2 illustrates this framework.

### 5.2.3 Searchlight

One of the most appealing methods for localization of informative regions was proposed by Kriegeskorte et al. (2006). This approach, known as Searchlight, defines a small spherical region and performs a classification analysis using the voxels contained in the sphere. The resulting performance measure (usually the accuracy) is then assigned to the central voxel. This sphere is moved across the brain, yielding an accuracies map once all voxels have been the central voxel of the sphere (see Figure 5.3). One of the main advantages of Searchlight is that it does not require an *a priori* specification of a region.

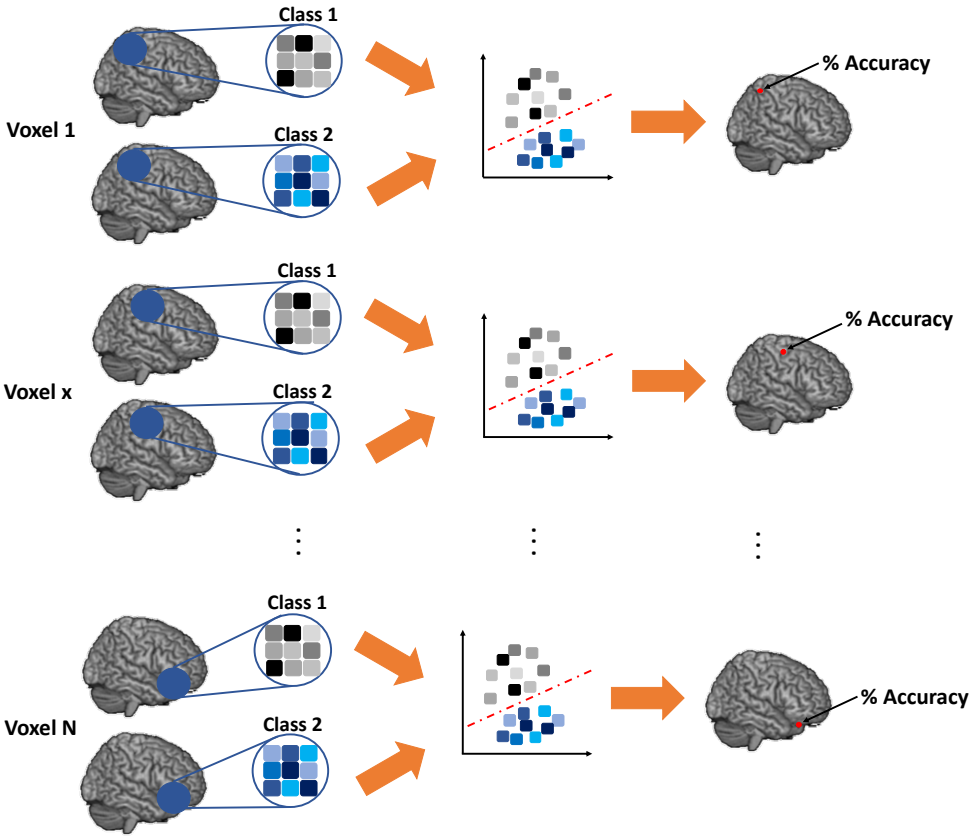


**Figure 5.3:** Schema of searchlight analysis. The sphere sweeps all the positions in the brain performing a classification in each one of them, assigning the resulting accuracy to the central voxel. Once all voxels have been the center of a sphere, an accuracies map is obtained.

This approach explores all the positions of the brain, so that results are not limited to a certain region. Searchlight minimizes the effects of the curse of dimensionality associated with whole-brain approaches with the definition of the sphere. There are many studies that have successfully employed Searchlight since it offers results that are potentially easy to interpret (Chen et al., 2017; Cichy et al., 2016; Coutanche et al., 2011; González-García et al., 2017; Qiao et al., 2017).

Nonetheless, Searchlight has some limitations to consider. Its performance in terms on accuracy and number of significant voxels depends on the size of the sphere (we explore this dependence in Chapter 7). Another important issue is the fact that the resulting accuracy is linked to the central voxel of the sphere. This methodology obviates that only a few voxels within the sphere usually contain truly information, considering voxels as significant when they actually are not (see Figure 3 in Etzel et al., 2013 for an extreme example of distorted results). An additional problem is related to computational cost. Each Searchlight requires as many classifications as voxels are in the brain. This time

is even larger when permutation tests are applied to assess the statistical significance of the results. When this procedure is employed in combination with grid search, the computational cost exponentially increases. This approach evaluates the performance of the classifier for different values of the hyparameters associated with the classifier, selecting those that lead to best results.

## 5.3 Support Vector Machine

Previous section has summarized the three main alternatives to select the voxels that are employed in the subsequent classification. In this section we provide an overview of one of the most commonly used classifiers in fMRI analysis: SVM (Cristianini and Shawe-Taylor, 2000). This approach employs a hyperplane to separate different classes (two, in a binary classification). Since different hyperplanes lead to different solutions, SVM selects the one that maximizes the distance between the hyperplane and the nearest data points of each class. This distance is known as margin, and the nearest data points are usually termed support vectors. From a mathematical perspective, it is possible to specify an SVM classification (Bennett and Blue, 1998; Burges, 1998) rule $f$ by a pair of $(\mathbf{w},\mathbf{x})$, from equation:

$$f(\mathbf{x}_i) = \langle \mathbf{w}, \mathbf{x}_i \rangle + b \tag{5.1}$$

where $\mathbf{w}$ is the weight vector, $\mathbf{x}_i$ is the feature vector and $b$ is the error term. Thus, a point $\mathbf{x}$ is classified as positive if $f(\mathbf{x}) > 0$ or negative if $f(\mathbf{x}) < 0$. This can be seen as an optimization problem: the resulting decision function is based on a rule that maximizes the geometrical margin between the two classes, and this solution is obtained by minimizing the classification error (Boser et al., 1992). This is given by:

$$\begin{aligned} &\frac{1}{2}\|\mathbf{w}_m\|^2 + C\sum_i \xi_i \qquad \text{subject to} \\ &y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i \qquad \forall_i \xi_i \geq 0 \;\; \forall_i \end{aligned} \tag{5.2}$$

where $C$ is a parameter known as penalty for misclassification, or cost parameter. This parameter is introduced for situations in which it is not possible to obtain a perfect classification during the training of the classifier. The parameter $C$ allows a few misclassifications among the training subset in order to guarantee a good generalization performance in the test subset. It also minimizes the overfitting of the training data by controlling regularization: this reduces the probability of adjusting the classifier

to the features of the training sample, which can lead to a decrease in the resulting accuracy when the test sample is evaluated. A low value of $C$ lets a large number of misclassifications, leading to a result with high bias and low variance. Moreover, a large $C$ highly penalizes for misclassified data, so that the classifier employs more complex decision functions that let a perfect classification.

From Equation 5.2, the solution to the optimization problem can be written as:

$$\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i \tag{5.3}$$

after applying the Lagrangian multipliers. We have assumed that data are linearly separable, but this is rarely met. To solve this problem, a mapping from the original space to a new one is usually applied, which generally allows a linear separation of the data. Nonetheless, it is not simple to know the exact transformation between the two spaces. Previous studies have employed the kernel trick (Mika et al., 1999; Schölkopf, 2001), a mechanism that applies the transformation mentioned before by using the inner product, in which is termed a kernel function (Min and Lee, 2005; Scholkopf and Smola, 2001).

Substituting the value of $\mathbf{w}$ in Equation 5.1, it is possible to rewrite the decision function in its dual form as:

$$f(\mathbf{x}_i) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b \tag{5.4}$$

where $\alpha_i$ and $b$ represent the coefficients to be learned from the examples and $K(\mathbf{x}, \mathbf{x}_i)$ is the kernel function characterizing the similarity between samples $\mathbf{x}$ and $\mathbf{x}_i$, where $\mathbf{x}_i$ are the support vectors. The use of a kernel function decreases the complexity of the optimization problem since it does not depend on the feature space but on its dimensionality. The kernel operator plays an important role in a classification task. In fact, variations in the kernel lead to different shapes of the hyperplane, which in turn can lead to a different classification performance. We describe in next section the most representative kernel functions in SVM classifiers.

### Linear Kernel

As its name suggests, this kernel uses a linear combination of the features to build the corresponding classifier. This is the default version of SVM, and employs the inner product as a measure of similarity between two values. The mathematical expression that defines a linear kernel is given by:

$$K(\mathbf{x},\mathbf{y}) = \langle \mathbf{x},\mathbf{y} \rangle \tag{5.5}$$

The only parameter that can be controlled in a linear SVM classifier is the one inherent to the support vector machine algorithm: the cost for misclassification ($C$). This parameter defines the margin between the hyperplane and the support vectors. Small values allows for large margins, whereas large values lead to small margins (Cherkassky et al., 1999; Cherkassky and Ma, 2004). Figure 5.4 shows the effect that different values of parameter $C$ have in classification. In this example, we employed simulated data that are not linearly separable. Low values like $C = 0.001$ or $C = 0.01$ lead to an extreme case in which the algorithm classifies all samples as belonging to Class 1. Since both classes were balanced, the misclassification cannot be due to a difference in the number of samples of the two classes, but to the low value of the penalization term. Setting larger values (bottom of Figure 5.4) establishes a separation hyperplane between the two classes. However, there are no differences between using a $C = 1$ or $C = 10$. In this specific scenario, modifying the value of $C$ will not lead to an increase in performance because classes are not linearly separable. Thus, a different kernel should be employed to yield a perfect classification.

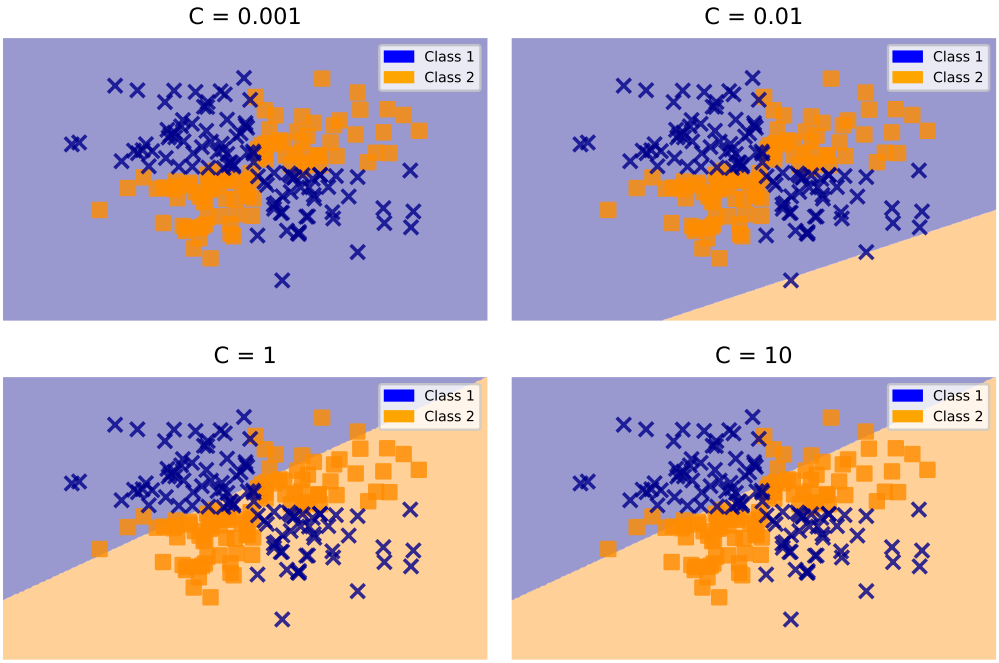


**Figure 5.4:** Influence of the cost parameter (C) in the decision function of a linear classifier.

### Polynomial Kernel

The polynomial kernel also employs an inner product, but includes additional parameters to the standard linear kernel, as follows:

$$K(\mathbf{x},\mathbf{y}) = (\langle \mathbf{x},\mathbf{y} \rangle + 1)^q,\ q \in N \tag{5.6}$$

where $q$ is the degree of the polynomial. When $q = 1$, this function is equivalent to the linear kernel (left side of the top of Figure 5.5), whereas $q = 2$ would lead to a quadratic kernel (right side of the top of Figure 5.5). The degree parameter controls the flexibility of the decision boundary: kernels with larger degrees yield a more flexible decision boundary. It is crucial to find the optimum degree of the polynomial that fits best the data. In our simulated example, none of the kernels lead to a perfect classification. However, the different decision functions derived from each classifier demonstrate the large influence that this parameter has in the results.

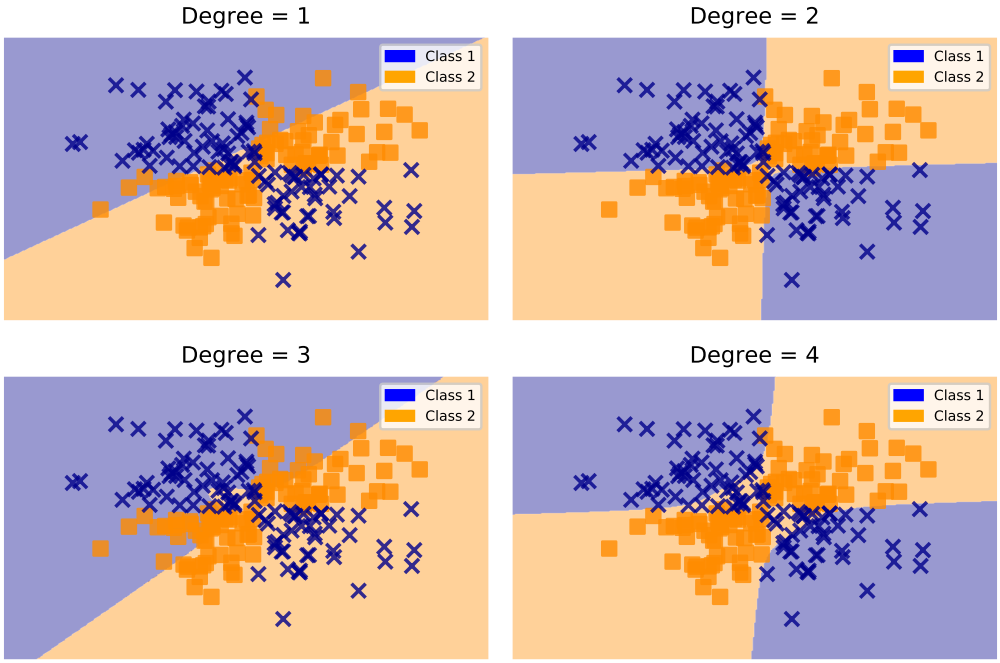


**Figure 5.5:** Influence of the degree parameter in the decision function of a polynomial classifier.

### Radial Basis Function Kernel

Another version of the SVM classifier that creates nonlinear decision boundaries is the Radial Basis function (RBF). This approach represents the similarity measure as a

decaying function of the distance between two vectors, computed as the squared norm of their distance. This kernel is defined as:

$$K(\mathbf{x},\mathbf{y}) = e^{-\gamma\|\mathbf{x}-\mathbf{y}\|^2},\ \sigma > 0$$
$$\text{where } \gamma = \frac{1}{2\sigma^2} \tag{5.7}$$

Thus, if the two vectors are similar, the $||\mathbf{x}-\mathbf{y}||$ will be small. The RBF function follows a bell-shaped curve. The $\gamma$ parameter represents the width of the bell-shaped curve. This means that large values of $\gamma$ (which correspond to low values of $\sigma$) will lead to a narrow bell, whereas small values of $\gamma$ (large values of $\sigma$) will yield wide bells. Figure 5.6 illustrates the influence of the $\gamma$ parameter in the decision hyperplane. A small value of $\gamma = 0.01$ leads to a hyperplane very similar to the one obtained by a linear classifier (see bottom of Figure 5.4). As $\gamma$ increases, the classifier adapts better to the idiosyncrasies of each class. In fact, a $\gamma = 0.1$ highly improves the accuracy of $\gamma = 0.01$, and a value of $\gamma = 1$ yields better results that the previous two. Finally, an algorithm with $\gamma = 10$ demonstrates a good performance, close to a perfect classification (bottom of Figure 5.6).

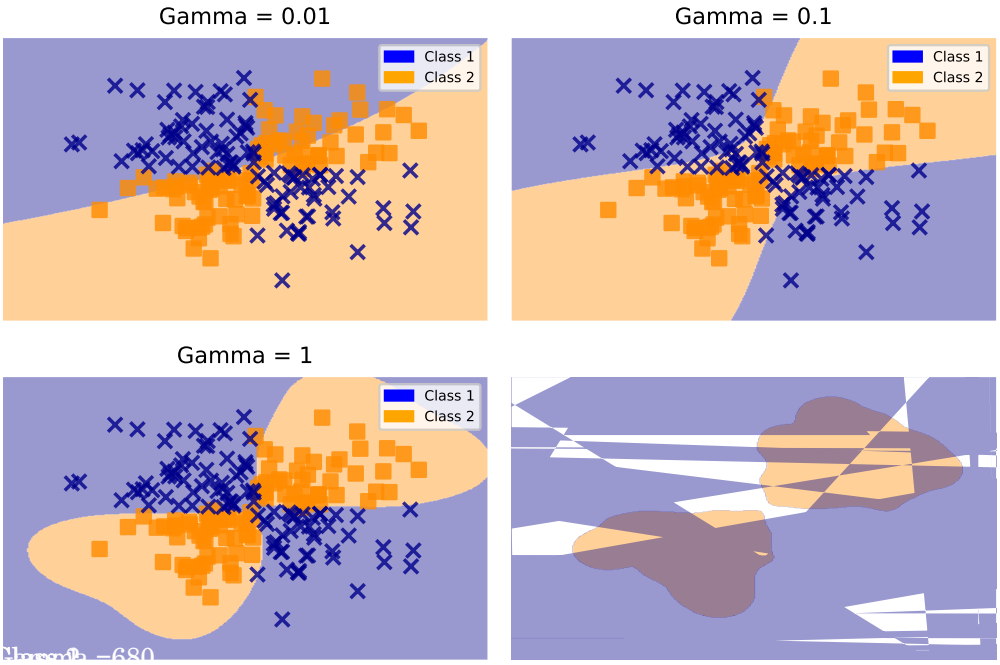


**Figure 5.6:** Influence of the gamma parameter in the decision function of an RBF classifier.

## 5.4 Cross-Validation

Machine learning algorithms aim at learning the information contained in an input dataset to classify new observations. Once they have identified the relevant information to build the decision function, their performance is evaluated. Data employed for evaluation have to be different than those used for learning. As we mentioned before, neuroimaging datasets usually have a limited sample size where only a tens of subjects are available. This problem is even harder in psychological experiments where within-subject classification is performed. In this case, the images from each subject are split in two different subsets, yielding a much more limited sample size than in between-subject classification, where the classifier learns from data of different subjects.

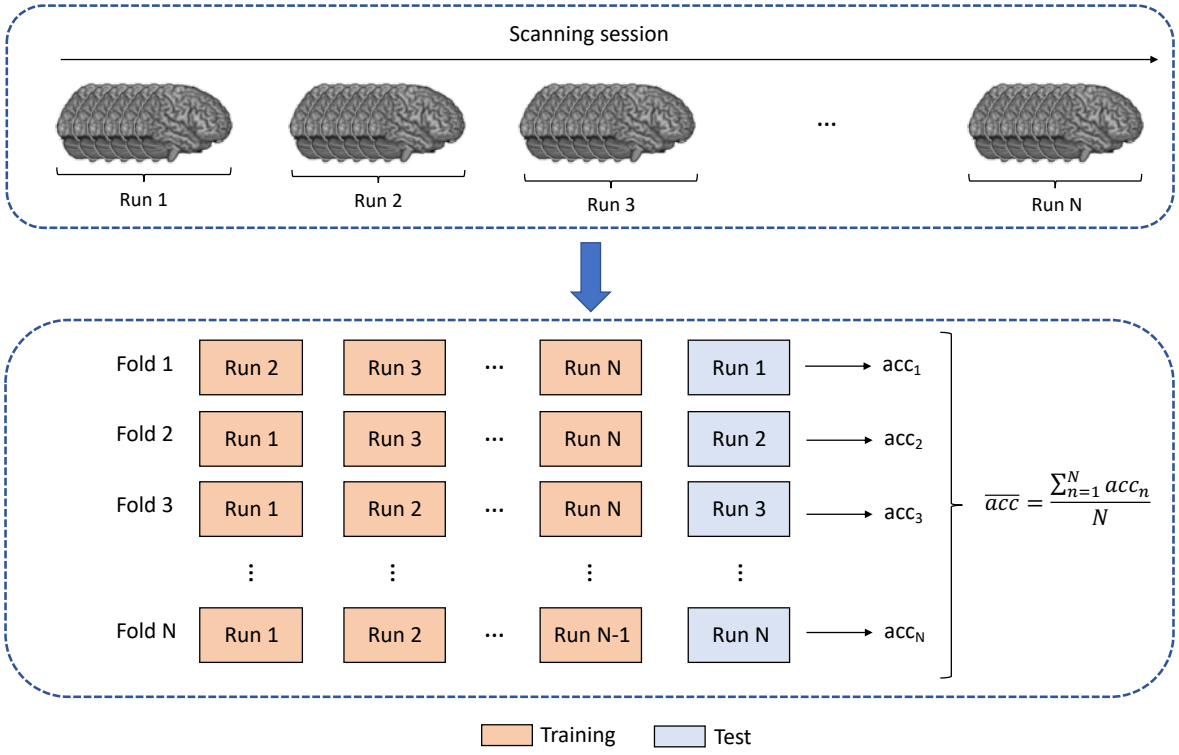


**Figure 5.7:** Diagram of the LORO cross-validation scheme employed in most fMRI analysis.

The sample size issue tends to be solved by employing cross-validation. This mechanism iteratively divides the sample in different subsets so that part of these subsets is used to train the classifier and the rest is employed to evaluate the previous learning. The resulting performance is then evaluated as the average of the performance obtained in each one of these subsets. The simplest cross-validation scheme is known as $k$-fold, where $k$ is the number of subsets the dataset is divided into. In this case, $k$-*1* subsets are

used to train the classification algorithm and the remaining one is used to evaluate its performance. It seems clear that data used for training and tests must be independent to assure that the resulting performance is not biased. In fMRI experiments, the scanning session is usually divided into different runs to account for this independence. From the total number of runs (ranging from 8 to 12 in most studies), all but one is used to train the classifier and the remaining one is employed for test. This scheme is known as leave-one-run-out (LORO) cross-validation (see Figure 5.7). Some studies concluded that $k$-fold is most adequate because it achieves a good balance between bias and performance (Pereira et al., 2009; Varoquaux et al., 2017). Moreover, LORO overcomes the limited sample size in fMRI experiments by providing a larger ratio between training and test images.

## 5.5 Measures of performance

After cross-validation, the classifier performance is measured to evaluate its ability to extract relevant information that distinguishes between two different classes. Different measures can be derived from the confusion matrix (see Table 5.1), which relates all the possible combinations of a classifier output. A classifier can consider than an image belongs to the class positive or negative. Moreover, the actual label of the image can be positive or negative. Thus, true positive and true negative correspond to correct predictions, whereas false positive and false negative are associated with incorrect predictions. From these four possibilities derived from the confusion matrix, we can define three different measures known as accuracy, sensitivity and specificity:

$$\begin{aligned} acc &= \frac{TP+TN}{TP+FP+TN+FP} \\ sens &= \frac{TP}{TP+FN} \\ spec &= \frac{TN}{TN+FP} \end{aligned} \tag{5.8}$$

Sensitivity provides information about Type I errors, whereas specificity is associated with Type II errors. Regarding accuracy, it is the standard performance measure used in identification scenarios. However, there are some situations in which the dataset is not balanced, i.e. the number of images in each class is not the same. This can bias the classifier, considering that most images in the test set belong to the majority class because of this difference in the sample size. To address this issue, the balanced accuracy

**Table 5.1:** Confusion matrix.

| | | **Predicted** | |
|---|---|---|---|
| | | **Positive** | **Negative** |
| **Actual** | **Positive** | True Positive | False Negative |
| | **Negative** | False Positive | True Negative |

is employed, which is given by the average of the accuracy obtained for each class separately. For a binary classification, the balanced accuracy is given by:

$$b_{acc} = \frac{1}{2}\left(\frac{TP_1 + TN_1}{TP_1 + FP_1 + TN_1 + FP_1} + \frac{TP_2 + TN_2}{TP_2 + FP_2 + TN_2 + FP_2}\right) = \frac{1}{2}(acc_1 + acc_2) \quad (5.9)$$

We can conclude from Equation 5.9 that accuracy and balanced accuracy agree when the accuracy obtained from each class separately is the same. These measures provide a measure of the classification performance, which can be very informative in some scenarios. However, in Cognitive Neuroscience a value of accuracy is not informative *per se*. Instead, the statistical significance of the resulting accuracy is crucial in this scenario, and a proper evaluation of this significance is of primary interest.

CHAPTER 6

# STATISTICAL SIGNIFICANCE

The use of methods originally developed for computer science has revolutionized the neuroscience field. Despite the large procedural differences between this framework and classical univariate analyses, the ultimate aim of neuroscientists has not changed: localizing the brain involved in different functions. To do so, it is necessary to find differences in activity associated with different experimental conditions and then evaluate if those differences are significant. However, fMRI studies require conducting a large number of independent statistical tests (one for each voxel of the image), which can inflate the false positives rate and invalidate any inference. Besides, the statistical method employed to assess significance has a large impact on the results. Several studies have proposed different approaches (Gilron et al., 2017; Kahnt et al., 2010; Nichols and Holmes, 2002) and examined their suitability in neuroscientific contexts (Arco et al., 2018; Schreiber and Krekelberg, 2013). The aim of this chapter is to present the mechanisms employed for reducing the number of false positives in fMRI data in addition to provide an overview of the different statistical approaches used in this context.

## 6.1 The multiple comparisons problem

Statistical methods play a crucial role in neuroscientific studies, but this incidence is even more relevant in fMRI studies. This is mainly due to the large number of statistical tests that is usually required . There are other contexts in which it is only necessary to assess

if the resulting accuracy of a classification task is significant or not. In this case, only one statistical test is required. However, fMRI studies aim at finding differences at the voxel level, so that a different statistical test for each voxel of the brain is applied. A voxel is marked as significant if the associated $p$-value is lower than the threshold previously established. A conventional threshold is $p < 0.05$, which indicates the probability to declare a voxel as significant by pure chance. This means that if we repeat the same test 100 times, 5 voxels would be wrongly marked as significant on average. The likelihood of obtaining at least one false positive at $p = 0.05$ is given by:

$$L(n) = 1 - (0.95)^n \tag{6.1}$$

where $n$ is the number of independent tests performed. fMRI images have tens of thousands voxels, so that applying tens of thousands statistical tests would almost guarantee the appearance of false positives. In case of an fMRI volume of 50000 voxels, 2500 false positives would arise, distorting the results and invalidating any inference. One possible solution is to use a more stringent threshold. For example, a $p = 0.01$ threshold would decrease the number of false positives, but it would not be enough since 500 voxels would surpass the significance threshold by pure chance.

### 6.1.1 Family-wise Error Rate

There are different solutions to control for multiple comparisons, based on estimating the number of resulting false positives. The most common method used in fMRI analyses is termed family-wise error rate (FWE). This approach computes the probability that at least one false positive appears in a family of tests. Thus, FWE controls the probability that a false positive appears in any of the multiple tests simultaneously performed and provides the value of the threshold to guarantee a specific false-positives rate. One of the most famous alternatives based on FWE is the Bonferroni correction (Friston et al., 1996). This method considers that a proper significance global threshold is derived from the division of the standard threshold $p = 0.05$ by the number of significance tests performed. This can be expressed as follows:

$$P(T_i \geq u_\alpha | H_0) \leq \frac{\alpha}{m} \tag{6.2}$$

where $T_i$ is the value of the test statistic at voxel $i$, $u_\alpha$ is the target threshold for each individual test that provides the desired $\alpha$ correction, $H_0$ is the null hypothesis and $m$ is the number of independent tests performed. Following the previous example, it

would be necessary to apply a threshold corresponding to an $\alpha$-level of 0.000001 for each independent test in order to control the FWER at $\alpha = 0.05$ while performing 50000 tests. Bonferroni correction controls successfully for false positives but it tends to be too conservative. Employing too stringent significance levels increases the false negatives rate, which can lead to a large reduction in sensitivity. Thus, Bonferroni is not the optimal choice for family-wise errors correction in fMRI studies (Han and Glenn, 2018; Lindquist and Mejia, 2015; Nichols and Hayasaka, 2003).

Random Field theory (RFT) is another method for controlling the FWER that overcomes the limitation of sensitivity. RFT relies on a mathematical methodology that assumes that statistical maps have an underlying smoothness, and based on these assumptions, selects the statistical threshold that leads to a proper FWER. Although it is not possible to previously quantify the amount of smoothness in an image, RFT estimates it from the number of resolution elements (resels) that an image has. The concept of resel (Worsley et al., 1992) indicates the number of independent observations in an image (in this context, the number of possible statistical tests) and can be computed as:

$$R = \frac{V}{FWHM_x\, FWHM_y\, FWHM_z} \tag{6.3}$$

Thus, the number of resels depends on two variables: the number of voxels of the image ($V$) and the full width at half maximum (FWHM) along the three dimensions ($FWHM_x$,$FWHM_y$ and $FWHM_z$). The FWHM values represent the estimated smoothness of the fMRI data, which is a combination of their intrinsic smoothness derived from different variables such as reconstruction parameters and physiological noise. Another important concept is the Euler characteristic (EC, Brett et al., 2003), a feature that relates the number of clusters than can be considered as significant for a given statistical threshold. The expected value of EC is given by:

$$E[EC] = R(4log_e2)(2\pi)^{-\frac{3}{2}} Z_t e^{-\frac{1}{2}Z_t^2} \tag{6.4}$$

where $R$ is the number of resels and $Z_t$ is the threshold of the statistic value ($Z$). Setting the expected of EC to the standard value of 0.05 means that the remaining clusters of an image would have a maximum probability of 0.05 of occurring by chance, obtaining the threshold required for controlling the FWER. Previous studies have evaluated the performance of RFT and they found different results depending on the features of the fMRI dataset. Factors like smoothness and sample size have a large importance in RFT, leading to conservative results when both parameters are large (Lindquist and Mejia, 2015; Nichols and Hayasaka, 2003).

Holmes et al. (1996) provided an alternative to RFT based on permutation tests. This approach relies on minimal assumptions and its use in classification contexts has been validated theoretically (Golland and Fischl, 2003). The concept of permutations is relatively simple: they aim at testing the dependence between the labels and the experimental data. To do so, labels are shuffled and classification is performed, evaluating the performance of the classifier in terms of accuracy. This procedure is repeated a large number of times, yielding an empirical distribution of the accuracies. The probability of obtaining a certain accuracy is assessed by comparing the accuracy obtained after training the classifier with the actual labels and the empirical distribution (see Figure 6.1). The subsequent $p$-value can be computed as follows:

$$p = \frac{1+n}{N} \tag{6.5}$$

where $n$ is the number of accuracies from the empirical distribution that surpass the actual accuracy and $N$ is the number of samples used to build the empirical distribution. To evaluate the significance of a certain accuracy, it is necessary to compare the $p$-value associated with that accuracy with a significance threshold previously established (e.g. $p < 0.01$). We can conclude that an accuracy is significant if the associated $p$-value is lower than the significance threshold.

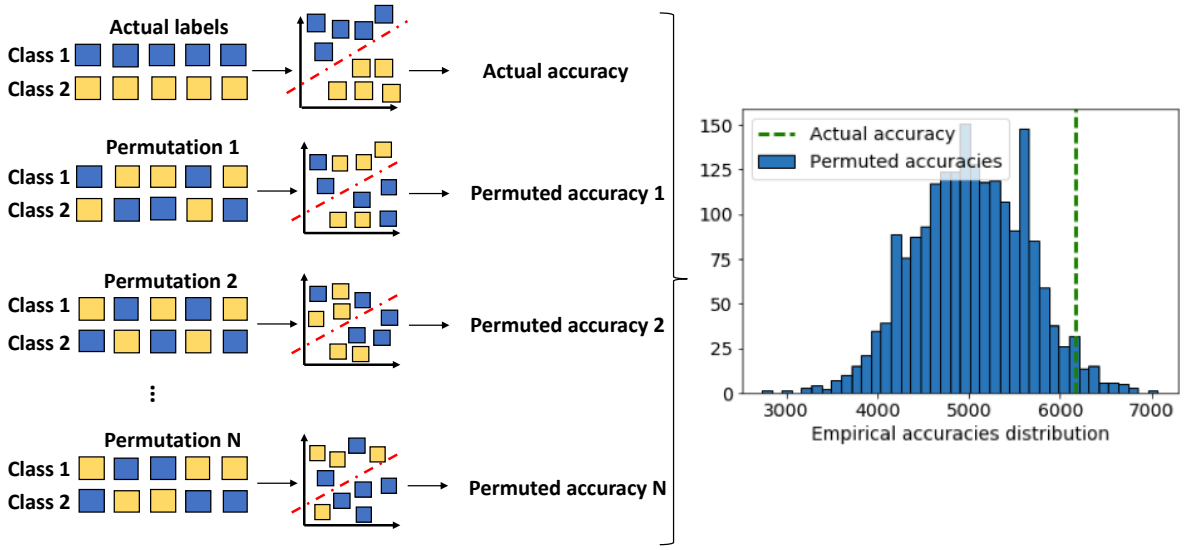


**Figure 6.1:** Diagram of the permutations approach (left). The classifier is trained with the permuted labels to build the empirical distribution (right). The actual accuracy corresponds to the one obtained after training the classifier with the actual labels.

### 6.1.2 False Discovery Rate

False Discovery Rate (FDR) is a relatively recent alternative to control for false positives (Benjamini and Hochberg, 1995). Unlike FWE, this approach does not focus on the probability of obtaining one false positive but computing the ratio of false positives among significant tests. From a total number of tests considered significants ($T$), only $S$ are really significant, so that $NS$ are the non significant tests that have been marked as significant. It is not possible to know the exact proportion of $S$ and $NS$ tests, but FDR computes the expected value of false positives, as follows:

$$FDR = E\left(\frac{NS}{T}\right) \tag{6.6}$$

Although the value of the $NS/T$ ratio is just an estimation, this mechanism controls on average the maximum FDR desired value, $FDR \leq p$, ranging from 0 to 1. Then, this method computes the corresponding $p$-values associated with a specific FDR rate (see Benjamini and Hochberg, 1995 for a detailed explanation of this process). Unlike FWE, this approach adapts to different levels of activity in the data, which means that the ability of this method of controlling for false positives varies depending on the features of each dataset. For large levels of activity, FDR can lead to very liberal statistical corrections. On the other hand, small levels of activity can induce the use of too-conservative thresholds, similar to Bonferroni in an extreme case (Lindquist and Mejia, 2015). There are not differences between FWE and FDR according to the results of previous studies, although FDR seems to be more exigent than FWE (Lindquist and Mejia, 2015; Nichols and Hayasaka, 2003).

### 6.1.3 Cluster-extent based thresholding

This approach relies on evaluating the minimum cluster size (number of contiguous voxels) that exceeds a previously established statistical threshold. As a result, this method does not control the probability that each voxel in the cluster is a false positive but controls the false positive probability of the whole cluster (Hayasaka and Nichols, 2003). This procedure usually consists of two steps. First, a primary threshold at the voxel level is set to define clusters that surpass the corresponding threshold. This threshold is arbitrarily selected, which can have a large influence in the results (Friston et al., 1994). Second, a cluster-level extent threshold is employed to compute the minimum cluster size that a set of contiguous voxels must have to be considered significant. This size is obtained from the distribution of cluster sizes under the null hypothesis of none of

the voxels within a cluster are activated. There are several mathematical procedures to compute this threshold, but most of them are based on the aforementioned RFT (Worsley et al., 1998), Monte Carlo simulations (Forman et al., 1995) or permutation tests (Nichols and Holmes, 2002).

Under some circumstances, Bonferroni and other correction methods based on RFT show a very limited sensitivity derived from an increase in the false negative rate. However, cluster-extent based thresholding exhibits a large sensitivity regardless of the context. Moreover, this approach takes into account that contiguous voxels in MR images are not independent, especially when these images have been previously smoothed (Heller et al., 2006; Wager et al., 2007). Nonetheless, this alternative also has some drawbacks. It describes the probability of finding a set of contiguous voxels of a certain size, and not the likelihood at a specific location within the cluster. This can limit the spatial specificity when clusters have a large size (Nichols, 2012; Woo et al., 2014): a significant cluster that covers two different brain regions should not be interpreted as voxels in both regions contain information. This can be a consequence of setting a liberal primary threshold, which can increase the false positives rate and affects the spatial localization and the subsequent interpretation (Woo et al., 2014). This issue can be addressed by using a conservative value for the primary threshold (e.g. $p < 0.001$), at the expense of limiting the sensitivity.

## 6.2 Group-level analysis

In neuroscientific studies, one of the main goals is to identify the regions involved in a brain function. Detecting differences in a certain region can inform about its involvement in a complex process. However, these differences have to be statistically significant across the different subjects evaluated. In a decoding analysis, the classification is usually performed at the subject level, obtaining a weights/accuracy map for each subject. Then, the average map is computed between the different subjects in addition to assess the subsequent statistical significance. It is common to employ a one-sample $t$-test to compare the mean value of the sample with a widely established mean value. This is performed voxel by voxel, so that the mean value of the sample is compared with the average accuracy obtained by pure chance, i.e. 50% of accuracy in a binary classification. Mathematically,

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} \qquad s_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{6.7}$$

where $\bar{x}$ is the sample mean, $\mu$ is the mean expected under the null hypothesis, $s_{\bar{x}}$ is the estimated standard error of the mean, $\sigma$ is the sample standard deviation and $n$ is the sample size. Employing $t$-tests requires making several assumptions that are not always met, which can invalidate its use from a theoretical perspective (Haynes, 2015). Specifically, the use of these tests for classification assumes that the distribution of the accuracies has a Gaussian shape. Besides, $t$-tests usually consider that Gaussian and T distributions are exactly the same, but this is not true for all circumstances. Figure 6.2 shows large similarities between the two distributions, but also evidences that obtaining a good fit depends on the degrees of freedom. Previous studies have evaluated the effect of the sample size in the distribution, concluding that a small number of samples can increase the number of false positives. In fact, Eklund et al. (2016) highlighted the need of addressing the limitations of this method and finding a more conservative approach to be used in neuroimaging.

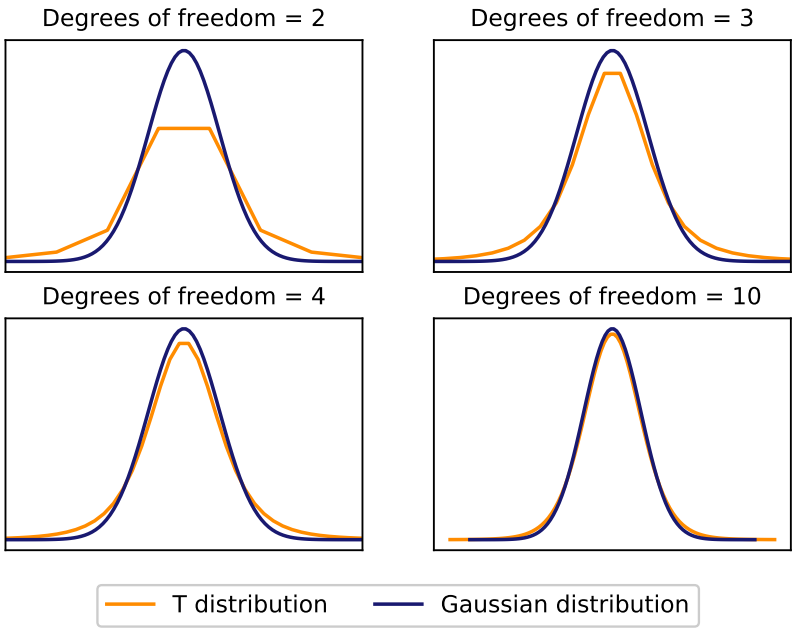


**Figure 6.2:** *T*-tests usually assume that data follow a Gaussian distribution.

Stelzer et al. (2013) proposed a framework that overcomes the limitations of $t$-tests. This method (from now on, Stelzer's) combines result maps from each subject following a Monte Carlo method and computes the $p$-value associated with each cluster size at the group level. Unlike $t$-tests, this approach does not make any assumption about the smoothness of the images. Besides, since Stelzer's relies on permutation tests, it does not

assume a specific shape for the distribution of the data, but computes empirically the distribution. At the subject level, Stelzer et al. (2013) recommend to shuffle the labels associated with the experimental conditions 100 times, and perform the subsequent classification analyses with each one of the 100 subsets of labels. In the original study, they employed this alternative in combination with Searchlight. However, it is possible to use this method with any classification approach that leads to a map of information (for example, a weight map). A this point, 100 information maps (accuracy or weights) would be computed for each subject. The next step is to normalize all these maps to a standard space to register all brains into the same coordinate system. Then, the different maps are combined following a Monte Carlo resampling with replacement procedure (Forman et al., 1995): a permuted map is randomly picked from each subject and averaged across voxels, yielding a permuted group map. This procedure is repeated a high number of times (100000 in the original study) to build the empirical chance distribution for each voxel. As explained in Section 6.1.1, the performance obtained when the classifier is trained with the actual labels is compared with the empirical distribution to evaluate its significance.

From the 100000 permuted maps at the group level, this method also builds an empirical distribution of the cluster sizes. A cluster is defined following a 6 connectivity-scheme, so that a number of contiguous voxels are considered a cluster if they share a face, but not an edge or a vertex (Stelzer et al., 2013). This cluster search is also applied to the group map derived from the actual labels. The $p$-value of a cluster of a size $s$ is given by:

$$p_{cluster} = \sum_{s'>s}^{\infty} H_{cluster}(s') \tag{6.8}$$

where $H_{cluster}$ is the normalized histogram of cluster sizes in the empirical distribution. All the $p$-values associated with each cluster size is then FWE-corrected to correct for multiple comparisons at the cluster level. Figure 6.3 illustrates all steps followed by Stelzer's method.

Another method based on permutation is termed Threshold-free cluster enhancement (TFCE, Smith and Nichols, 2009). This approach eliminates the need of defining a primary threshold and the detrimental effect that arbitrariness can have in the significant results. This method applies a transformation to the images (in most cases, group accuracies), changing the value of voxels of the images and facilitating the discrimination between significant and non significant voxels. In each image, there are groups of contiguous voxels that are candidates to be a cluster. From these two parameters (intensities

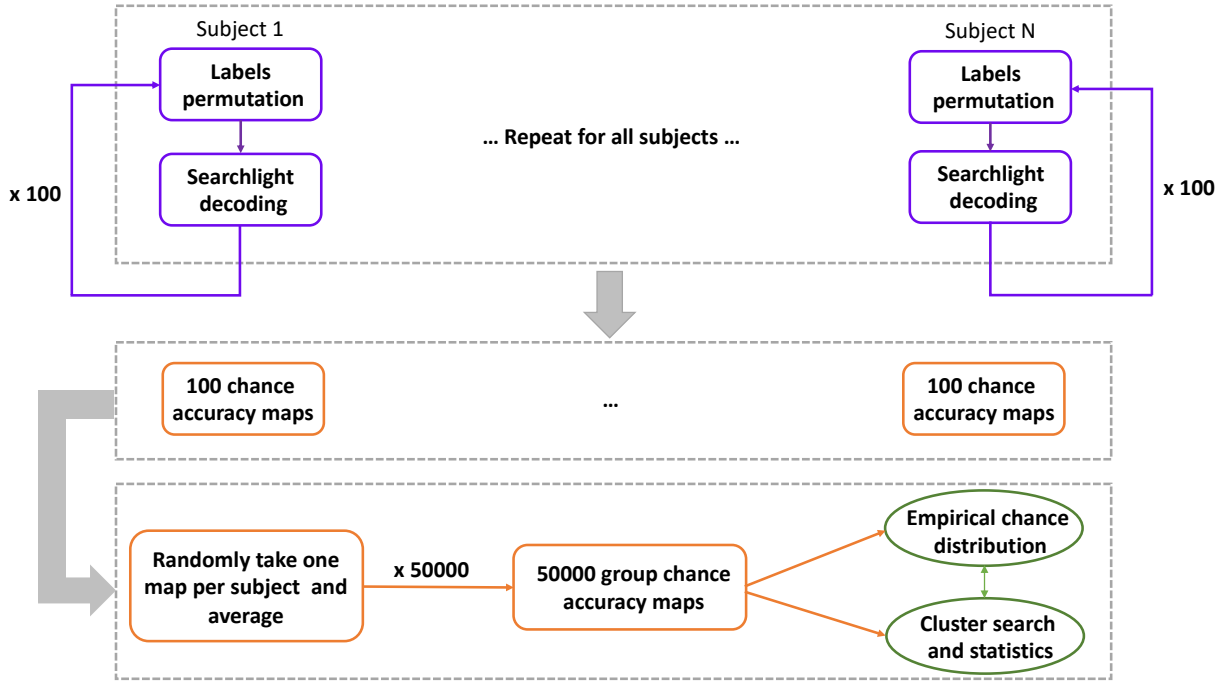


**Figure 6.3:** Schematic representation of Stelzer's method.

of the voxels and extension of the possible cluster), there is a large number of possible combinations. However, there are two extreme situations: the signal in a group of voxels is strong (large intensity) but it is extremely concentrated (small number of voxels). In contrast, there can be weak signal (small intensities) but much more widespread. The aim of TFCE is to level these situations so that it is equally likely to reach significance for both cases. The intensity of the voxel after applying the TFCE transformation is given by:

$$TFCE(p) = \int_{h=h_0}^{h_p} e(h)^E h^H dh \tag{6.9}$$

where $h_0$ is usually set to 0, $e$ is the extent of the cluster that contains voxel $p$, $h$ is the primary threshold, and $E$ and $H$ are usually set to 0.5 and 2, respectively. The TFCE score associated with each voxel is computed as the sum of the product between the extent of the candidate cluster and the different primary thresholds. Once the TFCE score is computed for the actual accuracies image, it is also computed for a large number of permutations to assess the statistical significance of the resulting clusters. The obtained $p$-values are FWE-corrected to control for false positives. Figure 6.4 shows an illustration of this framework.

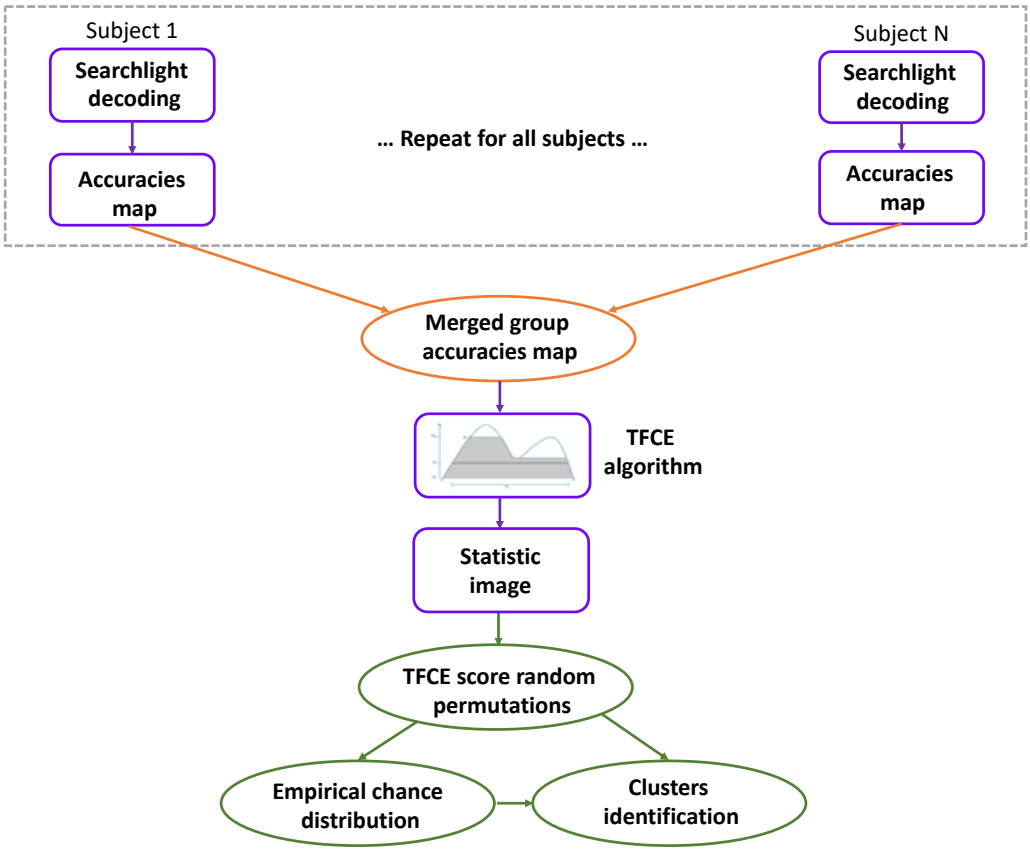


**Figure 6.4:** Schematic representation of the TFCE approach.

# Part II

# Contributions of this thesis

CHAPTER 7

# EFFECT OF THE CLASSIFICATION ALGORITHM IN THE PERFORMANCE OF FMRI ANALYSIS

As previous chapters described, MVPA aims at identifying the information contained in patterns of neural activity. To do so, MVPA performs a classification task to learn the relationship between the features of the images and the experimental conditions. The way this classification is done is crucial to lead to to a model that properly classifies between the different conditions, so that this stage is one of the most important in the decoding framework (see Figure 7.1). This chapter evaluates the effect of different parameters in the performance of the most common fMRI classification method, Searchlight. First, we evaluated the influence of the Searchlight size in the subsequent decoding results. Second, different classification algorithms were employed to identify the one that maximizes performance. Finally, we performed a grid search over the hyperparameters associated with each classification kernel to explore the dependence between these parameters and the decoding results, in terms of accuracy and number of significant voxels. Figure 7.2 shows a detailed schema of the system evaluated in this Chapter.

## 7.1 Introduction

Pattern-information-analyses are increasingly common in the study of distributed representations with fMRI (González-García et al., 2017; Haxby et al., 2001; Haynes and

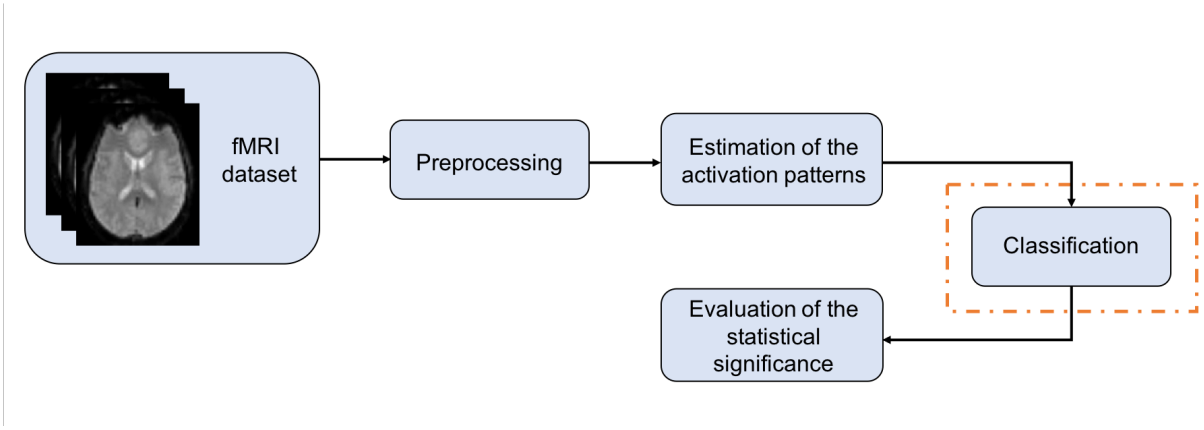


**Figure 7.1:** Illustration of the general framework in fMRI classification.

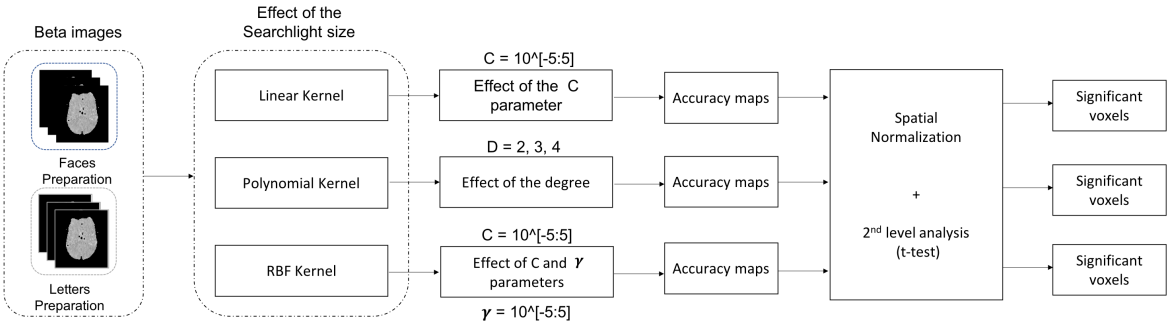


**Figure 7.2:** Schema of the classification framework evaluated in this chapter. Once beta maps were estimated, a Searchlight analysis was performed over these images. Performance was evaluated for a wide range of Searchlight sizes and three different classification kernels. Additionally, the effect of different hyperparameters associated with each classifier was assessed. The resulting Searchlight accuracy maps were then spatially normalized and entered into a second-level analysis to evaluate the statistical differences at the group level.

Rees, 2005; Mur et al., 2009), allowing for the identification of experimental conditions from their corresponding activation patterns with a multivariate classifier. A significant above-chance-level decoding accuracy could indicate that the response patterns contain information about the experimental condition. However, the sensitivity of this multivariate approach highly depends on the classifier used. Classifiers differ in the shapes of the decision boundaries employed to separate the different classes. Linear classifiers use hyperplanes, whereas non-linear classifiers employ nonplanar boundaries (see section 5.3 for a detailed explanation). Performance depends on how well the classifier's assumptions fit the properties of the data distribution and the features of fMRI data (number of voxels considered, number of response patterns to train the classifier, etc).

Previous studies have evaluated the performance of different classification algorithms

for fMRI data. Cox and Savoy (2003) employed SVMs with linear and polynomial kernels for classifying different visual objects in a block-design experiment. Although the polynomial kernel showed an acceptable performance, the linear classifier performed considerably better. LaConte et al. (2005) also reported a superior performance of linear SVMs compared to nonlinear SVMs in block-designs. However, the experimental design has a strong effect on the separability of the activation patterns and therefore, in decoding performance. Misaki et al. (2010) analyzed fMRI response patterns in a rapid event-related experiment and concluded that most classifiers obtained good results. Nevertheless, the linear version of SVM outperformed the rest of the classification algorithms.

Despite the previously mentioned differences in the experimental design, linear classifiers seem to be the optimal choice when trying to decode fMRI signal. However, it is necessary to note that decoding performance is highly influenced by the differences at the neural level of the response patterns. Previous studies that performed a comparison between different classifiers aimed to find differences in the activity patterns associated with different visual stimuli, which usually lead to large differences in the neural activity. There are other psychological studies that try to find subtler differences at the neural level, with the subsequent decrease in decoding performance. Nonetheless, little is known about the relative performance of different classifiers in this kind of scenarios. To address this issue, we evaluate different kernels based on SVM classifier and compare their performance. Variations in the kernel lead to different shapes of the hyperplane and thus different performance (see section 5.3). Specifically, we employ a linear classifier, a Polynomial kernel (with different degrees) and an RBF kernel. Moreover, we study the effects of the different hyperparameters associated with each algorithm in the decoding performance in terms of accuracy and sensitivity (number of significant voxels). As explained in 5.3, the cost parameter in an SVM classifier defines the margin between the hyperplane and the support vectors. Small values allows for large margins, whereas large values lead to small margins. Moreover, the degree parameter of polynomial kernels controls the flexibility of the decision boundary: kernels with larger degrees yield a more flexible decision boundary. Finally, the $\gamma$ parameter of the RBF kernel represents the width of the bell-shaped curve: large values lead to a narrow bell whereas small values lead to a wide bell. Hence, these parameters modify the shape of the hyperplane used to separate the different classes, and can have a large influence in results.

In Chapter 5.2.3, we introduced the concept of Searchlight, which has become one of the most popular methods for fMRI classification. This multivariate approach identifies

locally informative areas with larger power and flexibility than mass-univariate analyses (Kriegeskorte et al., 2006). Searchlight approaches produce maps of accuracy by measuring the information contained in small spherical regions. The resulting performance is associated with the central voxel of the sphere despite being derived from the information present in the whole sphere.

Some appealing aspects of Searchlight include its locally-multivariate nature (no *a priori* region specification is needed) and its ability to minimize the curse of dimensionality associated with whole-brain approaches (the spheres have usually a small size). However, it also has limitations that can affect the interpretability of the results. The number of significant voxels tend to grow as the sphere size increases (Etzel et al., 2013). Setting a large radius leads to a larger number of voxels to be used by the classifier and usually results in a larger number of significant voxels, which can be due to a boost in sensitivity but also to an increase in false-positives rate. Moreover, the larger the sphere, the more similar Searchlight is to a whole-brain classification, with the corresponding problems associated with the curse of dimensionality that this technique entails. It seems well-established to define a small sphere to perform Searchlight analyses (Kriegeskorte et al., 2006), although there is not a standard size to be used. It is not simple to find an optimum value that strikes a balance between sensitivity and a small false-positives rate.

## 7.2 Materials and Methods

This section describes the experimental task that participants performed and the differences in activity we aimed at classifying. Once a representative set of subjects was recruited, fMRI data were collected. A series of preprocessing steps was necessary to correct some distortions produced during image acquisition. This section also provides an explanation of the transformations applied prior to the classification analysis.

### 7.2.1 Participants

Twenty-two students from the University of Granada (M = 23, 7 men) took part in the experiment and received an economic remuneration (20-25 euros, according to performance). All of them were right-handed with normal to corrected-to-normal vision, no history of neurological disorders, and signed a consent form approved by the local Ethics Committee.

### 7.2.2 Image acquisition

fMRI data were acquired using a 3T Siemens Trio scanner at the Mind, Brain and Behavior Research Centre (CIMCYC) in Granada (Spain). Functional images were obtained with a T2*-weighted EPI sequence, with a TR of 2210 ms. Forty descendent slices with a thickness of 2.3 mm (20% gap) were obtained (TE = 23 ms, flip angle = 70°, voxel size of 3 $mm^3$). The event-related experiment was performed in a run consisting of 1240 volumes. After the functional sessions, a structural image of each participant with a high-resolution T1-weighted sequence (TR = 1900 ms; TE = 2.38 ms; flip angle = 9°, voxel size of 1 $mm^3$) was acquired.

### 7.2.3 Design

The task comprised a total of 160 trials. The color of the subsequent fixation cross (blue or green) signaled whether participants had to follow the instruction (80 trials) and thus prepare to implement it with a novel grid of stimuli or, alternatively, whether they had to ignore it (80 trials) and expect one of the eight practiced grids. For these practiced grids, participants had to respond based on the knowledge acquired during the learning session. For the novel grids, participants had to respond according to the conditional structure employed in each instruction, such as "If there are three contiguous blue vowels of the same size, press A, if not, press L".

The associations between type of trial (novel, practiced), category and response options were counterbalanced across participants. The duration of the fixation cross indicating the type of trial, as well as inter-trial intervals, was jittered to allow the deconvolution of instruction- and grid-related signals. The pseudorandom duration of the preparation interval allowed the disambiguation of this stage from the encoding and implementation. Each trial comprised the following events (see Figure 7.3): a 2.5 s instruction, a colored fixation cross (mean 6.25 s, range 4-8.5 s), a 2 s grid and an inter-trial interval displaying a black fixation cross (mean 6.25 s, range 4-8.5 s). On average, a trial lasted 10.750 s. The total fMRI task lasted 45 min approximately. In this chapter, we focused on the preparation interval, the stage after the instruction and previous to the grid. Specifically, we aimed at finding differences at the neural level when the instruction referred to faces (once the grid appeared, the participant had to focus on faces, ignoring letters) *vs* referred to letters (once the grid appeared, the participant had to focus on letters, ignoring faces).

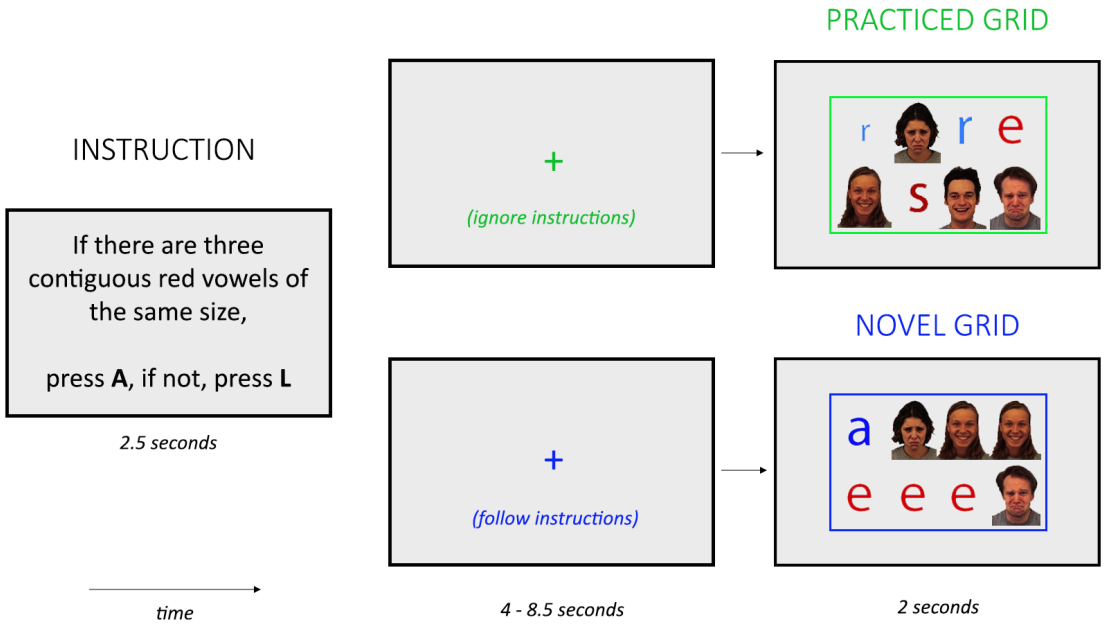


**Figure 7.3:** Behavioral paradigm.

### 7.2.4 Preprocessing

Once fMRI data were acquired, a serie of preprocessing steps was applied to remove artifacts (see Chapter 3). The first 4 volumes were discarded to allow for saturation of the signal, so that the preprocessing methods were applied only to the remaining ones. Then, the effect of motion in the fMRI signal was corrected by employing a spatial transformation based on 6 parameters. Next, a slice-timing procedure was applied to the resulting images to correct the timing variation between the acquisition of different slices. Afterwards, T1 images were coregistered with the resulting functional images. To better preserve the spatial configuration of activations in individual subjects, images were not smoothed nor spatially normalized into a common space (Hendriks et al., 2017; Misaki et al., 2013). We used SPM12 (Wellcome Centre for Human Neuroimaging, 2018) to perform all these preprocessing steps. Figure 7.4 shows an illustration of the complete preprocessing framework.

### 7.2.5 Estimation of the activation patterns

The preprocessed images were not directly used as input of the classifier due to the sluggishness of the BOLD signal. As described in Chapter 4, a GLM is usually employed to compute the beta maps that contain an estimation of the activation patterns. Figure

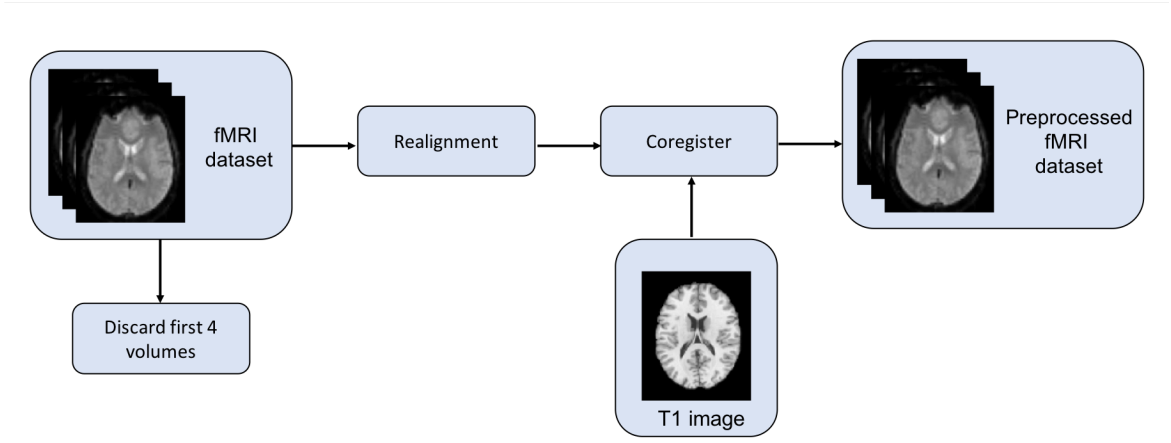


**Figure 7.4:** Schema of the fMRI preprocessing framework.

7.5 shows a diagram of how beta images were computed for each participant. In each GLM, different regressors were defined to group activity derived by different types of events. Specifically, the model included regressors for the three stages evaluated in the processing of instructions: encoding (faces/letters instruction), preparation (jitter of novel faces/letters task; jitter of practised task), and implementation (novel grid of faces/letters task). For example, the regressor "Jitter Faces Novel" computed the activity patterns associated with the preparation stage when participants had to focus on the faces content and the task was novel. We employed a Least-Squares Unitary (LSU) approach to estimate the activity patterns. This means that for each experimental run, activity elicited by events of the same condition was averaged. These regressors were convolved with the standard hemodynamic response function (see section 2.4.4). The preparation stage was modelled as the duration of the jittered interval between the instruction and the implementation, whereas instructions and grids were modelled as an impulse function (Dirac delta), i.e. with zero duration, as explained in Henson (2005).

### 7.2.6 Searchlight

Once beta maps were computed, a Searchlight analysis was performed over these images, for which we employed The Decoding Toolbox (TDT, Hebart et al., 2015). As explained in section 5.2.3, in this approach a spherical region is defined and the classification is performed employing only the voxels contained in the sphere. This sphere sweeps all the positions in the brain, yielding an accuracy map where the value of each voxel corresponds to the accuracy obtained when the voxel was the center of the sphere. We employed different radii of the sphere, ranging from 4 to 30 mm, to evaluate the effect

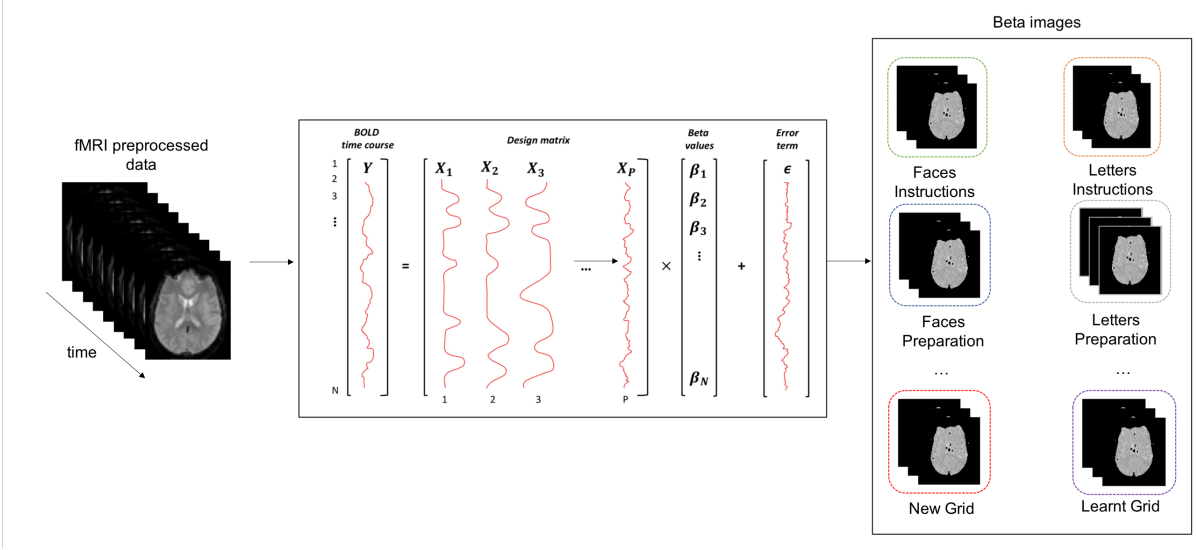


**Figure 7.5:** Diagram of the extraction of the beta images, which are used as the input of the classifier.

that dimensionality of input data has in the results. Table 7.1 summarizes the number of voxels contained in the sphere for each size. We also aimed to study the effect of different classifiers in the results. For this reason, we used different kernels based on SVM as explained in Section 5.3: linear, polynomial (degrees 2, 3 and 4; from now on, Poly2, Poly3 and Poly4, respectively) and RBF. The degree parameter changes the shape of the decision boundary. A polynomial kernel of degree 1 leads to a linear separation, which is equivalent to the linear kernel. For this reason, we employed kernels of larger degrees, which allows more flexible decision boundaries. Besides, the RBF kernel relies on a different function that leads to a different shape of the decision boundary. We aimed at evaluating the optimal shape that allows for separation between the different classes: neural activity when the instruction referred to faces *vs* when referred to letters.

We used a leave-one-run-out cross-validation scheme (Coutanche et al., 2011; Haynes and Rees, 2006; Lee et al., 2011; Reddy et al., 2010; Wolbers et al., 2011) to evaluate the performance of the different classifiers in each Searchlight sphere. In this scheme, the classifier is trained with the beta maps from all but one run, whereas the maps of the remaining one are used to test the performance of the algorithm. Once Searchlight was performed for each participant, the resulting accuracy maps were spatially normalized into a common space to compare the results of different subjects. Then, a $t$-test was used to assess the significance of each voxel of the Searchlight map. Cluster-wise inference with a cluster-defining threshold of $p < 0.001$ was used to identify the significant clusters. The resulting $p$-values were then FWE-corrected to address the multiple comparisons

problem.

**Table 7.1:** Number of voxels contained in a Searchlight sphere for different radii sizes.

| Size (mm) | Voxels | Size (mm) | Voxels | Size (mm) | Voxels |
|---|---|---|---|---|---|
| R = 4 | 9 | R = 13 | 196 | R = 22 | 843 |
| R = 5 | 13 | R = 14 | 218 | R = 23 | 984 |
| R = 6 | 15 | R = 15 | 290 | R = 24 | 1100 |
| R = 7 | 38 | R = 16 | 334 | R = 25 | 1223 |
| R = 8 | 44 | R = 17 | 397 | R = 26 | 1336 |
| R = 9 | 68 | R = 18 | 493 | R = 27 | 1501 |
| R = 10 | 92 | R = 19 | 559 | R = 28 | 1677 |
| R = 11 | 110 | R = 20 | 643 | R = 29 | 1874 |
| R = 12 | 147 | R = 21 | 744 | R = 30 | 2047 |

## 7.3 Results

In this section, we report the results in terms of accuracy and number of significant voxels obtained by the five different classifiers used. Additionally, we evaluated the effect of different Searchlight sizes in the results, as well as the influence of the hyperparameters associated with each classifier. Accuracies are only reported when the classifier detected statistically significant voxels. In this case, the accuracy was computed as the average of the accuracies from the Searchlight map that were significant.

### 7.3.1 Influence of the Searchlight size

We first focus on comparing the results obtained by different Searchlight sizes. The first noticeable feature is that accuracy remained stable across the different values of the Searchlight radius for all kernels, except for the Poly4. However, sensitivity for all kernels seems to be highly affected by the number of voxels considered during the classification. For the linear classifier, no clusters survived the statistical threshold for the smallest radius (4, 5 and 6 mm). These sphere sizes contained 9, 13 and 15 voxels, respectively, so that only a few voxels were employed as input of the classifier. These results indicate that none of the voxels of the resulting accuracy maps surpassed the significance threshold when a Searchlight sphere of those sizes was used. This tendency changes for larger sphere radii: the number of significant voxels rose for each increase in the Searchlight size. However, employing too large spheres reduces sensitivity: increasing the radius did not result in a boost in sensitivity from a certain size, and this size was different from each classification kernel. Figure 7.6 shows the performance obtained by

**Table 7.2:** Comparison of the performance of the different classifiers in terms of accuracy and significant voxels for different Searchlight sizes. The average results for each size were computed for those classifiers that found significant results. Similarly, the average accuracy and sensitivity obtained by each classifier were computed only for those sizes that led to significant results.

| Size (mm) | Linear | | Poly2 | | Poly3 | | Poly4 | | RBF | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | Sensitivity | Accuracy | Sensitivity | Accuracy | Sensitivity | Accuracy | Sensitivity | Accuracy | Sensitivity | Accuracy | Sensitivity |
| R = 4 | - | 0 | - | 0 | - | 0 | - | 0 | - | 0 | - | 0 |
| R = 5 | - | 0 | - | 0 | - | 0 | - | 0 | - | 0 | - | 0 |
| R = 6 | - | 0 | - | 0 | - | 0 | - | 0 | - | 0 | - | 0 |
| R = 7 | - | 0 | 56.65 | 209 | - | 0 | 53.64 | 73 | - | 0 | 55.14 | 141 |
| R = 8 | 58.19 | 85 | 56.56 | 191 | 58.19 | 85 | 53.89 | 66 | - | 0 | 56.7 | 107 |
| R = 9 | 57.99 | 366 | 56.69 | 275 | 56.63 | 61 | 53.76 | 80 | - | 0 | 56.26 | 196 |
| R = 10 | 58.45 | 982 | 56.68 | 310 | 56.49 | 79 | 54.18 | 88 | 57.28 | 224 | 56.62 | 337 |
| R = 11 | 58.62 | 1533 | 56.72 | 416 | 58.62 | 1533 | 54.3 | 89 | 57.29 | 139 | 57.11 | 742 |
| R = 12 | 58.53 | 1955 | 56.98 | 709 | 58.53 | 1955 | - | 0 | 57.13 | 370 | 57.79 | 797 |
| R = 13 | 58.64 | 2235 | 57.15 | 665 | **58.64** | 2235 | - | 0 | 57.27 | 482 | 57.93 | 1404 |
| R = 14 | 58.63 | 2368 | 57.19 | 729 | 58.62 | 2368 | - | 0 | 57.26 | 535 | 57.93 | 1500 |
| R = 15 | 58.6 | 2596 | 57.29 | 631 | 58.52 | 3035 | - | 0 | 57.29 | 661 | 57.93 | 1731 |
| R = 16 | 58.48 | 2948 | 57.15 | 493 | 58.6 | 3146 | - | 0 | **57.31** | 758 | 57.89 | 1836 |
| R = 17 | 58.33 | 2952 | 57 | 573 | 58.34 | 2952 | - | 0 | 57.26 | **903** | 57.73 | 1845 |
| R = 18 | 58.15 | 3035 | 51.38 | 3035 | 58.15 | 3035 | - | 0 | 57.15 | 766 | 56.2 | 2468 |
| R = 19 | 58.03 | 3146 | 57.25 | 3349 | 58.03 | 3146 | - | 0 | 57.11 | 785 | 57.61 | 2607 |
| R = 20 | 57.95 | **3566** | 57.14 | 3249 | 57.95 | **3566** | 50.3 | 3146 | 57.07 | 751 | 56.08 | 2855 |
| R = 21 | 57.78 | 3447 | 57.78 | **3447** | 57.78 | 3447 | 50.4 | **3566** | 57.05 | 667 | 56.16 | 2914 |
| R = 22 | 57.78 | 3349 | 57.77 | 3349 | 57.77 | 3349 | 54.31 | 3447 | 57.1 | 360 | 56.95 | 2771 |
| R = 23 | 57.8 | 3249 | 57.79 | 3249 | 57.79 | 3249 | **54.36** | 3319 | - | 0 | 56.94 | 3267 |
| R = 24 | 57.74 | 2781 | 57.74 | 2781 | 57.74 | 2781 | 50.37 | 3249 | - | 0 | 55.9 | 2898 |
| R = 25 | 57.68 | 2445 | 57.68 | 2445 | - | 0 | 50.37 | 2781 | - | 0 | 55.24 | 2557 |
| R = 26 | 57.81 | 2119 | 57.81 | 2119 | - | 0 | 50.38 | 2445 | - | 0 | 55.33 | 2228 |
| R = 27 | 58.22 | 1790 | 58.22 | 1790 | - | 0 | 50.39 | 2119 | - | 0 | 55.61 | 1900 |
| R = 28 | 58.7 | 1624 | 58.7 | 1624 | - | 0 | 50.49 | 1790 | - | 0 | 55.96 | 1679 |
| R = 29 | 58.71 | 1643 | 58.71 | 1643 | - | 0 | 50.48 | 1624 | - | 0 | 55.97 | 1637 |
| R = 30 | **59** | 1671 | **59** | 1671 | - | 0 | 50.45 | 1643 | - | 0 | 56.15 | 1662 |
| Average | 58.25 | 2256 | 57.21 | 1623 | 58.02 | 2354 | 52 | 1845 | 57.2 | 569.3 | 57.2 | 569.3 |

the five classifiers for all sphere sizes evaluated in terms of significant voxels (top) and accuracy (bottom). It seems clear that results are highly influenced by the Searchlight size, especially in the number of the resulting significant voxels. However, the maximum sensitivity is obtained for similar sphere sizes. Specifically, spheres of 20, 21, 20, 21 and 17 mm led to the maximum sensitivity values for Linear, Poly2, Poly2, Poly4 and RBF kernels, respectively. Table 7.2 summarizes the results obtained by each classifier for the different Searchlight sizes evaluated, providing an average of both accuracy and sensitivity for each size and kernel employed.

Figure 7.7 shows the changes in the significant clusters obtained for four different Searchlight sizes when the linear classifier was used. It is remarkable that clusters' size change as the sphere size does. Clusters marked as significant in the four different scenarios were small for the Searchlight of 9 mm, and tended to grow for the 14 and 24-mm spheres. Similarly to the decrease in sensitivity previously mentioned, clusters were smaller for the largest sphere (29 mm). It is worth mentioning that the Searchlight size does not only influence the size of the resulting clusters but their localization as well. For the intermediate values (14 and 24 mm), Searchlight identified significant clusters that were not found with the smallest and largest values. This confirms that the

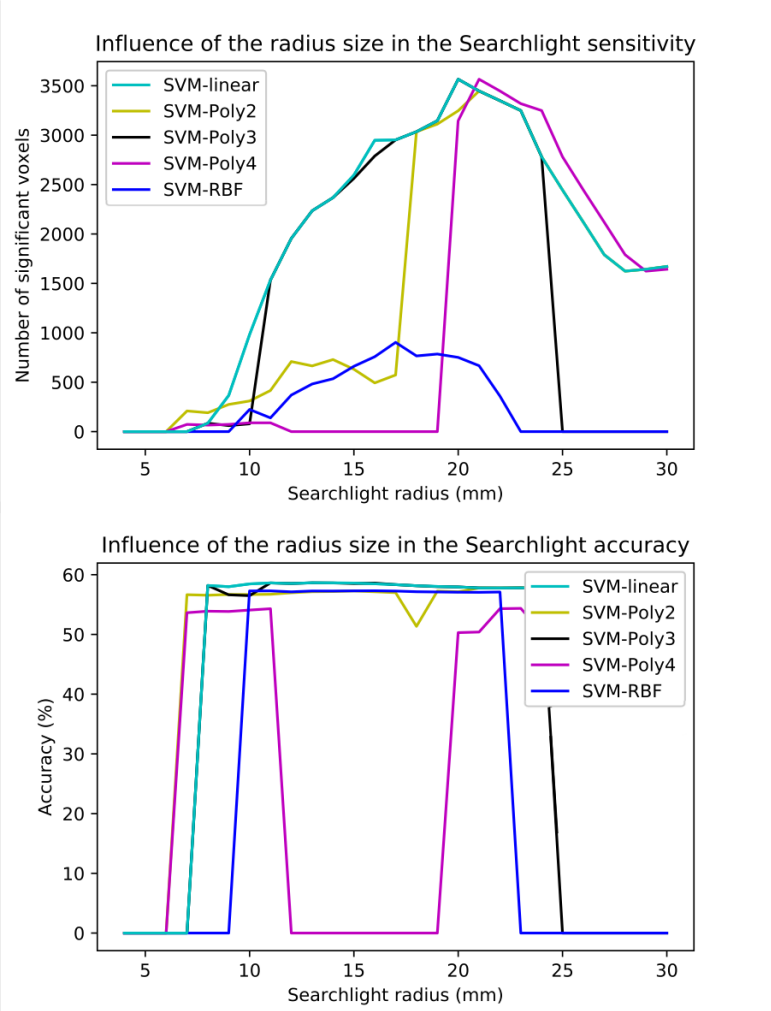


**Figure 7.6:** Influence of the Searchlight size in the performance of different classification kernels.

dimensionality of the input data in a Searchlight analysis has an important effect in the subsequent results.

Regarding the Poly2 classifier, significant clusters were found from 7 to 30-mm Searchlights radii. From Figure 7.6 we can infer that there were three main stages: in the first one, from a radius of 7 mm to a radius of 20, the number of significant voxels increased for small changes in the Searchlight size. From 21 mm to 23 mm, results reached a maximum, and for sizes larger than 23 mm, sensitivity decreased to a final number of 1671 voxels. Results for the Poly3 classifier reveal the previously mentioned tendency of the linear classifier: sensitivity increased until a maximum and then started to decrease. However, this reduction in the number of significant voxels occurs for a smaller sphere size than for the linear classifier (from a radius of 25 mm) and more

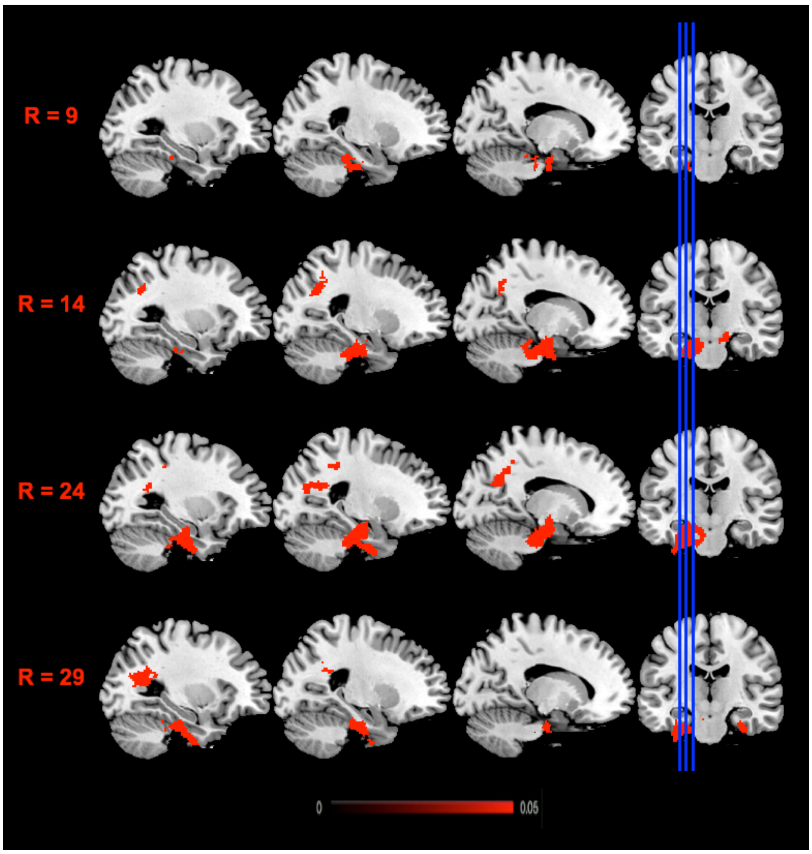


**Figure 7.7:** Significant results obtained for different Searchlight sizes in combination with the linear classifier.

abruptly (sizes larger than 25 mm did not find any significant voxel). Moreover, the Poly4 classifier shows a limited sensitivity for the smallest values (from 8 to 11 mm), whereas no significant voxels were found for sphere sizes ranging from 12 to 19 mm of radius. For larger values, it shows a large sensitivity, with the classical increase until certain values and the decrease once the maximum has been reached. Results obtained for the last classifier (RBF) show a limited sensitivity for intermediate values, whereas Searchlight with lowest sphere sizes or largest did not mark any voxel as significant.

### 7.3.2 Influence of the kernel of the classifier

After describing the general behavior of the classifiers, we can take another look at the differences between them. In terms of sensitivity (number of significant voxels), the linear classifier performed consistently better than the other approaches for all sizes (see top of Figure 7.6). In fact, this method obtained significant clusters for all Searchlight sizes except for the four smallest. Regarding the polynomial kernels, the Poly2 yielded

the same results as the linear classifier for large sphere sizes (from 21 mm to 30 mm). However, for smaller sizes, sensitivity drastically decreased. Similarly, the algorithms based on polynomial of third and fourth order show a reduced sensitivity for a large range of values: small and large spheres for the Poly3 and intermediate values for the Poly4 classifier. Using as a reference the linear classifier, the three polynomial kernels followed a similar tendency: an increase in performance, then a stability and a final decrease. The main differences between them is the range of Searchlight sizes from they started to increase and from they reduced their performance. The Poly2 kernel shows a slow increase, but exactly the same decreasing stage as the linear one. Moreover, the Poly3 classifier has a similar tendency than the linear for small sizes, but the performance of the Poly 3 decayed before the one of the linear kernel. Regarding the Poly4, it shows a larger sensitivity than the linear kernel for large sphere sizes, but its sensitivity only increased when employing a large Searchlight size. The sensitivity of the RBF kernel was drastically limited compared to other classifiers, invalidating its use for small, intermediate or large Searchlight sizes.

With reference to accuracy, all classifiers except the RBF one obtained similar results (see bottom of Figure 7.6). The abrupt drop in accuracies from 60% to %0 that this figure shows is due to only accuracies of classifiers that found significant voxels are reported. Hence, an accuracy of 0% means that no significant voxels were found when a specific Searchlight radius was used. There are not large differences in accuracies for those Searchlight sizes in which all classifiers detect significant voxels, as Table 7.2 shows.

### 7.3.3 Influence of the cost parameter in linear SVM

We will now describe how performance of the linear classifier evolved when varying the cost parameter (C) in the range $[10^{-5} - 10^{5}]$. Results exhibit large similarities in the results obtained for different values of C. Regarding sensitivity, the classification performed with the smallest C value ($10^{-5}$) led to no significant clusters, whereas the second one yielded 3511 significant voxels (see Table 7.3). From the third smallest value to the maximum, we did not find any difference in the significant results associated with them (see top of Figure 7.8). With reference to accuracy, the maximum value (57.95 %) was obtained for six of the eleven values evaluated. These results confirm the robustness of the linear SVM classifier for most values of the cost parameter (see bottom of Figure 7.8). Most importantly, this robustness is also obtained in the localization of informative regions. For most values evaluated, the linear classifier identified the same informative regions. As Figure 7.9 shows, results are similar except for the smallest value (C=0.0001).

**Table 7.3:** Influence of the cost parameter in the performance of the linear SVM classifier.

| Linear classifier | | |
|---|---|---|
| R = 20mm | Accuracy | Sensitivity |
| C = 0.00001 | - | 0 |
| C = 0.0001 | 56.88 | 209 |
| C = 0.001 | 53.02 | 3511 |
| C = 0.01 | 53.02 | 3566 |
| C = 0.1 | 50.12 | 3566 |
| C = 1 | 57.95 | 3566 |
| C = 10 | 57.95 | 3566 |
| C = 100 | 57.95 | 3566 |
| C = 1000 | 57.95 | 3566 |
| C = 10000 | 57.95 | 3566 |
| C = 100000 | 57.95 | 3566 |

Although results obtained for a cost parameter of C=0.01 and the largest values are not exactly the same, differences are minimal, as displayed in Figure 7.9.

### 7.3.4 Influence of the cost and gamma parameters in the RBF classifier.

The tuning of the cost and gamma parameters of the RBF classifier was done for the radius size that led to the maximum sensitivity: R=17 mm. The values of both cost and gamma parameters ranged from $10^{-5}$ to $10^{5}$. Table 7.4 and Figure 7.10 show the number of significant voxels and the average accuracy for each pair of cost-gamma parameters. The maximum accuracy and sensitivity values (66.07% and 1028 voxels, respectively) were obtained for $C$ = 1000 and $\gamma$ = 0.001. We can observe an increase in accuracy compared to the one obtained when the grid search was not performed. Specifically, for R=17 mm the RBF classifier yielded an accuracy of 57.26%. We also found a boost in sensitivity, from 903 to 1028 voxels, but considerably smaller compared to the increase in accuracy. Despite the improvement of the results that this grid search entails, results are worse in terms of accuracy and significant voxels compared to the linear classifier.

## 7.4 Discussion

In this study, we evaluated the influence of the Searchlight size in the classification results for different kernels based on SVM classifier (linear, polynomial and RBF). Moreover, we optimized the hyperparameters associated with each kernel to assess

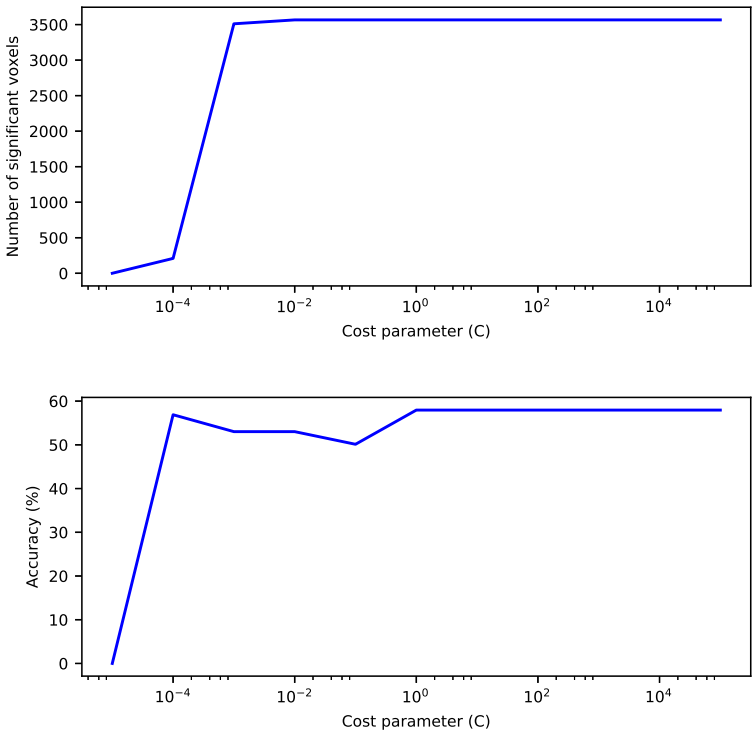


**Figure 7.8:** Influence of the cost parameter in the performance of the linear SVM classifier.

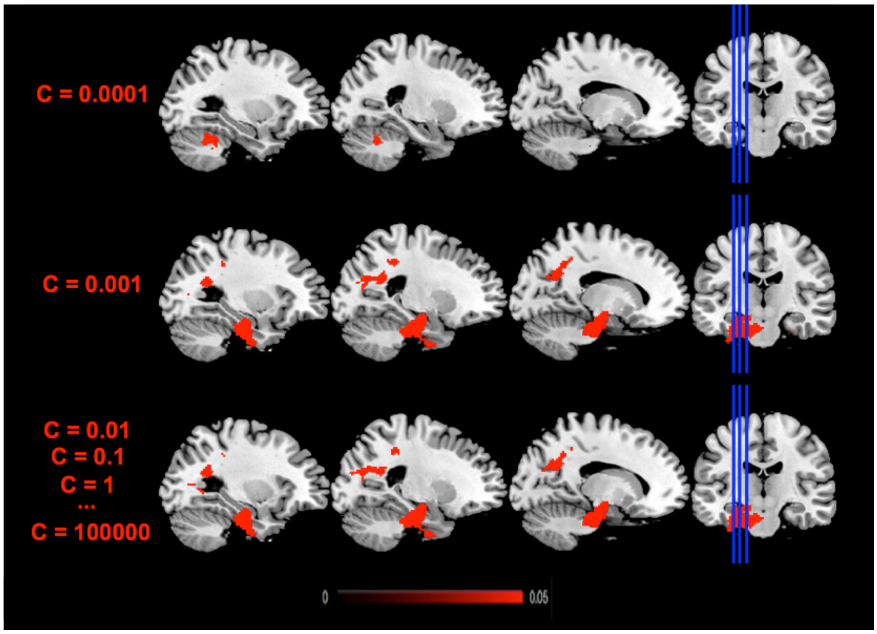


**Figure 7.9:** Significant results obtained by the linear SVM classifier for diffferent values of the cost parameter.

**Table 7.4:** Influence of the cost and gamma parameters in the performance of the RBF classifier. We only report the average accuracy for the significant voxels.

| RBF classifier | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Parameters | | $\gamma$ = 0.00001 | $\gamma$ = 0.0001 | $\gamma$ = 0.001 | $\gamma$=0.01 | $\gamma$=0.1 | $\gamma$=1 | $\gamma$=10 | $\gamma$=100 | $\gamma$=1000 | $\gamma$=10000 | $\gamma$=100000 |
| C= 0.00001 | Accuracy | - | - | 66.07 | 65.93 | - | - | - | - | - | - | - |
| | Sensitivity | 0 | 0 | 339 | 506 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C= 0.0001 | Accuracy | - | - | 66.07 | 61 | 66.12 | - | - | - | - | - | - |
| | Sensitivity | 0 | 0 | 339 | 653 | 480 | 0 | 0 | 0 | 0 | 0 | 0 |
| C= 0.001 | Accuracy | - | - | 66.07 | 60.76 | 65.88 | - | - | - | - | - | - |
| | Sensitivity | 0 | 0 | 339 | 589 | 490 | 0 | 0 | 0 | 0 | 0 | 0 |
| C= 0.01 | Accuracy | - | - | 66.07 | 61.04 | 54.89 | - | - | - | - | - | - |
| | Sensitivity | 0 | 0 | 339 | 616 | 490 | 0 | 0 | 0 | 0 | 0 | 0 |
| C= 0.01 | Accuracy | - | - | 66.07 | 61.2 | 66 | - | - | - | - | - | - |
| | Sensitivity | 0 | 0 | 339 | 597 | 494 | 0 | 0 | 0 | 0 | 0 | 0 |
| C= 1 | Accuracy | - | - | 66.07 | 61.2 | 66 | - | - | - | - | - | - |
| | Sensitivity | 0 | 0 | 339 | 597 | 494 | 0 | 0 | 0 | 0 | 0 | 0 |
| C= 10 | Accuracy | - | - | 66.07 | - | - | - | - | - | - | - | - |
| | Sensitivity | 0 | 0 | 626 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C= 100 | Accuracy | - | - | 66.07 | - | 55.76 | - | - | - | - | - | - |
| | Sensitivity | 0 | 0 | 339 | 0 | 494 | 0 | 0 | 0 | 0 | 0 | 0 |
| C= 1000 | Accuracy | - | - | 66.07 | - | - | - | - | - | - | - | - |
| | Sensitivity | 0 | 0 | 1028 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C= 10000 | Accuracy | - | - | 66.07 | - | - | - | - | - | - | - | - |
| | Sensitivity | 0 | 0 | 626 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C= 100000 | Accuracy | - | - | 66.07 | - | - | - | - | - | - | - | - |
| | Sensitivity | 0 | 0 | 626 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

their influence in the accuracy and sensitivity (number of significant voxels) of the classification algorithm. There is a direct relationship between the Searchlight size and the resulting number of significant voxels. Results show, in overall, three main stages. First, we found a boost in sensitivity while increasing the radius of the Searchlight, which could be explained by the nature of the Searchlight itself. It is more likely that a group of informative voxels is contained within the sphere if this sphere has a large size. The accuracy of a classification performed within a certain sphere is associated with the central voxel. This results in a large number of spheres that overlap the informative regions, increasing the number of voxels marked as informative. This finding is consistent with Etzel et al. (2013). They showed that the number of voxels considered informative by the Searchlight tends to grow as its radius increases, even when the size of the informative region stays fixed. This phenomenon has been termed needle-in-the-haystack-effect, and Viswanathan et al. (2012) demonstrated an extreme example.

In the second stage, the number of significant voxels remain stable despite increasing the Searchlight size. We suggest that this can be due to the curse of dimensionality: when a large number of features is used as input of the classifier, it is difficult to extract the informative ones that lead to a good performance. In fact, the larger the radius, the more similar to a whole-brain analysis, which explains the decrease in the number of significant voxels for the largest sizes (third stage). Our results are coherent with those obtained by Oosterhof et al. (2011). They found a saturation effect for larger radii

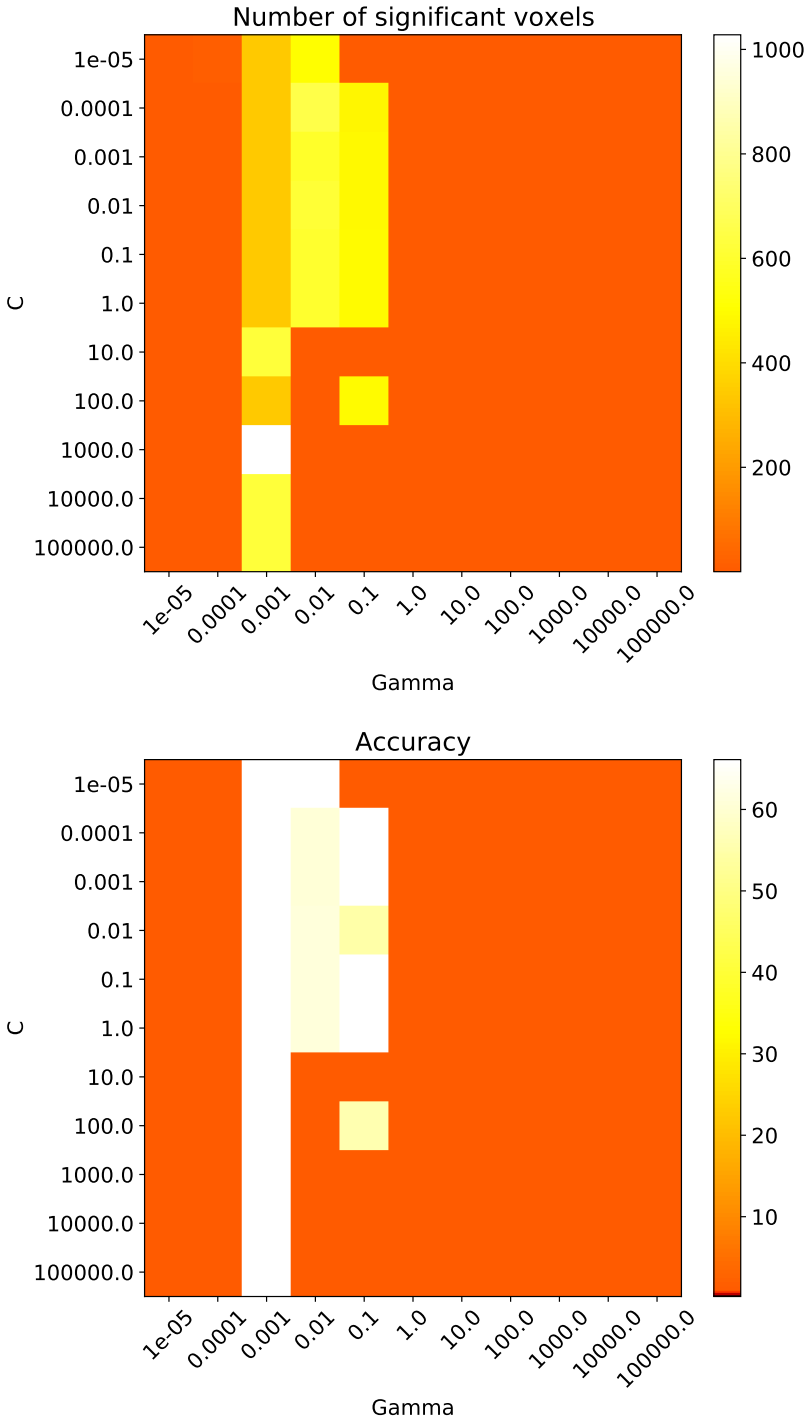


**Figure 7.10:** Influence of the gamma and C parameters in the performance of the RBF kernel.

because large Searchlights tend to select more uninformative voxels. Furthermore, Mur et al. (2009) had previously suggested that there is typically a decrease in performance as the number of voxels rises.

Another crucial aspect is related to the classifier employed in each Searchlight sphere. Overall, the linear SVM performed better than the nonlinear classifiers. These differences were especially large in terms of significant voxels rather than in accuracy: a classifier that obtains a large sensitivity does not necessarily lead to a high accuracy, and viceversa (Chapter 9 will explore this dichotomy). An algorithm can find a large number of significant voxels based on classification accuracies just above chance (Bhandari et al., 2018), yielding typical values of 53%-55% for binary classifications (Bode and Haynes, 2009; Nelissen et al., 2013; Woolgar et al., 2011). We found a superior performance of the linear classifier for all Searchlight sizes. These findings are in line with results obtained by Misaki et al. (2010), where linear classifiers performed significantly better than non-linears for different regions (early visual cortex and human inferior temporal cortex) and different areas sizes within these regions. These results suggest that non-linear kernels may not be ideal for classifying fMRI data because they are computationally expensive and difficult to interpret (Kamitami and Tong, 2005). Moreover, fMRI experiments do not usually lead to large datasets that take advance of the more complex boundaries that these algorithms employ. This is probably the main reason for its worse performance in previous research (Cox and Savoy, 2003; LaConte et al., 2005).

Nonetheless, our results do not only reveal a decrease in sensitivity for non-linear classifiers, but their inability to detect any significant voxel for some Searchlight sizes. An important cause for these findings is the classification task evaluated in this chapter. It seems obvious that decoding performance is intimately linked to the difficulty of the classification. Previous studies that compared linear and non-linear kernels focused on distinguishing between different visual stimuli, which generates large differences at the neural level. However, we aimed at discriminating between much subtler changes in activity. Specifically, our classifier was able to detect differences in the activity patterns when a person prepared to respond to complex grids from information provided by letters *vs* faces. Thus, these differences are not as large as those generated by the visual representation of different objects.

It is also noticeable the weak influence that the cost parameter has in the decoding performance of the linear classifier. Classification analyses usually require optimizing the hyperparameters associated with the kernel used and select those that lead to best results. This procedure is computationally expensive *per se*, as well as Searchlight, so it

is common in the literature not to perform this optimization and select a conservative value for this parameter (Gilbert and Fung, 2018; González-García et al., 2017; Ontivero-Ortega et al., 2017; Rundle et al., 2018). Our results reveal that both accuracy and sensitivity remain stable for most of the values evaluated. In fact, variations in the linear classifier are only obtained for the extreme values of the cost parameter. Regarding the RBF kernel, only a narrow range of $C$ and $\gamma$ combinations lead to no significant clusters. However, performing a grid search did not improve the limited sensitivity of this kernel. Thus, it seems reasonable to use a linear classifier when trying to decode information from fMRI data, especially when differences in neural activity are small.

Results obtained in this chapter provide a useful information about a crucial step in fMRI decoding: classification. Specifically, we have shown that the linear classifier performs consistently well for different dimensionality of the input data. Most fMRI studies aim at detecting the relevant brain regions during a specific mental operation, for which Searchlight is usually employed. However, this chapter highlights the dependence between the number of voxels that Searchlight marks as significant and its size. Employing large spheres can distort the results, which can invalidate any inference of the role of a region in a cognitive function. We have not discussed any information about the brain regions identified during these analyses because it was outside the scope of this chapter. In fact, conclusions derived from this experiment pave the way for the subsequent analysis described in next chapters.

CHAPTER 8

# ESTIMATION OF NEURAL ACTIVATION PATTERNS

In Chapter 7, we showed the large influence that the election of the classifier has in decoding fMRI data. However, classification is only one of the stages in fMRI analyses and there are other steps in the classification framework that have a large effect on the results. One of them is the estimation of the activation patterns. Previous research has compared how different methods compute the activity elicited by each trial of the experiment separately. However, paradigms frequently aim at isolating the activity of different events within the same trial, which suffers from significantly high signal overlap. The effect of alternative methods in this type of experimental design is not clear. In this chapter, we aimed at evaluating the performance of three different approaches in a context where a sustained activity had to be isolated from a zero-duration event, in addition to a classic block design and an event-related design. Moreover, we examined the suitability of the parametric and the two non-parametric approaches described in Chapter 6 to evaluate the significance of the results obtained with the different estimation methods. Figure 8.1 illustrates the global framework of fMRI classification, highlighting the two stages this chapter focuses on. Moreover, Figure 8.2 shows a detailed schema of the system evaluated in this chapter.

## 8.1 Introduction

The basis of multivariate analysis is to explore the information contained in patterns of neural activity as a classification task: the algorithm extracts the features from the input

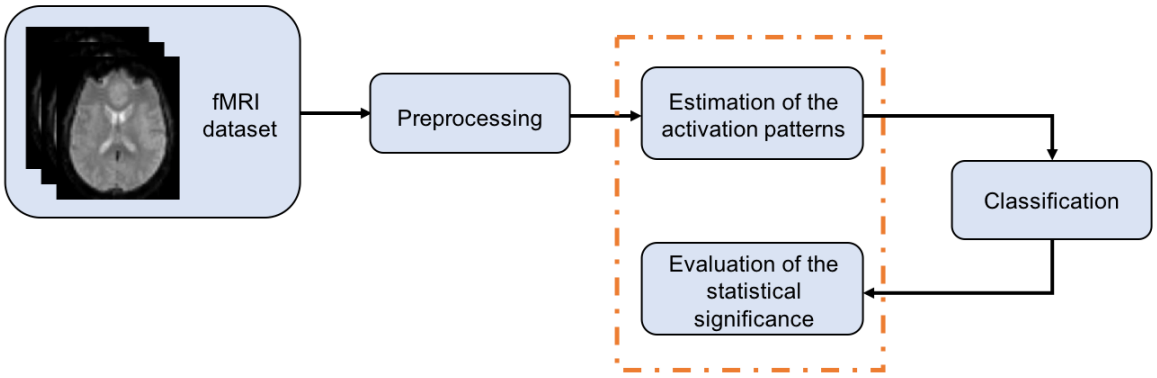


**Figure 8.1:** Illustration of the general framework in fMRI classification. This chapter focuses on two main goals: first, optimizing the way activation patterns are estimated, since these images are then fed to the classifier. Second, obtaining the most accurate method to assess the statistical significance of the results.

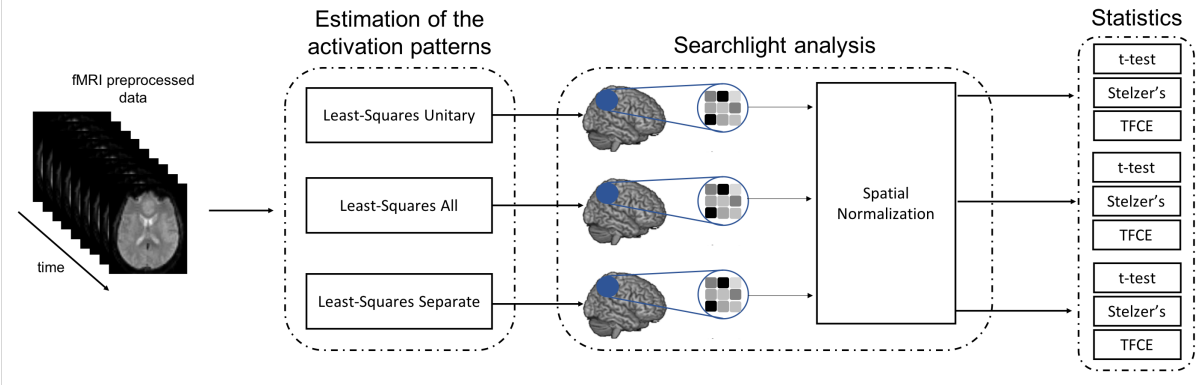


**Figure 8.2:** Overview of the system evaluated in this chapter. The images used as input of the classifier are computed with three pattern estimation methods. Then, a Searchlight analysis is performed over these beta maps. The resulting accuracy maps are spatial normalized, and finally, the significance of each voxel is assessed with three statistical approaches: a $t$-test and two non-parametric methods: Stelzer's and TFCE.

images and learns the relationship between the features and the experimental conditions. Then, based on this learning, the classifier predicts the experimental conditions to which new images belong using only their activation patterns. Since the classifier uses this information as input, the result of the classification process is affected by the quality of the patterns and the way they are estimated. The sluggishness of BOLD signal adds difficulty to this classification endeavor: during an experimental condition, the BOLD response increases about 2 seconds after neural activity, peaking at about 6-8 seconds later and returning to baseline approximately at 20 seconds (Logothetis, 2004). In block designs, isolating the relevant signal is relatively straightforward. This is similar to slow-event related designs where the ISI is longer than the duration of the BOLD. However, when the ISI is short (such as in rapid event-related designs), there is a large overlap between trials, and it is harder to estimate each individual contribution to the hemodynamic response.

Most fMRI analyses use linear convolution models like the GLM to extract estimates of responses to different event types (Friston et al., 1998), where the model estimation is carried out voxelwise and the BOLD time series is the dependent variable. As explained in Chapter 3, the parameters of the GLM are computed by minimizing the squared errors across scans between the estimated timeseries of each voxel and the actual timeseries. It is worth remembering the sluggishness of the BOLD response: the signal measured at a time point $t$ is not only due to the experimental condition at time $t$ but to a linear combination of the previous experimental conditions. For this reason, it is computed the contribution of each condition to the model in each temporal moment that minimizes the error between the actual signal and the estimated one. This procedure is performed voxelwise, so that a different estimation is made for each voxel in the brain. The resulting vector of activation estimates (the beta image) represents the contribution of each experimental condition to the fMRI signal along time. Based on GLM, there are different alternatives in the way activation patterns are estimated, and they have a large influence in the results. Section 8.3 explains the three estimation methods evaluated in this work.

Another crucial step in fMRI classification is the evaluation of the statistical significance of decoding accuracies. The large number of voxels in fMRI analyses results in massive statistical tests that need to be corrected for multiple comparisons. Cluster-level inference has become the most popular method due to its larger sensitivity compared to voxel-level inference. As the name suggests, this method does not estimate the false-positive probability of isolated voxels, but evaluates if a cluster is significant as a whole.

To do so, this approach assumes that there is a correlation between adjacent voxels, so that the signal in each voxel is not completely independent of its neighbors. As explained in Chapter 6, cluster-level inference consists of two stages: first, a primary threshold at the voxel level is employed to obtain those that surpass a certain statistical $p$-value. The election of the threshold is arbitrary in some way (Friston et al., 1994), and what is more important, results can highly vary depending on the threshold considered. This threshold is computed based on theoretical methods such as RFT (Worsley et al., 1998), Monte Carlo simulations (Forman et al., 1995) or non-parametric approaches (Nichols and Holmes, 2002).

Previous studies have shown that RFT corrections tend to be too conservative as well as prone to false positives (Eklund et al., 2016). However, the key problem for applying RFT in classification-based analysis is that the distribution of the accuracies is unknown, and they are assumed to be normally distributed. As an alternative, statistical significance can be evaluated by non-parametric approaches based on permutation testing, which does not require any assumption except exchangeability (see Section 6.1.1). The basic principles of permutation testing are simple (Brammer et al., 1997; Bullmore et al., 1999; Chen et al., 2011; Nichols and Holmes, 2002; Pereira and Botvinick, 2011; Winkler et al., 2014), and previous research has theoretically evaluated their use in classification analyses (Golland and Fischl, 2003). Based on this concept, Stelzer et al. (2013) proposed a framework to derive a cluster size $p$-value at the group level, employing a Monte Carlo method to combine individual results. To compute the cluster-defining primary threshold, this method builds an empirical distribution for each voxel separately, minimizing the consequences related to spatial inhomogeneities that a global accuracy threshold would have. An alternative solution was proposed by Smith and Nichols (2009), the so-called TFCE. This algorithm transforms the value of each voxel to a weighted score of the surrounding voxels, summarizing the cluster-wise evidence at each voxel. However, its most interesting contribution is that it does not require setting a cluster-defining primary threshold, eliminating the arbitrariness on this election and the subsequent impact on the results. Section 8.4 includes a detailed explanation of the three statistical methods employed in this chapter.

## 8.2 Materials

This section describes the different datasets employed during the analyses carried out in this chapter. Each dataset is based on an experiment that studied different Cognitive

Neuroscience questions. We used these datasets because their experimental designs led to a different amount of overlap between the stimuli. Thus, we studied the optimal estimation of the activation patterns in different contexts (high, medium and low overlap of signal).

### 8.2.1 Ultimatum Game dataset (UG)

#### Participants

Twenty-four students from the University of Granada (M = 21.08, SD = 2.92, 12 men) took part in the experiment and received an economic remuneration (20-25 euros, according to performance). All of them were right-handed with normal to corrected-to-normal vision, no history of neurological disorders, and signed a consent form approved by the local Ethics Committee.

#### Image acquisition

fMRI data were acquired using a 3T Siemens Trio scanner at the Mind, Brain and Behavior Research Centre (CIMCYC) in Granada (Spain). Functional images were obtained with a T2*-weighted echo planar imaging (EPI) sequence, with a TR of 2000 ms. Thirty-two descendent slices with a thickness of 3.5 mm (20% gap) were obtained (TE = 30 ms, flip angle = 80°, voxel size of 3.5 $mm^3$). The sequence was divided in 8 runs, consisting of 166 volumes each. After the functional sessions, a structural image of each participant with a high-resolution T1-weighted sequence (TR = 1900 ms; TE = 2.52 ms; flip angle = 9°, voxel size of 1 $mm^3$) was acquired.

We used SPM12 (Wellcome Centre for Human Neuroimaging, 2018) to preprocess and analyze the neuroimaging data. The procedure employed is the same as described in Chapter 7.3. The first 3 volumes were discarded to allow for saturation of the signal. Images were realigned and unwarped to correct for head motion, followed by slice-timing correction. Afterwards, T1 images were coregistered with the realigned functional images. To better preserve the spatial configuration of activations in individual subjects, images were not smoothed nor spatially normalized into a common space. Figure 8.3 includes a visual representation of the preprocessing framework.

#### Design

Participants played the role of the responder in a modified Ultimatum Game (for theoretical background, which is not the focus of the present chapter, see Gaertig et al.,

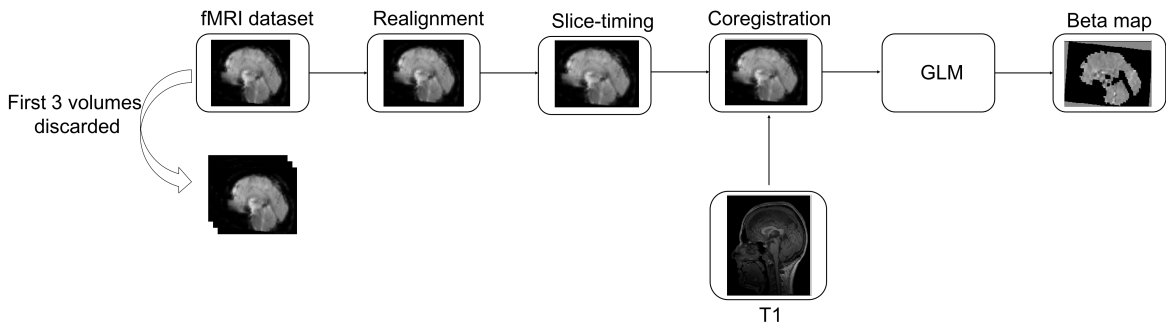


**Figure 8.3:** Schema of the preprocessing framework.

2012), deciding whether to accept or reject monetary offers made by different partners. If they accepted the offer, both parts earned their respective splits, whereas if they rejected it, neither of them earned money from that exchange. Offers consisted in splits of 10 Euros, which could be fair (5/5, 4/6) or unfair (3/7, 2/8, 1/9). The number on the left was always the amount of money given to the participant, and the one on the right was the one proposed by the partners for themselves. Figure 8.4 includes a schema of the experimental paradigm.

The task contained two events in each trial, first a word (positive, negative or neutral in valence) and second two numbers, to which participants had to respond (Figure 8.4 includes a schema of the experimental paradigm). These two numbers corresponded to the offer that participants received, from which they decided to collaborate or not based on the fairness/unfairness of the offer. They performed a total of 192 trials, arranged in 8 runs (24 trials per run), in a counterbalanced order across participants. Each trial started with the word for 1000 ms, followed by a jittered interval lasting 5500 ms on average (4-7 s, +/0.25°). Then, the numbers appeared for 500 ms followed by a second jittered interval (5500 ms on average; 4-7 s, +/0.25°). The first event was modelled as the duration of the word and the variable jittered interval, yielding a global duration ranging from 5 to 8 seconds. The second event was modelled as an impulse function (Dirac delta), i.e. with zero duration, as explained in Henson (2005). Participants read an adjective with a certain valence, and then they used this information to prepare to respond to the offer (second event). Thus, there was a preparatory process that led to a sustained activity along time. However, the second event captured a completely different process. Once participants made a decision (cooperate or not depending on the fairness of the offer), the process ended. A large body of literature shows that preparatory processes extend in time (e.g. Bode and Haynes, 2009; González-García et al., 2017, 2016; Sakai, 2008) whereas responding to a brief target does not (see the temporal duration of the potentials in Moser

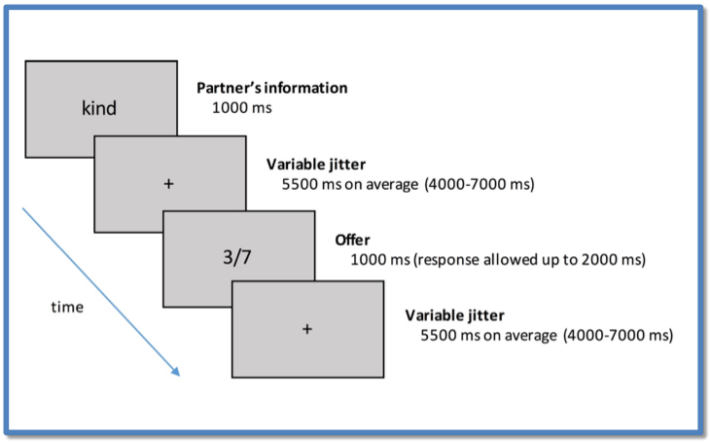


**Figure 8.4:** Schematic display of the paradigm employed in UG dataset.

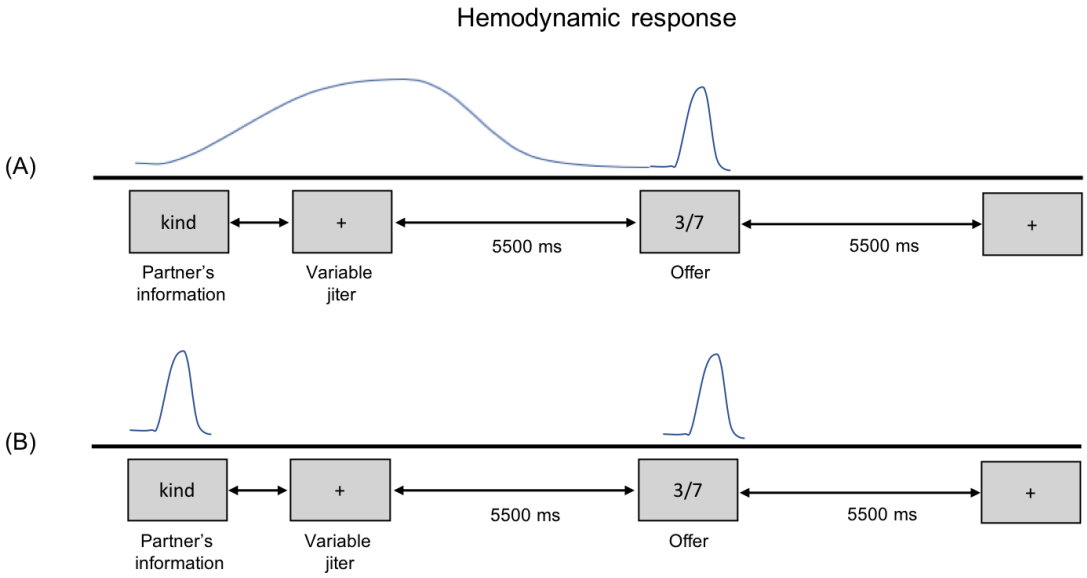


**Figure 8.5:** The regressor associated with the first event (adjective) was modelled in two different ways: First, as the duration of the word and the variable jittered interval; Second, as an impulse function (Dirac delta).

et al., 2014). This has been also measured by other neuroimaging methods, such as the CNV ERP potential (Di Russo et al., 2017). For this theoretical cognitive reason, this second event was modelled with zero duration. Besides, we ran an additional analysis to evaluate whether modelling the first event as an impulse function (zero duration) influenced the results by reducing the collinearity of the regressors between the first and second events in a trial. Moreover, the beginning of runs and the inter-trial jittered intervals served as the implicit baseline. The whole fMRI session lasted 41 minutes approximately

We focused on two different classification analyses, one for each part of the trial. We first aimed at discriminating the positive *vs*. negative valence of the words (e.g. Lindquist et al., 2015; from now on, *valence* classification) that were equated in number of letters, frequency of use and arousal (Gaertig et al., 2012). Then, we aimed to discriminate between fair and unfair offers (*fairness* classification). The total number of images available for the classification procedure varied according to the method used to estimate the patterns. We employed three different methods: Least-Squares Unitary (LSU), Least-Squares All (LSA) and Least-Squares Separate (LSS). The different number of beta weights that each alternative computes is due to the basis they rely on, which are explained in detail in section 8.3. Briefly, LSU averages the activity elicited by all trials of the same type for each experimental run. Thus, a beta map is computed for each condition and run. Moreover, LSA and LSS compute the activity elicited by each trial separately, so that the number of beta maps is equal to the number of trials or events of interest in each experimental run. In the *valence* classification, LSU yielded 8 images per condition, one for each run. LSA and LSS obtained the same number as positive/negative trials in the experiment (64 of each category, per participant). Regarding the *fairness* classification, LSU yielded again 8 images per condition. Moreover, LSA and LSS obtained 96 images for each condition and participant (see Table 8.1).

### 8.2.2 Faces and Objects Representations dataset (FaOR)

This dataset comprises a block-design fMRI experiment. In this case, the overlap of signal is much lower than in the UG dataset, where activity associated with different events within each trial is isolated. We used data of six participants from the study published by Haxby et al. (2001), which has served as example fMRI dataset several times (e.g. Hanson et al., 2004; O'Toole et al., 2007). Neural responses were measured with gradient echoplanar imaging on a GE 3T scanner (General Electric, Milwaukee, WI) [repetition time (TR) = 2500 ms, 40 3.5-mm-thick sagittal images, field of view (FOV) = 24 cm, echo time (TE) = 30 ms, flip angle = 90°] while they performed a one-back repetition detection task. High-resolution T1-weighted spoiled gradient recall (SPGR) images were obtained for each subject to provide detailed anatomy (124 1.2-mm-thick sagittal images, FOV = 24 cm).

The dataset consists of 12 runs where the participants viewed grayscale images of eight object categories: faces, houses, cats, bottles, scissors, shoes, chairs and scrambled images. Each run began and ended with 12-s rest and contained eight blocks of 24-s duration, one for each category, separated by 12-s of rest. Stimuli were presented for

**Table 8.1:** Average number of beta maps obtained by each pattern estimation method and dataset used, for each classification problem evaluated

| | UG | | FaOR | FI |
|---|---|---|---|---|
| Method | Valence | Fairness | Faces vs Houses | Familiarity |
| LSU | 8 | 8 | 12 | 11 |
| LSA | 64 | 96 | 12 | 176 |
| LSS | 64 | 96 | 12 | 176 |

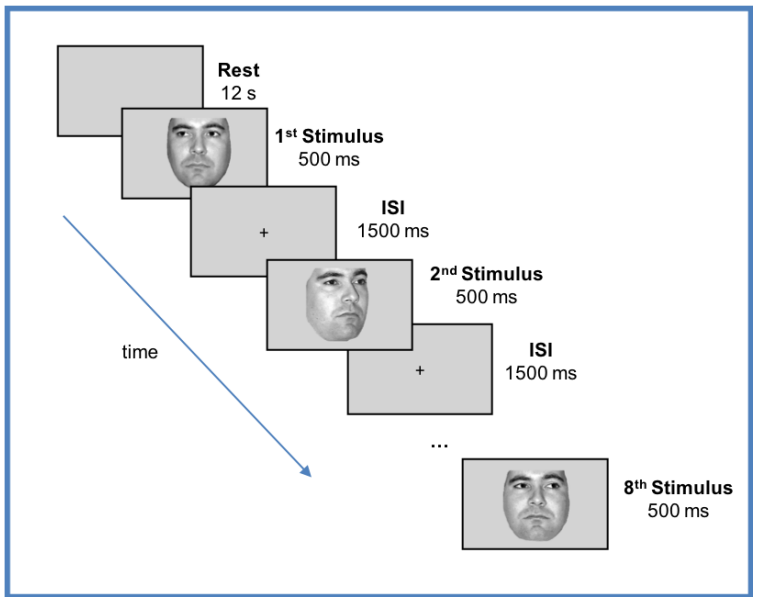


**Figure 8.6:** Example of a block in the experimental design of FaOR dataset. Each block along the experiment follows the same structure but changing the category of the stimulus.

500 ms with an interstimulus interval of 1500 ms (see Figure 8.6). We focused on the faces *vs*. houses classification, although the rest of the stimuli were also included in the GLM to preserve the implicit baseline. Since only one block for each stimulus type was presented in each run, LSU and LSA were equivalent. Although the LSS estimation was developed for event-related designs, we implemented a blocked-version of the LSS approach by iteratively fitting a new GLM for each block. For each model, the target condition is associated with one regressor, and the rest are associated with one error regressor. Thus, there are 8 models for each run, one for category. All methods yielded the same number of estimates to train the algorithm: 1 per run and condition.

### 8.2.3 Faces Identification dataset (FI)

The block-design of the FaOR dataset is considerably different than the one in UG dataset. For this reason, we include an additional dataset based on an event-related

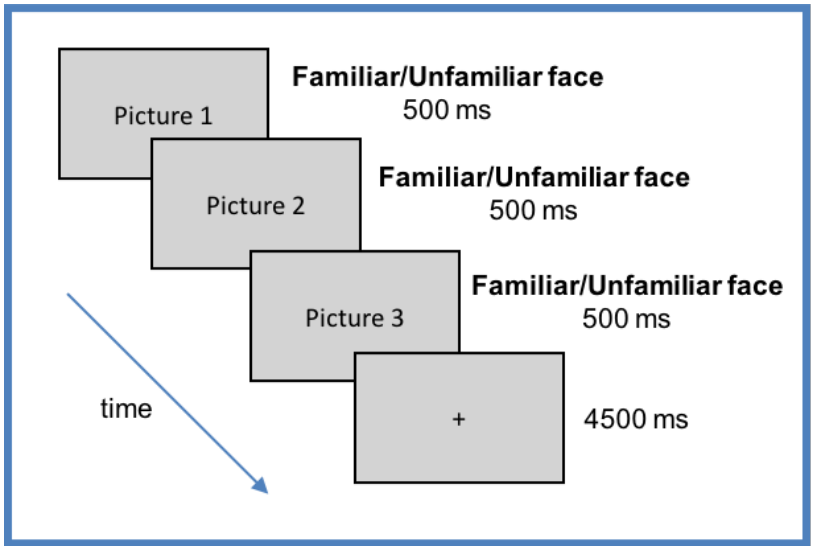


**Figure 8.7:** During each trial, three images of the same identity (normal trial) or two different identities (odball trial) were presented.

design as an intermediate scenario to evaluate the performance of the different methods in the wider range of contexts as possible. To do so, we used data from 33 participants of a recent study published by Visconti di Oleggio Castello et al. (2017). The full database was openly available in Datalad repository (http://datalad.org). Brain images were acquired using a 3T Philips Achieva Intera scanner with a 32-channel head coil [repetition time (TR) = 2000 ms, 35 3-mm-thick axial images, field of view (FOV) = 24 cm, echo time (TE) = 35 ms, flip angle = 90°]. A single high-resolution T1-weighted (TE/TR = 3.7/8.2 ms) anatomical scan was acquired with a 3D-TFE sequence. Preprocessing was carried out following the same procedure used for UG dataset. For a more detailed explanation see the original work (Visconti di Oleggio Castello et al., 2017).

The dataset consists of 11 runs where the participants viewed pictures portraying different familiar and unfamiliar identities: four faces of friends, four unknown faces, and the participant's own face. A trial consisted of three different images of the same individual (normal trial) or two different identities (odball trials), each presented for 500 ms with no gap, followed by a 4500 ms inter-trial interval displaying a white fixation cross (see Figure 8.7). Each trial was modelled with a duration of 1.5 seconds, as it was done in the original paper (Visconti di Oleggio Castello et al., 2017). The order of the events was pseudo-randomized to approximate a first-order counterbalancing of conditions. A functional run contained 48 trials: four trials for each of the nine individuals (four familiar, four unfamiliar and self), four blank trials, four oddball and four buffer trials (three at the beginning and one at the end). Each run had 10 seconds of fixation at

the beginning to stabilize the BOLD signal and at the end (to collect the response to the last trials). We focused on discriminating the neural activity associated with familiar *vs*. unfamiliar faces although the rest of the stimuli were also included in the GLM to preserve the implicit baseline. Eleven beta estimates per condition were obtained by LSU, whereas LSA and LSS yielded 176.

## 8.3 Pattern estimation methods

Once we have introduced the different datasets employed in this chapter, this section describes the methods used for the estimation of the activation patterns: LSU, LSA and LSS. These approaches are employed to compute the contribution of each experimental condition to the acquired fMRI signal. As explained in previous sections, images from the MR scanner in a fMRI study can not be directly used as the input images to train the classifier since each volume contains information from different experimental conditions. Thus, it is crucial to estimate images that take into account the sluggishness of the BOLD signal.

### 8.3.1 Least-Squares Unitary

Previous studies have explored different methods to obtain activation estimates in event-related designs (Abdulrahman and Henson, 2016; Mumford et al., 2012). The most common is the so-called 'Least-Squares Unitary' (LSU), in which all trials of the same type (e.g. experimental conditions) are collapsed into one single regressor, relegating trial variability to the GLM error term. Following the example proposed in Figure 8.8 (top of the figure), LSU tries to estimate the activity elicited by the blue vowel along time, for which a single regressor is employed. Similarly, a different regressor is used to estimate the average activity associated with a male face for each run. Thus, LSU yields a beta image for each condition and run, so that the number of images available to feed the classifier is limited by the number of runs the experiment is divided into.

### 8.3.2 Least-Squares All

Other studies have focused on obtaining single-trial parameter estimates. The most straightforward approach is known as beta-series regression (Rissman et al., 2004), in which a different regressor is used for each trial. Following the notation in Mumford et al. (2012), we from now on denote it as 'Least-Squares All' (LSA). Figure 8.8 (bottom)

shows a visual representation of how this method works. LSA estimates the contribution to the hemodynamic signal that the activity elicited by each blue vowel or male face has, unlike LSU that computes the average of each condition along the run. LSA yields a beta image for each trial of the experiment, increasing considerably the number of samples for the subsequent classification.

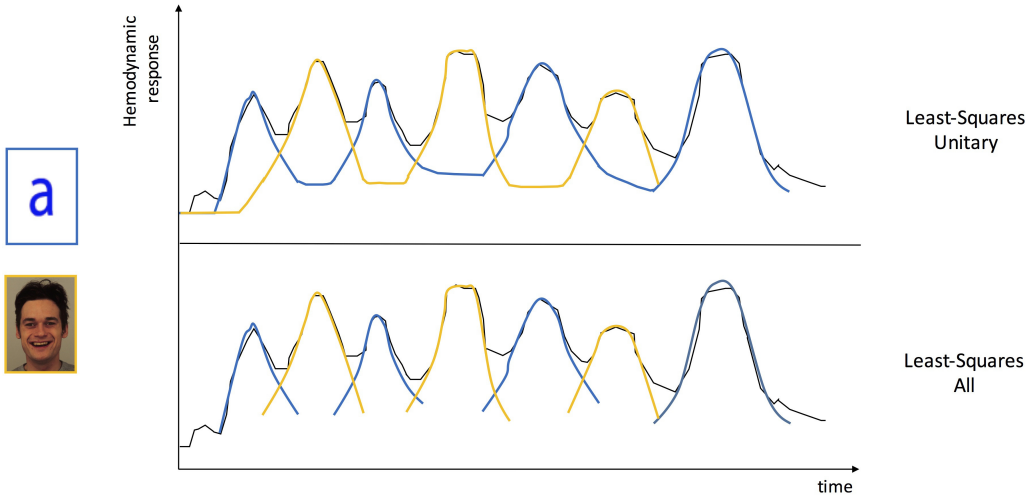


**Figure 8.8:** Comparison of two different approaches for pattern estimate. At the top of the figure, LSU, in which all the trials of the same type for each run are collapsed into the same regressor. At the bottom, LSA, based on estimating one model in which each event is modelled as a separate regressor. LSU can yield less noisy estimates because of the averaging of all the stimuli of the same type within a run, but the amount of resulting estimates to train the classifier is limited to the number of runs the experiment is divided into.

### 8.3.3 Least-Squares Separate

A large number of images to train the classifier can be beneficial from the machine learning standpoint. However, the are other parameters that affect beta estimation that need to be considered. One of these variables is the ISI, i.e. the time between two consecutive trials. When this time is short, the regressors become highly correlated, which can inflate the variance of the resulting LSA-estimates and the subsequent classification accuracies (Mumford et al., 2014). To address this drawback, Turner (2010) introduced an alternative method known as 'Least-Squares Separate' (LSS), based on iteratively fitting a new GLM for each trial. There are different variants on this approach depending on the number of regressors defined. In the simplest one, LSS-1, there is a parameter for the target trial and another single nuisance parameter for the rest (see Figure 8.9). In LSS-2, each GLM includes three regressors: the first one, for the target trial; the second

for the rest of the trials of the same type as the target, whereas the third is used for the trials of a different type. It is thus possible to define as many nuisance parameters as trial types (e.g. LSS-N), although LSS-1 (from now on, LSS) is commonly used due to its simplicity and high performance (Abdulrahman and Henson, 2016; Turner et al., 2012).

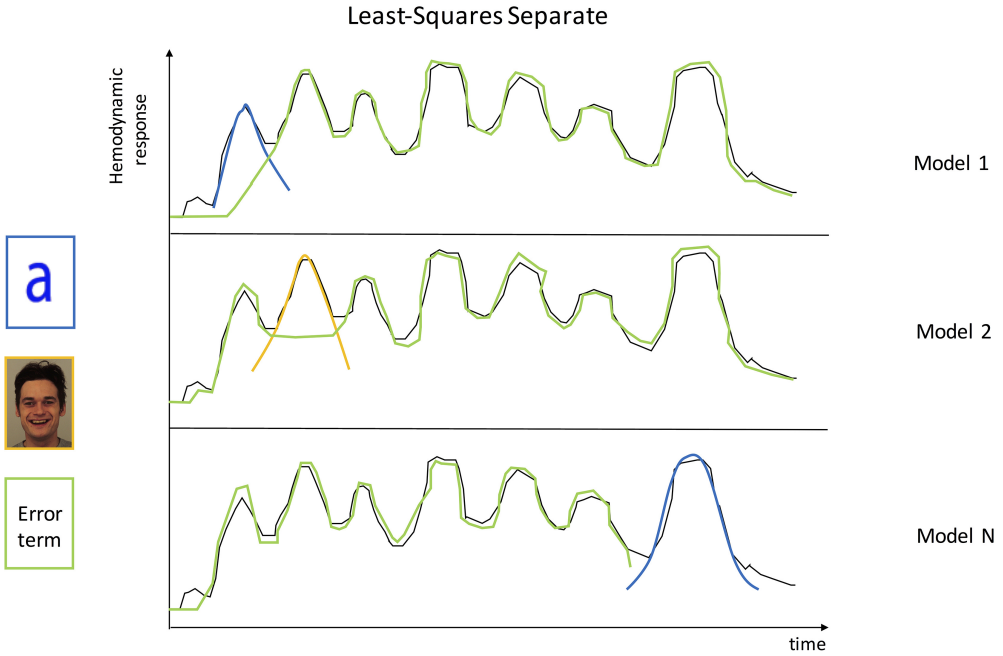


**Figure 8.9:** LSS iteratively fits a new GLM for each unique event with two predicted BOLD time courses: one for the target event and a nuisance parameter estimate which represents the activation for the rest of the events. LSS estimates as many models as the total number of regressors, and in each one only two of them are included: one for the event of interest and a nuisance parameter estimate which stands for the activation for the rest of the events.

The advantages of single-trial estimates are reflected in the fields of neuroscience and also machine learning. Regarding the first one, a good example is the study of working memory, where classic models assume that information is stored via persistent neural activity (Sreenivasan et al., 2014). Whereas averaging across trials may cancel out the noise and improve the signal-to-noise ratio, trial-wise averaging may also remove important coding signals (e.g. Stokes and Spaal, 2016). From the machine learning standpoint, estimating one beta map per trial yields a larger number of images to train the classifier, which improves generalization. Pereira et al. (2009) discussed the important tradeoff between having many noisy examples (e.g. one per trial) or fewer, cleaner ones (e.g. one of each class per run), as a result of averaging images of the same class. Although there is not a fixed number of examples necessary to train the classifier, the more the better. Hebart et al. (2016) showed that run-wise beta estimates can be

more accurate than single-trial ones, which can potentially lead to higher accuracies (Ku et al., 2008) or slightly improve power (Allefeld and Haynes, 2014). However, according to Pereira et al. (2009), at least a few tens of examples in each class are needed to properly estimate the parameters of the classifier, so LSS or LSA would be the most recommended option.

### 8.3.4 Searchlight analysis

Once beta maps are computed, the classifier can learn the relationship between these images and the experimental conditions. The number of beta maps is much lower than the number of voxels that each image contains, so that it is hard to find a generalizable solution for the classification problem. To address this issue, we employed a Searchlight approach across the whole brain (Kriegeskorte et al., 2006), a method that reduces drastically the dimensionality of the input data. Specifically, we created spherical regions of 12-mm radius and only voxels contained in it are evaluated by the classifier. This size was chosen according to previous studies that showed a systematic decrease in performance when a too much large size is selected (e.g. Arco et al., 2016; Chen et al., 2011). Moreover, results obtained in Chapter 7 highlighted that the number of significant voxels exponentially increases for a certain range of Searchlight size, which can lead to distort results and considering voxels as significant that actually are not (Etzel et al., 2013) when large spheres are employed.

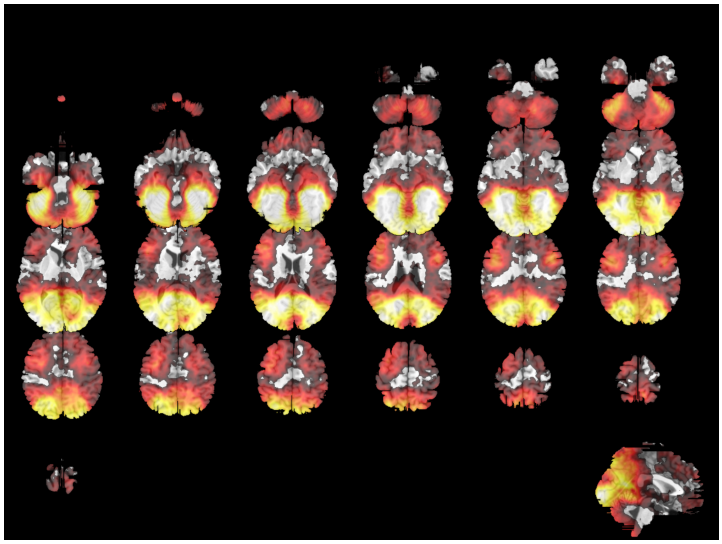

**Figure 8.10:** Searchlight accuracy map obtained for the FaOR dataset. Only above chance accuracies are displayed.

A linear kernel based on SVM was used to classify the images due to the superior performance shown in Chapter 7 and in previous studies (Misaki et al., 2010; Pereira et al., 2009). A leave-one-run-out scheme was used to cross-validate the performance of the classifier (Coutanche et al., 2011; Haynes and Rees, 2006; Lee et al., 2011; Reddy et al., 2010; Wolbers et al., 2011). In this scheme, the classifier is trained with images from all but one run, whereas the patterns of the remaining run are used to test the performance of the algorithm. The number of images available for the training/testing process highly depends on the dataset used, the pattern estimation method employed and the classification problem evaluated (see Table 8.1).

Figure 8.10 shows a map of the group accuracies for the FaOR dataset. Largest accuracies are found in the occipital pole since in this experiment the aim was to identify information patterns associated with different visual stimuli.

## 8.4 Statistical significance of decoding accuracies

Searchlight provides an accuracy map for each individual. However, results derived for a single subject are not informative enough to make any conclusion. Thus, results have to be evaluated at the group level. Brains of different individuals vary in terms of size and shape, so that the resulting accuracy maps are then spatially normalized into a common space. Those voxels with a decoding accuracy better than chance would demonstrate that there is information in the data related to the experimental condition under study, which would increase our knowledge about the neural mechanisms associated with a certain cognitive function. However, the differences of these accuracies from chance level have to be statistically significant. This section describes the approaches employed to assess the statistical significance of Searchlight accuracies.

### 8.4.1 *t*-test and Gaussian Random Field Theory

The first method evaluated is based on Gaussian RFT, a mathematical framework that finds the specific threshold for a smooth statistical map that meets the required family-wise error rate (Brett et al., 2003). The smoothness of a statistical image is usually known, but RFT makes an estimation based on the number of resolution elements (resels) that an image has. Moreover, the probability that a cluster of a certain size is found by chance is derived from the EC, a property related to the probability that a

number of clusters is considered significant when a certain statistical threshold is used (see Section 6.1.1 for a better understanding of these concepts.

We employed the functions provided by the SPM12 package (Wellcome Centre for Human Neuroimaging, 2018) to apply this method. The procedure followed was the same for all the datasets evaluated. After computing the decoding accuracy map for each subject, all maps were normalized to a standard EPI. Then, a voxel-wise $t$-test against the theoretical chance (0.5 in our binary-classification analyses) was applied to these normalized maps. We employed a cluster-defining primary threshold of $p < 0.001$ (uncorrected), which was later used to find significant clusters (FWE corrected, $p < 0.05$) on the resulting map.

### 8.4.2 Stelzer's

Some of the assumptions that the $t$-test and the RFT make are not always met, which would bias the results. For this reason, we employed two non-parametric approaches (see section 6.1.3 for a detailed description) to evaluate how results change in the different datasets described in section 8.2 when these methods are used to assess statistical significance. The first of these alternatives, Stelzer's, combines results from each subject with a Monte Carlo method and based on that, derives a cluster size $p$-value at the group level. This approach is based on permutation tests, which unlike RFT, rely on minimal assumptions. Specifically, a within-subject searchlight analysis was performed shuffling the labels corresponding to the two experimental conditions to distinguish from. We carried out this step 100 times per participant, yielding 100 permuted accuracy maps. Then, these maps were spatially-normalized to a standard EPI image to register images of different subjects into the same coordinate system. A map from each participant was randomly picked following a Monte Carlo resampling with replacement (Forman et al., 1995), averaging the values voxel-wise and obtaining a permuted group map. This procedure was carried out 50000 times, yielding 50000 group permuted maps. To evaluate the significance of each voxel, it is necessary to compare the null distribution with the real accuracy. This accuracy is obtained by training the classifier with actual true labels, and averaging the resulting maps across subjects (from now on, the real group map). For a cluster-defining primary threshold of $p$-value = 0.001 and a distribution of 50000 samples, a voxel will be significant when no more than 50 voxels of the empirical distribution have a larger value than the value of the real group map. To compute the specific $p$-value for a voxel $x$, we employed the following equation:

$$p_{voxel}(x) = \frac{1 + n(x)}{1 + N} \tag{8.1}$$

Once the image has been thresholded at the voxel-level (applying the cluster-defining primary threshold), an empirical distribution of the cluster sizes of the 50000 permuted maps is built to compute the required family-wise error rate at the cluster-level. A set of contiguous voxels are considered a cluster if they share a face, but not an edge or a vertex, in which Stelzer et al. (2013) defines as a 6-connectivity scheme. This cluster search is also applied to the real group map, so that only the clusters which surpass the cluster-level extent threshold are considered significant (see Section 6.2). Once each cluster size has an associated $p$-value, an FWE correction ($p = 0.05$) is applied on all clusters $p$-values to correct for multiple comparisons at the cluster level.

Figure 8.11 shows an example of the group distribution of the accuracies in one voxel for UG (*valence* classification) and FaOR datasets. Training with permuted labels results in accuracies around chance level (50%) in most of the permutations. The green vertical line indicates the significance threshold at which a given accuracy is considered significant, whereas the black one shows the accuracy level obtained after training the classifier with the true labels. It is worth noting that accuracies are not homogeneous across the brain, but they depend on the region from which information is being decoded. For this reason, it is remarkable that this method computes a different empirical distribution for each voxel separately. We employed custom code to carry out all the described stages of Stelzer's method.

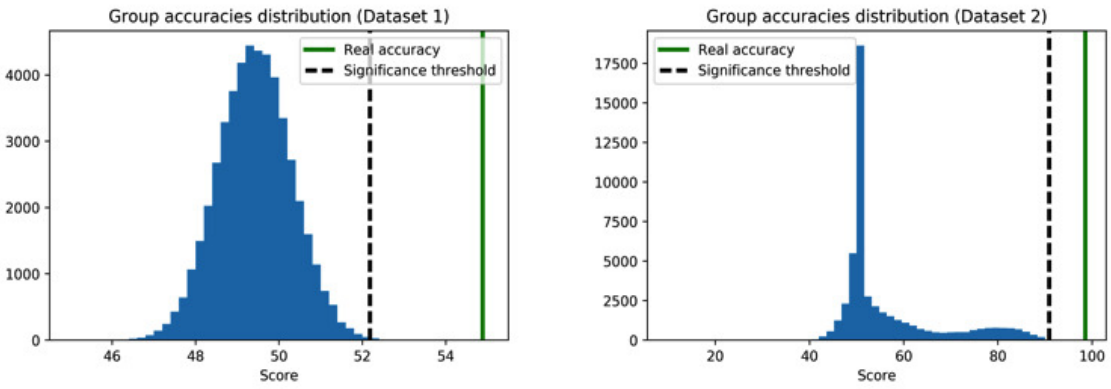


**Figure 8.11:** Distributions of group permuted accuracies in one voxel for UG (left) and FaOR datasets (right). While in UG most accuracies are around chance level, in the second one the number of voxels that surpass the threshold is much larger.

### 8.4.3 TFCE

The last method used was TFCE, included in the FMRIB Software Library (FSL; `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki`). The basis of this method is to transform images to facilitate the discrimination between significant and non significant voxels (see Section 6.2). This transformation relies on the concept that in each image, there are sets of contiguous voxels which are candidates to belong to a cluster. There are two possible extreme scenarios: the first one is that the intensity of the voxels is large (high statistical values) but they are locally distributed. However, it is also possible that the signal is weak (low statistical values) but spatially extended. The main aim of TFCE is to level these two situations so that both are equally likely to be a significant cluster.

On our analysis following this approach, the accuracy maps for all participants were entered into a second-level analysis, where a one-sample $t$-test was used to contrast conditions. To assess significance at the population level, permutation tests were applied. On each permutation, the signs of the individual accuracy maps were randomly flipped and a new $t$-test was performed. This was repeated 50000 times, obtaining an empirical null distribution of $t$-values. The TFCE transformation was later applied to find significant clusters (FWE-corrected, $p = 0.05$).

## 8.5 Results

In this section, we report the results from the three datasets evaluated in this study (1: two events of different duration in each trial, 2: block design, 3: event related design with events of the same non-zero duration) estimated with LSU, LSA and LSS and statistically tested with parametric ($t$-test) and non-parametric (Stelzer's and TFCE) approaches. Additionally, we evaluated two different ways of modelling the two events in UG dataset to study how the duration of the events influenced the results. In the first one, the duration of the jittered interval that separates the two events was added to the first event (words). The alternative was to model both events (words and numbers) as impulse functions, i.e. with zero duration.

### 8.5.1 Comparison of different pattern estimation models (LSU, LSA and LSS)

We first focused on comparing the three pattern estimation methods in four different scenarios: *i*) a paradigm with two events of different durations per trial (event-related

**Table 8.2:** Comparison of the clusters distribution by the different pattern estimation methods and statistical tests in the *valence* classification after modelling the words with the duration of the jittered interval and as zero-duration events.

| Least-Squares Unitary (LSU) | | | | | | |
|---|---|---|---|---|---|---|
| Results | Duration | | | Impulse | | |
| | t-test | Stelzer's | TFCE | t-test | Stelzer's | TFCE |
| Number of clusters | 0 | 0 | 0 | 0 | 0 | 0 |
| Average cluster size | 0 | 0 | 0 | 0 | 0 | 0 |
| Significant voxels | 0 | 0 | 0 | 0 | 0 | 0 |
| Least-Squares All (LSA) | | | | | | |
| Results | t-test | Stelzer's | TFCE | t-test | Stelzer's | TFCE |
| Number of clusters | 0 | 0 | 0 | 0 | 0 | 0 |
| Average cluster size | 0 | 0 | 0 | 0 | 0 | 0 |
| Significant voxels | 0 | 0 | 0 | 0 | 0 | 0 |
| Least-Squares Separate (LSS) | | | | | | |
| Results | t-test | Stelzer's | TFCE | t-test | Stelzer's | TFCE |
| Number of clusters | 4 | 3 | 5 | 1 | 1 | 0 |
| Average cluster size | 30 | 329 | 24 | 52 | 54 | 0 |
| Significant voxels | 122 | 987 | 120 | 52 | 54 | 0 |

design) where the individual contribution of both events was computed, UG, *ii*) same paradigm but modelling the two events with zero duration, UG, *iii*) a block-design from the pioneering study of Haxby et al. (2001), FaOR, and *iv*) an event-related design from a recently published study (Visconti di Oleggio Castello et al., 2017) where all trials were modelled with the same duration, FI.

For the *valence* classification in UG, results in terms of cluster detection and number of significant voxels are summarized in Table 8.2. No significant voxels were found when LSU or LSA were applied regardless of the statistical method used and the way that events were modelled. However, LSS was able to estimate properly the activation patterns associated with each experimental condition in this context of high overlap. Specifically, the LSS approach uncovered a set of informative regions when the first event was modelled with the duration of the jittered interval, as Figure 8.12 reports. It is remarkable that LSS leads to a superior performance over the other two pattern estimation methods regardless the way statistical significance is assessed. However, when the first event is modelled as an impulse function, informative regions were hardly detected even though LSS was used to estimate the activation patterns. Figure 8.13 displays a comparison between the informative regions obtained when words were modelled as duration/impulse events.

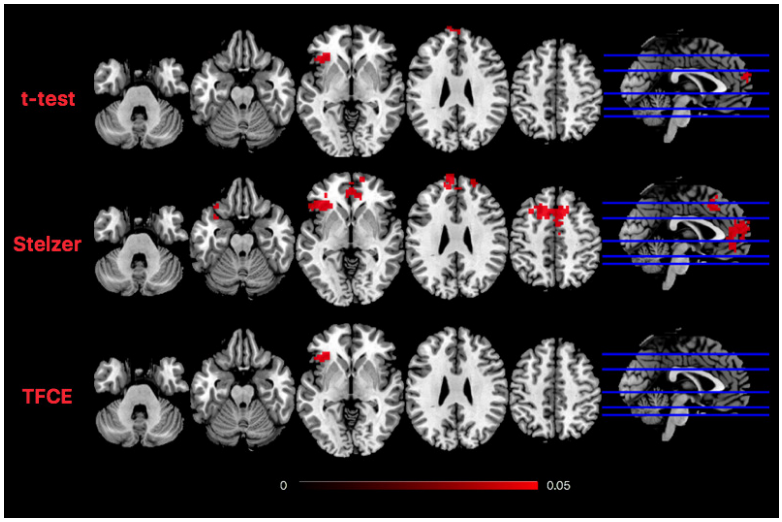


**Figure 8.12:** Significant results obtained by the LSS method when discriminating word valence in UG dataset, modelling the words with its corresponding duration. Results from the *t*-test and TFCE are practically the same, both in location and number of significant voxels. On the contrary, Stelzer's method yields the significant regions obtained by the other methods while increasing the number of significant voxels, showing higher sensitivity.

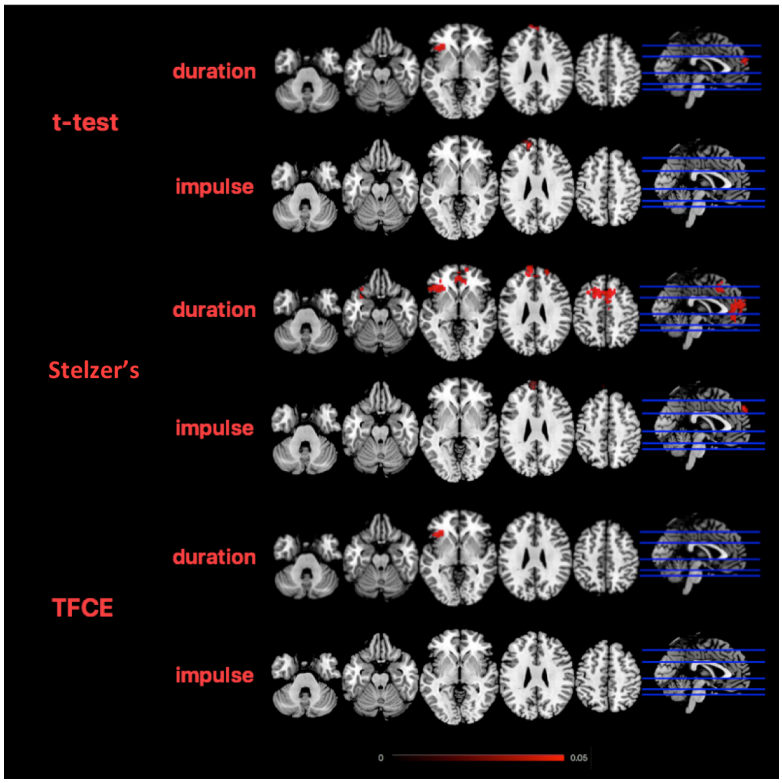


**Figure 8.13:** Comparison of the results obtained in the *valence* classification after modelling the words as duration/impulse events for all statistical methods.

In the *fairness* classification, LSA was the only method that did not obtain any significant result. Table 8.3 summarizes the results for the *fairness* classification, illustrating very similar results for the two ways of modelling the first event. Results obtained when the first event was modelled with the duration of the jittered interval are shown in Figure 8.14. Moreover, Figure 8.15 depicts the results obtained when the first event was modelled as a zero-duration event. As we can see in these figures, results are essentially the same. Thus, it seems that the way the first event is modelled has a minimum influence in the estimability of the activation patterns corresponding to the second event and in the subsequent classification based on these estimations.

**Table 8.3:** Comparison of the clusters distribution by the different pattern estimation methods and statistical tests in the *fairness* classification after modelling the words with the duration of the jittered interval and as zero-duration events.

| Least-Squares Unitary (LSU) | | | | | | |
|---|---|---|---|---|---|---|
| Results | Duration | | | Impulse | | |
| | t-test | Stelzer's | TFCE | t-test | Stelzer's | TFCE |
| Number of clusters | 3 | 1 | 1 | 2 | 1 | 1 |
| Average cluster size | 628 | 15422 | 13909 | 1058 | 13832 | 14399 |
| Significant voxels | 1883 | 15422 | 13909 | 2116 | 13832 | 14399 |
| Least-Squares All (LSA) | | | | | | |
| Results | t-test | Stelzer's | TFCE | t-test | Stelzer's | TFCE |
| Number of clusters | 0 | 0 | 0 | 0 | 0 | 0 |
| Average cluster size | 0 | 0 | 0 | 0 | 0 | 0 |
| Significant voxels | 0 | 0 | 0 | 0 | 0 | 0 |
| Least-Squares Separate (LSS) | | | | | | |
| Results | t-test | Stelzer's | TFCE | t-test | Stelzer's | TFCE |
| Number of clusters | 2 | 1 | 1 | 1 | 1 | 1 |
| Average cluster size | 4469 | 17620 | 16790 | 9742 | 16342 | 13584 |
| Significant voxels | 8938 | 17620 | 16790 | 9742 | 16342 | 13584 |

Regarding FaOR dataset, all pattern estimation methods showed a larger sensitivity than with UG. The three estimation approaches led to very similar results in terms of number of significant voxels detected, as Table 8.4 shows. In fact, this context seems to be not affected by the way activation patterns are estimated. Moreover, the large similarity in the results obtained by the three methods is also found in terms of localization as they identified the same informative clusters. Figure 8.16 illustrates how similar are the results obtained by all the approaches employed. Informative areas obtained by LSU and LSS are only reported because in this particular case LSU is equivalent to LSA, since for each run, there was an only run for each experimental condition.

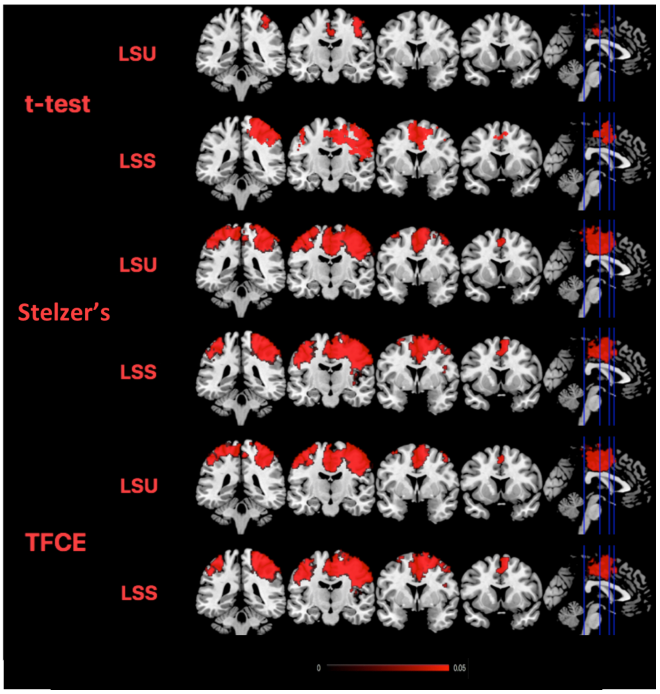


**Figure 8.14:** Significant results obtained by the different pattern estimation and statistical methods in the *fairness* classification modelling the words with its corresponding duration.

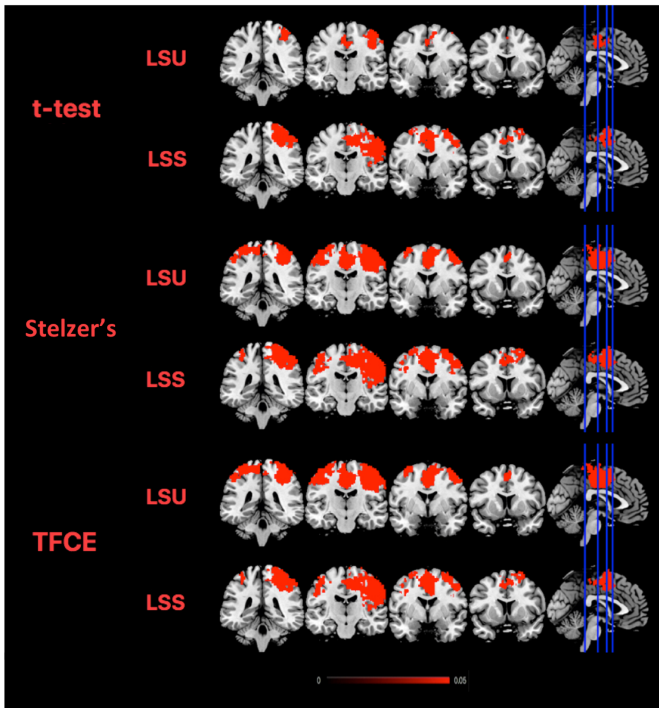


**Figure 8.15:** Significant results obtained by the different pattern estimation and statistical methods in the *fairness* classification of UG where words were modelled as zero-duration events.

Both LSU and LSA allowed an accurate estimation of the neural activity in FI dataset, unlike in UG where no significant voxels were found by these methods. In fact, differences appeared between the three estimation approaches, but results are more similar than in the UG dataset. This can be due to the nature of the experiment itself. The experiment conducted in FI dataset aimed at finding differences at the trial level and not to isolate the neural activity of different events within each trial (as in UG dataset), which is considerably harder. Table 8.4 shows a summary of the informative clusters obtained by the different methods employed, whereas Figure 8.17 illustrates the localization of these informative regions.

**Table 8.4:** Summary of the clusters distribution by the different pattern estimation methods and statistical tests in FaOR and FI datasets.

| Least-Squares Unitary (LSU) | | | | | | |
|---|---|---|---|---|---|---|
| Results | FaOR | | | FI | | |
| | t-test | Stelzer's | TFCE | t-test | Stelzer's | TFCE |
| Number of clusters | 4 | 1 | 1 | 5 | 2 | 3 |
| Average cluster size | 1821 | 9881 | 7717 | 527 | 2511 | 748 |
| Significant voxels | 7283 | 9881 | 7717 | 2635 | 5021 | 2244 |
| Least-Squares All (LSA) | | | | | | |
| Results | t-test | Stelzer's | TFCE | t-test | Stelzer's | TFCE |
| Number of clusters | 4 | 1 | 1 | 2 | 2 | 1 |
| Average cluster size | 1821 | 9881 | 7717 | 1476 | 1383 | 4489 |
| Significant voxels | 7283 | 9881 | 7717 | 2952 | 2766 | 4489 |
| Least-Squares Separate (LSS) | | | | | | |
| Results | t-test | Stelzer's | TFCE | t-test | Stelzer's | TFCE |
| Number of clusters | 4 | 1 | 1 | 2 | 3 | 1 |
| Average cluster size | 1831 | 9906 | 7692 | 1424 | 1463 | 4551 |
| Significant voxels | 7321 | 9906 | 7692 | 2847 | 4387 | 4551 |

### 8.5.2 t-test *vs*. non-parametric methods

We next employed the three methods described in Section 8.4, that is, the *t*-test, Stelzer's and TFCE, to assess significance of the obtained results. Figure 8.12 shows the significant results obtained by each of them when the LSS estimation method was employed in the *valence* classification of UG, when the first event was modelled with its corresponding duration. Here, the *t*-test and TFCE yielded essentially the same results in terms of number of voxels marked as significant and their spatial distribution, but largely differed from Stelzer's. In fact, this method obtained approximately 8 times more significant

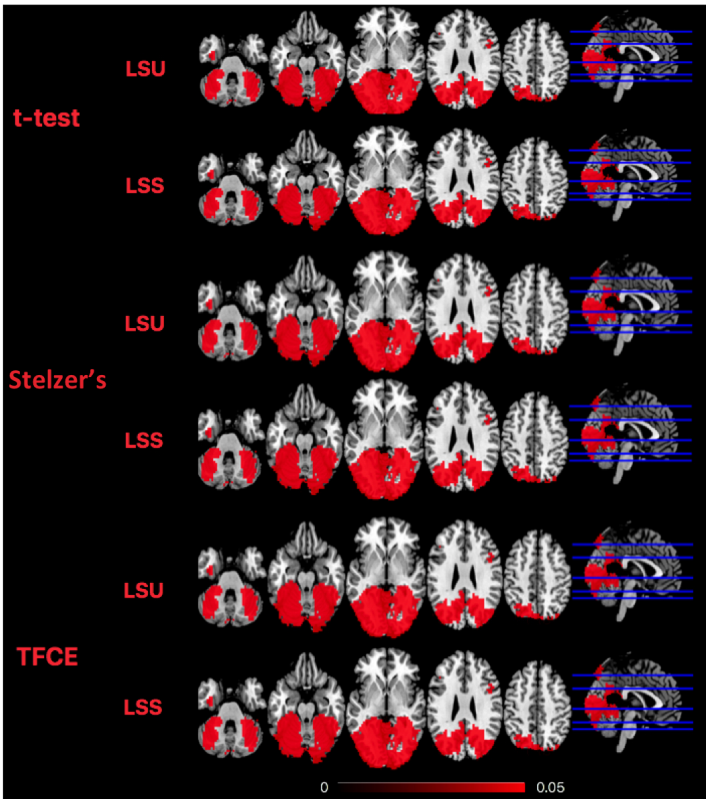


**Figure 8.16:** Significant results obtained by the different pattern estimation methods and techniques for evaluating the statistical significance in Haxby's experiment. LSA is equivalent to LSU in this case, so only results for LSU and LSS are presented.

voxels than the others. All clusters found by the *t*-test and TFCE were also included in Stelzer's, but their spatial extent was larger in the latter. Besides, significant results obtained by Stelzer's are very similar to those obtained by the *t*-test when no correction is applied, as Figure 8.18 shows. Moreover, Figure 8.19 illustrates the distribution of the significant voxels across the most informative regions. When the first event was modelled with zero duration, TFCE did not yield any significant result, but the *t*-test and Stelzer's obtained exactly the same informative cluster. Figure 8.13 shows these results and illustrates the differences between the two ways of modelling the first event.

In the *fairness* classification (first event modelled with its corresponding duration), the differential sensitivity between the *t*-test and Stelzer's was also obtained, but in this case, TFCE yielded very similar results to Stelzer's instead than to the *t*-test (localization of the significant regions are shown in Figure 8.14 whereas the clusters distribution in the most informative regions is depicted in Figure 8.20). Modelling the first event as zero duration did not change much the results, as Figure 8.15 shows. It is important to highlight that when any of the non-parametric approaches was used, the difference

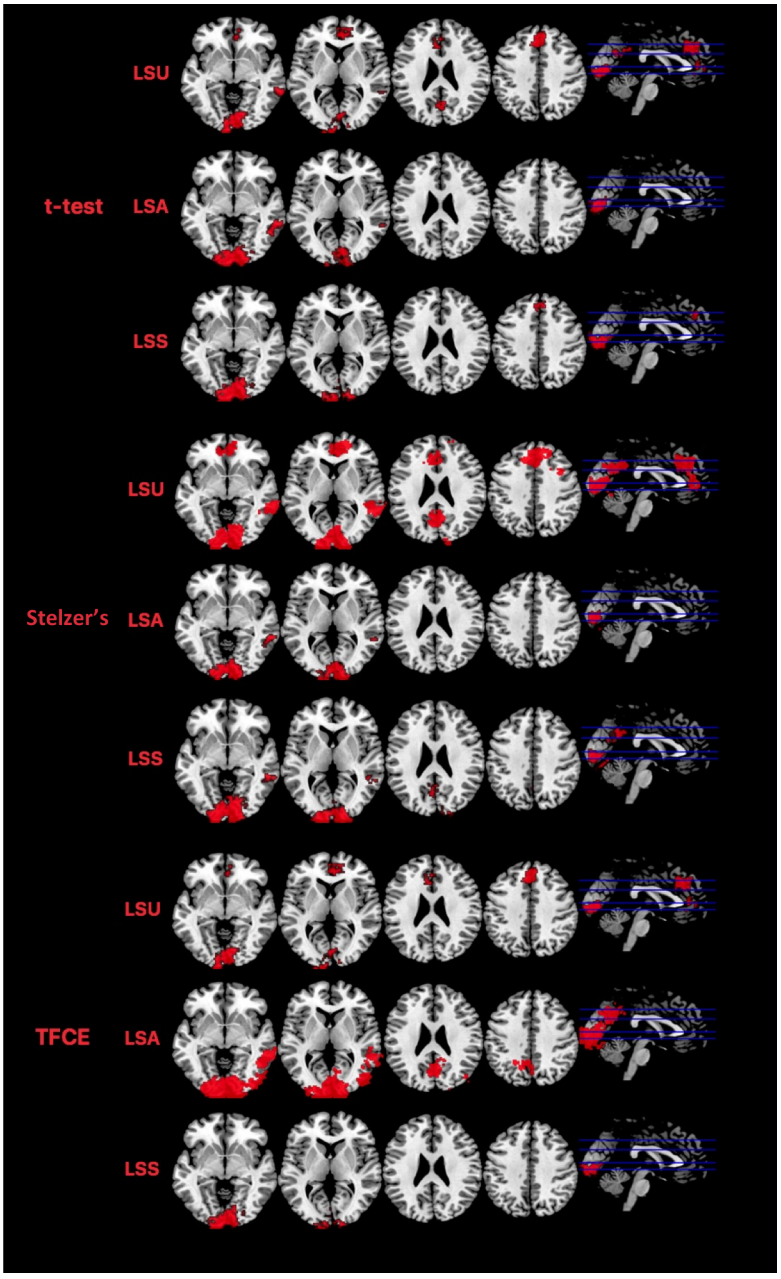


**Figure 8.17:** Significant results obtained by the different pattern estimation methods and techniques for evaluating the statistical significance in FI dataset.

in the informative regions obtained by the LSU and LSS methods was minimum. We further discuss the implications of this finding in Section 8.6.

Figure 8.16 reveals the differences between the three approaches for FaOR dataset, whereas Figure 8.21 shows how information is distributed across different regions. Similarly to UG, Stelzer's shows larger sensitivity regardless of the estimation method used, (see Table 8.4). Moreover, the location of the significant voxels is quite similar

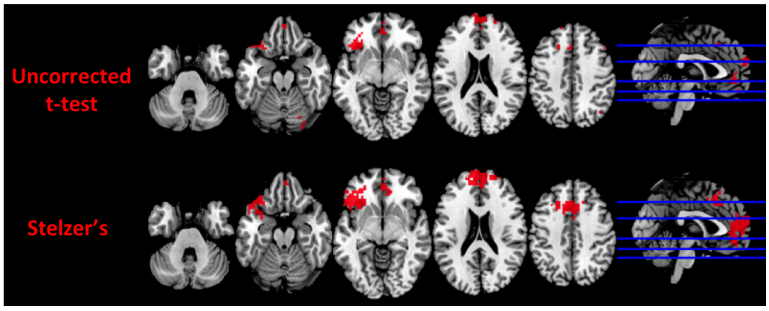


**Figure 8.18:** Comparison of the uncorrected results from the *t*-test ($p$ <0.001) and the significant voxels obtained by Stelzer's in UG dataset (modelling the duration of the words). The distribution of the voxels is similar in both cases, so that differences may rely on the inability to surpass the statistical threshold when the *t*-test is applied.

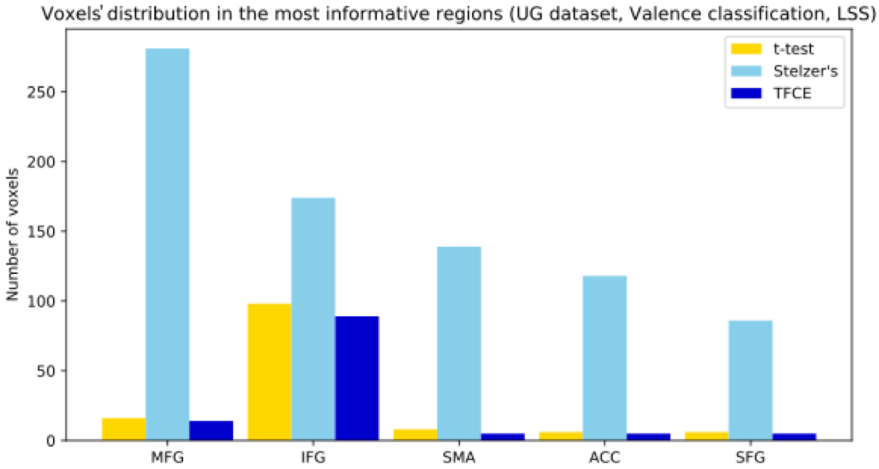


**Figure 8.19:** Voxels distribution in the most informative regions for the *valence* classification of UG, modelling the first event with its corresponding duration. IFG = Inferior Frontal Gyrus; ACC = Anterior Cingulate Cortex; SFG = Superior Frontal Gyrus.

across the three approaches: they found a single massive significant cluster, slightly larger in case of TFCE and with a 35% of more significant voxels in the case of Stelzer's in comparison with the *t*-test. This superior sensitivity of non-parametric methods is also observed in FI dataset (see Table 8.4), whereas the most informative brain regions are summarized in Figure 8.17.

## 8.6 Discussion

In this chapter, we evaluated different pattern estimation methods in a context where a sustained activity had to be isolated from a zero-duration event, in addition to a classic

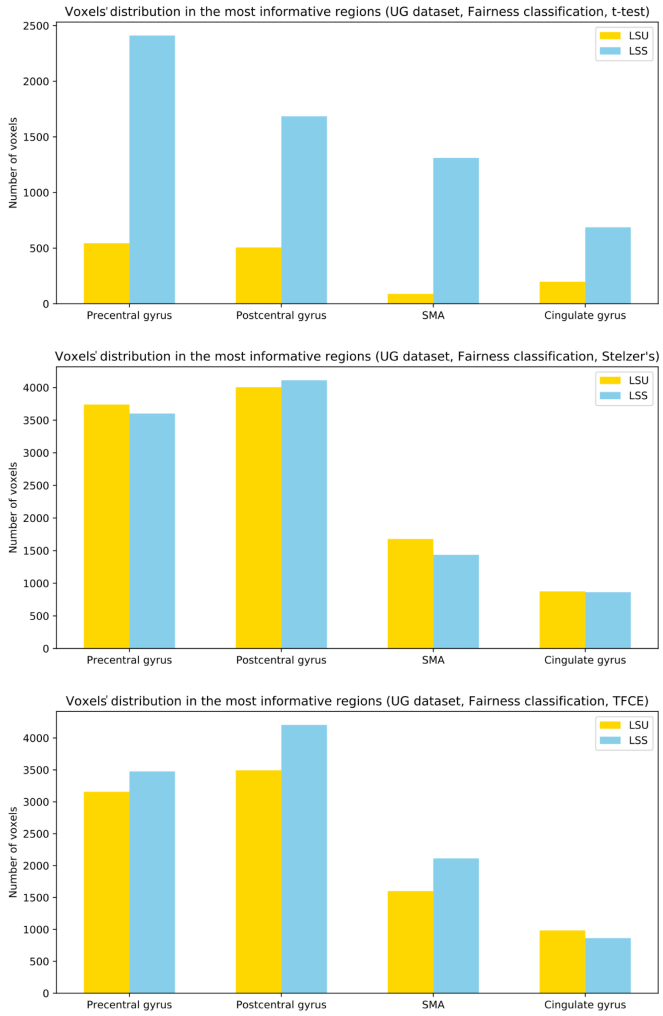


**Figure 8.20:** Voxels distribution in the most informative regions for the *fairness* classification of UG, modelling the first event with its corresponding duration. Region SMA = Supplementary Motor Area.

block design and an event-related design. In the first one, the LSS method resulted the optimal option, whereas in the other two contexts, the differences between the estimation methods were smaller. We also examined the suitability of parametric and the non-parametric approaches to evaluate the significance of the results obtained with the different estimation methods. Stelzer's showed a large sensitivity for all the contexts evaluated, especially when the differences in the activity associated with each class were much subtler.

In UG dataset, we found large differences in performance across the pattern estimation methods, particularly for the *valence* classification. These differences were present in the two ways of modelling the first event. Estimating responses through LSS allowed

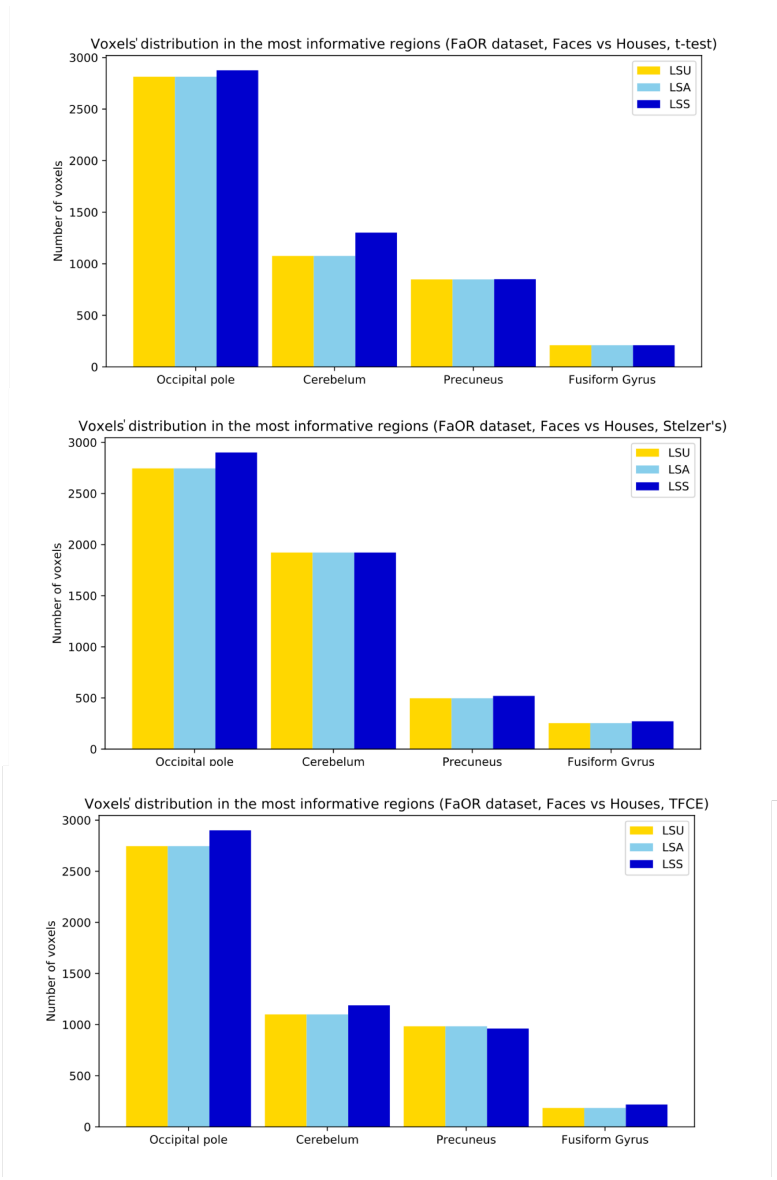


**Figure 8.21:** Voxels distribution in the most informative regions for each statistical and pattern estimation method in FaOR dataset. Results are similar for all the methods employed.

us to detect the involvement of a coherent set of brain regions, whereas using LSU and LSA did not yield significant results. Previous studies showed that the performance of LSA and LSS (Abdulrahman and Henson, 2016) is affected by parameters such as the ISI, noise and trial variability. Collinearity is another element that plays a crucial role in the estimation of neural activity. The difficulty of applying decoding analyses in our paradigm is not due to a short interval between consecutive trials, but to the lack of separation between the activity associated with each event within a trial. When the words were modelled as extended events, in accordance to the sustained preparatory ac-

tivity that they generate, the LSS approach reached its maximal sensitivity. Collinearity cannot be reduced to the way in which events are modelled, as it is also highly affected by the cognitive nature of the process underlying the events. It is worth highlighting that, to the best of our knowledge, this is the first time that these estimation methods are compared in a setting like this. Our results are coherent with findings of previous studies. Analyses carried out by Mumford et al. (2012) concluded that LSS outperforms LSA in high collinearity settings, as it does not employ any regularization strategy. Besides, it is worth remembering that this method was developed due to the poor performance of LSA in rapid event-related designs.

The previous conclusions are strengthened by the fact that the two ways of modelling the first event lead to very different results. Considering the words as zero-duration events yielded no significant results for LSU/LSA in the *valence* classification, and only when *t*-test and Stelzer's were applied to the LSS estimation a small cluster was found. Collinearity cannot be reduced to the way in which events are modelled, as it is also highly affected by the cognitive nature of the process underlying the events. Participants read an adjective with a certain valence, and then they use this information to prepare to respond to the offer (second event). Thus, there is a preparatory process that leads to a sustained activity along time. However, the second event captures a completely different process. Once participants make a decision (cooperate or not depending on the fairness of the offer), the process ends. A large body of literature shows that preparatory processes extend in time (e.g. Bode and Haynes, 2009; González-García et al., 2017, 2016; Sakai, 2008) whereas responding to a brief target does not (see the temporal duration of the potentials in Moser et al., 2014). This has been also measured by other neuroimaging methods, such as the CNV ERP potential (Di Russo et al., 2017). For this theoretical cognitive reason, this second event is modelled with zero duration.

The analyses of the second event of UG (e.g. the *fairness* classification) yielded significant results for the three pattern estimation methods, unlike the *valence* classification where only LSS was sensitive enough. Besides, the influence of the different ways of modelling the first event into the estimability of the second event was minimal since results are essentially the same in both contexts. The key of this finding is the classification problem itself. Neural activity differentially associated with valence is hard to obtain, as shown by recent metaanalytic approaches (Lindquist et al., 2015), whereas the fairness of an offer generates large differences and thus it is easier for the LSU approach to make an accurate estimation. Regarding LSA, we mentioned above the large collinearity between the first event (adjective) and the second (offer), so it was

highly expected that LSA did not find any informative regions in neither the *valence* nor the *fairness* classification. This raises the intriguing possibility that in contexts where most of the strategies fail to detect differential activity, LSS might be sensitive to small variations.

In FaOR dataset we found large similarities in the results obtained by all pattern estimation methods. A block for each object category was presented only once in each run, which means that no average was applied across experimental conditions of the same type. This yields the same number of beta maps for all classifiers, so that the disadvantages of LSA from a machine learning standpoint are not met. Besides, block settings are not propitious for a better performance of LSS since the overlap of signals is much lower than in event-related designs. Another reason for this similarity is the large perceptual difference in the neural activity elicited by each type of stimulus (faces and houses), so that it is straightforward for a classifier to build a decision hyperplane that properly separates the corresponding activation patterns.

We used FI dataset to evaluate the performance of the different pattern estimation methods in a context more similar to our experiment than FaOR dataset. In FI, all pattern estimation methods were able to extract significant regions. Besides, these regions are quite similar regardless of the method used. It is remarkable that LSA allows a good estimation in this setting. There is an important difference in the experimental design that can explain this result: in FI all events represent faces: participants evaluate if these faces are familiar/unfamiliar and respond according to that, which involves a brief activity. However, in UG dataset participants read an adjective with a certain valence, and according to this valence, they prepare to respond to an offer. Thus, there is a preparatory process that leads to a sustained activity along time.

As a further goal, we aimed at testing the adequacy of different statistical approaches. For the *valence* classification of UG, we only obtained significant results when the LSS method was employed, for the two different durations assigned to the first event. The significance maps are essentially the same after applying $t$-test and TFCE, both in the number of significant voxels and in their location. Moreover, Stelzer's resulted in a larger sensitivity than the other methods, yielding eight times more significant voxels. Figure 8.18 compares the uncorrected results for the $t$-test (voxel-level threshold: $p < 0.001$, but uncorrected for multiple comparisons) with the corrected results obtained by Stelzer's. In this case, there is much more coherence between both methods regarding the number of voxels and, crucially, their location. In fact, the three clusters that Stelzer's marked as significant are found with the uncorrected $t$-test as well. Therefore, rather than

being less sensitive to false positives, Stelzer's method seems to efficiently detect true data that otherwise do not surpass the statistical threshold. There are several studies that support that non-parametric approaches are able to simultaneously improve the sensitivity while precisely controlling for false positives (e.g. Eklund et al., 2016; Nichols and Hayasaka, 2003; Silver et al., 2011; Stelzer et al., 2013; Winkler et al., 2014). In addition and most interestingly, the largest cluster uncovered by LSS in the *valence* classification resides in the Medial Frontal Cortex (see Figure 8.19) and includes the peak of maximum differences between positive and negative valence observed in the metaanalyses published by Lindquist et al. (2015) (MNI = [9, 39, -9], see Figure 1). Thus, this close correspondence speaks strongly in favor of the higher sensitivity of the method.

Moreover, our study is the first to compare Stelzer's and TFCE methods. Although both use permutation testing for evaluating significance, the way in which they implement permutations may lead to the large differences observed. One of the most appealing aspects of Stelzer's is that it takes into account the spatial inhomogeneities of the image. In fact, the scheme used by this approach is equivalent to computing a significance threshold for each voxel separately. This controls the false-positives rate in non-informative voxels and avoids being too conservative in the informative ones (Stelzer et al., 2013), which may lend it more sensitive in event-by-event estimations. An encouraging finding is that there is large spatial overlap between the regions that TFCE and Stelzer's mark as significant. Specifically, all significant voxels in TFCE are also considered significant by Stelzer's, but the latter adds voxels to the previously identified clusters (see Figure 8.12). We found even more similarities between Stelzer's and TFCE in the *fairness* classification. In fact, the way these voxels are distributed is almost identical as Figure 8.14 reveals. Most information is encoded in the Pre/Postcentral gyrus, the SMA (Supplementary Motor Area) and the Cingulate Gyrus, as Figure 8.20 shows. These areas are consistent with previous experiments based on the Ultimatum Game (UG), Corradi-Dell'Acqua et al. (2013). For a more detailed explanation of this task and the concordance between the informative regions and our results, see the meta-analysis by Gabay et al. (2014).

As predicted, similarities between the different statistical methods were larger in FaOR. Regarding the *t*-test and TFCE, the spatial distribution of the voxels was essentially the same, with a slight boost of 5% in the number of significant voxels when the latter was applied. Moreover, Stelzer's yielded 35% more significant voxels, but all the additional ones marked as significant were adjacent to the clusters obtained by the other two methods. Figure 8.21 highlights the regions where the information is

mainly distributed and its variability over different statistical methods, much smaller than in UG dataset. Results are essentially the same for each pattern estimation and statistical method, principally in the occipital pole and the fusiform gyrus. Stelzer's yielded more informative voxels in the cerebelum, but the *t*-test and TFCE were more sensitive in the precuneus. It is important to point out the much larger increase in sensitivity that Stelzer's yielded in UG in comparison with FaOR. One possibility is that noise was differently distributed in both designs and generated a differential tendency to false-positives. The jitter between experimental conditions in UG dataset and the fact that we were isolating different events within a trial may be the reason why a more adequate statistical method leads to larger improvement of sensitivity in this dataset compared to a block design (FaOR). We highlight the importance of this finding since although Stelzer's showed a larger sensitivity in all contexts, it was even higher than the other two methods in the most difficult case, when the overlap and collinearity between conditions were highest. The nature of the classification *per se* may also be of importance in this difference. Whereas the classic block design from Haxby et al. (2001) contrasted two stimuli with large perceptual and phylogenetic differences (e.g. Kanwisher and Yovel 2006), the classification employed in UG dataset compared the same physical stimuli (words), equated in length (number of letters), frequency of use and arousal levels. In addition, whereas the brain networks involved in face processing are different from those activated by houses (Haxby et al., 2014), isolating regions with a differential involvement in valence processing is much harder (e.g. Lindquist et al., 2012, 2015).

Results in FI dataset show a great similarity between the two methods based on permutations, more sensitive than the *t*-test as in previous datasets. In fact, they are more similar to those obtained in a block-design (FaOR) than in the event-related of UG dataset. Specifically, the occipital pole, followed by the MFG (Medial Frontal Gyrus) and the MTG (Middle Temporal Gyrus) are the most informative regions (see Figure 8.22), which are consistent with the original study (Visconti di Oleggio Castello et al., 2017). It is important to mention that the additional mechanisms that we have employed to ascertain that results in all the analyses conducted are trustworthy. The first one is the proper selection of a searchlight size. Experiments carried out in Etzel et al. (2013) showed that the number of voxels considered informative in a searchlight map tends to grow as the searchlight radius increases, even when the size of the informative region stays fixed. Thus, the larger the searchlight size, the more likely to obtain false positives. This is consistent with findings in Stelzer et al. (2013), where false positives were boosted for a searchlight diameter of 11 voxels. For our analyses, we chose an intermediate

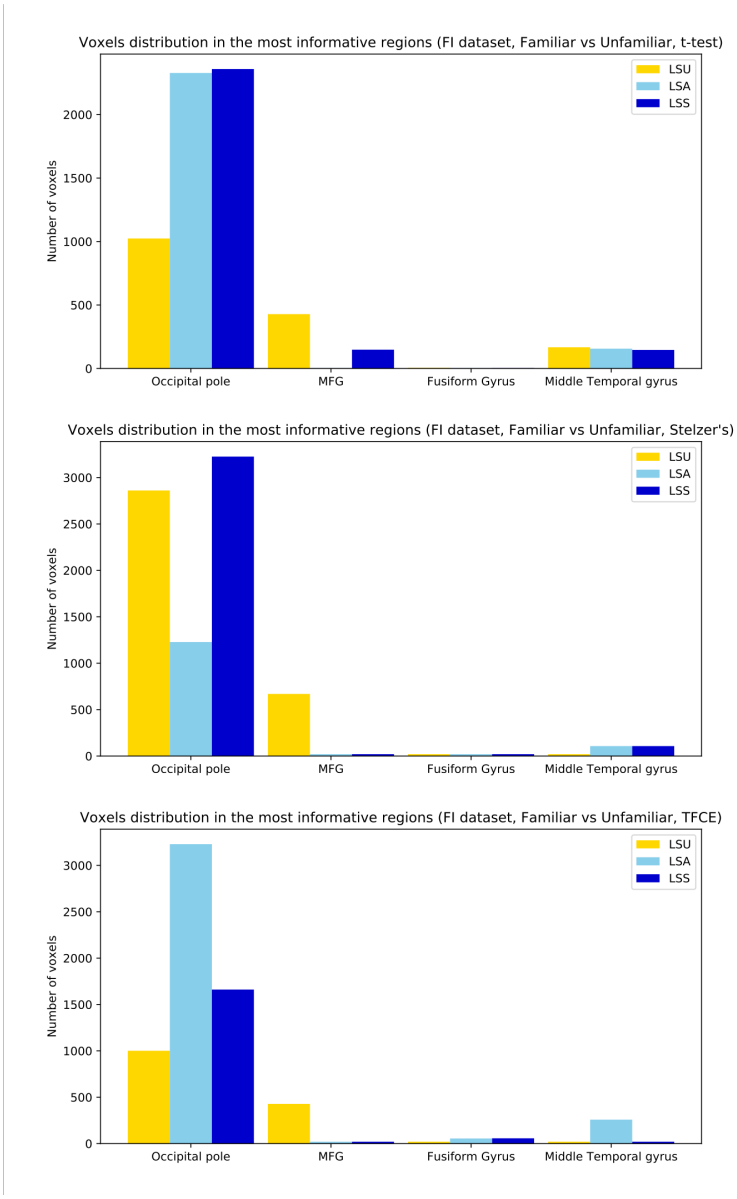


**Figure 8.22:** Voxels distribution in the most informative regions for each statistical and pattern estimation method in FI. MFG = Medial Frontal Gyrus.

value of 8-voxels searchlights to strike a balance between sensitivity and specificity (Arco et al., 2018; Chen et al., 2011). Additionally, we selected a conservative value for the initial-cluster forming threshold to control false positives. The use of a liberal value can have detrimental effects on false positives, location and even interpretation of neural mechanisms (Woo et al., 2014). Likewise, Stelzer et al. (2013) fully studied the relationship between this parameter and the results obtained and they highly recommend the election of a $p$-value ranging from 0.005 to 0.001. We chose the most conservative value ($p$=*0.001*), prioritizing the control of false positives over sensitivity.

This chapter highlights the importance of the estimation of the activation patterns in the subsequent classification analysis. Classification in contexts with high signal overlap and collinearity is not possible unless the estimation method provides a reliable estimate of the activation patterns. Analyses performed in this chapter evidence that the way statistical significance is assessed has a large influence in decoding results. In the contexts evaluated, non-parametric approaches seem to be more sensitive than parametric methods. Besides, Stelzer's computes a significance threshold for each voxel separately, taking into account that fMRI decoding can be more difficult in some brain regions than in others.

CHAPTER 9

# ATLAS-BASED METHODS FOR IDENTIFICATION OF INFORMATIVE BRAIN REGIONS

Previous chapters have evaluated the performance of different methods of the decoding framework: from the "Classification" stage (Chapter 7) to the optimal estimation of the activation patterns in different contexts (Chapter 8). All the analyses performed in these chapters employed Searchlight as a tool to identify informative brain regions. In this chapter, we aimed at evaluating the performance of different alternatives based on atlas in two different classification contexts (large and subtle differences in the activity associated with each condition). Figure 9.1 illustrates the global framework of fMRI classification, highlighting the "Classification" stage this chapter focuses on. Moreover, Figure 9.2 shows a detailed schema of the system evaluated in this Chapter.

## 9.1 Introduction

Previous chapters have shown the ability of multivariate methods to extract information from fMRI activation patterns, even when the differences in the contrast evaluated are very subtle. One of the main advantages of this kind of analyses is the multiple contexts where they can be applied. Several studies have employed these methods in clinical contexts, providing tools for computer-aided diagnosis of different neurological disorders, such as Alzheimer's (Arco et al., 2015), Parkinson's (Choi et al., 2017), epilepsy (Del Gaizo

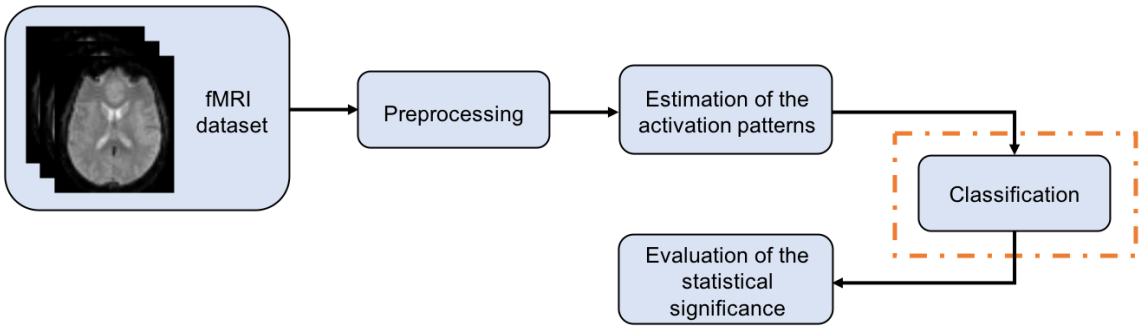


**Figure 9.1:** Illustration of the general framework in fMRI classification. This chapter focuses on the Classification stage.

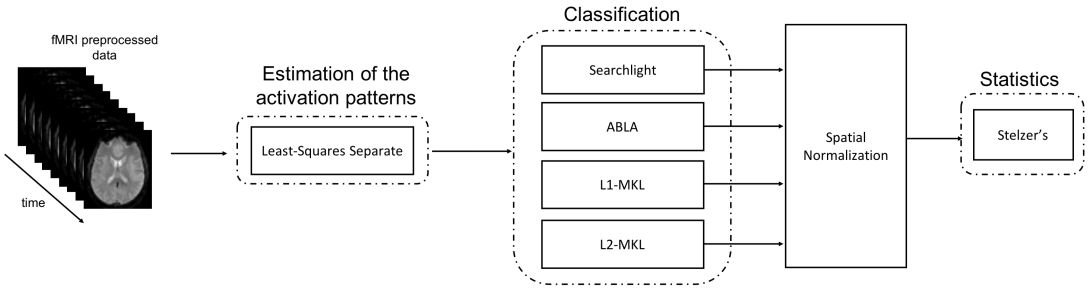


**Figure 9.2:** Overview of the system evaluated in this chapter. Once fMRI volumes are preprocessed, the images used as input of the classifier are computed with the LSS method. Then a classification analysis is performed (Searchlight and three atlas-based approaches) over these beta maps. The resulting accuracy/weight maps are then spatial normalized, and finally, the significance of each voxel is assessed with the Stelzer's approach.

et al., 2017) or brain computer interfaces in quadriplegic patients (Blankertz et al., 2007; Nurse et al., 2015). Here, obtaining the maximum decoding performance is the main aim, whereas the source of information is not of interest. On the other hand, machine learning methods are also used to study the brain regions involved in different cognitive operations (Haxby et al., 2014), and here the main goal is not prediction itself. Hebart and Baker (2017) remarked the importance of differentiating multivariate decoding for prediction and for interpretation as two independent frameworks. In the interpretation context, MVPA provides larger sensitivity than classic univariate approaches (Haynes and Rees, 2005; Norman et al., 2006), as it localizes where information is contained based on the distribution of spatial patterns. However, MVPA brings some crucial points to be considered: is it possible to use these techniques in a different context from which they were developed for? If so, would it be necessary to modify the existing algorithms to accomplish the new goals? Finding the most adequate method for each specific context is

of vital importance, and in the current chapter we aimed to compare the sensitivity of different approaches and to propose some variations to assess the suitability of these methods in the field of Cognitive Neuroscience.

From the identification perspective, classification is simplest when performed in an ROI based on *a priori* knowledge. The accuracy of the algorithm highly depends on how well the regional hypothesis fits the observed data. Haxby et al. (2001) demonstrated that the representations of faces and objects were differentially distributed in the ventral temporal cortex, whereas Haynes and Rees (2005) showed an orientation-selective processing in the primary visual cortex (V1). Other studies detected distributed patterns of activity in the visual cortex (Cox and Savoy, 2003; Kamitami and Tong, 2005), whereas Poldrack (2007) highlighted the Type I error reduction when a statistical test was applied to each ROI. However, when there is not a straightforward hypothesis regarding the regions involved in specific computations, the whole brain may need to be explored. The main drawback of whole-brain analyses is related to the curse of dimensionality: in fMRI studies, there are usually many more features (e.g. voxels) than samples (e.g. images or volumes), which complicates the definition of a classification model to separate the two classes (Fort and Lambert-Lacroix, 2005). Alternatively, feature-selection methods select a subset of informative features (e.g. voxels in fMRI) that will be the input to the subsequent classification. As an example, *t*-tests can be used to restrict the voxels fed to the classifier to those that differ between the classes. Previous studies have employed this method to localize the regions associated with different pathologies and psychological contexts (Arco et al., 2015; Balci et al., 2008; De Martino et al., 2008; Haynes and Rees, 2005; see Mwangi et al., 2014, for a detailed review).

One of the most appealing approaches for identification of cognitive informative regions is the Searchlight technique (Kriegeskorte et al., 2006). As explained in Chapter 5, this method offers results potentially easier to interpret due to its larger spatial precision without the need to define specific ROIs. Many studies have successfully used this technique (e.g. Chen et al., 2017; Cichy et al., 2016; Coutanche et al., 2011; González-García et al., 2017; Loose et al., 2017; Qiao et al., 2017). However, it also has some disadvantages and limitations to consider. Searchlight performance depends on the size of the sphere as previous studies showed (Arco et al., 2016; Etzel et al., 2013; although see Chapter 7). This method also tends to mark voxels as significant only because they are at the center of a sphere that contains informative voxels, leading to somewhat distorted results (see Figure 3 in Etzel et al., 2013 for an extreme example). Another problem is its high computational cost. Each Searchlight analysis entails a massive

number of classifications (one for each voxel of the brain), increasing the computational time compared to other simpler approaches. This time cost increases exponentially when different values of the parameters associated with the classifier are evaluated to find the one with the largest performance (grid search) and also when permutation tests are used to evaluate the statistical significance of the results.

There are other alternatives based on atlas that do not suffer from this large computational cost. This is the case of Multiple Kernel Learning (Lanckriet et al., 2004), a method that uses *a priori* templates of brain organization to guide the decision of the classifier. Specifically, this approach extracts information from brain parcellations provided by an atlas to maximize the performance of the classification algorithm, and ranks the regions according to their importance in the decision. A crucial aspect is the two-level hierarchical model that this approach entails. The regions used for classification have an associated weight, which indicates their contribution to the model. Voxels within each region have a similar weight value. Thus, MKL offers information both at the region and at the voxel level. Previous studies have used this method in the context of neuroimaging, e.g. discrimination between Parkinson's neurological disorders (Adeli et al., 2017; Filippone et al., 2012), identification of attention deficit hyperactivity disorder (ADHD) patients (Dai et al., 2012; Qureshi et al., 2017) and localization of informative regions (Schrouff et al., 2018). This approach leads to a sparse solution, which means that only a subset of regions is selected to contribute to the decision function (similarly to feature-selection methods). However, this decreases its ability to detect informative regions, which is not recommended when identification of informative areas is the main aim. Schrouff et al. (2013a) proposed another method based on local averages of the weights from each region defined in an atlas. This is known as Atlas-based local averaging (ABLA). First, a whole-brain classification is performed, leading to a weight map summarizing the contribution of each voxel. Then, the weights defined in each region of the atlas are averaged and normalized by the size of the region. This yields a score of the informativeness of each region. Hence, this approach builds only one classification model since the summary of the weights is done *a posteriori*. In contrast, MKL combines the different regions of the atlas as part of the learning process, so that using a different atlas will result in a different classification, with the subsequent increase in computational cost compared to ABLA.

Previous research has usually employed atlas-based methods in classification contexts, where the main aim is to obtain the largest accuracy possible. However, the validity of these approaches in an identification scenario, where the goal is to find the informative

brain regions during a certain mental operation, is yet unknown. Therefore, in this chapter, we evaluated the performance of different atlas-based approaches in an fMRI experiment, in two contexts with differential changes in neural activity. To do so, we modified the MKL and ABLA methods to better fit the requirements of an identification context instead of a classification one. Specifically, we included an L2-version of MKL, which avoids sparsity by allowing all regions of the corresponding atlas to contribute to the model. We compared the results obtained by MKL and ABLA methods to those obtained by Searchlight, as this approach is mainstream in recent neuroimaging research. In our study, we employed nine different atlases to examine how different brain parcellations influenced the identification of informative regions of MKL and ABLA.

## 9.2 Materials

We employed UG dataset to test the reliability of the different classification approaches (sensitivity and overlap of the significant regions with those obtained by Searchlight). This dataset has been previously described in Section 8.2.1. We focused on two different classification analyses. First, we aimed at discriminating between the neural activity associated with accepting vs. rejecting offers (from now on, *decision* classification). The hand used to respond was counterbalanced across participants, which means that odd subjects used the right/left hand to accept/reject an offer, whereas in even subjects the order was the opposite. Second, we focused on distinguishing the positive vs. negative valence of the words (the *valence* classification performed in Chapter 8). We employed a Least-Squares Separate (LSS) model to obtain an accurate estimation of the neural activity (Turner et al., 2012). Previous studies have shown that this is the best approach for isolating the activity in contexts like this experiment (Abdulrahman and Henson, 2016), where overlap and collinearity are large (see also results obtained in Chapter 8).

## 9.3 Atlases

In this study, we used 9 atlases to assess the reliability of the informative regions obtained by the three atlas-based classification methods. They differ in three main aspects: the information that they use to cluster the brain regions (anatomical, functional or multimodal), the number of resulting regions (from 12 to 400) and the algorithms that implement the parcellation (a wide spectrum, from the $k$-means clustering to Bayesian

models). Figure 9.3 illustrates the parcellations proposed by the different atlases, which are described in detail in the following paragraphs.

### 9.3.1 BASC Cambridge

This atlas was computed from group brain parcellations generated by the BASC (Bootstrap Analysis of Stable Clusters) method, an algorithm based on $k$-means clustering to identify brain networks with coherent activity in resting-state fMRI (Bellec et al., 2010). These networks were generated from the Cambridge sample from the 1000 Functional Connectome Project (Liu et al., 2009). Based on this framework, different atlases were built depending on the number of networks defined (Urchs et al., 2015). In this study, we used four versions with 12, 20, 36 and 64 regions.

### 9.3.2 AICHA

This atlas covers the whole cerebrum and is based on resting-state fMRI data acquired in 281 individuals (Joliot et al., 2015), and also relies on $k$-means clustering. One interesting feature is that it accounts for homotopy, relying on the assumption that a region in one hemisphere has a homologue in the other hemisphere. This leads to 192 homotopic region pairs (122 gyral, 50 sulcal and 20 gray nuclei).

### 9.3.3 Brainnetome

Fan et al. (2016) introduced an atlas based on connectivity using in vivo diffusion MRI (dMRI) and fMRI data acquired in 40 subjects. It divides the human brain into 210 cortical and 36 subcortical regions, providing detailed information based on both anatomical and functional connectivity. The number of regions was computed with a cross-validation procedure to maximize consistency across subjects (Fan et al., 2014; Liu et al., 2013). All functional data, connections and brain parcellations are freely available at http://atlas.brainnetome.org.

### 9.3.4 Yeo2011

This atlas used a clustering algorithm to parcellate the cerebral cortex into networks of functionally coupled regions, employing fMRI data from 1000 subjects. The method employed assumes that each vertex of the cortex belongs to a single network (see Yeo et al., 2011). There are two versions available depending on the number of networks

considered (7 or 17). We employed the latter for the subsequent analysis as it offers a more detailed parcellation of the brain. This atlas is preinstalled in Lead-dbs toolbox (http://www.lead-dbs.org).

### 9.3.5 Harvard-Oxford

Clustering in this atlas was performed with the automatic algorithm presented in Desikan et al. (2006), which subdivides structural magnetic resonance data of the human cerebral cortex into gyral based ROI. Its validity was evaluated by computing correlation coefficients and mean distances between these results and manually identified cortical ROIs. Forty-eight cortical regions were obtained from data of 37 subjects. The resulting atlas is freely distributed with FSL (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/).

### 9.3.6 Schaefer

This atlas adds novel parcellations and a larger precision to the brain networks published in Yeo et al. (2011) by using a local gradient approach to detect abrupt transitions in functional connectivity patterns (Schaefer et al., 2018). These transitions are likely to reflect cortical areal boundaries defined by histology or visuotopic fMRI. The resulting parcellations were generated from resting-state fMRI based on 1489 participants (see original paper for further details). There are several versions of this atlas depending on the number of regions the brain is divided into (400, 600, 800 or 1000), but we selected the first one to maintain reasonable speed on computation analyses.

## 9.4 Methods

In this study, we considered four different algorithms based on linear classifiers. First, the atlas-based local averaging method (ABLA) presented in Schrouff et al. (2013a). Second, an L1-MKL version of the algorithm introduced in Rakotomamonjy et al. (2008) and implemented in the PRoNTo toolbox (Schrouff et al., 2013b). Third, a modification of the L1-MKL to use an L2-nom instead of an L1 (from now on, L2-MKL) to avoid the sparsity that L1 leads to and the subsequent decrease in detecting informative regions. Finally, we used a Searchlight approach as a contrast common reference.

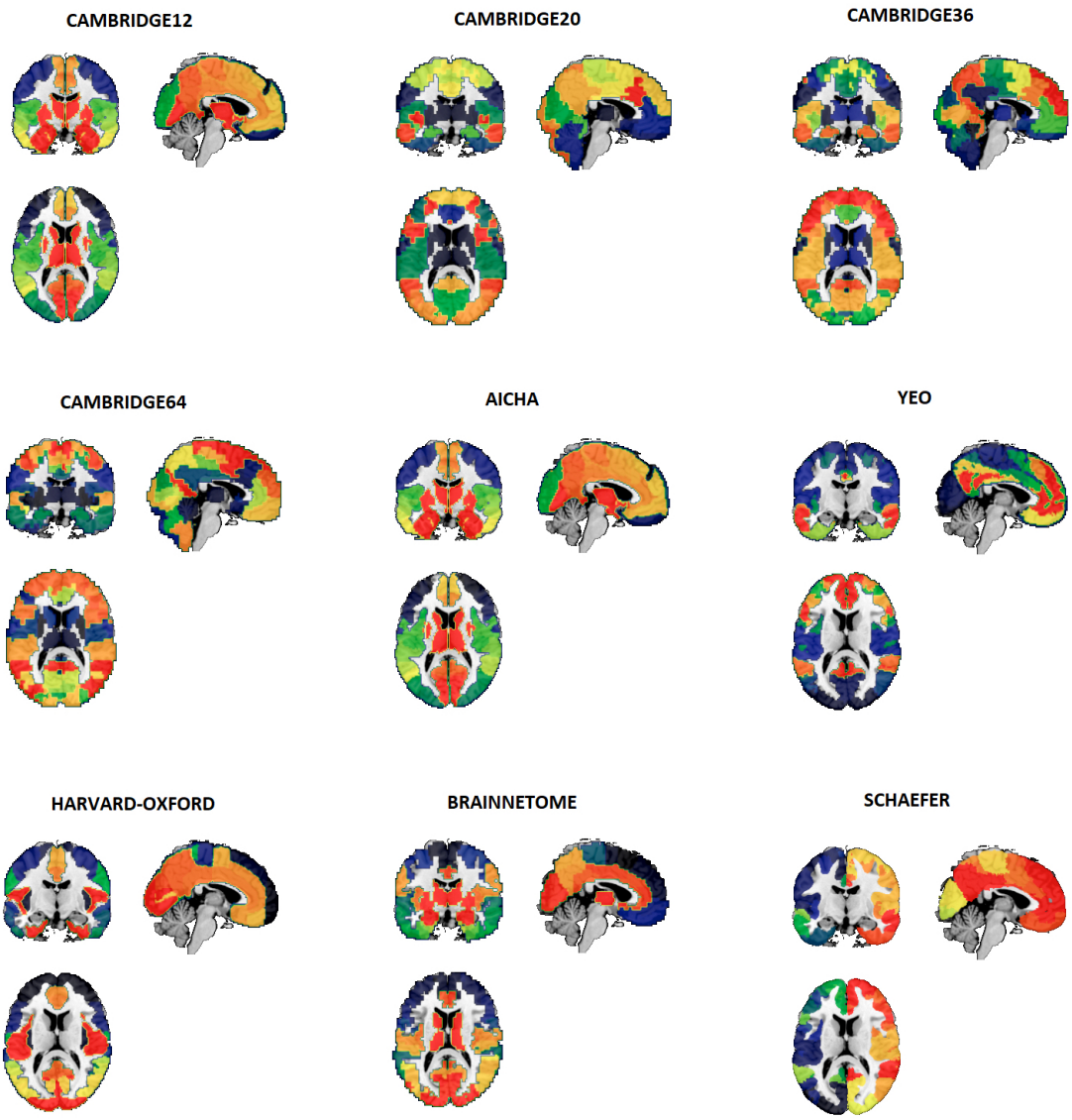


**Figure 9.3:** Brain parcellations derived from each of the 9 anatomical/functional atlases employed

### 9.4.1 Atlas-based local averaging

This method is used after performing a whole-brain analysis in which all voxels of the brain are used as input to the classification algorithm. A linear classifier leads to a weight map where each value corresponds to the contribution of each voxel to the decision function. ABLA computes a normalization of the average weight for each region of an atlas that summarizes the importance of this region in a certain classification context (see Figure 9.4). The classifier employed was the SVM, and Section 5.3 includes a detailed description of its mathematical basis.

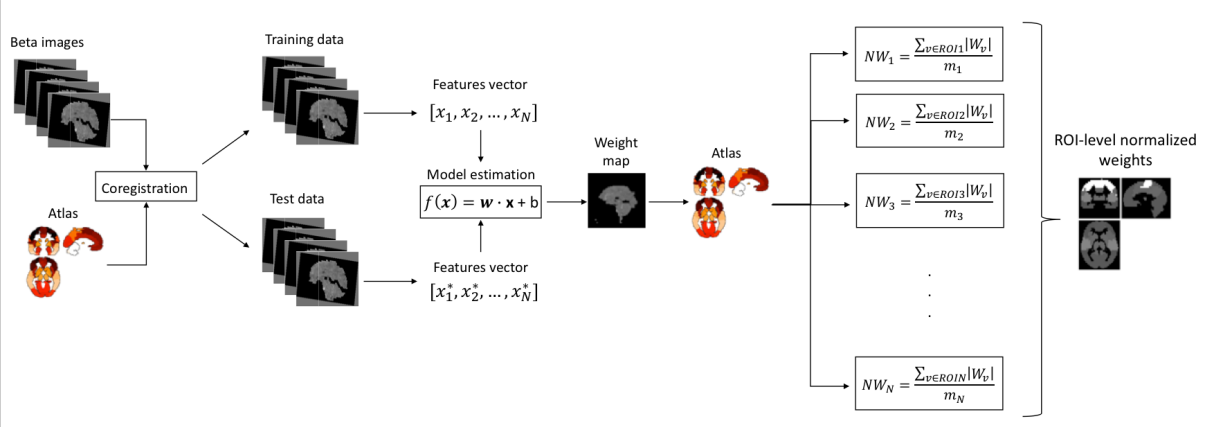


**Figure 9.4:** Schema of the ABLA method.

Once the significant model was obtained, we extracted the weight maps that guided the decision of the classifier. Then, we computed the normalized weight for each region in the atlas as the average of the absolute value of the weights contained in each region, as explained in Schrouff et al. (2013a). Equation 9.1 summarizes mathematically this computation:

$$NW_{ROI} = \frac{\sum_{v \epsilon ROI} |W_v|}{m_{ROI}} \tag{9.1}$$

with $v$ representing the index of a voxel in the weight map, $W_v$ its weight and $m_{ROI}$, the number of voxels in region ROI.

### 9.4.2 Multiple Kernel Learning

This method combines the information from the different brain regions of an atlas to build the classification model, in contrast to the use in ABLA of the corresponding brain organization *a posteriori*. Specifically, MKL combines different kernels and optimizes

their contribution to the model to obtain the highest performance. As a result, this approach offers information at two levels: regions and each voxel within them (see Figure 9.5). Mathematically, the decision function is computed as a linear combination of all these basis kernels as stated in Lanckriet et al. (2004):

$$K(\mathbf{x},\mathbf{x}') = \sum_{m=1}^{M} d_m K_m(\mathbf{x},\mathbf{x}') \qquad \text{with} \\ d_m \geq 0,\ \sum_{m=1}^{M} d_m = 1 \tag{9.2}$$

where $M$ is the total number of kernels. The decision function of the MKL problem is very similar to SVM (described in Chapter 5) but adding the sum of the different kernels from the corresponding atlas:

$$f(\mathbf{x}_i) = \sum_m \langle \mathbf{w}_m, \mathbf{x}_i \rangle + b \tag{9.3}$$

The MKL version considered in this study is based on the formulation presented in Rakotomamonjy et al. (2008), where a solution is obtained by solving the following optimization problem:

$$\text{minimize } \frac{1}{2}\sum_m \frac{1}{d_m}\|\mathbf{w}_m\|^2 + C\sum_i \xi_i \quad \text{subject to} \\ y_i(\sum_m \langle \mathbf{w}_m, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i \quad \forall_i \xi_i \geq 0 \quad \forall_i \sum_m d_m = 1,\ d_m \geq 0\ \forall_m \tag{9.4}$$

where $d_m$ is the contribution to the decision function of each region (see Rakotomamonjy et al., 2008 for a detailed explanation).

This MKL variation optimizes, in a simultaneous manner, the contribution to the decision function of every voxel within a region and the contribution of the region as a whole, in a two-level hierarchical model. In addition, the L1-norm (Tibshirani, 1996) constraint on $d_m$ enforces sparsity on some kernels, resulting in a zero-contribution of these regions: information from a region that is present in another is automatically discarded in one of them. Mathematically:

$$S = \sum_{i=1}^{n} |y_i - f(x_i)| \tag{9.5}$$

Thus, the L1-norm is based on minimizing the sum of the absolute differences between the target value ($y_i$) and the estimated values ($f(\mathbf{x}_i)$). This hierarchical model leads to two different weight maps: one that summarizes the contribution to the model of each region (region level), and another that provides the contribution of each voxel within

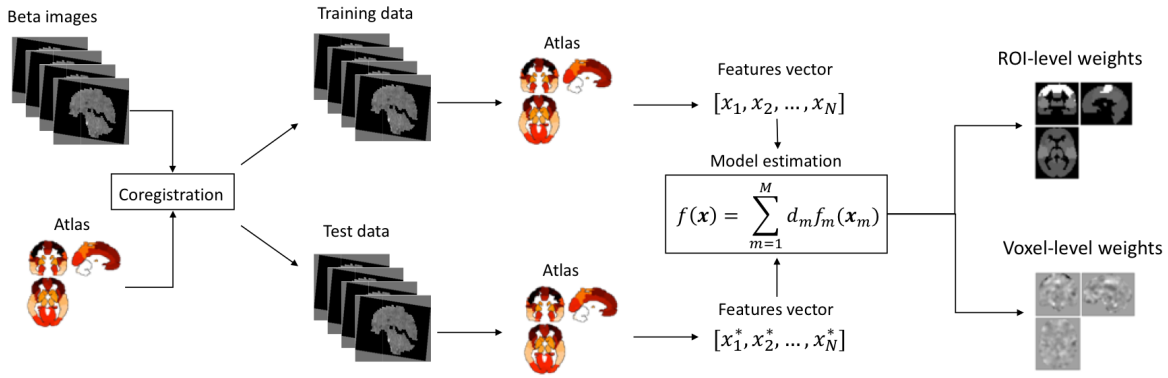


**Figure 9.5:** Schema of the MKL method.

its corresponding region (voxel-level). The sparsity that this method entails can be very interesting in classification problems (Arco et al., 2015; Khedher et al., 2017; Plant et al., 2010), but it can also potentiate the instability of the selected regions and decrease the sensitivity in identification contexts (Baldassarre et al., 2017). For this reason, we applied a different version of MKL based on L2-norm instead of L1. In this case all regions defined by the atlas are used to build the model. Mathematically:

$$S = \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 \tag{9.6}$$

Thus, the L2-norm relies on minimizing the sum of the square of the differences between the target value ($y_i$) and the estimated values ($f(\mathbf{x}_i)$).

In both versions of the MKL, we applied two preprocessing steps before classification: first, we applied a mean-centering to all kernels from each region of the atlas, a very common step in machine learning. This operation relies on subtracting the voxel-wise mean for each voxel across samples, which is computed on the training data to maintain the independence between the training and test subsets. Then, we normalized the kernel dividing each sample by its norm. Regions from which kernels are computed usually have different sizes, and larger regions would have a larger contribution to the model simply because of its larger size. This operation guarantees that all regions have an equal chance regardless of their sizes.

### 9.4.3 Searchlight

This method has been described in previous chapters, and was introduced by Kriegeskorte et al. (2006) to identify the location of the neural activity that contains information about

a given classification. In each sphere, we employed an SVM classifier with a linear kernel due to its simplicity and the high performance reported by previous studies (Misaki et al., 2010; Pereira et al., 2009). We used a 12-mm radius sphere to strike a balance between sensitivity and spatial precision: smaller sizes may not detect some informative voxels whereas larger values can boost false-positives rates (Arco et al., 2016; Chen et al., 2011).

### 9.4.4 Performance and statistical significance

We performed a nested cross-validation to train the model and optimize the hyperparameters of the classifier (cost parameter, C), both in the ABLA and in the two MKL versions: L1-MKL and L2-MKL. In these situations, the C hyperparameter range was $[10^{-5} : 10^{5}]$. Regarding Searchlight, we used a standard cost parameter of C=1 for each SVM classifier due to the high performance that it provides according to previous studies (e.g. Chanel et al., 2016; Dosenbach et al., 2010; Fan et al., 2008). The dataset comprised an fMRI experiment divided into 8 independent runs. To maintain the independence between training and testing, we used a *leave-one-run-out* cross-validation for the external loop (all methods) and the internal loop (MKL and SVM). This means that in the Searchlight approach, 7 runs were employed to train the classifier, using the remaining one for testing. In MKL and SVM, six runs were used for training, the seventh for validation and the last one for testing. We computed the balanced accuracy within participants to evaluate the performance of the model. For a binary classification, the balance accuracy is computed as the average of the accuracy obtained in the images belonging to each experimental condition individually, which increases the robustness of the performance evaluated when there is a different number of images of each class.

Statistical significance was assessed with the method proposed by Stelzer et al. (2013), with a slight difference when the procedure was applied to Searchlight or the atlas-based approaches (see Figure 9.6 and 9.7). Unlike Searchlight, Atlas-based methods perform a whole-brain classification, obtaining a single global accuracy. Moreover, weights reflect how spatial information is distributed across voxels. Hence, whereas in Searchlight the significance was computed from accuracy maps, in the other methods weight maps were used instead (see also e.g. Haufe et al., 2014; Schrouff et al., 2018). First, the labels of the images were randomly shuffled. Then, the corresponding classification method (ABLA, MKL or Searchlight) was applied. This procedure was repeated 100 times in a within-subject classification, resulting in 100 permuted accuracy/weight maps per participant (accuracy for Searchlight and weight for the rest). A map from each individual was randomly picked following a Monte Carlo resampling with replacement

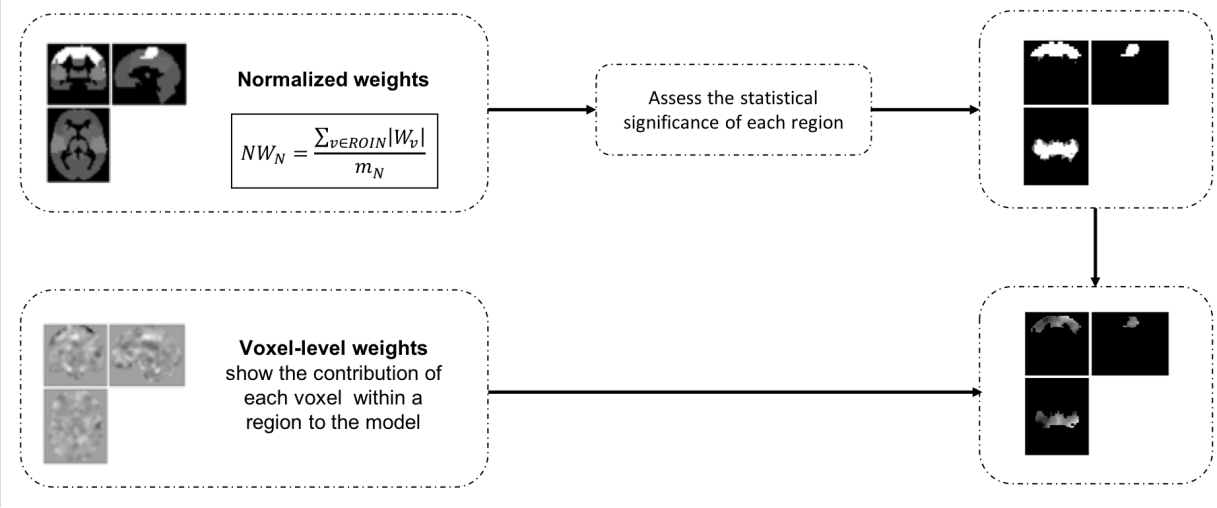


**Figure 9.6:** ABLA computes the normalized weight of each region of the atlas as a measure of their importance to the classification model. Next, statistical significance is evaluated to identify the brain regions that are really informative.

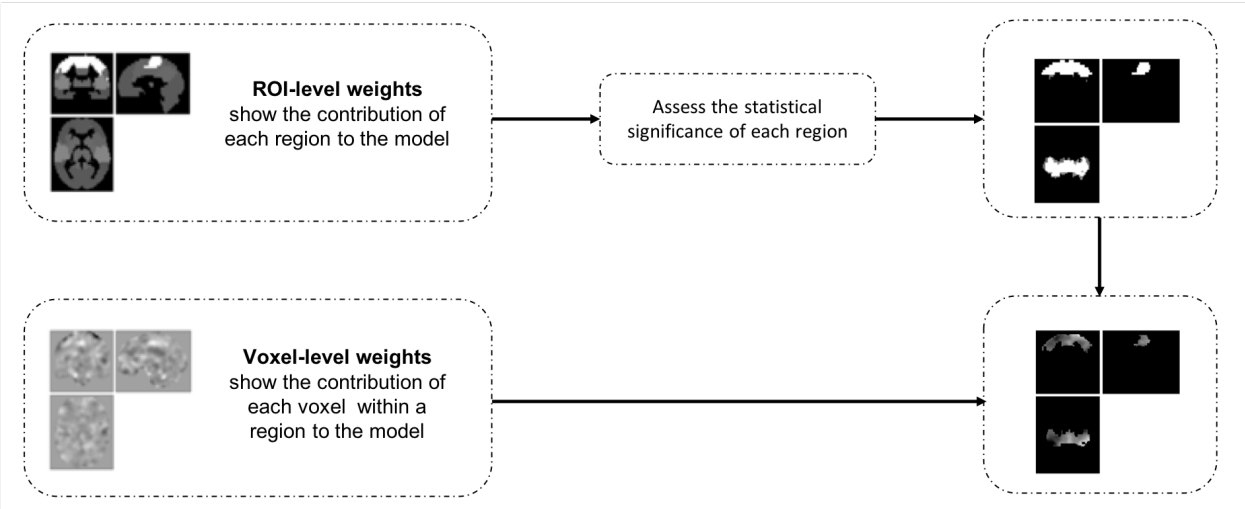


**Figure 9.7:** MKL results show the contribution to the model of each region of the atlas in addition to the weights of the voxels within each region. Then, statistical significance is evaluated at the region level.

(Forman et al., 1995), averaging the permuted maps and obtaining a permuted group map. This procedure was carried out 50000 times to build an empirical chance distribution. A voxel/region was considered significant if no more than 50 samples of the empirical distribution had a larger value than the one obtained without shuffling the labels, which corresponds to a cluster-defining primary-threshold of p=0.001 (50/50000). Once the image was thresholded, an empirical distribution of the cluster sizes of the 50000 permuted maps was built to compute the required family-wise error rate at the cluster

level. After associating a p-value to each cluster, an FWE correction was applied (p=0.05) on all-cluster p-values to correct for multiple comparisons at the cluster level.

### 9.4.5 Comparison of different atlases

Following the procedure proposed by Schrouff et al. (2018), we computed the Pearson correlation between the weight maps obtained by the different atlases. Since ABLA organizes the weights a posteriori in regions from a whole-brain classification, it is only possible to compute this correlation for L1-MKL and L2-MKL. To do so, we calculated the overlap between the significant voxels obtained by each atlas, leading to a value ranging from 0 to 1. We employed permutation tests to assess the significance of the correlation coefficients using a similar framework as described in Section 9.4.4.

## 9.5 Results

In this section, we report the results obtained by the three approaches evaluated in this study: Atlas-based local averaging (ABLA), and the two versions of Multiple Kernel Learning (L1-MKL and L2-MKL). We compared the weight maps of these three methods with the accuracies map obtained by Searchlight by computing the overlap between significant voxels. Moreover, for L1 and L2-MKL we show the stability of the selected regions across atlases by computing a correlation between their overlapping-significant weight maps, using permutation tests to assess the significance of these correlations. We did not compute this correlation for the ABLA method because weights are exactly the same for all atlases. Additionally, we include the results obtained by these methods in two classification contexts (*decision* and *valence*) that lead to large or subtle differences between the conditions contrasted, to test the generalizability of the results of the different approaches.

### 9.5.1 Influence of the classification methods

We first focus on comparing the results obtained by ABLA, L1-MKL and L2-MKL in the *decision* classification. Table 9.1 summarizes these results in terms of accuracy and overlap between the significant regions identified by each method and those obtained by Searchlight (SL). The accuracies discussed in this section correspond to the ones obtained in the maximum overlap scenario, which does not mean that these accuracies were the absolute maximum itself. We further discuss the implications of this finding

**Table 9.1:** Summary of the results obtained for the different methods and atlases in the *decision* classification.

| Methods | Atlas | Accuracy (%) | Significant voxels | SL voxels defined in atlas | Overlap with SL (%) | Regions | Significant regions |
|---|---|---|---|---|---|---|---|
| ABLA | Camb12 | 81.51 | 4704 | 4302 | 61.69 | 12 | 1 |
| L1-MKL | Camb12 | 86.2 | 4704 | 4302 | 61.69 | 12 | 1 |
| L2-MKL | Camb12 | 72.43 | 2654 | 4302 | 61.69 | 12 | 1 |
| ABLA | Camb20 | 81.51 | 3692 | 4302 | 58 | 20 | 1 |
| L1-MKL | Camb20 | 85.02 | 3692 | 4302 | 58 | 20 | 1 |
| L2-MKL | Camb20 | 65.21 | 2598 | 4302 | 58 | 20 | 1 |
| ABLA | Camb36 | 81.51 | 982 | 4302 | 21.36 | 36 | 1 |
| L1-MKL | Camb36 | 89.37 | 982 | 4302 | 21.36 | 36 | 1 |
| L2-MKL | Camb36 | 74.74 | 1000 | 4302 | 21.36 | 36 | 1 |
| ABLA | Camb64 | 81.51 | 3740 | 4302 | 63.34 | 64 | 1 |
| L1-MKL | Camb64 | 84.62 | 982 | 4302 | 21.36 | 64 | 1 |
| L2-MKL | Camb64 | 70.31 | 7613 | 4302 | 21.36 | 64 | 1 |
| ABLA | AICHA | 81.51 | 1802 | 3291 | 48 | 192 | 5 |
| L1-MKL | AICHA | 86.76 | 636 | 3291 | 19.23 | 192 | 1 |
| L2-MKL | AICHA | 69.53 | 2867 | 3291 | 60.13 | 192 | 11 |
| ABLA | Yeo2011 | 81.51 | 2731 | 3137 | 56.39 | 17 | 1 |
| L1-MKL | Yeo2011 | 87.34 | 2731 | 3137 | 56.07 | 17 | 1 |
| L2-MKL | Yeo2011 | 71.35 | 2731 | 3137 | 56.39 | 17 | 1 |
| ABLA | Harvard-Oxford | 81.51 | 4609 | 3389 | 70.58 | 48 | 2 |
| L1-MKL | Harvard-Oxford | 85.02 | 4609 | 3389 | 70.58 | 48 | 2 |
| L2-MKL | Harvard-Oxford | 70.65 | 6531 | 3389 | 77.93 | 48 | 5 |
| ABLA | Brainnetome | 81.51 | 2051 | 3057 | 47 | 246 | 9 |
| L1-MKL | Brainnetome | 78.77 | 904 | 3057 | 21.56 | 246 | 4 |
| L2-MKL | Brainnetome | 66.53 | 1129 | 3057 | 28.39 | 246 | 5 |
| ABLA | Schaefer | 81.51 | 1558 | 3137 | 42.9 | 400 | 23 |
| L1-MKL | Schaefer | 77.84 | 465 | 3137 | 14.5 | 400 | 6 |
| L2-MKL | Schaefer | 71.61 | 1926 | 3137 | 51.35 | 400 | 28 |

in Section 9.6. The first approach, ABLA, yielded a maximum overlap of 70.58%, and a corresponding accuracy of 81.51%. L1-MKL led to the same maximum overlap value, 70.58%, but a higher corresponding accuracy compared to ABLA: 85.02%. On the other hand, L2-MKL obtained a maximum overlap of 77.93%, whereas the accuracy was 70.65% after employing this approach.

We will now describe the results obtained in the *valence* classification. In this context, the ABLA method obtained a maximum overlap of 41.49%, with a corresponding accuracy of 51.77%. We assessed the significance of the accuracy by employing the non-parametric method described in section 9.4.4. This last value is considerably lower than the one obtained in the *decision* classification and it likely reflects the subtle differences in the neural activity associated with the valence of a word. We observed that after applying the L1-MKL method, only one of the nine atlases employed led to a significant region that overlapped with Searchlight. However, the small size of this region (only 0.14% of the significant voxels obtained by L1-MKL overlapped with Searchlight results) highlights the inadequacy of this method to identify significant regions in a context like the *valence*

one. With reference to L2-MKL, the maximum overlap slightly increased (3.81%), with a corresponding accuracy of 49.14%. Since the value of the overlap is considerably small as well, conclusions derived for L1-MKL results can be also applied to L2-MKL. All the results obtained by the three classification methods in the *valence* classification are summarized in Table 9.2.

**Table 9.2:** Summary of the results obtained for the different methods and atlases in the *valence* classification.

| Methods | Atlas | Accuracy (%) | Significant voxels | SL voxels defined in atlas | Overlap with SL (%) | Regions | Significant regions |
|---|---|---|---|---|---|---|---|
| ABLA | Camb12 | 51.77 | 2095 | 911 | 41.49 | 12 | 1 |
| L1-MKL | Camb12 | 48.44 | 0 | 911 | 0 | 12 | 0 |
| L2-MKL | Camb12 | 50.71 | 0 | 911 | 0 | 12 | 0 |
| ABLA | Camb20 | 51.77 | 0 | 911 | 0 | 20 | 0 |
| L1-MKL | Camb20 | 48.18 | 0 | 911 | 0 | 20 | 0 |
| L2-MKL | Camb20 | 50.74 | 0 | 911 | 0 | 20 | 0 |
| ABLA | Camb36 | 51.77 | 0 | 911 | 0 | 36 | 0 |
| L1-MKL | Camb36 | 49.74 | 0 | 911 | 0 | 36 | 0 |
| L2-MKL | Camb36 | 49.22 | 406 | 911 | 0 | 36 | 1 |
| ABLA | Camb64 | 51.77 | 341 | 911 | 7.14 | 64 | 1 |
| L1-MKL | Camb64 | 46.88 | 549 | 911 | 0 | 64 | 1 |
| L2-MKL | Camb64 | 51.75 | 0 | 911 | 0 | 64 | 0 |
| ABLA | AICHA | 51.77 | 663 | 729 | 20.58 | 192 | 5 |
| L1-MKL | AICHA | 47.1 | 780 | 729 | 0 | 192 | 5 |
| L2-MKL | AICHA | 52.83 | 35 | 729 | 0 | 192 | 1 |
| ABLA | Yeo2011 | 51.77 | 0 | 709 | 0 | 17 | 0 |
| L1-MKL | Yeo2011 | 46.15 | 0 | 709 | 0 | 17 | 0 |
| L2-MKL | Yeo2011 | 50.97 | 1010 | 709 | 1.7 | 17 | 1 |
| ABLA | Harvard-Oxford | 51.77 | 439 | 715 | 21.4 | 48 | 1 |
| L1-MKL | Harvard-Oxford | 47.18 | 145 | 715 | 0 | 48 | 1 |
| L2-MKL | Harvard-Oxford | 48.33 | 0 | 715 | 0 | 48 | 0 |
| ABLA | Brainnetome | 51.77 | 438 | 738 | 7.99 | 246 | 3 |
| L1-MKL | Brainnetome | 43.34 | 349 | 738 | 0 | 246 | 3 |
| L2-MKL | Brainnetome | 51.19 | 137 | 738 | 0 | 246 | 1 |
| ABLA | Schaefer | 51.77 | 61 | 708 | 4.8 | 400 | 1 |
| L1-MKL | Schaefer | 46.24 | 123 | 708 | 0.14 | 400 | 2 |
| L2-MKL | Schaefer | 49.14 | 302 | 708 | 3.81 | 400 | 6 |

### 9.5.2 Influence of the atlases

Previous section has depicted the influence of the different classification methods in the results. On the other hand, this section will describe the effect of the different brain parcellations in the identification of brain informative regions for the three approaches used. In the first context (the *decision* classification), ABLA marked as informative similar regions regardless of the atlas employed. This similarity can be shown in Figure 9.8, where the spatial distribution of the significant regions across atlases is reported. Despite most of atlases led to the same results, we found a variability in terms of overlap

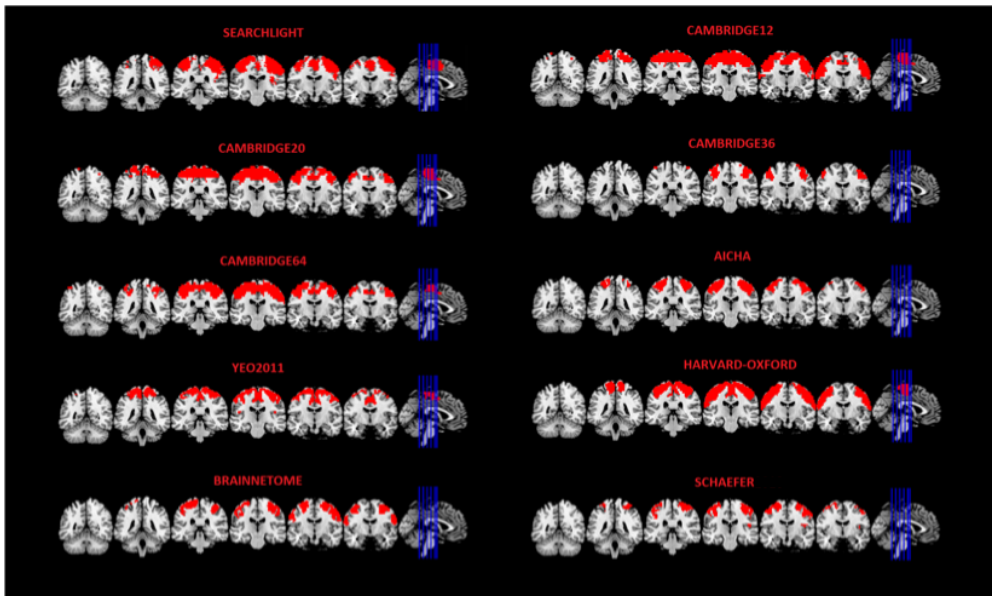


**Figure 9.8:** Significant voxels obtained by the Searchlight approach and the ABLA method for all the atlases used in the *decision* classification.

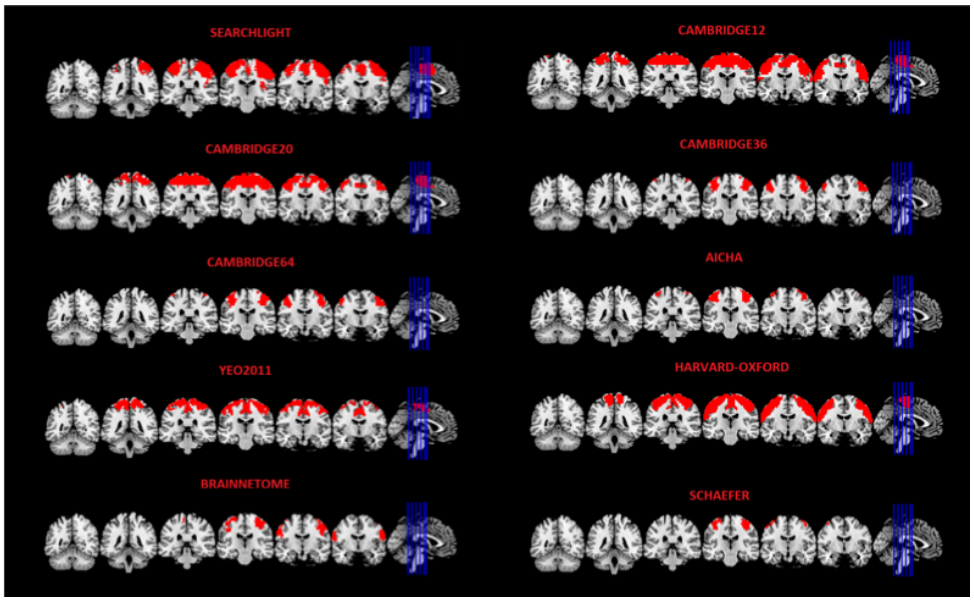


**Figure 9.9:** Significant voxels obtained by the Searchlight approach and the L1-MKL method for all the atlases used in the *decision* classification.

between the different atlases. Specifically, the largest overlap score with Searchlight was obtained by the Harvard-Oxford atlas (70.58%), whereas the minimum value was derived from the Camb36 division of the brain (21.36%). Results for the rest of the atlases are summarized in Table 9.1.

When applying L1-MKL, the largest overlap value was obtained by the same atlas

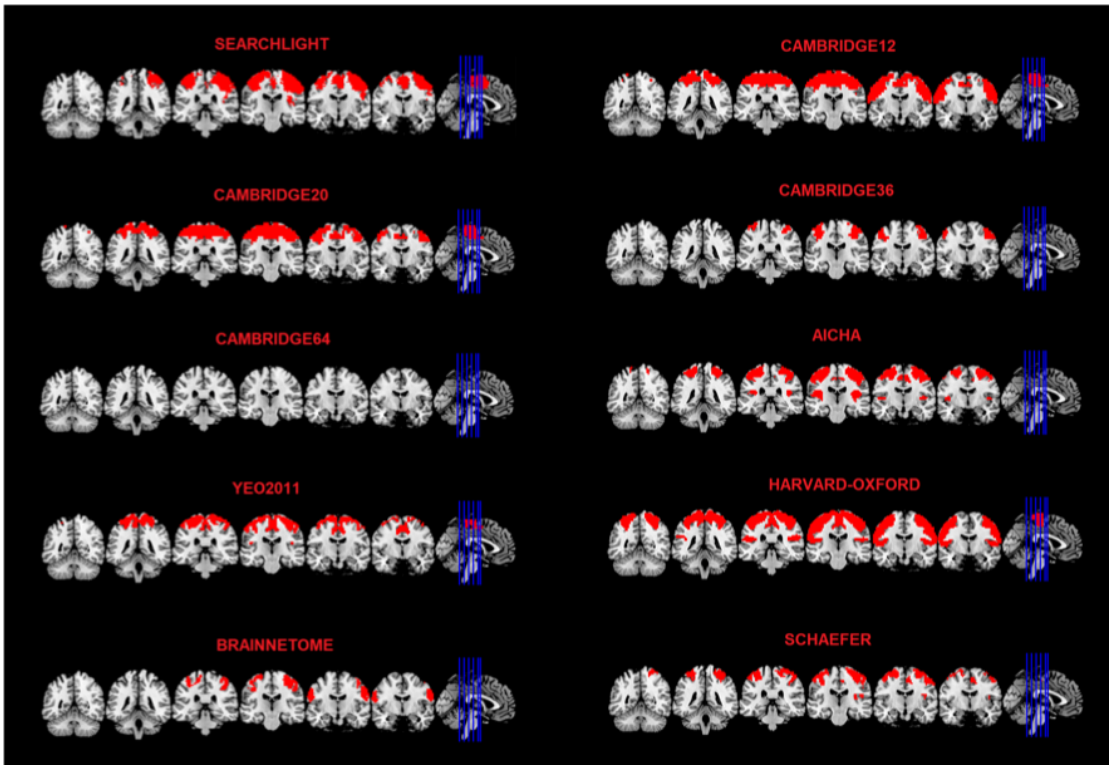


**Figure 9.10:** Significant voxels obtained by the Searchlight approach and the L2-MKL method for all the atlases used in the *decision* classification.

as with ABLA: the Harvard-Oxford, with a 70.58%. However, the minimum overlap corresponded to Schaefer atlas (14.5%). It seems that this method is more affected than ABLA by the different brain parcellations. As Figure 9.9 shows, the distribution of the significant regions is similar for all atlases, but in this case, sensitivity is lower than ABLA for most atlases. Quantitative results for each brain parcellation are summarized in Table 9.1. For the last classification method used, L2-MKL, the parcellation derived from the Camb64 atlas yielded the largest accuracy and minimum overlap score (74.74% and 21.36%, respectively). This finding remarks that maximum overlap and accuracy is not usually simultaneously obtained. On the other hand, Harvard-Oxford led to a good accuracy value and the largest overlap (70.65% and 77.93%, respectively). Figure 9.10 shows how informative regions vary for the different atlases, whereas Table 9.1 includes numerical results about accuracy, overlap and number of significant voxels for all brain parcellations.

Regarding the *valence* context, results were highly affected by the atlas used. We found a large consistency in the significant regions obtained by ABLA and Searchlight when the Cambridge12 atlas was employed. However, this atlas was not the only one that led to a good performance. Specifically, the brain parcellations provided by AICHA

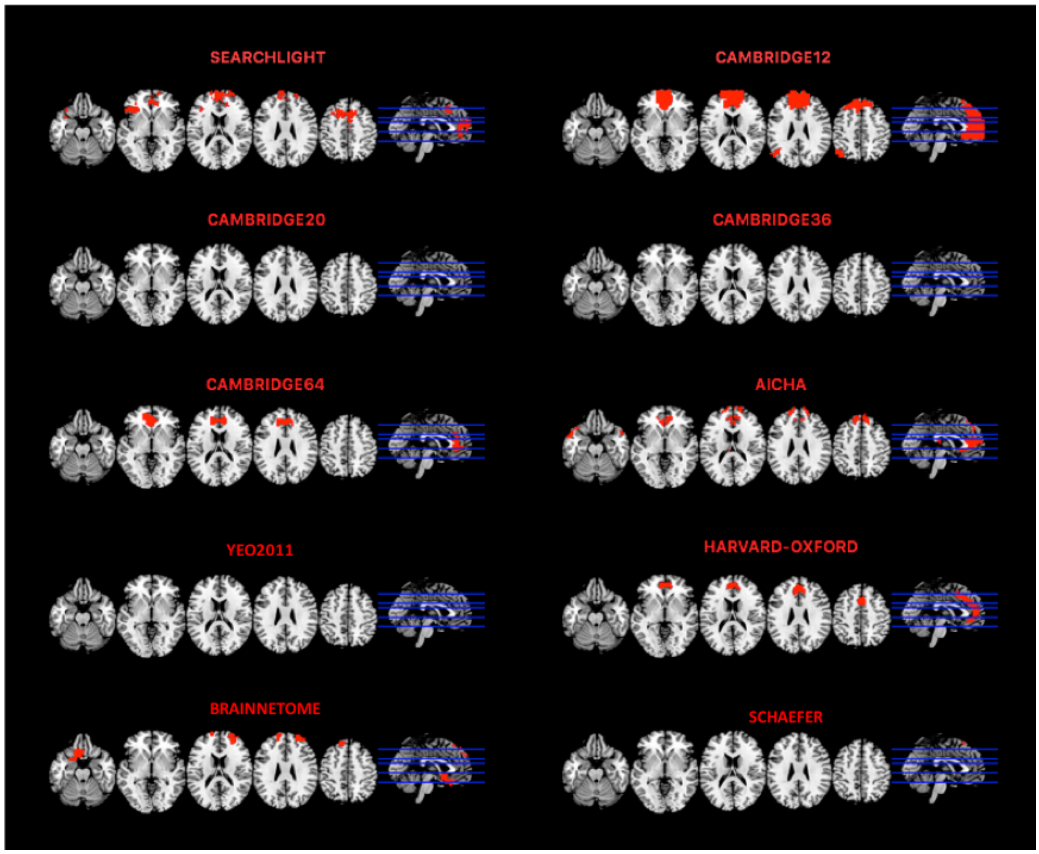


**Figure 9.11:** Significant voxels obtained by Searchlight and ABLA method for all the atlases used in the *valence* classification.

and Harvard-Oxford also identified informative regions similar to those obtained by Searchlight. Most importantly, these regions contain areas that have been reported by previous research, (e.g. ventromedial prefrontal cortex, Lindquist and Mejia, 2015), which supports the reliability of the results. Figure 9.11 includes the significant results associated with the nine different atlases and the ABLA approach in this *valence* classification, whereas Table 9.2 provides additional information to these results.

Unlike ABLA, the two methods based on MKL hardly detected reliable information for all the atlases employed. With reference to L1-MKL, each brain parcellation led to a completely different distribution of the informative voxels. However, none of the nine atlases that we employed yielded an accuracy that surpassed the chance level, so that the subsequent model did not provide useful information about where the information regions were located. Results were very similar when L2-MKL was employed. In this case, models derived from some atlases surpassed the chance level, but they were not able to identify the regions that truly contained information. Figures 9.12 and 9.13 show the informative voxels that both L1-MKL and L2-MKL yielded, respectively, whereas 9.2

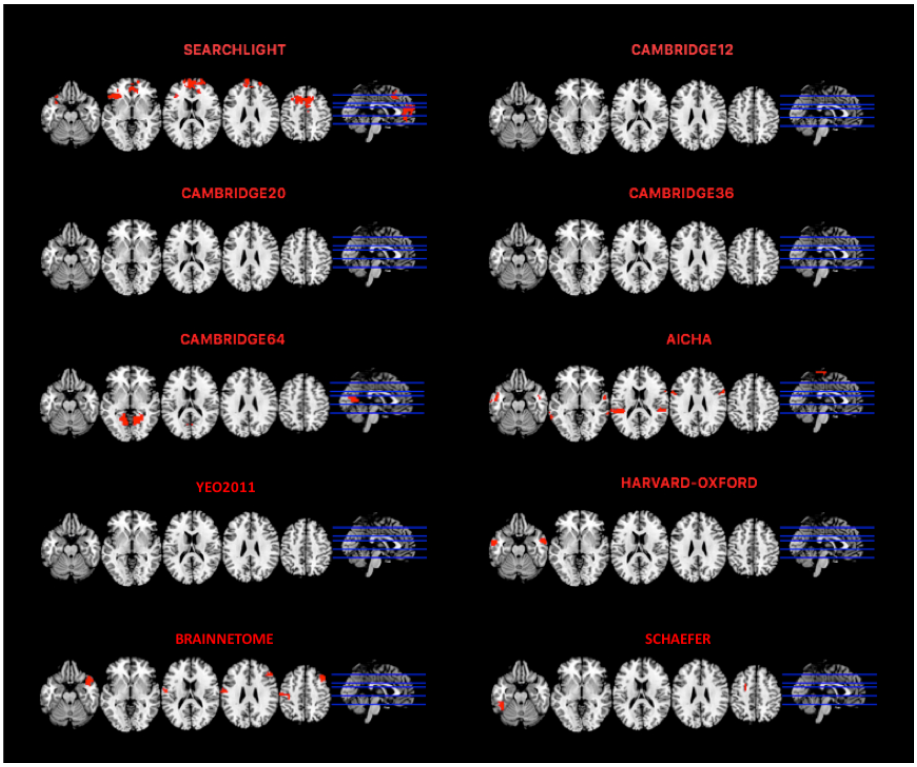


**Figure 9.12:** Significant voxels obtained by Searchlight and L1-MKL methods for all the atlases used in the *valence* classification.

summarizes the results obtained by these two approaches.

### 9.5.3 Stability of the weights across atlases

We compared the weight maps across the different atlases for L1-MKL and L2-MKL, in the two classification contexts. Due to the nature of the ABLA method, weights are the same for all the atlases since different brain parcellations are used once the model is built. For this reason, we only computed the stability of the subsequent weights for the two approaches based on MKL. In the *decision* classification, the correlation values obtained by the first 6 atlases (Camb12, Camb20, Camb36, Camb64, AICHA and Yeo2011) ranging from 0.882 to 0.974 when the L1-MKL was used. The weight maps derived from the Harvard-Oxford atlas also yielded a large similarity to these 6 atlases, but this correlation decreased when the Brainnetome atlas was employed. By contrast, the Schaefer atlas led to very different weights compared to any of the other atlases. These results suggest that, for this contrast, the decision function derived from L1-MKL is based on the same voxels. Moreover, the contribution of these voxels to the classifier decision is stable for all brain parcellations proposed by each atlas. Table 9.3

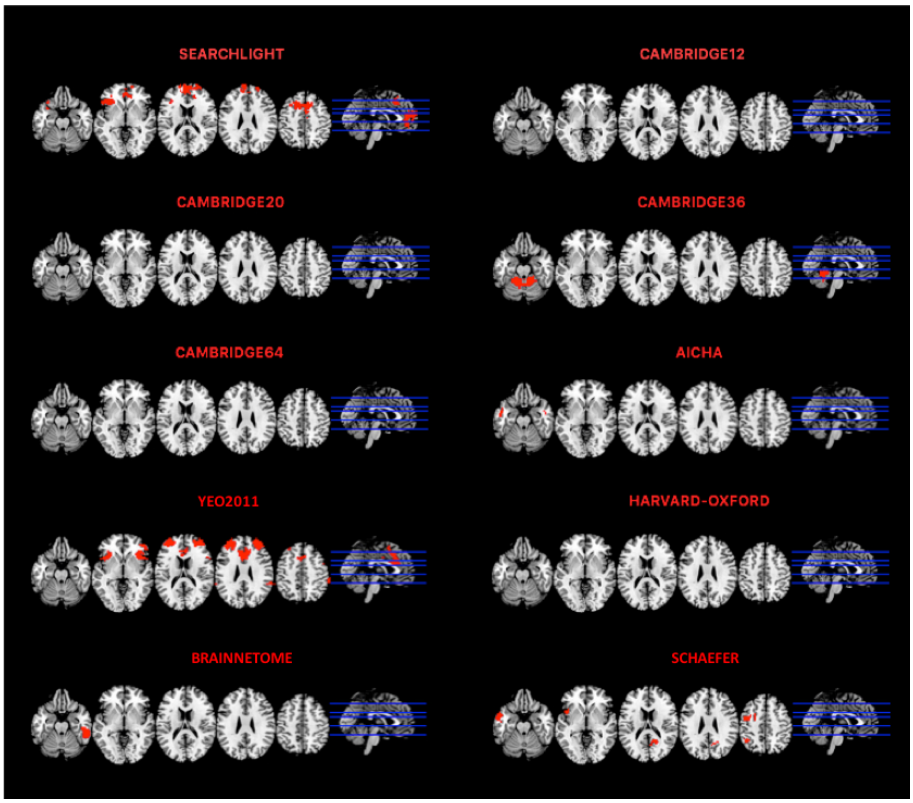


**Figure 9.13:** Significant voxels obtained by Searchlight and L2-MKL method for all the atlases used in the *valence* classification.

**Table 9.3:** Correlation between the different atlases after applying the L1-MKL method in the *decision* classification.

| L1-Multi Kernel Learning | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Atlas | Camb12 | Camb20 | Camb36 | Camb64 | AICHA | Yeo2011 | Harvard-Oxford | Brainnetome | Schaefer |
| Camb12 | 1 | 0.974 | 0.906 | 0.936 | 0.889 | 0.937 | 0.539 | 0.383 | 0.144 |
| Camb20 | 0.974 | 1 | 0.908 | 0.951 | 0.934 | 0.947 | 0.564 | 0.552 | 0.125 |
| Camb36 | 0.906 | 0.908 | 1 | 0.975 | 0.933 | 0.911 | 0.497 | 0.568 | 0.1 |
| Camb64 | 0.936 | 0.951 | 0.975 | 1 | 0.963 | 0.933 | 0.542 | 0.573 | 0.081 |
| AICHA | 0.889 | 0.934 | 0.933 | 0.963 | 1 | 0.882 | 0.61 | 0.549 | 0.088 |
| Yeo2011 | 0.937 | 0.947 | 0.911 | 0.933 | 0.882 | 1 | 0.566 | 0.528 | 0.193 |
| Harvard-Oxford | 0.539 | 0.564 | 0.497 | 0.542 | 0.61 | 0.566 | 1 | 0.172 | 0.051 |
| Brainnetome | 0.383 | 0.552 | 0.568 | 0.573 | 0.549 | 0.528 | 0.172 | 1 | 0.109 |
| Schaefer | 0.144 | 0.125 | 0.1 | 0.081 | 0.088 | 0.193 | 0.051 | 0.109 | 1 |

summarizes the correlation values obtained between each pair of atlases, whereas Figure 9.14 shows a visual representation of this large similarity. Each color represents the similarity between the weights associated with one atlas and those derived by the rest of the atlases. According to the colorbar used, the more yellow, the larger the correlation (i.e. the more similar are the weights). On the other hand, blue values indicate that weights derived from two atlases are considerably different. The values of the diagonal are set to 1 (yellow) because the correlation of a vector of weights with itself is maximum.

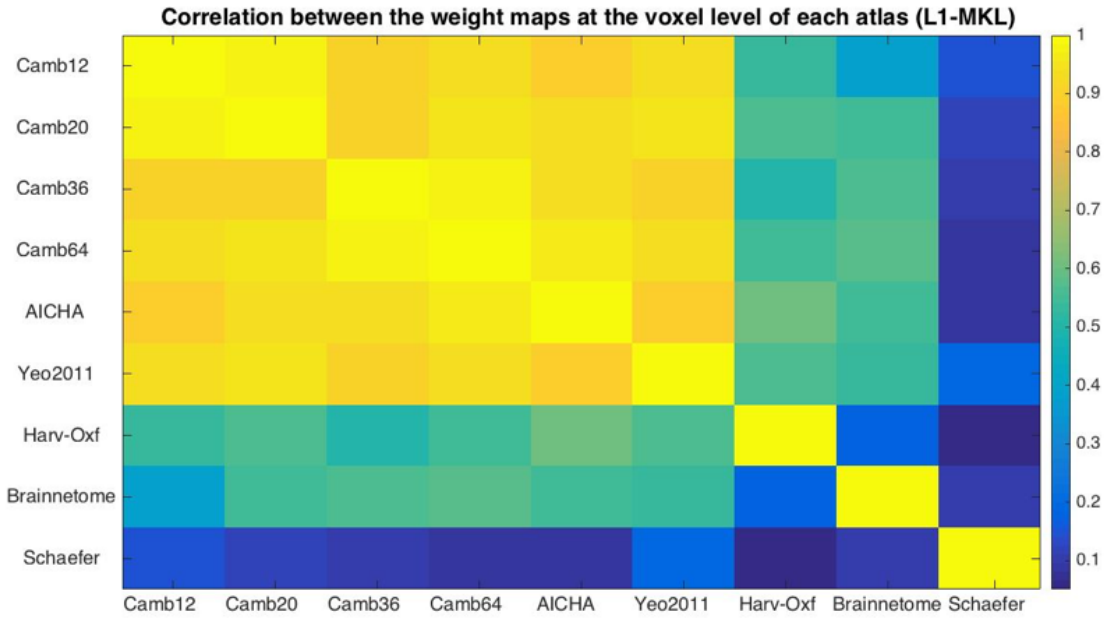


**Figure 9.14:** Correlation between the weight maps of each atlas at the voxel level in the *decision* classification for the L1-MKL method.

**Table 9.4:** Correlation between the significant weight maps across the different atlases after applying the L2-MKL method in the *decision* classification.

| L2-Multi Kernel Learning | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Atlas | Camb12 | Camb20 | Camb36 | Camb64 | AICHA | Yeo2011 | Harvard-Oxford | Brainnetome | Schaefer |
| Camb12 | 1 | 0.927 | 0.94 | 0.926 | 0.845 | 0.768 | 0.837 | 0.72 | 0.829 |
| Camb20 | 0.927 | 1 | 0.958 | 0.973 | 0.776 | 0.71 | 0.795 | 0.64 | 0.762 |
| Camb36 | 0.94 | 0.958 | 1 | 0.981 | 0.795 | 0.721 | 0.789 | 0.663 | 0.776 |
| Camb64 | 0.926 | 0.973 | 0.981 | 1 | 0.805 | 0.738 | 0.798 | 0.67 | 0.783 |
| AICHA | 0.845 | 0.776 | 0.795 | 0.805 | 1 | 0.948 | 0.94 | 0.904 | 0.973 |
| Yeo2011 | 0.768 | 0.71 | 0.721 | 0.738 | 0.948 | 1 | 0.897 | 0.857 | 0.945 |
| Harvard-Oxford | 0.837 | 0.795 | 0.789 | 0.798 | 0.94 | 0.897 | 1 | 0.849 | 0.945 |
| Brainnetome | 0.72 | 0.64 | 0.663 | 0.67 | 0.904 | 0.857 | 0.849 | 1 | 0.889 |
| Schaefer | 0.829 | 0.762 | 0.776 | 0.783 | 0.973 | 0.945 | 0.945 | 0.889 | 1 |

Regarding L2-MKL, it yielded very similar weight maps regardless of the atlases used. It is worth noting the large correlation between each pair of atlases (see Table 9.4), even with the Schaefer atlas that yielded very different weights when L1-MKL was applied. We can see in Figure 9.15 how similar the different weights are: only maps provided by Yeo2011 and Brainnetome are slightly less similar to those obtained by the four Cambridge atlases, whereas both show a large correlation with the others. The rest of the atlases present correlation values close to 1. Different atlases led to very similar results, highlighting the robustness of L2-MKL in the identification of informative regions. Moreover, this finding shows the low influence that the brain parcellation has in the results, which validates the use of these atlas-based methods even without a prior

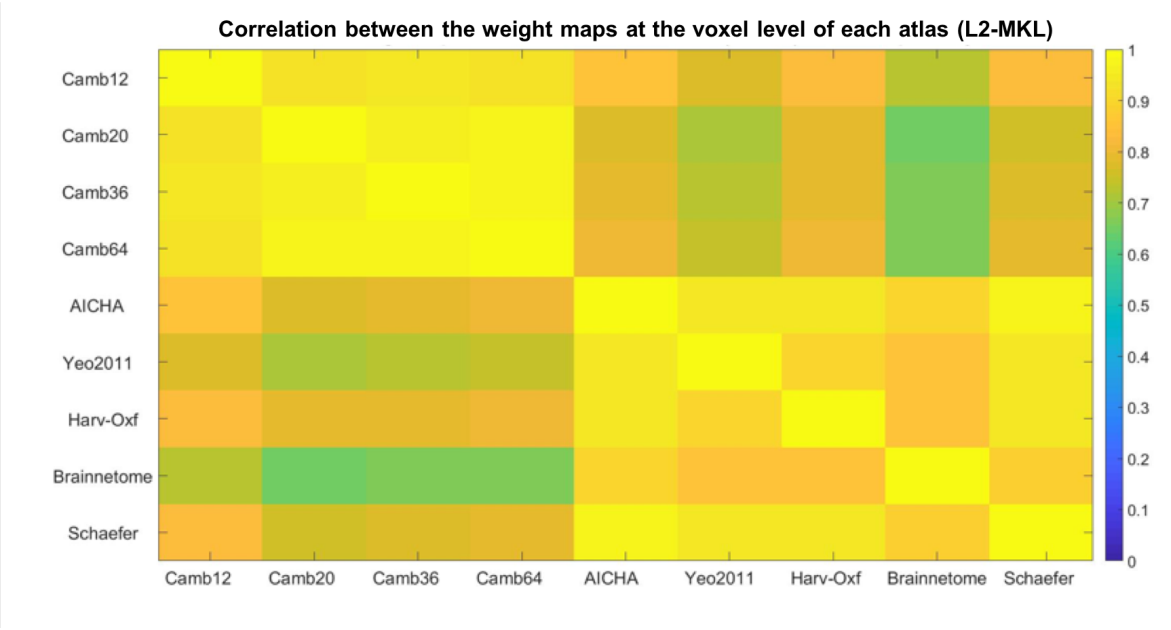


**Figure 9.15:** Correlation between the weight maps of each atlas at the voxel level in the *decision* classification for the L2-MKL method.

**Table 9.5:** Correlation between the significant weight maps across the different atlases after applying the L1-MKL method in the *valence* classification.

| L1-Multi Kernel Learning | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Atlas | Camb12 | Camb20 | Camb36 | Camb64 | AICHA | Yeo2011 | Harvard-Oxford | Brainnetome | Schaefer |
| Camb12 | 1 | - | - | - | - | - | - | - | - |
| Camb20 | - | 1 | - | - | - | - | - | - | - |
| Camb36 | - | - | 1 | - | - | - | - | - | - |
| Camb64 | - | - | - | 1 | - | - | - | - | - |
| AICHA | - | - | - | - | 1 | - | 0.217 | 0.428 | - |
| Yeo2011 | - | - | - | - | - | 1 | - | - | - |
| Harvard-Oxford | - | - | - | - | 0.217 | - | 1 | - | - |
| Brainnetome | - | - | - | - | 0.428 | - | - | 1 | - |
| Schaefer | - | - | - | - | - | - | - | - | 1 |

hypothesis about the brain organization in a specific process.

For the *valence* classification, the localization of the informative regions was so variable that results derived from most atlases did not overlap the ones obtained by the others. For this reason, we could compute the correlation between AICHA, Harvard-Oxford and BN for L1-MKL because they were the only atlases that shared some informative voxels, yielding a maximum overlap of 0.428 (see Table 9.5). Results obtained by L2-MKL also showed a reduced overlap between the weight maps and we could only correlate the significant results of AICHA, Yeo2011 and Schaefer atlases. In this case, the maximum correlation was obtained by Yeo2011 and Schaefer, yielding a value of 0.99 (see Table 9.6). Nevertheless, this value was obtained from a small region since significant results provided by these two atlases were considerably different. We further

**Table 9.6:** Correlation between the significant weight maps across the different atlases after applying the L2-MKL method in the *valence* classification.

| L2-Multi Kernel Learning | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Atlas | Camb12 | Camb20 | Camb36 | Camb64 | AICHA | Yeo2011 | Harvard-Oxford | Brainnetome | Schaefer |
| Camb12 | 1 | - | - | - | - | - | - | - | - |
| Camb20 | - | 1 | - | - | - | - | - | - | - |
| Camb36 | - | - | 1 | - | - | - | - | - | - |
| Camb64 | - | - | - | 1 | - | - | - | - | - |
| AICHA | - | - | - | - | 1 | - | - | - | 0.975 |
| Yeo2011 | - | - | - | - | - | 1 | - | - | 0.99 |
| Harvard-Oxford | - | - | - | - | - | - | 1 | - | - |
| Brainnetome | - | - | - | - | - | - | - | 1 | - |
| Schaefer | - | - | - | - | 0.975 | 0.99 | - | - | 1 |

discuss these results in Section 9.6.

### 9.5.4 Directionality of the weights

In the *decision* classification, we evaluated not only the localization of the informative weights but their sign. Due to the nature of the contrast, it was expected that weights were organized according to their sign in a specific hemisphere for each group of subjects. Figure 9.16 shows the distribution of the significant voxels depending on the sign of their weights for the ABLA method, comparing them with results obtained by univariate analysis. It is remarkable that participants who accepted the offer with the right hand and rejected it with the left hand (odd group) show a cluster of positive weights in the left hemisphere and a cluster of negative weights in the right hemisphere. On the other hand, these results are shifted when results from even participants were evaluated: the right hemisphere contains weights associated with accepting an offer, whereas the left hemisphere shows the negative weights. These results are consistent with those obtained by univariate analysis. Specifically, results from the odd group are quite similar from the Acceptance>Reject contrast, and results from the even group have a lot of similarities with the Acceptance<Reject contrast.

Results for both MKL methods are very similar to the ones obtained by ABLA. Regardless of the differences in the spatial localization of the information (already commented in previous sections), the weights follow the same distribution as ABLA. For those participants that accepted the offer by employing their right hand, the weights in the right hemisphere are positive, whereas the same hemisphere in the group of people that used their left hand to accept the offer contains negative weights. The signs of the significant voxels for the L1-MKL and L2-MKL methods are exhibited in Figures 9.17 and 9.18, respectively.

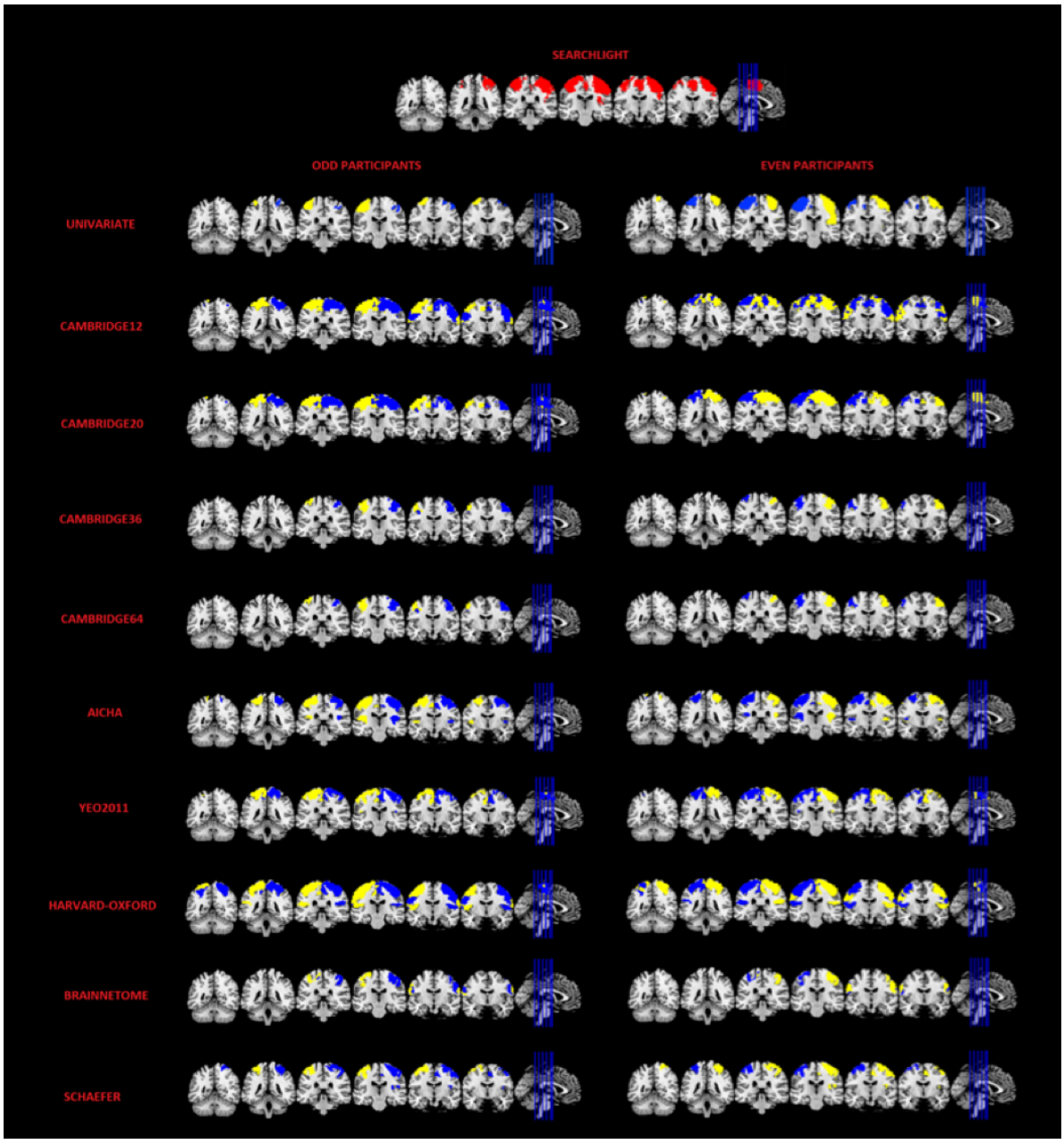


**Figure 9.16:** Summary of the results obtained for the *decision* classification by Searchlight, ABLA and univariate approaches. The latter two show large differences between the two groups considered (odd/even participants). Searchlight only provides information about the significance of each voxel itself, so that no separation between groups was considered.

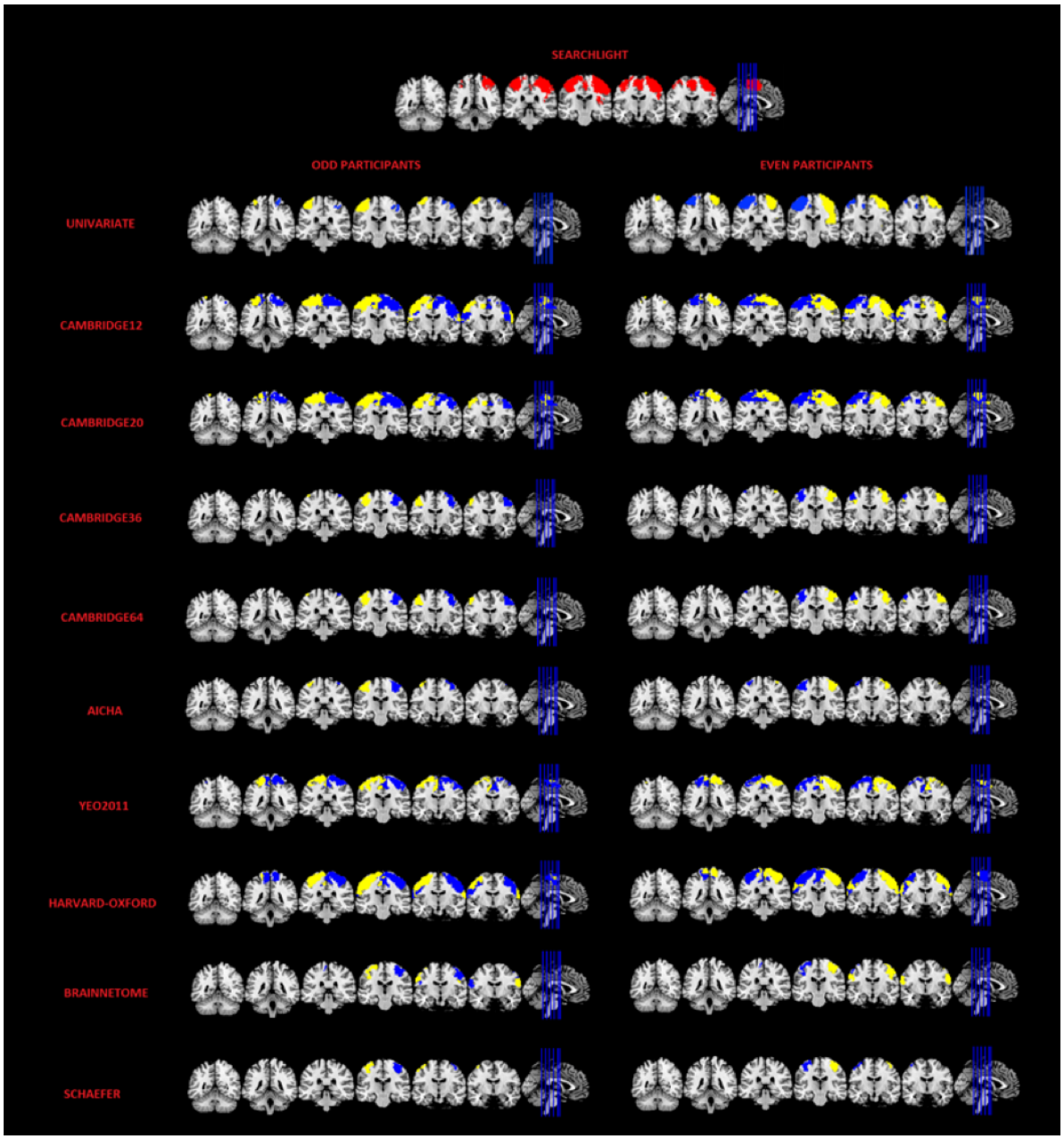


**Figure 9.17:** Summary of the results obtained for the decision classification by Searchlight, L1-MKL and univariate approaches. The latter two show large differences between the two groups considered (odd/even participants). Searchlight only provides information about the significance of each voxel itself, so that no separation between groups was considered.

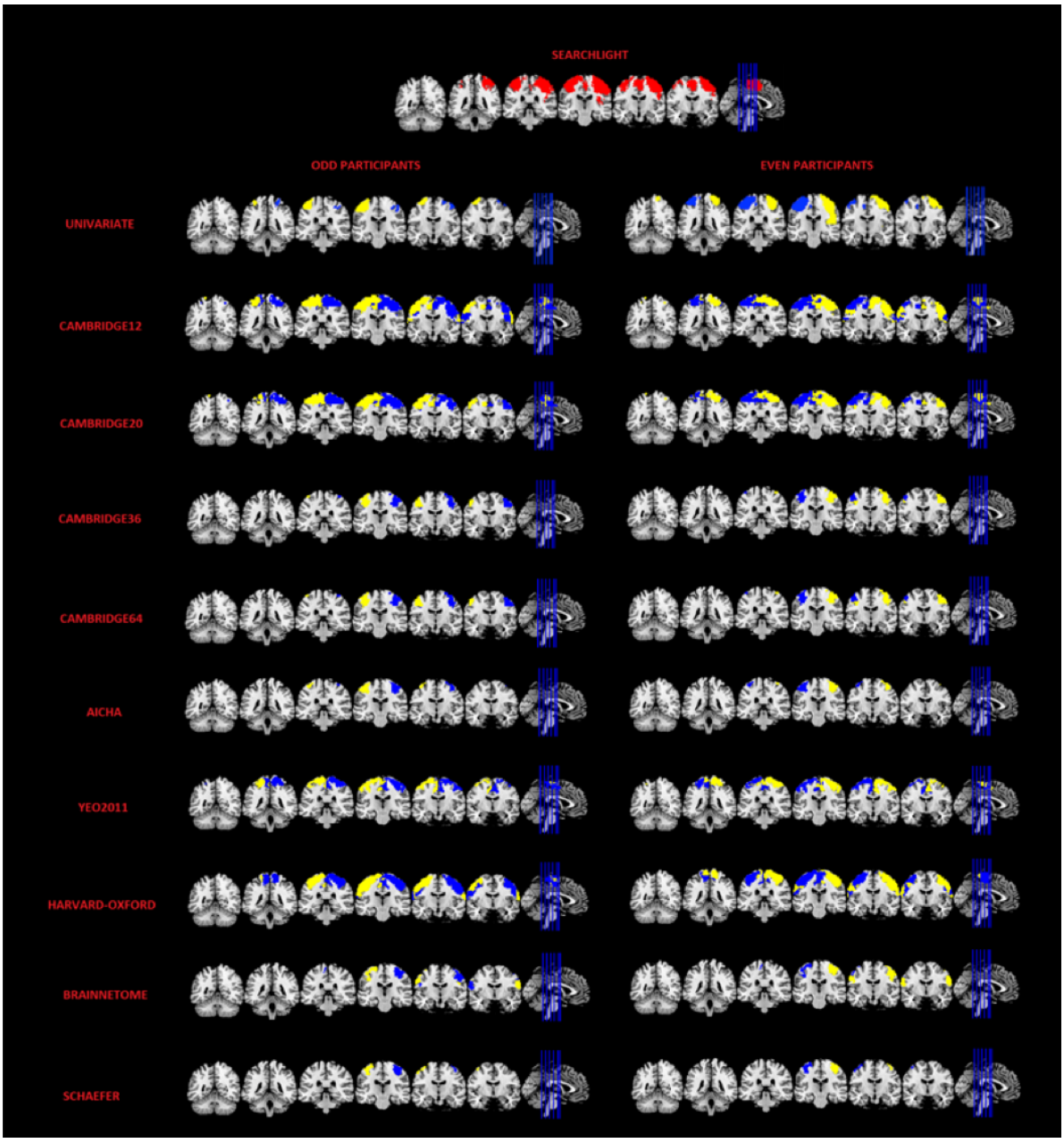


**Figure 9.18:** Summary of the results obtained for the decision classification by Searchlight, L2-MKL and univariate approaches. The latter two show large differences between the two groups considered (odd/even participants). Searchlight only provides information about the significance of each voxel itself, so that no separation between groups was considered.

## 9.6 Discussion

In this chapter, we evaluated atlas-based methods, alternative to Searchlight, to localize the informative regions involved in cognitive functions. We extracted the weight maps from three atlas-based classification approaches (ABLA, L1-MKL and L2-MKL) and evaluated the statistical significance of each region. We used these methods in two different contexts. In the first one, where the two classes generated large differences in the observed pattern, L2-regularization resulted the best option for identification purposes. Moreover, atlas-based approaches showed a large stability in the informative regions found regardless of the atlas employed, which highlights the adequacy of these methods. In contrast, when the differences in the activity associated with each class were much subtler, only the ABLA approach showed certain stability in the informative regions across the atlases. However, both L1-MKL and L2-MKL were highly affected by the specific brain organization reflected in the atlases.

### 9.6.1 Influence of the classification methods

We have found that maximum accuracy and overlap do not usually concur, especially when detecting subtle differences in the observed pattern. In the *decision* classification, we identified differences across the methods in terms of overlap and accuracy. L1-MKL usually obtained a larger accuracy than ABLA and L2-MKL for the different atlases, but a lower spatial overlap with Searchlight results. We can separate the different approaches in two groups: on the one hand, ABLA and L2-MKL; on the other, L1-MKL. The reason for this difference is the regularization used by each method: while ABLA and L2-MKL use an L2-norm regularization, L1-MKL employs an L1. L1-norm provides sparse solutions since it only selects a subset of regions that contain predictive information, while the rest are automatically driven to zero. This can be helpful from the classification standpoint since it leads to larger accuracies: when a lower number of features are considered, the dimensionality of the data is reduced, which facilitates finding the optimal solution to the classification problem. However, our results show that the model with the largest overlap is not usually the most accurate. This is consistent with previous studies, e.g. the extreme case reported by Sona et al. (2007). They proposed a model for decoding subjective perception of participants from their neural data while viewing movie segments. They found a framework that yielded a large performance, but the regions that guided the classifier were partially contained in the ventricles and other regions with large physiological noise. This means that their algorithm performed

consistently well in the classification task, but it did not provide any useful information for a better understanding of the human brain. Our results support the need of clearly separating the use of multivariate decoding for prediction and for identification (Hebart and Baker, 2017) in addition to highlight the importance of selecting the methods that best fit the desired aim.

In the *valence* classification, we also found differences across the methods in terms of overlap and accuracy, but in this case these differences were even larger. ABLA was the only method that obtained some overlap with the voxels marked as informative by Searchlight, whereas L1-MKL and L2-MKL hardly detected those significant regions. The key of this finding is the classification problem itself. Evaluating whether a participant responded with the right/left hand to a stimulus generates large differences in the observed pattern and it is easy for a classifier to find a hyperplane that maximizes the separation of the two classes. On the other hand, isolating regions with a differential involvement in valence processing is much harder, as shown by recent metanalytic approaches (Lindquist and Mejia, 2015). Moreover, previous research has suggested that decoding accuracies from the PFC are lower than in other regions like the visual sensory cortex. This could reflect an important role of neural coding in the PFC. In fact, most studies evaluated in Bhandari et al. (2018) yielded a just above-chance accuracy (see Figure 1 in Bhandari et al., 2018), similar to the accuracy obtained by ABLA in the *valence* classification (51.77%).

It is worth noting that permutation tests were employed to assess the significance. We set a strict value for the cluster-defining primary-threshold ($p$=0.001) and we computed an empirical distribution of the clusters sizes. Then, we performed an FWE-correction to the $p$-values associated with each cluster to account for the multiple comparisons problem. In addition, the low value of the accuracy can also be influenced by the way information is represented in the brain. As an example, Dubois et al. (2015) could not retrieve information about face identity employing fMRI despite single-unit recordings demonstrated that this information was represented in the underlying neuronal populations. Hence, the level of accuracy of the classification in decoding fMRI data depends in part on how the underlying neural activity is organized. The highest accuracy possible given a specific contrast in a certain brain region can be far from maximal, and thus a low but statistically significant value can still be theoretically relevant (Bhandari et al., 2018; Hebart and Baker, 2017). Regarding the large difference in sensitivity obtained by the two approaches, the number of significant voxels depends on the brain regions involved in the cognitive function under study. In the *decision* classification, the number

of regions (mainly motor-related), as well as their sizes, is large. Moreover, the informative regions identified in the valence classification are related to valence-dependent behaviour (vmPFC; Kuzmanovic et al., 2018). The cluster contained in the vmPFC has a considerably smaller size and can hardly be properly isolated in an atlas.

Our results show that ABLA provides a larger overlap than all the other methods in the two classification problems, especially in the *valence* one. In this case, most MKL results lead to a below-chance accuracy in addition to a poor or null overlap with Searchlight results. This discrepancy must be due to the different framework of ABLA. Both L1-MKL and L2-MKL consider the regions provided by the atlas to build the model as part of the learning process. Hence, if the parcellations derived from the atlas do not properly delimitate the different regions involved in the context under study, the resulting model would be suboptimal since it is based on non-valid assumptions. On the other hand, ABLA builds the classification model from a whole-brain approach, which means that the atlas parcellations do not have any influence in the learning process. Instead, the brain organization is incorporated after building the model to summarize how informative each region is. For this reason, ABLA leads to a better performance in conditions of subtle or small differences between the conditions, although it is supposed to have a lower ability to detect informative regions compared to methods based on MKL when the atlas leads to a realistic approximation of the brain subdivisions.

### 9.6.2 Influence of the atlases

Results show that specific brain parcellations of each atlas impact the spatial accuracy of the different methods only when differences in the observed data are small, but not when these are large. In the *decision* classification, there was a large consistency among the significant regions obtained by all methods across the different atlases. These results carry important implications. Atlas-based approaches are assumed to have a large dependency on the way brain parcellations are computed. Our results evidence that atlases can be of help to identify informative regions even when the brain parcellations provided by the atlases do not perfectly delimitate the regions involved in the cognitive function under study. However, according to the results obtained in the *valence* classification, there are other contexts where these atlases are not accurate enough to guarantee a good performance in the identification of the sources of information. This is probably related to the size and the specific shape of the region involved in a certain cognitive function, such as the aforementioned vmPFC associated with the *valence* classification. The only region that ABLA marked as significant in the Camb12

parcellation is the one that contains the vmPFC, which shows that this method was able to identify where the information was located. Nevertheless, this region has a massive size in this atlas, and these atlas-based methods consider each region as a whole, and thus a large number of voxels are marked as significant only because they are in the same region as the one that is really informative. However, using atlases with more subdivisions implies that regions are much smaller. Here, the parcellation proposed by the atlas is a good match to the spatial organization of a small structure such as the vmPFC, leading to a higher sensitivity and spatial accuracy.

The number of subdivisions of an atlas also influenced the performance of the three algorithms evaluated. In the *decision* classification, the optimum value in terms of overlap was obtained by the 36 regions that the Camb36 atlas is divided into. We hypothesize that the number of regions is also important to obtain these results. Using an atlas with few subdivisions means that it is more likely to find an informative region, despite the small ratio between the voxels that are really significant and the total number of voxels that comprise the region. Instead, a large number of parcellations means that the classifier has to be much subtler in the identification of informative regions. The parcellations derived from Schaefer add larger precision and subdivisions to the brain networks published by the Yeo2011 atlas. However, results show a better performance in terms of sensitivity when the simplest approach was used. This strongly indicates that using atlases that do not reflect an accurate representation of the brain organization is similar to choosing a large Searchlight sphere where only a few voxels within this sphere are informative (Etzel et al., 2013). Using a large radius increases the probability of marking as significant voxels that are not, increasing the false-positives rate.

### 9.6.3 Stability of the regions across atlases

We have found a large correlation between the significant weight maps obtained by different atlases in the *decision* classification. The magnitude of the weight of a specific voxel quantifies the contribution of this voxel to the model, and its sign informs us about its relationship to the *Accept* or to the *Reject* class. For the L1-MKL approach, we found large correlation values for all atlases except for the Schaefer one. Hence, the resulting weights associated with each model are very similar, which highlights the stability of the classification methods regardless of the atlas used. Interestingly, we found the largest correlations in the weight maps obtained by the four Cambridge atlases, which are all derived from the same clustering algorithm (BASC). This result supports the idea that the mathematical framework employed to delimitate the different brain

regions is important since it can influence the success of the subsequent analyses. On the other hand, the poor performance of L1-MKL when the Schaefer atlas is used can be due to the conjunction of a sparse method and an atlas with a large number of regions, as mentioned in the previous section. It is important to note that our results do not invalidate the use of ambitious atlases aiming at obtaining a detailed parcellation of each cortical region. However, if these parcellations are not informative enough about how information is represented in the brain (unless an atlas is computed separately for each participant from his/her neural data: Gordon et al., 2017; Schaefer et al., 2018; Wang et al., 2015), sparse solutions are not recommended. Unlike L1-MKL, L2-MKL obtained a large correlation score between each pair of atlases (Figure 9.15 and Table 9.4). Thus, L2-MKL adapts to different idiosyncrasies and leads to a common solution for different brain mappings. This means that the weight maps that guide the classification are essentially the same regardless of the atlas used so that it is possible to successfully employ this approach even without a clear hypothesis about the brain organization in a specific context.

Nevertheless, these conclusions are only valid when there are large differences in the observed pattern associated with the two classes to distinguish from. Our findings in the *valence* classification differ substantially from those obtained in the *decision* classification. L1-MKL results (summarized in Table 9.5 and Figure 9.14) show that we could hardly compute the correlation between two pairs of atlases: the first one, AICHA and Harvard-Oxford; the second, AICHA and BN. In addition, none of the significant results provided by these atlases share any voxel with the Searchlight results, which illustrates that weight maps are similar from a mathematical perspective, but make a null contribution to the neuroscience standpoint. Results obtained by L2-MKL are summarized in Table 9.6 and conclusions derived from them are essentially the same than L1-MKL. We could only compute the correlation between two pairs of atlases: Schaefer-AICHA and Schaefer-Yeo2011. From these three atlases, Schaefer is the one that leads to some overlap with the Searchlight approach: 3.81%. However, none of these significant voxels are shared by AICHA and Yeo2011. This reflects that the two versions of MKL are not able to identify small informative regions in contexts where differences in the observed pattern of the two conditions are minimum.

### 9.6.4 Directionality of the weights

One of the main advantages of using weights instead of accuracy is the directionality that they provide. We have evaluated the sign of each weight within the significant

regions for each of the three atlas-based methods for the *decision* classification, where it is easy to evaluate whether the sign of the weight matches the psychological prediction. In this experiment, participants used one hand to accept an offer and the other one to reject it. The decoding analysis should mark as informative motor-related since these differentiate the conditions. However, it is worth remembering that the hand employed was counterbalanced across participants. We obtained exactly the expected results: the three approaches led to a map in which weights were organized according to their sign. For odd participants, regions associated with the acceptance of an offer (use of the right hand) were localized in the left hemisphere, with a positive sign. On the other hand, regions that contained information when the offer was rejected (left hand) were found in the right hemisphere, with a negative weight. More importantly, the informative regions for even participants shifted: positive weights were found in the right hemisphere, whereas weights with a negative sign were found in the left hemisphere. These results are very similar to those obtained by the univariate approach (see Figure 9.16): regions with a larger activation when participants accept/reject an offer match the sign of the weights of the different multivariate methods. However, atlas-based approaches use normalized data, which eliminates the differences in the global activation levels associated with each condition. Thus, these methods identify areas that show a different spatial distribution of the information, while the univariate approach purely relies on differences in the activation level. Despite this similarity, it is worth noting that multivariate approaches show a large sensitivity compared to a univariate. We have employed a classification context where differences at the cluster level can also be picked up by univariate methods to highlight one of the main advantages of atlas-based approaches. Similarly to Searchlight, these techniques are able to extract information from fine-grained differential activation patterns, which results in a boost in sensitivity compared to univariate analysis. Figure 9.16 reveals the differences in sensitivity between multivariate and univariate methods in the *decision* context, whereas differences between these two approaches in the valence classification can be found in previous studies (see Figure 10 in Arco et al., 2018 and Figure 1 in Lindquist and Mejia, 2015).

This chapter has provided some alternative methods to Searchlight in the localization of informative brain regions. These approaches are based on atlas and show a large performance when there are large differences between the observed pattern associated with the two classes to distinguish from. Moreover, the use of weight maps provides additional information to accuracies, combining the sensitivity of decoding analyses

and the directionality of univariate results. However, MKL methods are highly affected by the discrepancy of the shape of brain regions containing information and the one proposed by the atlases when the differential observed pattern is much lower. Hence, ABLA is the only approach that identifies informative regions in this difficult context.

# Part III

# General discussion and conclusions

CHAPTER 10

# GENERAL DISCUSSION AND CONCLUSIONS

The different contributions provided by this thesis have already been discussed in detail in each chapter. In this last chapter, we will discuss the contributions that this work makes to neuroimaging and machine learning

## 10.1 General discussion

Machine learning (ML) has revolutionized the world and changed the way we interact with it. Nowadays, there are hundreds of applications in our daily life that make use of this technology. The ability of these algorithms to compute large amounts of operations in real time has represented a paradigm shift in different contexts. In economy, the application of this technology to predict the future of international markets is exponentially increasing. Trading bots are able to make important decisions in a few miliseconds, and they have a growing impact on economic growth and international commerce. In security, devices extract unique characteristics of people based on biometric data (facial features or fingerprint) to provide access to a system. One of the most interesting applications derives from the use of ML for personalized medicine. The ability of these methods to extract patterns of information from a large amount of genes provides a powerful tool to develop medical treatment to the need of each individual. Similarly, the use of ML as an automatic tool for diagnosis of neurological disorders has led to an important development in the knowledge of these diseases.

In Cognitive Neuroscience, this methodology has led to a more accurate and powerful

understanding of the human brain. The superior sensitivity of MVPA in addition to its ability to analyze how information is spatially distributed across different voxels has proven to be more informative about the functional organization of cortex than univariate analysis. Despite the appealing aspects of multivariate analysis in fMRI decoding, there are many difficulties in the proper application of these methods that are usually underestimated. The most important one relies on the difference between the purpose for which this approach was initially developed and its main use in the Cognitive Neuroscience. Most applications aim at obtaining a classification model that predicts with the largest accuracy. However, in Cognitive Neuroscience the goal is to study different functions of the human brain. Thus, obtaining the largest accuracy is not of first interest, but to identify the regions involved in a certain cognitive process. This has two important consequences. First, the ML methods employed in other contexts cannot be adequate for identification purposes. Second, the methodology used in univariate analysis as well as the interpretation of results differs considerably for MVPA. Hence, it seems clear the need of reducing the gap between the frameworks used in contexts as different as Computer Science and Cognitive Neuroscience, developing approaches that overcome the issues of this last field.

One of the methods especially created for fMRI decoding is Searchlight. The first aim of this thesis was to study the effect of different factors in the performance of this approach. Our results demonstrated the strong relationship between the number of voxels marked as significant and the size of the Searchlight. First, the number of voxels marked as informative tend to grow as the Searchlight radius increases. Then, the number of significant voxels remain stable despite increasing the Searchlight size. Finally, sensitivity decreases for largest sizes. We also measured the influence of the type of the classifier in Searchlight results. ML applications as the ones previously described have large amounts of data to extract the information patterns. In these contexts, non-linear classifiers such as neural networks or algorithms based on deep learning can build classification models with a high generalization ability for unseen data. Nonetheless, datasets in fMRI are usually small and they have much more features than samples. This explains the superior performance of the linear classifier in this context.

Regardless the classification algorithm used, ML frameworks optimize the hyperparameters associated with the classifier to find the one that leads to the best performance. From the computer science perspective, employing a single value for these parameters can lead to suboptimal decision boundaries. However, in Searchlight analysis this optimization is not always performed because of the large computational cost that this

process entails. Results obtained in this thesis show that, in the context evaluated, the number of significant voxels and accuracy only lead to large changes when extreme values for these parameters were employed. Thus, conservative values of these parameters can be successfully used without biasing the results.

One of the main important differences between the use of ML methods in Computer Science and Cognitive Neuroscience is the number of images employed to train the classifier. From the Computer Science standpoint, at least a few tens of images are required to guarantee that the classifier learns actual information and not noise. However, in Cognitive Neuroscience is not uncommon to employ a few images. In fact, when activation patterns are averaged for each condition and run, the number of images for each class of the subsequent classification is given by the number of runs the experiment is divided into. The number of runs employed in a standard fMRI experiment ranges from 8 to 12, which means that 8-12 images for each class are available to compute the classification model. This is clearly insufficient for most ML applications, and the use of this reduced number in MVPA is derived from univariate analysis. The use of single-trial estimates in this thesis was not only due to optimize the contribution to the fMRI signal of adjacent events, but to increase the number of images available to train the classifier. Despite the different goal of MVPA compared to other ML applications, the methods employed are based on ML algorithms, and they have to meet the requirements that guarantee the reliability of the results.

Results obtained in Chapter 9 suggest the large contribution that methods originally developed for other ML applications can have in the Cognitive Neuroscience field. Most studies that employ MKL aimed at maximizing the clasification accuracy. In fact, the type of regularization that this approach uses enforces sparsity, which means that part of the input features is eliminated to lead to the best performance. However, the two-level hierarchy that this models yields (region and voxel level) can provide a very useful information about the brain regions involved in a certain process. The modification of the regularization (L2-norm instead to L1-norm) decreases the decoding performance, but increases the number of informative regions marked as significant and the overlap with Searchlight results. This evidences two important things. First, methods that are not commonly used in fMRI analyses can provide useful information. Second, to do so, it is probably required to modify these algorithms to maximize their identification ability.

Despite the high performance of the classification methods employed in Chapter 9 in contexts that lead to large differences between the experimental conditions, when these differences are subtle, the performance of these approaches decreases. This strongly

indicates that using atlases that do not reflect an accurate representation of the brain organization leads to non optimal results. Future studies are needed to widen the findings of this thesis by evaluating the performance of these methods when the brain parcellations are specifically computed for each participant, which may substantially improve the neuroanatomical functional precision. This combination might boost their sensitivity and widen their adequacy in different contexts, especially where an accurate parcellation is crucial.

## 10.2 General conclusions

Finally, as a summary of this thesis, we provide the following conclusions:

- We have proposed different methods to improve the performance in decoding fMRI data. To do so, we focused on three main aspects: First, an optimal estimation of the activation patterns that successfully computes the contribution of each stimulus to the hemodynamic response. Second, a classification algorithm that allows the identification of informative brain regions. Third, a proper evaluation of the statistical significance of the decoding performance.

- The linear classifier proposed in Chapter 5 has shown a large robustness in the identification of informative regions for a wide range of feature dimensions. However, the number of significant voxels tend to grow as the radius of the Searchlight increases, so that large sphere sizes can overestimate the number of informative voxels.

- The iterative estimation of the method employed in Chapter 8 is able to isolate the contribution to the hemodynamic response of adjacent events in a scenario with high overlap of signal and where differences between the conditions compared are very subtle.

- Non-parametric methods based on permutation testing highly improve sensitivity in the contexts evaluated, as it has been proven in Chapter 8. They compute more accurately the significance thresholds, detecting true informative data that otherwise do not surpass these thresholds.

- Atlas-based methods proposed in Chapter 9 are found to be very powerful and extendable tool for the identification of informative brain regions in fMRI analyses.

They have showed their ability to provide a similar sensitivity than other multivariate approaches in conjunction to the directionality of univariate traditional techniques. This directionality may enhance our knowledge about the brain since it informs about the contribution of each voxel/region to the classification decision, ranking the regions according to their importance in this decision.

- Our results reported in Chapter 9 evidence that atlases can be of help to identify informative regions even when the brain parcellations provided by the atlases do not perfectly delimitate the regions involved in the cognitive function under study. Nonetheless, contexts with minimal differences in neural activity are not accurate enough to guarantee a good performance in the identification of the sources of information.

- Information derived from just above-chance accuracies should not be rejected if these accuracies are significant. We should expect high values of accuracy when contrasting stimuli that lead to large differences in the activity pattern. If these differences are minimal, the accuracy will be small. Moreover, even when other measures of the electro-physiological responses evidence the presence of information, fMRI decoding can fail because of the way the underlying neuronal populations are oganized.

# RESUMEN AMPLIO EN ESPAÑOL

## Motivación

En los últimos años, el uso de la resonancia magnética funcional (RMf) ha aumentado de manera exponencial en el ámbito de la investigación en Neurociencia Cognitiva. Gracias a su precisión espacial, se ha convertido en una herramienta de gran utilidad para conocer en mayor profundidad el funcionamiento del cerebro humano. Esta técnica no invasiva ofrece una medida indirecta de los cambios producidos en la actividad neural en un contexto determinado, lo que permite estudiar la participación de diferentes regiones en determinadas funciones.

Los estudios desarrollados suelen tener unas bases comunes. Primero, se recluta una muestra representativa. Posteriormente, estos participantes realizan un experimento psicológico diseñado con el propósito de estudiar algún proceso o función cognitiva determinada, aunque existen otros estudios donde se mide la actividad del cerebro en estado de "reposo", sin la realización de ninguna tarea. A continuación, se evalúan si existen diferencias a nivel de grupo entre condiciones experimentales, de manera que puedan identificarse las regiones de interés en cada una de ellas.

La aproximación clásica para el análisis de este tipo de imágenes se conoce como análisis univariado. Esta técnica estudia la señal contenida en cada vóxel por separado, o bien la actividad promedio en una determinada región, y evalúa si existen diferencias entre dos condiciones experimentales (Friston et al., 1995; Worsley and Friston, 1995). Sin embargo, hay situaciones en las que aunque no haya diferencias en cada vóxel por separado, sí que las hay respecto a cómo se distribuye la actividad a través de los vóxeles. El análisis de patrones multivariados (MVPA, por sus siglas en inglés), es capaz de detectar estas diferencias, aumentando de manera notable la sensibilidad en comparación con los análisis clásicos. De hecho, esta aproximación permite descubrir diferencias tan sutiles que hasta entonces habían pasado totalmente desapercibidas (Haxby et al., 2001; Norman et al., 2006) empleando análisis univariados.

Este tipo de técnicas, basadas en *machine learning*, ya se habían utilizado previamente en aplicaciones tan variadas como el pronóstico del tiempo (Krasnopolsky and Fox-Rabinovitz, 2006) o la predicción del desarrollo de la economía (Lin et al., 2012). También se habían usado en el análisis de imágenes de RM estructural como una herramienta automática para el diagnóstico de diferentes trastornos neurológicos y psiquiátricos (Adeli et al., 2017; Arco et al., 2015; Choi et al., 2017; Del Gaizo et al., 2017; Khedher et al., 2017; Plant et al., 2010; Salvatore et al., 2014). En este ámbito clínico, el principal objetivo es obtener un modelo que prediga con la precisión más alta si una persona padece una enfermedad o no. Sin embargo, en Neurociencia Cognitiva el principal objetivo es estudiar las diferentes funciones del cerebro humano. Por lo tanto, los sistemas de clasificación utilizados en este contexto deben estar encaminados en detectar las regiones involucradas en un determinado proceso cognitivo, y no tanto en maximizar la precisión. Esta es una tarea de gran complejidad, dado que las diferencias entre condiciones experimentales están en muchas ocasiones enmascaradas por neuronas con diferentes patrones de respuesta en el mismo vóxel. A pesar de ello, el clasificador es capaz de identificar estas subpoblaciones de neuronas y decodificar la información (Norman et al., 2006).

Una de las mayores dificultades en el análisis de imágenes de RMf viene dada por la lentitud de la señal BOLD (Blood-Oxygen-Level-Dependent, en inglés). Esta señal proporciona una medida indirecta de la actividad neuronal basada en la oxigenación de la sangre: cuando hay un aumento de la actividad neuronal, se consume más oxígeno.

Esta señal alcanza su máximo entre 6 y 8 segundos después del inicio de la actividad neural, y tarda en torno a 16 segundos en volver a línea de base (Logothetis, 2003, 2004; Zaidi et al., 2018). Sin embargo, el tiempo entre los diferentes estímulos de un experimento psicológico, conocido como *inter-stimulus-interval* (ISI, por sus siglas en inglés), es normalmente mucho más corto que 16 segundos (González-García et al., 2017; Palenciano et al., 2018; Visconti di Oleggio Castello et al., 2017). Esto significa que la señal medida por el escáner de RM en cada instante no se debe a la actividad neural de un único estímulo, sino a una combinación de estímulos previos. Cuanto más corto sea este intervalo, más difícil será estimar de manera adecuada la contribución de cada estímulo a la señal hemodinámica y, por consiguiente, la posterior clasificación (Abdulrahman and Henson, 2016; Mumford et al., 2014; Turner et al., 2012).

Dada la gran complejidad que esta clasificación entraña, los análisis se realizan de manera independiente para cada persona. En cambio, las diferencias se evalúan a nivel de grupo, con el objetivo de comprobar que los resultados obtenidos para cada

sujeto son consistentes en toda la muestra. Para ello no es necesario que el clasificador alcance valores de precisión altos, sino que el promedio de esta difiera del azar de manera significativa. Los métodos paramétricos se han usado en cientos de estudios de RMf (Forman et al., 1995; Friston et al., 1994; Hayasaka and Nichols, 2003; Misaki et al., 2010; Woo et al., 2014). Sin embargo, recientemente se ha demostrado que su uso para evaluar la significatividad de la precisión de un clasificador puede no ser apropiado, ya que parten de ciertas asunciones que no siempre se cumplen (Eklund et al., 2016; Stelzer et al., 2013). Como alternativa a estos, los métodos no paramétricos basados en permutaciones calculan de manera empírica la distribución de los datos, en lugar de asumir que siguen una distribución Gaussiana. Esto permite que puedan potencialmente aumentar la sensibilidad de los análisis estadísticos controlando de una manera más fiable el número de falsos positivos (Smith and Nichols, 2009; Stelzer et al., 2013).

Dada la gran relevancia que tiene la información espacial en Neurociencia Cognitiva, todo el sistema de clasificación debe orientarse al objetivo de la identificación de regiones informativas. Aunque un clasificador obtenga una precisión significativa, no es de mucha utilidad desde el punto de vista psicológico si no es capaz de proporcionar las regiones cerebrales de las que ha extraído dicha información. Por lo tanto, todas las etapas del sistema de clasificación deben preservar la información espacial, lo que limita enormemente las técnicas que pueden emplearse. Por ejemplo, en *machine learning*, es habitual emplear técnicas de extracción de características para la reducción de la dimensionalidad de los datos. Estos métodos emplean transformaciones geométricas desde el espacio original hasta un nuevo espacio transformado. A pesar de que pueden mejorar el rendimiento del clasificador, eliminan toda información espacial, por lo que su uso en contextos psicológicos debe hacerse junto a otro tipo de análisis que preserve dicha información.

# Objetivos

El objetivo de esta tesis es comparar diferentes métodos en varias etapas del sistema de clasificación y proporcionar la solución que lleve a un mayor rendimiento en múltiples contextos: en escenarios complejos donde cada ensayo contiene varios eventos de diferente duración, en diseños de eventos con un intervalo de longitud intermedia entre lo diferentes ensayos, en diseños de bloques, etc. Para ello, se llevarán a cabo dos estudios principales con los siguientes propósitos: aislar de manera precisa la contribución de cada evento a la señal BOLD en cada uno de estos escenarios y desarrollar nuevos métodos de

clasificación que proporcionen información adicional acerca de las regiones involucradas en un determinado proceso psicológico. De esta manera, podemos definir los siguientes objetivos:

1. Comparar el rendimiento de diferentes clasificadores y encontrar el que obtenga una mayor sensibilidad cuando el solapamiento entre eventos cercanos es mayor.

2. Evaluar diferentes métodos para una estimación óptima de los patrones de activación en los escenarios mencionados anteriormente.

3. Comparar diferentes métodos para la evaluación de la significatividad estadística en esos contextos, con el objetivo de controlar el número de falsos positivos ofreciendo simultáneamente la mayor sensibilidad posible.

4. Desarrollar nuevas estrategias que maximicen la detección de regiones cerebrales informativas, proporcionando nuevas perspectivas acerca de cómo se distribuye la información dentro de esas regiones.

Para lograr todos estos objetivos, se han llevado a cabo los siguientes estudios:

1. Un análisis en profundidad del método de clasificación más común en imágenes de RMf: *Searchlight* (nombre original en inglés). Se ha evaluado la variabilidad de las regiones informativas dependiendo de diversos factores como la dimensionalidad de los datos de entrada, el algoritmo de clasificación utilizado y los hiperparámetros asociados al mismo.

2. La búsqueda de una estimación óptima de los patrones de activación. En concreto, se han usado tres alternativas. En la primera, todos los ensayos del mismo tipo pertenecientes al mismo *run* del experimento son colapsados en un único regresor. En el segundo, se utiliza un regresor diferente para cada ensayo. El último se basa en un proceso iterativo en la que la actividad debida a cada ensayo se estima en un modelo diferente. Cada modelo tiene dos regresores: uno para el ensayo objetivo y otro para el resto.

3. La comparación de tres métodos estadísticos (uno paramétrico y dos no paramétricos) para evaluar la significatividad de las precisiones resultantes del proceso de clasificación.

4. El desarrollo de distintos métodos de clasificación basados en atlas que proporcionan una medida alternativa del rendimiento del clasificador. Esta medida, basada en los pesos de un clasificador lineal, es mucho más informativa que la precisión desde el punto de vista de la Neurociencia Cognitiva.

# Contribuciones

Los capítulos previos a los que detallan las contribuciones realizadas en esta tesis contienen la información necesaria para entender la relevancia de las mismas.

En primer lugar, se describen las bases físicas de la resonancia magnética, así como las operaciones matemáticas que permiten la generación de la imagen. Posteriormente, se describe el preprocesado necesario para el análisis de GLM. A continuación se lleva a cabo una descripción del estado del arte en el uso de técnicas basadas en *machine learning* para el análisis de imágenes de RMf. En los siguientes capítulos se describen las diferentes alternativas para la clasificación en los análisis multivariados de RMf. Por último, se explican la importancia de los tests para evaluar la significatividad estadística en este tipo de escenarios, así como las aproximaciones más utilizadas en este ámbito.

## Algoritmo de clasificación óptimo para el análisis de imágenes de RMf

En el Capítulo 7, nos centramos en el método de clasificación de imágenes de resonancia magnética funcional más utilizado en la actualidad. Esta alternativa, conocida como *Searchlight*, permite identificar con gran precisión las regiones involucradas en un determinado proceso. Nuestros resultados muestran que el rendimiento (medido en términos de precisión y número de vóxeles significativos) depende principalmente de dos factores: el tamaño del *Searchlight* y el algoritmo de clasificación. En primer lugar, conforme va aumentando el tamaño del *Searchlight*, también lo hace el número de vóxeles significativos. Sin embargo, a partir de un umbral determinado, la sensibilidad permanece estable, disminuyendo cuando el tamaño es demasiado grande. Por otro lado, la precisión permanece estable independientemente del tamaño del *Searchlight*.

También evaluamos distintos algoritmos de clasificación, todos ellos basados en *Support Vector Machines* (máquinas de vectores de soporte, en español). Concretamente, utilizamos un kernel lineal, otros tres polinómicos (de segundo, tercer y cuarto grado) y un último basado en el kernel conocido como *Radial Basis Function* (RBF). Los resultados

mostraron un rendimiento mucho mayor en el caso del algoritmo lineal. De hecho, fue el único capaz de encontrar diferencias significativas para todos los tamaños de *Searchlight*. El rendimiento de los kernels polinómicos dependía en gran medida del tamaño del *Searchlight*, mientras que el RBF no fue capaz de identificar ningún vóxel significativo para la gran mayoría de tamaños.

Por último, en este capítulo estudiamos la influencia de los hiperpárametros del clasificador en el rendimiento del mismo. En general, aplicar una optimización de los mismos (proceso conocido como *grid search*, en inglés) no supuso una mejora ni en precisión ni en sensibilidad. Este hallazgo es muy importante ya que la búsqueda de la configuración óptima de un clasificador es un proceso muy costoso desde el punto de vista computacional. De hecho, el *Searchlight* también lo es, por lo que que es muy importante encontrar evidencia de que utilizando los parámetros por defecto es posible obtener un rendimiento considerable.

## Estimación precisa de los patrones de activación neural

Para el Capítulo 8, proponemos distintos métodos de estimación de patrones de activación. Estos métodos estiman la contribución que tiene cada evento del diseño experimental en los cambios producidos en la señal hemodinámica medidos por el escáner de RM. Para ello, utilizamos un contexto con gran colinearidad y solapamiento entre distintos regresores muy cercanos en el tiempo, lo que dificulta en gran medida la separabilidad. En concreto, el objetivo era distinguir la actividad asociada a distintos eventos dentro de un mismo ensayo. Además, también evaluamos el rendimiento de las diferentes alternativas en otros dos contextos con menor solapamiento: un diseño de bloques y un diseño de eventos con un intervalo mayor entre los ensayos.

Nuestros resultados muestran que en el contexto más adverso solo un método consiguió llevar a cabo de manera exitosa una estimación de la señal. Esta aproximación, denominada *Least-Squares Separate*, estima un modelo lineal general (GLM, en inglés) de manera independiente para cada uno de los regresores del mismo, lo que permite un aumento considerable en la precisión de la estimación. En los otros dos escenarios, los demás métodos de estimación también son capaces de aislar la contribución de cada evento de manera satisfactoria, especialmente en el diseño de bloques, donde la colinearidad y el solapamiento son más bajos.

El segundo gran objetivo de este capítulo es el de evaluar hasta qué punto los métodos paramétricos son adecuados para evaluar la significatividad de las precisiones obtenidas por un clasificador. Estos métodos se basan en los llamados *Random Field Theory*, y

asumen que las precisiones siguen una distribución Gaussiana. Por el contrario, los métodos no paramétricos no realizan ninguna asunción y calculan de manera empírica la distribución de las precisiones. Nuestros resultados muestran que los acercamientos no paramétricos permiten detectar información que de otra forma no superaría el umbral estadístico necesario para ser considerado como significativo. De hecho, en los contextos evaluados estos métodos consiguen un aumento considerable en la sensibilidad comparados con las técnicas paramétricas.

## Métodos basados en atlas para la identificación de regiones cerebrales informativas

En el Capítulo 9, propusimos los métodos basados en atlas como una alternativa al *Searchlight* en la identificación de regiones cerebrales informativas. Estos métodos ofrecerían un menor coste computacional y una medida del rendimiento del clasificador que amplía nuestro conocimiento acerca de cómo se distribuye dicha información en las regiones involucradas en un proceso cognitivo. Para ello, utilizamos tres métodos diferentes basados en atlas en un contexto donde el objetivo principal no es obtener una precisión alta, sino detectar las regiones cerebrales que contienen información durante una determinada operación mental. Además, también evaluamos la influencia que tienen en los resultados las diferentes parcelaciones del cerebro propuestas por cada atlas, así como el número de regiones contenidas en los mismos.

Nuestros resultados revelaron que no siempre se obtienen de manera simultánea una máxima precisión y sensibilidad, especialmente cuando el objetivo es detectar diferencias muy sutiles. Esto quiere decir que un algoritmo puede obtener una precisión muy alta y que, sin embargo, no aporte ninguna información de utilidad que permita un mejor entendimiento del cerebro humano. Este es el caso de aquellos algoritmos que emplean la regularización *L1*, ya que un número elevado de regiones son automáticamente descartadas para la posterior clasificación debido a la naturaleza del propio algoritmo. En cambio, una regularización *L2* ofrece unos resultados mucho más útiles para un contexto de identificación, a pesar de que en este caso la precisión sea menor.

Los experimentos desarrollados en este capítulo permitieron observar que los algoritmos de clasificación propuestos presentan un buen rendimiento independientemente del atlas utilizado cuando las diferencias a nivel neural son grandes. Sin embargo, si las diferencias son muy sutiles, los atlas no son los suficientemente precisos para garantizar una buena identificación de las regiones que contienen información.

Por último, un hallazgo muy interesante es que la precisión del clasificador puede no ser una medida óptima de la información contenida en una región cerebral. Los algoritmos propuestos en este capítulo emplean los pesos del clasificador en lugar de la precisión. Esto permite que los métodos obtengan una sensibilidad similar al *Searchlight* y una direccionalidad propia de los métodos univariados. Además, los pesos no solo ofrecen información acerca de la significatividad de un vóxel o región sino también de su importancia en la decisión llevada a cabo por el clasificador.

## Conclusiones

Por último, resumimos las contribuciones de esta tesis en una serie de conclusiones:

- Hemos propuesto diferentes métodos para mejorar el rendimiento de los análisis multivariados en imágenes de resonancia magnética funcional. Para ello, nos hemos centrado principalmente en tres aspectos. Primero, un análisis en profundidad del método de clasificación conocido como Searchlight, evaluando la influencia de diversos factores en el rendimiento del mismo. Segundo, una estimación óptima de los patrones de activación para calcular de manera precisa la contribución de cada estímulo a la respuesta hemodinámica. Tercero, un algoritmo de clasificación que permite identificar las regiones cerebrales que contienen información en un contexto donde las diferencias a nivel neural son muy pequeñas. Cuarto, un método adecuado para evaluar la significatividad estadística de las precisiones obtenidas por los algoritmos de clasificación.

- El clasificador lineal propuesto en el Capítulo 5 muestra una gran robustez en la identificación de regiones informativas para un rango amplio de dimensionalidad en los datos de entrada. Sin embargo, el número de vóxeles significativos tiende a crecer conforme lo hace el tamaño del *Searchlight*, por lo que un tamaño de esfera demasiado grande puede sobreestimar el número real de vóxeles informativos.

- La estimación iterativa del método empleado en el Capítulo 8 es capaz de aislar la contribución a la señal hemodinámica de eventos cercanos en el tiempo en un escenario con una alta superposición y colinearidad, donde las diferencias a nivel neural entre las distintas condiciones evaluadas eran muy sutiles.

- Los métodos no paramétricos basados en tests de permutaciones producen una gran mejora en la sensibilidad en los contextos evaluados en la presente tesis, como queda reflejado en el Capítulo 8. La mejor estimación de los umbrales de

significatividad permite identificar regiones informativas que con los métodos paramétricos no superarían dicho umbral.

- Los métodos basados en atlas propuestos en el Capítulo 9 permiten una identificación precisa de las diferentes regiones informativas involucradas en la función cognitiva de interés. En concreto, proporcionan una sensibilidad similar a otras técnicas como el *Searchlight*, así como la direccionalidad de las técnicas univariadas tradicionales. Esta direccionalidad proporciona una información de gran utilidad ya que permite ordenar las regiones involucradas en un determinado contexto en función de su importancia en el proceso de clasificación.

- Los resultados obtenidos en el Capítulo 9 evidencian que los atlas pueden ser de gran ayuda para identificar regiones informativas incluso cuando las parcelaciones cerebrales propuestas por dichos atlas no proporcionan una delimitación precisa. Sin embargo, en aquellos contextos donde las diferencias a nivel neural son muy sutiles, estos métodos no garantizan una localización correcta de la información.

- La información espacial derivada de precisiones justo por encima del azar no debe ser rechazada siempre y cuando dichas precisiones sean significativas. Es de esperar encontrar valores de precisión altos en aquellos casos en los que el contraste evaluado genere unas diferencias perceptuales grandes. Si estas diferencias son mínimas, la precisión será baja. De hecho, incluso cuando otras medidas de respuestas electro-fisiológicas evidencian la presencia de información, es posible que los análisis multivariados de imágenes de resonancia magnética no puedan clasificar por encima del azar debido a la forma en que las diferentes poblaciones neuronales subyacentes están representadas.

# BIBLIOGRAPHY

Abdulrahman, H., Henson, R. N., 2016. Effect of trial-to-trial variability on optimal event-related fMRI design: Implications for Beta-series correlation and multi-voxel pattern analysis. NeuroImage 125, 756 – 766. doi: https://doi.org/10.1016/j.NeuroImage.2015.11.009.

Adeli, E., Wu, G., Saghafi, S., An, L., Shi, F., Shen, D., 2017. Kernel-based Joint Feature Selection and Max-Margin Classification for Early Diagnosis of Parkinson's Disease. Scientific Reports 7, 41069. doi: https://doi.org/10.1038/srep41069.

Aitken, A. C., 1936. On Least Squares and Linear Combination of Observations. Proceedings of the Royal Society of Edinburgh 55, 42–48. doi: https://doi.org/10.1017/S0370164600014346.

Allefeld, C., Haynes, J.-D., 2014. Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. NeuroImage 89, 345 – 357. doi: https://doi.org/10.1016/j.NeuroImage.2013.11.043.

Altaf, T., Anwar, S. M., Gul, N., Majeed, M. N., Majid, M., 2018. Multi-class Alzheimer's disease classification using image and clinical features. Biomedical Signal Processing and Control 43, 64 – 74. doi: https://doi.org/10.1016/j.bspc.2018.02.019.

Arco, J. E., González-García, C., Díaz-Gutiérrez, P., Ramírez, J., Ruz, M., 2018. Influence of activation pattern estimates and statistical significance tests in fMRI decoding analysis. Journal of Neuroscience Methods 308, 248–260. doi: https://doi.org/10.1016/j.jneumeth.2018.06.017.

Arco, J. E., González-García, C., Ramírez, J., Ruz, M., 2016. Comparison of different methods for brain decoding from fMRI beta maps. Poster presented at 22nd Annual Meeting of the Organization for Human Brain Mapping, Geneve, (Switzerland).

Arco, J. E., Ramírez, J., Puntonet, C. G., Górriz, J. M., Ruz, M., 2015. Short-term prediction of MCI to AD conversion based on longitudinal MRI analysis and neuropsychological tests. In: Innovation in Medicine Healthcare. pp. 385–394. doi: https://doi.org/10.1007/978-3-319-23024-5_35.

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. NeuroImage 38 (1), 95–113. doi: https://doi.org/10.1016/j.NeuroImage.2007.07.007.

Ashburner, J., Friston, K. J., 1999. Nonlinear spatial normalization using basis functions. Human Brain Mapping 7 (4), 254–266. doi: https://doi.org/10.1002/(sici)1097-0193(1999)7:4<254::aid-hbm4>3.0.co;2-g.

Attwell, D., Buchan, A. M., Charpak, S., Lauritzen, M., Macvicar, B. A., Newman, E. A., 2010. Glial and neuronal control of brain blood flow. Nature 468 (7231), 232–43. doi: https://doi.org/10.1038/nature09613.

Attwell, D., Iadecola, C., 2002. The neural basis of functional brain imaging signals. Trends in Neurosciences 25 (12), 621–5. doi: https://doi.org/10.1016/S0166-2236(02)02264-6.

Badre, D., Frank, M. J., 2011. Mechanisms of Hierarchical Reinforcement Learning in Cortico-Striatal Circuits 2: Evidence from fMRI. Cerebral Cortex 22 (3), 527–536. doi: https://doi.org/10.1093/cercor/bhr117.

Balci, S. K., Sabuncu, M. R., Yoo, J., Ghosh, S. S., Whitfield-Gabriel, S., Gabriel, J. D., Golland, P., 2008. Prediction of Successful Memory Encoding from fMRI Data. Medical Image Computing and Computer-Assisted Intervention 2008 (11), 97–104. doi: https://doi.org/10.1007/s12021-013-9204-3.

Beall, E. B., Lowe, M. J., 2014. SimPACE: generating simulated motion corrupted BOLD data with synthetic-navigated acquisition for the development and evaluation of SLOMOCO: a new, highly effective slicewise motion correction. NeuroImage 101, 21–34.

Bellec, P., Rosa-Neto, P., Lyttelton, O. C., Benali, H., Evans, A. C., 2010. Multi-level bootstrap analysis of stable clusters in resting-state fMRI. NeuroImage 51 (3), 1126–39. doi: https://doi.org/10.1016/j.NeuroImage.2010.02.082.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B 57 (1), 289–300. doi: https://doi.org/10.2307/2346101.

Bennett, C., Miller, M., Wolford, G., 07 2009. Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: an argument for multiple comparisons correction. NeuroImage 47. doi: https://doi.org/10.1016/S1053-8119(09)71202-9.

Bennett, K. P., Blue, J. A., 1998. A support vector machine approach to decision trees. In: 1998 IEEE International Joint Conference on Neural Networks Proceedings. Vol. 3. pp. 2396–2401. doi: https://doi.org/10.1109/IJCNN.1998.687237.

Berens, P., Logothetis, N., Tolias, A., 2010. Local field potentials, BOLD and spiking activity-relationships and physiological mechanisms. Nature Precedings http://precedings.nature.com/documents/5216/version/1/files/npre20105216-1.pdf.

Betancur, J., Commandeur, F., Motlagh, M., Sharir, T., Einstein, A. J., Bokhari, S., Fish, M. B., Ruddy, T. D., Kaufmann, P., Sinusas, A. J., Miller, E. J., Bateman, T. M., Dorbala, S., Carli, M. D., Germano, G., Otaki, Y., Tamarappoo, B. K., Dey, D., Berman, D. S., Slomka, P. J., 2018. Deep Learning for Prediction of Obstructive Disease From Fast Myocardial Perfusion SPECT: A Multicenter Study. JACC: Cardiovascular Imaging 11 (11), 1654 – 1663. doi: https://doi.org/10.1016/j.jcmg.2018.01.020.

Bhagalla, R., Kim, B., 2008. Spin saturation artifact correction using slice-to-volume registration motion estimates for fMRI time series. Medical Physics 35 (2), 424–34. doi: https://doi.org/10.1118/1.2826555.

Bhandari, A., Gagne, C., Badre, D., 2018. Just above Chance: Is It Harder to Decode Information from Prefrontal Cortex Hemodynamic Activity Patterns? Journal of Cognitive Neuroscience 30 (10), 1473–1498. doi: http://dx.doi.org/10.1162/jocn_a_01291.

Bishop, C. M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K., Curio, G., 2007. The non-invasive Berlin Brain-Computer Interface: fast acquisition of effective performance in untrained subjects. NeuroImage 37 (2), 539–50. doi: https://doi.org/10.1016/j.NeuroImage.2007.01.051.

Bode, S., Haynes, J. D., 2009. Decoding sequential stages of task preparation in the human brain. NeuroImage 45, 606–613. doi: https://doi.org/10.1016/j.NeuroImage.2008.11.031.

Bode, S., Stahl, J., 2014. Predicting errors from patterns of event-related potentials preceding an overt response. Biological Psychology 103, 357 – 369. doi: https://doi.org/10.1016/j.biopsycho.2014.10.002.

Boser, B. E., Guyon, I., Vapnik, V., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. pp. 144–152. doi: https://doi.org/10.1145/130385.130401.

Botvinick, M., Weinstein, A., Solway, A., Barto, A., 2015. Reinforcement learning, efficient coding, and the statistics of natural tasks. Current Opinion in Behavioral Sciences 5, 71 – 77. doi: https://doi.org/10.1016/j.cobeha.2015.08.009.

Brammer, M., Bullmore, E., Simmons, A., Williams, S., Grasby, P., Howard, R., Woodruff, P., Rabe-Hesketh, S., 1997. Generic brain activation mapping in functional magnetic resonance imaging: A nonparametric approach. Magnetic Resonance Imaging 15 (7), 763 – 770. doi: https://doi.org/10.1016/S0730-725X(97)00135-5.

Brett, M., Christoff, K., Cusack, R., Lancaster, J. L., 2001. Using the Talairach atlas with the MNI template. NeuroImage 13 (6), 85–85. doi: https://doi.org/10.1016/S1053-8119(01)91428-4.

Brett, M., Penny, W., Kiebel, S., 2003. An introduction to Random Field Theory. Http://www.fil.ion.ucl.ac.uk/spm/doc/books/hbf2/pdfs/Ch14.pdf.

Breuer, F. A., Moriguchi, H., Seiberlich, N., Blaimer, M., Jakob, P. M., Duerk, J. L., Griswold, M. A., 2008. Zigzag sampling for improved parallel imaging. Magnetic Resonance in Medicine 60, 474–478. doi: https://doi.org/10.1002/mrm.21643.

Brockway, J. P., 2000. Two functional magnetic resonance imaging f(MRI) tasks that may replace the gold standard, Wada testing, for language lateralization while giving additional localization information. Brain and Cognition 43 (1-3), 57–9.

Bron, E. E., Smits, M., Niessen, W. J., Klein, S., 2015. Feature Selection Based on the SVM Weight Vector for Classification of Dementia. IEEE Journal of Biomedical and Health Informatics 19 (5), 1617–1626. doi: https://doi.org/10.1109/JBHI.2015.2432832.

Buckner, R. L., Bandettini, P. A., O'Craven, K. M., Savoy, R. L., Petersen, S. E., Raichle, M. E., Rosen, B. R., 1996. Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. Proceedings of National Academy of Sciences 93 (25), 14878–14883. doi: https://doi.org/10.1073/pnas.93.25.14878.

Buckner, R. L., Goodman, J., Burock, M., Rotter, M., Koutstaal, W., Schachter, D., Rosen, B., Dale, A. M., 1998. Functional-anatomic correlates of object priming in humans revealed by rapid presentation event-related fMRI. Neuron 20 (2), 285–96. doi: https://doi.org/10.1016/S0896-6273(00)80456-0.

Bullmore, E. T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., Brammer, M. J., 1999. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. IEEE Transactions on Medical Imaging 18 (1), 32–42. doi: https://doi.org/10.1109/42.750253.

Burges, C. J. C., 1998. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2 (2), 121–167.

Burns, S. P., Xing, D., Shapley, R. M., 2010. Comparisons of the dynamics of LFP and MUA signas in macaque visual cortex. Journal of Neuroscience 30 (41), 13739–13749. doi: https://doi.org/10.1523/JNEUROSCI.0743-10.2010.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robison, E. S. J., Munafò, M. R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience 14, 365–376. doi: https://doi.org/10.1038/nrn3475.

Buxton, R. B., Wong, E. C., Frank, L. R., 1998. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. Magnetic Resonance in Medicine 39 (6), 855–64. doi: https://doi.org/10.1002/mrm.1910390602.

Calhoun, V., Golaxy, X., Pearlson, G., 2000. Improved fMRI slice timing correction: interpolation errors and wrap around effects. Proceedings of the 9th ISMRM Annual Meeting, 810Doi: https://doi.org/10.1073/pnas.191101098.

Chanel, G., Pichon, S., Conty, S., Berthoz, S., Chevallier, C., Grèzes, J., 2016. Classification of autistic individuals and controls using cross-task characterization of fMRI activity. NeuroImage: Clinical 10, 78–88. doi: https://doi.org/10.1016/j.nicl.2015.11.010.

Chen, G., Chen, J., 2015. A novel wrapper method for feature selection and its applications. Neurocomputing 159, 219–226. doi: https://doi.org/10.1016/j.neucom.2015.01.070.

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., Hasson, U., 2017. Shared memories reveal shared structure in neural activity across individuals. Nature Neuroscience 20 (1), 115–125. doi: https://doi.org/10.1038/nn.4450.

Chen, Y., Namburi, P., Elliott, L. T., Heinzle, J., Soon, C. S., Chee, M. W., Haynes, J. D., 2011. Cortical surface-based searchlight decoding. NeuroImage 56, 582–592. doi: https://doi.org/10.1016/j.NeuroImage.2010.07.035.

Chen, Z., Guo, Y., Zhang, S., Feng, T., 2019. Pattern classification differentiates decision of intertemporal choices using multi-voxel pattern analysis. Cortex 111, 183 – 195. doi: https://doi.org/10.1016/j.cortex.2018.11.001.

Cherkassky, V., , Mulier, F. M., Vapnik, V. N., 1999. Model complexity control for regression using VC generalization bounds. IEEE Transactions on Neural Networks 10 (5), 1075–1089. doi: https://doi.org/10.1109/72.788648.

Cherkassky, V., Ma, Y., 2004. Practical selection of SVM parameters and noise estimation for SVM regression. Neural Networks 17 (1), 113 – 126. doi: https://doi.org/10.1016/S0893-6080(03)00169-2.

Choi, H., Ha, S., Im, H. J., Paek, S. H., Lee, D. S., 2017. Refining diagnosis of Parkinson's disease with deep learning-based interpretation of dopamine transporter imaging. NeuroImage Clinical 10 (16), 586–594. doi: https://doi.org/10.1016/j.nicl.2017.09.010.

Cichy, R. M., Pantazis, D., 2017. Multivariate pattern analysis of MEG and EEG: A comparison of representational structure in time and space. NeuroImage 158, 441 – 454. doi: https://doi.org/10.1016/j.NeuroImage.2017.07.023.
URL http://www.sciencedirect.com/science/article/pii/S1053811917305918

Cichy, R. M., Pantazis, D., Oliva, A., 2016. Similarity-Based Fusion of MEG and fMRI Reveals Spatio-Temporal Dynamics in Human Cortex During Visual Object Recognition. Cerebral Cortex 26 (8), 3563–3579. doi: https://doi.org/10.1093/cercor/bhw135.

Cohen, M. S., 2000. Echo-plannar imaging (EPI) and functional MRI. Springer. ISBN: 978-354-06-7215-9.

Constable, R. T., Spencer, D. D., 2001. Repetition time in echo planar functional MRI. Magnetic Resonance in Medicine 46, 748–755. doi: https://doi.org/10.1002/mrm.1253.

Corradi-Dell'Acqua, C., Civai, C., Rumiati, R. I., Fink, G. R., 2013. Disentangling self- and fairness-related neural mechanisms involved in the ultimatum game: an fMRI study. Social Cognitive and Affective Neuroscience 8 (4), 424–431. doi: http://dx.doi.org/10.1093/scan/nss014.

Coutanche, M. N., Thompson-Schill, S. L., 2012. The advantage of brief fMRI acquisition runs for multi-voxel pattern detection across runs. NeuroImage 61, 1113–1119. doi: https://doi.org/10.1016/j.NeuroImage.2012.03.076.

Coutanche, M. N., Thompson-Schill, S. L., Schultz, R. T., 2011. Multi-voxel pattern analysis of fMRI data predicts clinical symptom severity. NeuroImage 57 (1), 113–123. doi: https://doi.org/10.1016/j.NeuroImage.2011.04.016.

Cox, D. D., Savoy, R. L., 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage. 19, 261–270. doi: http://dx.doi.org/10.1016/S1053-8119(03)00049-1.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press. doi: https://doi.org/10.1017/CBO9780511801389.

Dai, D., Wang, J., Hua, J., He, H., 2012. Classification of ADHD children through multimodal magnetic resonance imaging. Frontiers in Systems Neuroscience 6, 63. doi: https://doi.org/10.3389/fnsys.2012.00063.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E., 2008. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. NeuroImage 43 (1), 44–58. doi: https://doi.org/10.1016/j.NeuroImage.2008.06.037.

Deichmann, R., Josephs, O., Hutton, C., Corfield, D. R., Turner, R., 2002. Compensation of susceptibility-induced BOLD sensitivity losses in echo-planar fMRI imaging. NeuroImage 15 (1), 120–35. doi: https://doi.org/10.1006/nimg.2001.0985.

Del Gaizo, J., Mofrad, N., Jensen, J. H., Clark, D., Glenn, R., Helpern, J., Bonilha, L., 2017. Using machine learning to classify temporal lobe epilepsy based on diffusion MRI. Brain and Behavior 7 (10), e00801. doi: https://doi.org/10.1002/brb3.801.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., Killiany, R. J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31 (3), 968–80. doi: https://doi.org/10.1016/j.NeuroImage.2006.01.021.

D'Esposito, M., Deouell, L. Y., Gazzaley, A., 2003. Alterations in the BOLD fMRI signal with ageing and disease: a challenge for neuroimaging. Nature Reviews Neuroscience 4 (11), 863–72. doi: https://doi.org/10.1038/nrn1246.

Di Russo, F., Berchicci, M., Bozzacchi, C., Perri, R., Pitzalis, S., Spinelli, D., 2017. Beyond the "Bereitschaftspotential": Action preparation behind cognitive functions. Neuroscience & Biobehavioral Reviews 78, 57 – 81. doi: https://doi.org/10.1016/j.neubiorev.2017.04.019.

Dobbs, D., 2005. Fact or phrenology? Scientific American Mind 16, 24–31. doi: https://doi.org/10.1038/scientificamericanmind0405-24.

Donahue, M. J., Blicher, J. U., Østergaard, L., Feinberg, D. A., MacIntosh, B. J., Miller, K. L., Günther, M., Jezzard, P., 2009. Cerebral blood flow, blood volume, and oxygen metabolism dynamics in human visual and motor cortex as measured by whole-brain multi-modal magnetic resonance imaging. Journal of Cerebral Blood Flow & Metabolism 29, 1856–1866. doi: http://dx.doi.org/10.1038/jcbfm.2009.107.

Dosenbach, N. U., Nardos, B., Cohen, A. L., Fair, D. A., Power, J. D., Church, J. A., Nelson, S. M., Wig, G. S., Vogel, A. C., Lessov-Schlaggar, C. N., Barnes, K. A., Dubis, J. W., Feczko, E., Coalson, R. S., Pruett, J. R. J., Barch, D. M., Petersen, S. E., Schlaggar, B. L., 2010. Prediction of individual brain maturity using fMRI. Science 329 (5997), 1358–61. doi: https://doi.org/10.1126/science.1194144.

Draschkow, D., Heikel, E., Võ, M. L. H., Fiebach, F., Sassenhagen, J., 2018. No evidence from MVPA for different processes underlying the N300 and N400 incongruity effects in object-scene processing. Neuropsychologia 120, 9 – 17. doi: https://doi.org/10.1016/j.neuropsychologia.2018.09.016.

Dubois, J., de Berker, A. O., Tsao, D., 2015. Single-unit recordings in the macaque face patch system reveal limitations of fMRI MVPA. The Journal of Neuroscience 35 (6), 2791–2802. doi: https://doi.org/10.1523/JNEUROSCI.4037-14.2015.

Duong, T. Q., Kim, D. S., Uğurbil, K., Kim, S. G., 2001. Localized cerebral blood flow response at submillimeter columnar resolution. Proceedings of the National Academy of Sciences of the United States of America 98 (19), 10904–10909. doi: https://doi.org/10.1073/pnas.191101098.

Eklund, A., Nichols, T. E., Knutsson, H., 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. PNAS 113 (28), 7900–7905. doi: https://doi.org/10.1073/pnas.1602413113.

Etzel, J. A., Gazzola, V., Keysers, C., 2009. An introduction to anatomical ROI-based fMRI classification analysis. Brain Research 28 (1282), 114–25. doi: https://doi.org/10.1016/j.brainres.2009.05.090.

Etzel, J. A., Zacks, J. M., Braver, T. S., 2013. Searchlight analysis: promise, pitfalls, and potential. NeuroImage 37, 261–269. doi: https://doi.org/10.1016/j.NeuroImage.2013.03.041.

Evans, A. C., Collins, D. L., Mills, S. R., Brown, E. D., Kelly, R. L., Peters, T. M., 1993. 3D statistical neuroanatomical models from 305 MRI volumes. In: 1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference. pp. 1813–1817 vol.3. doi: https://doi.org/10.1109/NSSMIC.1993.373602.

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A. R., Fox, P. T., Eickhoff, S. B., Yu, C., Jiang, T., 2016. The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. Cerebral Cortex 26 (8), 3508–26. doi: https://doi.org/10.1093/cercor/bhw157.

Fan, L., Wang, J., Zhang, Y., Han, W., Yu, C., Jiang, T., 2014. Connectivity-based parcellation of the human temporal pole using diffusion tensor imaging. Cerebral Cortex 24 (12), 3365–78. doi: https://doi.org/10.1093/cercor/bht196.

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., Lin, C. J., 2008. LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research 9 (9), 1871–1874. doi: https://doi.org/10.1145/1390681.1442794.

Field, A. S., Yen, Y. F., Burdette, J. H., Elster, A. D., 2000. False cerebral activation on BOLD functional MR images: study of low-amplitude motion weakly correlated to stimulus. American Journal of Neuroradiology, 21 (8), 1388–96.

Filippone, M., Marquand, A., Blain, C., Williams, S., Mourão Miranda, J., Girolami, M., 2012. Probabilistic prediction of neurological disorders with a statistical assessment of neuroimaging data modalities. The Annals of Applied Statistics 6 (4), 1883–1905. doi: https://doi.org/10.1214/12-AOAS562.

Filosa, J. A., Bonev, A. D., Straub, S. V., Meredith, A. L., Wilkerson, M. K., Aldrich, R. W., Nelson, M. T., 2006. Local potassium signaling couples neuronal activity to vasodilation in the brain. Nature Neuroscience 9, 1397–1403. doi: https://doi.org/10.1038/nn1779.

Forman, S., Cohen, J., Fitzgerald, M., Eddy, M., Mintum, M., Noll, D., 1995. Improved assessment of significant activation in functional magnetic resonace imaging (fMRI): use of a cluster. Magnetic Resonance in Medicine 33, 636–647. doi: https://doi.org/10.1002/mrm.1910330508.

Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. "Who" Is Saying "What"? Brain-Based Decoding of Human Voice and Speech. Science 322 (5903), 970–973. doi: https://doi.org/10.1126/science.1164318.

Fort, G., Lambert-Lacroix, S., 2005. Classification using partial least squares with penalized logistic regression. Bioinformatics 21 (7), 1104–11. doi: https://doi.org/10.1093/bioinformatics/bti114.

Frank, M. J., Badre, D., 2011. Mechanisms of Hierarchical Reinforcement Learning in Corticostriatal Circuits 1: Computational Analysis. Cerebral Cortex 22 (3), 509–526. doi: https://doi.org/10.1093/cercor/bhr114.

Friston, K., Worsley, K., Frackowiak, R., Mazziotta, J., Evans, A. C., 1994. Assessing the significance of focal activations using their spatial extent. Hum Brain Mapping 1 (3), 210–20. doi: https://doi.org/10.1002/hbm.460010306.

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M. D., Turner, R., 1998. Event-related fMRI: characterizing differential responses. NeuroImage 7 (1), 30–40. doi: https://doi.org/10.1006/nimg.1997.0306.

Friston, K. J., Holmes, A., Poline, J. B., Price, C. J., Frith, C. D., 1996. Detecting activations in PET and fMRI: levels of inference and power. NeuroImage 40, 223–35. doi: https://doi.org/10.1006/nimg.1996.0074.

Friston, K. J., Holmes, A. P., Price, C. J., Büchel, C., Worsley, K. J., 1999. Multisubject fMRI studies and conjunction analyses. NeuroImage 10 (4), 385–96. doi: https://doi.org/10.1006/nimg.1999.0484.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., Frackowiak, R. S. J., 1995. Statistical parametric maps in functional imaging: a general linear approach. Human Brain Mapping 2 (3), 189–210. doi: https://doi.org/10.1002/hbm.460020402.

Gabay, A. S., Radua, J., Kempton, M. J., Mehta, M. A., 2014. The ultimatum game and the brain: A meta-analysis of neuroimaging studies. Neuroscience & Biobehavioral Reviews 47, 549 – 558. doi: http://dx.doi.org/10.1016/j.neubiorev.2014.10.014.

Gaertig, C., Moser, A., Alguacil, S., Ruz, M., 2012. Social information and economic decision-making in the ultimatum game. Frontiers in Neuroscience 6 (103). doi: https://doi.org/10.3389/fnins.2012.00103.

Georgiopoulos, C., Witt, S. T., Haller, S., Dizdar, N., Zachrisson, H., Engström, Larsson, E. M., 2018. Olfactory fMRI: Implications of stimulation length and repetition time. Chemical Senses 43 (6), 389–398. doi: https://doi.org/10.1093/chemse/bjy025.

Ghanbari, Y., Smith, A. R., Schultz, R. T., Verma, R., 2014. Identifying group discriminative and age regressive sub-networks from dti-based connectivity via a unified framework of non-negative matrix factorization and graph embedding. Medical Image Analysis 18 (8), 1337 – 1348. doi: https://doi.org/10.1016/j.media.2014.06.006.

Gilbert, S. J., Fung, H., 2018. Decoding intentions of self and others from fMRI activity patterns. NeuroImage 172, 278 – 290. doi: https://doi.org/10.1016/j.NeuroImage.2017.12.090.

Gilron, R., Rosenblatt, J., Koyejo, O., Poldrack, R. A., Mukamel, R., 2017. What's in a pattern? Examining the type of signal multivariate analysis uncovers at the group level. NeuroImage 146, 113–120. doi: https://doi.org/10.1016/j.NeuroImage.2016.11.019.

Goebel, R., Esposito, F., Formisano, E., 2006. Analysis of FIAC data with brainVoyager QX: From a single-subject to cortically aligned group GLM analysis and self-organizing group ICA. Human Brain Mapping 27 (5), 392–401. doi: https://doi.org/10.1002/hbm.20249.

Golland, P., Fischl, B., 2003. Permutation tests for classification: towards statistical significance in image-based studies. In: Information Processing in Medical Imaging. Vol. 18. pp. 330–341. doi: https://doi.org/10.1007/978-3-540-45087-0_28.

González-García, C., Arco, J. E., Palenciano, A. F., Ramírez, J., Ruz, M., 2017. Encoding, preparation and implementation of novel complex verbal instructions. NeuroImage 148, 264–273. doi: https://doi.org/10.1016/j.NeuroImage.2017.01.037.

González-García, C., Mas-Herrero, E., de Diego-Balaguer, R., Ruz, M., 2016. Task-specific preparatory neural activations in low-interference contexts. Brain Structure & Function 8, 3997–4006. doi: https://doi.org/10.1007/s00429-015-1141-5.

Gordon, E. M., Laumann, T. O., Gilmore, A. W., Newbold, D. J., Greene, D. J., Berg, J. J., Ortega, M., Hoyt-Drazen, C., Gratton, C., Sun, H., Hampton, J. M., Coalson, R. S., Nguyen, A. L., McDermott, K. B., Shimony, J. S., Snyder, A. Z., Schlaggar, B. L., Petersen, S. E., Nelson, S. M., Dosenbach, N. U. F., 2017. Precision functional mapping of individual human brains. Neuron 95 (4), 791–807. doi: https://doi.org/10.1016/j.neuron.2017.07.011.

Górriz, J. M., Ramírez, J., Illán, I. A., Martínez-Murcia, F. J., Segovia, F., Salas-Gonzalez, D., Ortiz, A., 2017. Case-based statistical learning applied to SPECT image classification. Vol. 10134. pp. 10134 – 10134 – 6. doi: https://doi.org/10.1117/12.2253853.

Grover, V. P., Tognarelli, J. M., Crossey, M. M., Cox, I. J., Taylor-Robinson, S. D., McPhail, M. J., 2015. Magnetic resonance imaging: principles and techniques: lessons for clinicians. Journal of Clinical and Experimental Hepatology 5 (3), 246–55. doi: https://doi.org/10.1016/j.jceh.2015.08.001.

Guggenmos, M., Sterzer, P., Cichy, R. M., 2018. Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. NeuroImage 173, 434 – 447. doi: https://doi.org/10.1016/j.NeuroImage.2018.02.044.

Guntupalli, J. S., Hanke, M., Halchenko, Y. O., Connolly, A. C., Ramadge, P. J., Haxby, J. V., 2016. A Model of Representational Spaces in Human Cortex. Cerebral Cortex 26 (6), 2919–2934. doi: https://doi.org/10.1093/cercor/bhw068.

Haller, S., Bartsch, A. J., 2009. Pitfalls in fMRI. European Radiology 19 (11), 2689–706. doi: https://doi.org/10.1007/s00330-009-1456-9.

Han, H., Glenn, A. L., 2018. Evaluating methods of correcting for multiple comparisons implemented in SPM12 in social neuroscience fMRI studies: an example from moral psychology. Social Neuroscience 13 (3), 257–267. doi: https://doi.org/10.1080/17470919.2017.1324521.

Handwerker, D. A., Ollinger, J. M., D'Esposito, M., 2004. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. NeuroImage 21 (4), 1639–51. doi: https://doi.org/10.1016/j.NeuroImage.2003.11.029.

Hanso, S. J., Halchenko, Y. O., 2008. Brain reading using full brain support vector machines for object recognition: there is no "face" identification area. Neural Computation 20 (2), 468–503. doi: https://doi.org/10.1162/neco.2007.09-06-340.

Hanson, S., Matsuka, T., V Haxby, J., 2004. Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: Is there a "face" area? NeuroImage 23, 156–66. doi: https://doi.org/10.1016/j.NeuroImage.2004.05.020.

Harrison, S. A., Tong, F., 03 2009. Decoding reveals the content of visual working memory in early visual areas. Nature 458, 632–5. doi: https://doi.org/10.1038/nature07832.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage 87, 96–110. doi: https://doi.org/10.1016/j.NeuroImage.2013.10.067.

Hauswald, A., Tucciarelli, R., Lingnau, A., 2018. MEG adaptation reveals action representations in posterior occipitotemporal regions. Cortex 103, 266 – 276. doi: https://doi.org/10.1016/j.cortex.2018.03.016.

Haxby, J. V., Connolly, A., Guntupalli, J. S., 2014. Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. Annual Review of Neuroscience 37, 435–456. doi: http://dx.doi.org/10.1146/annurev-neuro-062012-170325.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293 (5539), 2425–30. doi: https://doi.org/10.1126/science.1063736.

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Hanke, M., Ramadge, P. J., 2011. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72, 404–416. doi: https://doi.org/10.1016/j.neuron.2011.08.026.

Hayasaka, S., Nichols, T. E., 2003. Validating cluster size inference: random field and permutation methods. NeuroImage 20 (4), 2343–56. doi: https://doi.org/10.1016/j.NeuroImage.2003.08.003.

Haynes, J., Rees, G., 2006. Decoding mental states from brain activity in humans. Reviews in the Neurosciences. 7, 523–534. doi: https://doi.org/10.1038/nrn1931.

Haynes, J.-D., 2015. A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. Neuron 87 (2), 257 – 270. doi: https://doi.org/10.1016/j.neuron.2015.05.025.

Haynes, J. D., Rees, G., 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nature Neuroscience 8 (5), 686–91. doi: https://doi.org/10.1038/nn1445.

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., Passingham, R. E., 2007. Reading hidden intentions in the human brain. Current Biology 17 (4), 323 – 328. doi: https://doi.org/10.1016/j.cub.2006.11.072.

Hebart, M., Görgen, K., Haynes, J., 2015. The decoding toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. Frontiers in Neuroinformatics 8, Article 88. doi: https://doi.org/10.3389/fninf.2014.00088.

Hebart, M. N., Baker, C. I., 2017. Deconstructing multivariate decoding for the study of brain function. NeuroImage 15 (180), 4–18. doi: https://doi.org/10.1016/j.NeuroImage.2017.08.005.

Hebart, M. N., Schriever, Y., Donner, T. H., Haynes, J.-D., 2016. The Relationship between Perceptual Decision Variables and Confidence in the Human Brain. Cerebral Cortex 26 (1), 118–130. doi: https://doi.org/10.1093/cercor/bhu181.

Heller, R., Stanley, D., Yekutieli, D., Rubin, N., Benjamini, Y., 2006. Cluster-based analysis of fMRI data. NeuroImage 33 (2), 599–608. doi: https://doi.org/10.1016/j.NeuroImage.2006.04.233.

Hendriks, M., Daniels, N., Pegado, F., Op de Beeck, H., 2017. The Effect of Spatial Smoothing on Representational Similarity in a Simple Motor Paradigm. Frontiers in Neurology 8 (222), 355–361. doi: https://doi.org/10.3389/fneur.2017.00222.

Henson, R., 2005. Design efficiency in fMRI.
URL http://imaging.mrc-cbu.cam.ac.uk/imaging/DesignEfficiency#VII._Should_I_treat_my_trials_as_events_or_epochs_.3F

Henson, R., Friston, K., 2007. Convolution models for fMRI. Academic Press. ISBN = 978-012-37-2560-8, doi: https://doi.org/10.1016/B978-012372560-8/50014-0.

Herreras, O., 2016. Local field potentials: myths and misunderstandings. Frontiers in Neural Circuits 10 (101). doi: https://doi.org/10.3389/fncir.2016.00101.

Hillman, E. M. C., 2014. Coupling mechanism and significance of the BOLD signal: A status report. Annual Review of Neuroscience 37, 161–181. doi: https://doi.org/10.1146/annurev-neuro-071013-014111.

Hoeft, F., McCanliss, B. D., Black, J. M., Gantman, A., Zakerani, N., Hulme, C., Lyytinen, H., Whitfield-Gabriel, S., Glover, G. H., Reiss, A. L., Gabrieli, J. D., 2011. Neural systems predicting long-term outcome in dyslexia. Vol. 108. pp. 361–6. doi: https://doi.org/10.1073/pnas.1008950108.

Hoge, R. D., Atkinson, J., Gill, B., Crelier, G. R., Marrett, S., Pike, G. B., 1999. Linear coupling between cerebral blood flow and oxygen consumption in activated human cortex. Neurobiology 96 (16), 9403–9408. doi: https://doi.org/10.1073/pnas.96.16.9403.

Holden, M., 2008. A review of geometric transformations for nonrigid body registration. In: IEEE Transactions on Medical Imaging. pp. 111–128 vol.27. doi: https://doi.org/10.1109/TMI.2007.904691.

Holmes, A. P., Blair, R. C., Watson, J. D. G., Ford, I., 1996. Nonparametric analysis of statistic images from functional mapping experiments. Journal of Cerebral Blood Flow & Metabolism 16, 7–22. doi: https://doi.org/10.1097/00004647-199601000-00002.

Hu, X., Yacoub, E., 2012. The story of the initial dip in fMRI. NeuroImage 62, 1103–1108. doi: https://doi.org/10.1016/j.NeuroImage.2012.03.005.

Huettel, S. A., Song, A. W., McCarthy, G., 2009. Functional Magnetic Resonance Imaging. Sinauer Associates.

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., Smith, S. M., 2012. FSL. NeuroImage 62 (2), 782–790. doi: https://doi.org/10.1016/j.NeuroImage.2011.09.015.

Johnstone, T., Ores Walsh, K. S., Greischar, L. L., Alexander, A. L., Fox, A. S., Davison, R. J., Oakes, T. R., 2006. Motion correction and the use of covariates in multiple-subject fMRI analysis. Human Brain Mapping 27 (10), 779–88. doi: https://doi.org/10.1002/hbm.20219.

Joliot, M., Jobard, G., Naveau, M., Delcroix, N., Petit, L., Zago, L., Crivello, F., Mellet, E., Mazoyer, B., Tzourio-Mazoyer, N., 2015. AICHA: An atlas of intrinsic connectivity of homotopic areas. Journal of Neuroscience Methods 30 (254), 46–59. doi: https://doi.org/10.1016/j.jneumeth.2015.07.013.

Jolliffe, I., 2002. Principal component analysis. Springer Verlag. doi: https://doi.org/10.1007/b98835.

Kahnt, T., 2018. A decade of decoding reward-related fmri signals and where we go from here. NeuroImage 180, 324 – 333. doi: https://doi.org/10.1016/j.NeuroImage.2017.03.067.

Kahnt, T., Heinzle, J., Park, S. Q., Haynes, J. D., 2010. The neural code of reward anticipation in human orbitofrontal cortex. Proceedings of the National Academy of Sciences of the United States of America 107, 6010–6015. doi: https://doi.org/10.1073/pnas.0912838107.

Kamitami, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nature Neuroscience 8 (5), 679–85. doi: https://doi.org/10.1038/nn1444.

Kanwisher, N., Yovel, G., 2006. The fusiform face area: a cortical region specialized for the perception of faces. Philosophical Transactions of the Royal Society B: Biological Sciences 361 (1476), 2109–2128. doi: http://dx.doi.org/10.1098/rstb.2006.1934.

Khedher, L., Illán, I. A., Górriz, J. M., Ramírez, J., Brahim, A., Meyer-Baese, A., 2017. Independent Component Analysis-Support Vector Machine-Based Computer-Aided Diagnosis System for Alzheimer's with Visual Support. International Journal of Neural Systems 27 (3), 1650050. doi: https://doi.org/10.1142/S0129065716500507.

Klöppel, S., Abdulkadir, A., Jack, C. R. J., Koutsouleris, N., Mourão Miranda, J., Vemuri, P., 2012. Diagnostic neuroimaging across diseases. NeuroImage 61 (2), 457–63. doi: https://doi.org/10.1016/j.NeuroImage.2011.11.002.

Kolb, B., Whishaw, I. Q., 2003. Fundamentals of Human Neuropsychology. Worth Publishers. ISBN: 978-071-67-5300-1.

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., 2006. Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. Neural Networks 19 (2), 122 – 134. doi: https://doi.org/10.1016/j.neunet.2006.01.002.

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America 103 (10), 3863–8. doi: https://doi.org/10.1038/nn.2303.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., Bandettini, P. A., 2008. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60 (6), 1126 – 1141. doi: https://doi.org/10.1016/j.neuron.2008.10.043.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., Baker, C. I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nature Neurosci 12 (5), 535–40. doi: https://doi.org/10.1038/nn.2303.

Ku, S.-P., Gretton, A., Macke, J., Logothetis, N. K., 2008. Comparison of pattern recognition methods in classifying high-resolution BOLD signals obtained at high magnetic field in monkeys. Magnetic Resonance Imaging 26 (7), 1007 – 1014. doi: https://doi.org/10.1016/j.mri.2008.02.016.

Kuzmanovic, B., Rigoux, L., Tittgemeyer, M., 2018. Influence of vmPFC on dmPFC Predicts Valence-Guided Belief Formation. The Journal of Neuroscience 38 (37), 7996–8010. doi: https://doi.org/10.1523/JNEUROSCI.0266-18.2018.

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X., 2005. Support vector machines for temporal classification of block design fMRI data. NeuroImage 26 (2), 317–29. doi: https://doi.org/10.1016/j.NeuroImage.2005.01.048.

Lanckriet, G. R. G., Cristianini, N., Bartlett, P., El Ghaoui, L., Jordan, M. I., 2004. Learning the Kernel Matrix with Semidefinite Programming. Journal of Machine Learning Research 5, 27–72.

Lauritzen, M., Mathiesen, C., Schaefer, K., Thomsen, K. J., 2001. Neuronal inhibition and excitation, and the dichotomic control of brain hemodynamic and oxygen responses. NeuroImage 62 (2), 1040–50. doi: https://doi.org/10.1016/j.NeuroImage.2012.01.040.

Lee, Y.-S., Janata, P., Frost, C., Hanke, M., Granger, R., 2011. Investigation of melodic contour processing in the brain using multivariate pattern-based fMRI. NeuroImage 57 (1), 293 – 300. doi: https://doi.org/10.1016/j.NeuroImage.2011.02.006.

Lee, Y. S., Zreik, J. T., Hamilton, R. H., 2017. Patterns of neural activity predict picture-naming performance of a patient with chronic aphasia. Neuropsychologia 94, 52 – 60. doi: https://doi.org/10.1016/j.neuropsychologia.2016.11.010.

Lin, W., Hu, Y., Tsai, C., 2012. Machine Learning in Financial Crisis Prediction: A Survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42 (4), 421–436. doi: https://doi.org/10.1109/TSMCC.2011.2170420.

Lindquist, K., Wager, T., Kober, H., Bliss-Moreau, E., Barrett, L., 2012. The Brain Basis of Emotion: A Meta-Analytic Review. The Behavioral and Brain Sciences 35, 121–43. doi: http://dx.doi.org/10.1017/S0140525X11000446.

Lindquist, K. A., Satpute, A. B., Wager, T. D., Weber, J., Barrett, L. F., 2015. The brain basis of positive and negative affect: evidence from a meta-analysis of the human neuroimaging literature. Cerebral Cortex 26 (5), 1910–1922. doi: https://doi.org/10.1093/cercor/bhv001.

Lindquist, M. A., Loh, J. M., Atlas, L. Y., Waget, T. D., 2009. Modeling the hemodynamic response function in fMRI: Efficiency, bias and mis-modeling. NeuroImage 45 (1), S187–S198. doi: https://doi.org/10.1016/j.NeuroImage.2008.10.065.

Lindquist, M. A., Mejia, A., 2015. Zen and the Art of Multiple Comparisons. Psychosomatic Medicine 77 (2), 114–125. doi: https://doi.org/10.1097/PSY.0000000000000148.

Lindquist, M. A., Wager, T. D., 2008. Spatial smooghing in fMRI using prolate spheroidal wave functions. Human Brain Mapping 29 (11), 1276–87. doi: https://doi.org/10.1002/hbm.20475.

Liu, H., Qin, W., Qi, H., Jiang, T. an Yu, C., 2013. Connectivity- based parcellation of the human frontal pole with diffusion tensor imaging. Journal of Neuroscience 3, 6782–6790. doi: https://doi.org/10.1523/JNEUROSCI.4882-12.2013.

Liu, H., Stufflebeam, S. M., Sepulcre, J., Hedden, T., Buckner, R. L., 2009. Evidence from intrinsic activity that asymmetry of the human brain is controlled by multiple factors. Proceedings of the National Academy of Sciences of the United States of America 106 (48), 20499–503. doi: https://doi.org/10.1073/pnas.0908073106.

Liu, T. T., 2016. Noise contributions to the fMRI signal: An overview. NeuroImage 143, 141–151. doi: http://dx.doi.org/10.1016/j.NeuroImage.2016.09.008.

Logothetis, N. K., 2003. The underpinnings of the BOLD functional magnetic resonance imaging signal. Journal of Neuroscience 23 (10), 3963–3971. doi: https://doi.org/10.1523/JNEUROSCI.23-10-03963.2003.

Logothetis, N. K., 2004. Interpreting the BOLD signal. Annual Review of Physiology 66, 735–69. doi: https://doi.org/10.1146/annurev.physiol.66.082602.092845.

Logothetis, N. K., 2014. Neurovascular uncoupling: Much Ado about nothing. Frontiers in Neuroenergetics 2 (2). doi: https://doi.org/10.3389/fnene.2010.00002.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., Oeltermann, A., 2001. Neurophysiological investigation of the basis of the fMRI signal. Nature 412 (6843), 150–7. doi: https://doi.org/10.1038/35084005.

Loose, L. S., Wisniewski, D., Rusconi, M., Goschke, T., Haynes, J. D., 2017. Switch-Independent Task Representations in Frontal and Parietal Cortex. Journal of Neuroscience 37 (33), 8033–8042. doi: https://doi.org/10.1523/JNEUROSCI.3656-16.2017.

Maclaren, J., Herbst, M., Speck, O., Zaitsev, M., 2013. Prospective motion correction in brain imaging: a review. Magnetic Resonance in Medicine 69 (3), 621–36. doi: https://doi.org/10.1002/mrm.24314.

Mahmoudi, A., Takerkart, S., Regragui, F., Boussaoud, D., Brovelli, A., 12 2012. Review Article Multivoxel Pattern Analysis for fMRI Data: A Review. Computational and Mathematical Methods in Medicine 14, 961257. doi: https://doi.org/10.1155/2012/961257.

Marquand, A. F., Brammer, M., Williams, S. C., Doyle, O. M., 2014. Bayesian multi-task learning for decoding multi-subject neuroimaging data. NeuroImage 15 (92), 298–311. doi: https://doi.org/10.1016/j.NeuroImage.2014.02.008.

Martinez-Murcia, F. J., Górriz, J. M., Ramírez, J., 2017. Feature Extraction. American Cancer Society. pp. 1–9. doi: https://doi.org/10.1002/047134608X.W5506.pub2.

Martinez-Murcia, F. J., Górriz, J. M., Ramírez, J., Ortiz, A., 2016. A Structural Parametrization of the Brain Using Hidden Markov Models-Based Paths in Alzheimer's Disease. International Journal of Neural Systems 26 (7), 1650024. doi: https://doi.org/10.1142/S0129065716500246.

Martínez-Ramón, M., Koltchinskii, V., Heileman, G. L., Posse, S., 2006. fMRI pattern classification using neuroanatomically constrained boosting. NeuroImage 31 (3), 1129 – 1141. doi: https://doi.org/10.1016/j.NeuroImage.2006.01.022.

Masteron, R. A., Harvey, A. S., Archer, J. S., Lillyhite, L. M., Abbott, D. F., Scheffer, I. E., Jackson, G. D., 2010. Focal epileptiform spikes do not show a canonical BOLD response in patients with benign rolandic epilepsy (BECTS). NeuroImage 51 (1). doi: https://doi.org/10.1016/j.NeuroImage.2010.01.109.

Mathias, E. J., Kenny, A., Plank, M. J., David, T., 2018. Integrated models of neurovascular coupling and BOLD signals: Responses for varying neural activations. NeuroImage 174, 69–86. doi: https://doi.org/10.1016/j.NeuroImage.2018.03.010.

Mathiesen, C., Caesar, K., Akgören, N., Lauritzen, M., 1998. Modification of activity-dependent increases of cerebral blood flow by excitatory synaptic activity and spikes in rat cerebellar cortex. Journal of Physiology 512, 555–66. doi: https://doi.org/10.1111/j.1469-7793.1998.555be.x.

Mattia, M., Ferraina, S., Del Giudice, P., 2010. Dissociated multi-unit activity and local field potentials: A theory inspired analysis of a motor decision task. NeuroImage 52, 812–823. doi: https://doi.org/10.1016/j.NeuroImage.2010.01.063.

McKinnon, G. C., 1993. Ultrafast interleaved gradient-echo-planar imaging on a standard scanner. Magnetic Resonance in Medicine 30 (5), 609–616. doi: https://doi.org/10.1002/mrm.1910300512.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Mullers, K. R., 1999. Fisher discriminant analysis with kernels. In: Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No.98TH8468). pp. 41–48.

Min, J.-H., Lee, Y.-C., 2005. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Systems with Applications 28 (4), 603–614. doi: https://doi.org/10.1016/j.eswa.2004.12.008.

Misaki, M., Kim, Y., Bandettini, P., Kriegeskorte, N., 2010. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. NeuroImage 53 (1), 103–18. doi: https://doi.org/10.1016/j.NeuroImage.2010.05.051.

Misaki, M., Wen-Ming, L., Bandettini, P., 2013. The effect of spatial smoothing on fMRI decoding of columnar-level organization with linear support vector machine. Journal of Neuroscience Methods. 212 (22). doi: https://doi.org/10.1016/j.jneumeth.2012.11.004.

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M., Newman, S., 2004. Learning to Decode Cognitive States from Brain Images. Machine Learning 57 (1), 145–175. doi: https://doi.org/10.1023/B:MACH.0000035475.85309.1b.

Mohanty, R., Sinha, A. M., Remsik, A. B., Dodd, K. C., Young, B. M., Jacobson, T., McMillan, M., Thoma, J., Advani, H., Nair, V. A., Kang, T. J., Caldera, K., Edwards, D. F., Williams, J. C., Prabhakaran, V., 2018. Machine Learning Classification to Identify the Stage of Brain-Computer Interface Therapy for Stroke Rehabilitation Using Functional Connectivity. Frontiers in Neuroscience 12, 353. doi: https://doi.org/10.3389/fnins.2018.00353.

Molloy, E. R., Meyerand, M. E., Rasmus, M. B., 2014. The influence of spatial resolution and smoothing on the detectability of resting-state and task fMRI. NeuroImage 86, 221–230. doi: https://doi.org/10.1016/j.NeuroImage.2013.09.001.

Monti, M. M., 2011. Statistical analysis of fMRI time-series: a critical review of the GLM approach. Frontiers in Human Neuroscience 5 (28). doi: https://dx.doi.org/10.3389%2Ffnhum.2011.00028.

Moser, A., Gaertig, C., Ruz, M., 2014. Social information and personal interests modulate neural activity during economic decision-making. Frontiers in Human Neuroscience 8, 31. doi: https://doi.org/10.3389/fnhum.2014.00031.

Mosso, A., 1881. Ueber denkreislauf des Blutes im Menschlichen Gehirn. Verlag von Veit & Company.

Mourão-Miranda, J., Bokde, A. L., Born, C., Hampel, H., Stetter, M., 2005. Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. NeuroImage 28 (4), 980–95. doi: https://doi.org/10.1016/j.NeuroImage.2005.06.070.

Mourão-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., Brammer, M., 2006. The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. NeuroImage 33 (4), 1055 – 1065. doi: https://doi.org/10.1016/j.NeuroImage.2006.08.016.

Mumford, J. A., Davis, T., Poldrack, R. A., 2014. The impact of study design on pattern estimation for single-trial multivariate pattern analysis. NeuroImage 103 (Supplement C), 130 – 138. doi: https://doi.org/10.1016/j.NeuroImage.2014.09.026.

Mumford, J. A., Turner, B. O., Ashby, F. G., Poldrack, R. A., 2012. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. NeuroImage 59 (3), 2636 – 2643. doi: https://doi.org/10.1016/j.NeuroImage.2014.09.026.

Mur, M., Bandettini, P. A., Kriegeskorte, N., 2009. Revealing representational content with pattern-information fMRI-an introductory guide. Society of Cognitive Affective Neuroscience 4 (1), 101–109. doi: https://doi.org/10.1093/scan/nsn044.

Mur, M., Ruff, D. A., Bodurka, J., Bandettini, P. A., Kriegeskorte, N., 2010. Face-Identity Change Activation Outside the Face System: "Release from Adaptation" May Not Always Indicate Neuronal Selectivity. Cerebral Cortex 20 (9), 2027–2042. doi: https://doi.org/10.1093/cercor/bhp272.

Mwangi, B., Hasan, K. M., Soares, J. C., 2013. Prediction of individual subject's age across the human lifespan using diffusion tensor imaging: A machine learning approach. NeuroImage 75, 58 – 67. doi: https://doi.org/10.1016/j.NeuroImage.2013.02.055.

Mwangi, B., Tian, T. S., Soares, J. C., 2014. A review of feature reduction techniques in neuroimaging. Neuroinformatics 12 (2), 229–44. doi: https://doi.org/10.1007/s12021-013-9204-3.

Nakagawa, M., Nakaura, T., Namimoto, T., Iyama, Y., Kidoh, M., Hirata, K., Nagayama, Y., Oda, S., Sakamoto, F., Shiraishi, S., Yamashita, Y., 2019. A multiparametric MRI-based machine learning to distinguish between uterine sarcoma and benign leiomyoma: comparison with 18F-FDG PET/CT. Clinical Radiology 74 (2), 167.e1 – 167.e7. doi: https://doi.org/10.1016/j.crad.2018.10.010.

Nelissen, N., Stokes, M., Nobre, A. C., Rushworth, M. F. S., 2013. Frontal and Parietal Cortical Interactions with Distributed Visual Representations during Selective Attention and Action Selection. Journal of Neuroscience 33 (42), 16443–16458. doi: https://doi.org/10.1523/JNEUROSCI.2625-13.2013.

Nichols, T., Hayasaka, S., 2003. Controlling the family wise error rate in functional neuroimaging: a comparative review. Statistical Methods in Medical Research 12, 419–446. doi: https://doi.org/10.1191/0962280203sm341ra.

Nichols, T. E., 2012. Multiple testing corrections, nonparametric methods, and random field theory. NeuroImage 62 (2), 811–5. doi: https://doi.org/10.1016/j.NeuroImage.2012.04.014.

Nichols, T. E., Holmes, A. P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Human Brain Mapping 15 (1), 1–25. doi: https://doi.org/10.1002/hbm.1058.

Nieto-Castanon, A., Ghosh, S. S., Tourville, J. A., Guenther, F. H., 2003. Region of interest based analysis of functional imaging data. NeuroImage 19 (4), 1303–16.

Norman, K. A., Polyn, S. M., Detre, G. J., Haxby, J. V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences 10 (9), 424–30. doi: https://doi.org/10.1016/j.tics.2006.07.005.

Nurse, E. S., Karoly, P. J., Grayden, D. B., Freestone, D. R., 2015. A Generalizable Brain-Computer Interface (BCI) Using Machine Learning for Feature Discovery. PLOS ONE 10 (6), 1–22. doi: https://doi.org/10.1371/journal.pone.0131328.

Ogawa, S., Lee, T. M., 1990. Magnetic resonance imaging of blood vessels at high fields: in vivo and in vitro measurements and image simulations. Magnetic Resonance in Medicine 16 (1), 9–18. doi: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.1910160103.

Ontivero-Ortega, M., Lage-Castellanos, A., Valente, G., Goebel, R., Valdes-Sosa, M., 2017. Fast Gaussian Naïve Bayes for searchlight classification analysis. NeuroImage 163, 471–479. doi: http://dx.doi.org/10.1016/j.NeuroImage.2017.09.001.

Oosterhof, N. N., Wiestler, T., Downing, P. E., Diedrichsen, J., 2011. A comparison of volume-based and surface-based multi-voxel pattern analysis. NeuroImage 56, 593–600. doi: http://dx.doi.org/10.1016/j.NeuroImage.2010.04.270.

Orchard, J., Greif, C., Golub, G. H., Bjornson, B., Atkins, M. S., 2003. Simultaneous registration and activation detection for fMRI 22 (11), 1427–1435.

O'Toole, A. J., Jiang, F., Abdi, H., Penard, N., Dunlop, J. P., Parent, M. A., 2007. Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. NeuroImage 19 (11), 1735–1752. doi: https://doi.org/10.1162/jocn.2007.19.11.1735.

Pagnozzi, A. M., Conti, E., Calderoni, S., Fripp, J., Rose, S. E., 2018. A systematic review of structural MRI biomarkers in autism spectrum disorder: A machine learning perspective. International Journal of Developmental Neuroscience 71, 68 – 82. doi: https://doi.org/10.1016/j.ijdevneu.2018.08.010.

Palenciano, A. F., González-García, C., Arco, J. E., Ruz, M., 2018. Transient and sustained control mechanisms supporting novel instructed behavior. Cerebral cortex 4 (11), 863–72. doi: https://doi.org/10.1093/cercor/bhy273.

Parker, D., Xueqing, L., Qolamreza, R. R., 2017. Optical slice timing correction and its interaction with fMRI parameters and artifacts. Medical Image Analysis 35, 434–445. doi: http://dx.doi.org/10.1016/j.media.2016.08.006.

Penny, W., Friston, K., Ashburner, J., Kiebel, S., Nichols, T., 2006. Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press. ISBN = 978-008-04-6650-7.

Pereira, F., Botvinick, M., 2011. Information mapping with pattern classifiers: a comparative study. NeuroImage 56 (5). doi: https://doi.org/10.1016/j.NeuroImage.2010.05.026.

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45 (1 Suppl), S199–209. doi: https://doi.org/10.1016/j.NeuroImage.2008.11.007.

Plant, C., Teipel, S. J., Oswald, A., Böhm, C., Meindi, T., Mourao-Miranda, J., Bokde, A. W., Hampel, H., Ewers, M., 2010. Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. NeuroImage 50 (1), 162–74. doi: https://doi.org/10.1016/j.NeuroImage.2009.11.046.

Poldrack, R. A., 2007. Region of interest analysis for fMRI. Social Cognitive and Affective Neuroscience 2 (1), 67–70. doi: https://doi.org/10.1093/scan/nsm006.

Poldrack, R. A., Mumford, J. A., Nichols, T. E., 2011. Statistical modeling: single subject analysis. Cambridge University Press. ISBN = 978-052-15-1766-9, doi: https://doi.org/10.1017/CBO9780511895029.

Poline, J. B., Brett, M., 2012. The general linear model and fMRI: Does love last forever? NeuroImage 62, 871–880. doi: https://doi.org/10.1016/j.NeuroImage.2012.01.133.

Poustchi-Amin, M., Mirowitz, S. A., Brown, J. J., McKinstry, R. C., Li, T., 2001. Principles and Applications of Echo-planar Imaging: A Review for the General Radiologist. RadioGraphics 21 (3), 767–779. doi: https://doi.org/10.1148/radiographics.21.3.g01ma23767.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., Petersen, S. E., 2012. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. NeuroImage 59 (3), 2142–2154. doi: https://doi.org/10.1016/j.NeuroImage.2011.10.018.

Pruessmann, K. P., Weiger, M., Börnet, P., Boesiger, P., 2001. Advances in sensitivity encoding with arbitrary $k$-space trajectories. Magnetic Resonance in Medicine 46, 638–651. doi: https://doi.org/10.1002/mrm.1241.

Qiao, L., Zhang, L., Chen, A., Egner, T., 2017. Dynamic Trial-by-Trial Recoding of Task-Set Representations in the Frontoparietal Cortex Mediates Behavioral Flexibility. Journal of Neuroscience 37 (45), 11037–11050. doi: https://doi.org/10.1523/JNEUROSCI.0935-17.2017.

Qureshi, M. N. I., Oh, J., Min, B., Jo, H. J., Lee, B., 2017. Multi-modal, Multi-measure, and Multi-class Discrimination of ADHD with Hierarchical Feature Extraction and Extreme Learning Machine Using Structural and Functional Brain MRI. Frontiers in Systems Neuroscience 11, 157. doi: https://doi.org/10.3389/fnhum.2017.00157.

Raichle, M. E., 1983. Positron emission tomography. Annual Review of Neuroscience 6, 249–267. doi: http://dx.doi.org/10.1146/annurev.ne.06.030183.001341.

Raichle, M. E., 2009. A brief history of human brain mapping. Trends in Neurosciences 32 (2), 118–126. doi: https://doi.org/10.1016/j.tins.2008.11.001.

Rakotomamonjy, A., Bach, F. R., Stéphane, C., Grandvalet, Y., 2008. SimpleMKL. Journal of Machine Learning Research 9, 2491–2521.

Rasmussen, P. M., Madsen, K. H., Lund, T. E., Hansen, L. K., 2011. Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. NeuroImage 55 (3), 1120 – 1131. doi: https://doi.org/10.1016/j.NeuroImage.2010.12.035.

Rauch, A., Rainer, G., Logothetis, N. K., 2008. The effect of a serotonin-induced dissociation between spiking and perisynaptic activity on BOLD functional MRI. Proceedings of the National Academy of Sciences of the United States of America 105, 6759–6764. doi: https://doi.org/10.1073/pnas.0800312105.

Reddy, L., Tsuchiya, N., Serre, T., 2010. Reading the mind's eye: Decoding category information during mental imagery. NeuroImage 50 (2), 818 – 825. doi: https://doi.org/10.1016/j.NeuroImage.2009.11.084.

Rissman, J., Gazzaley, A., D'Esposito, M., 2004. Measuring functional connectivity during distinct stages of a cognitive task. NeuroImage 23 (2), 752 – 763. doi: https://doi.org/10.1016/j.NeuroImage.2004.06.035.

Röther, J., Knab, R., Hamzei, F., Fiehler, J., Reichenbach, J. R., Büchel, C., Weiller, C., 2002. Negative Dip in BOLD fMRI Is Caused by Blood Flow–Oxygen Consumption Uncoupling In Humans. NeuroImage 15 (1), 98 – 102. doi: https://doi.org/10.1006/nimg.2001.0965.

Roy, C. S., Sherrington, C. S., 1890. On the regulation of the blood supply of the brain. The Journal of Physiology 11, 85–108.

Rundle, M. M., Coch, D., Connolly, A. C., Granger, R. H., 2018. Dissociating frequency and animacy effects in visual word processing: An fMRI study. Brain and Language 183, 54 – 63. doi: https://doi.org/10.1016/j.bandl.2018.05.005.

Rzedzian, R. R., 1987. High speed, High resolution, spin-echo imaging by Mosaic scan and MESH. In: Sixth Annual Meeting of the Society of Magnetic Resonance in Medicine, SMRM. p. 51.

Sakai, K., 2008. Task set and prefrontal cortex. Annual Review of Neuroscience 31, 219–245. doi: https://doi.org/10.1146/annurev.neuro.31.060407.125642.

Salvatore, C., Cerasa, A. ana Castiglioni, I., Callivanone, F., Augimeri, A. an Lopez, M. a. A. G., Morelli, M., Gilardi, M. C., Quattrone, A., 2014. Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy. Journal of Neuroscience Methods 30 (222), 230–7. doi: https://doi.org/10.1016/j.jneumeth.2013.11.016.

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X. N., Holmes, A. J., Eickhoff, S. B., Yeo, B. T. T., 2018. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cerebral Cortex 28 (9), 3095–3114. doi: https://doi.org/10.1093/cercor/bhx179.

Schölkopf, B., 2001. The Kernel Trick for Distances. MIT Press. pp. 301–307.
URL http://papers.nips.cc/paper/1862-the-kernel-trick-for-distances.pdf

Scholkopf, B., Smola, A. J., 2001. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA. ISBN: 0262194759.

Schreiber, K., Krekelberg, B., 2013. The Statistical Analysis of Multi-Voxel Patterns in Functional Imaging. PLoS One 8 (7), e69328. doi: https://doi.org/10.1371/journal.pone.0069328.

Schrouff, J., Cremers, J., Garraux, G., Balsassarre, L., Mourão Miranda, J., Phillips, C., 2013a. Localizing and Comparing Weight Maps Generated from Linear Kernel Machine Learning Models. In: 2013 International Workshop on Pattern Recognition in Neuroimaging. pp. 124–127. doi: https://doi.org/10.1109/PRNI.2013.40.

Schrouff, J., Monteiro, J. M., Portugal, L., Rosa, M. J., Phillips, C., Mourão Miranda, J., 2018. Embedding anatomical or functional knowledge in whole-brain multiple kernel learning models. Neuroinformatics 16 (1), 117–143. doi: https://doi.org/10.1007/s12021-017-9347-8.

Schrouff, J., Rosa, M. J., Rondina, J. M., Marquand, A. F., Chu, C., Ashburner, J., Phillips, C., Richiardi, J., Mourão Miranda, J., 2013b. PRoNTo: Pattern recognition for neuroimaging toolbox. Neuroinformatics 11 (3), 319–337. doi: https://doi.org/10.1007/s12021-013-9178-1.

Segovia, F., Górriz, J., Ramírez, J., Salas-Gonzáez, D., Álvarez, I., 2013. Early diagnosis of Alzheimer's disease based on Partial Least Squares and Support Vector Machine. Expert Systems with Applications 40 (2), 677 – 683. doi: https://doi.org/10.1016/j.eswa.2012.07.071.

Segovia, F., Salas-González, D., Górriz, J. M., Ramírez, J., Martínez-Murcia, F. J., 2017. Analysis of 18F-DMFP-PET data using Hidden Markov Random Field and the Gaussian distribution to assist the diagnosis of Parkinsonism. Vol. 10134. pp. 10134 – 10134 – 7. doi: https://doi.org/10.1117/12.2250281.

Shin, C., 2000. Neurophysiologic basis of functional neuroimaging: animal studies. Journal of Clinical Neurophysiology 17 (1), 2–9.

Silver, M., Montana, G., Nichols, T. E., 2011. False positives in neuroimaging genetics using voxel-based morphometry data. NeuroImage 54 (2), 992 – 1000. doi: http://dx.doi.org/10.1016/j.NeuroImage.2010.08.049.

Simon, A. B., Buxton, R. B., 2015. Understanding the dynamic relationship between cerebral blood flow and the BOLD signal: Implications for quantitative functional MRI. NeuroImage 116, 158–167. doi: https://doi.org/10.1016/j.NeuroImage.2015.03.080.

Sladky, R., Friston, K. J., Tröstl, Cunnington, R., Moser, E., Windischberger, C., 2011. Slice-timing effects and their correction in functional fMRI. NeuroImage 58 (2-2), 588–594. doi: https://doi.org/10.1016/j.NeuroImage.2011.06.078.

Smith, S. M., Nichols, T. E., 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage 44 (1), 83–98. doi: https://doi.org/10.1016/j.NeuroImage.2008.03.061.

Soares, J. M., Magalhães, R., Moreira, P. S., Sousa, A., Ganz, E., Sampaio, A., Alves, V., Marques, P., Sousa, N., 2016. A Hitchhiker's Guide to Functional Magnetic Resonance Imaging. Frontiers in Neuroscience 10, 515. doi: https://doi.org/10.3389/fnins.2016.00515.

Sona, D., Veeramachaneni, S., Olivetti, E., Avesani, P., 2007. Inferring cognition from fmri brain images. In: Artificial Neural Networks – ICANN 2007. pp. 869–878. isbn="978-3-540-74695-9.

Soon, C. S., Brass, M., Heinze, H. J., Haynes, J. D., 2008. Unconscious determinants of free decisions in the human brain. Nature Neuroscience 11, 543–545. doi: https://doi.org/10.1038/nn.2112.

Sreenivasan, K. K., Curtis, C. E., D'Esposito, M., 2014. Revisiting the role of persistent neural activity during working memory. Trends in Cognitive Sciences 18 (2), 82 – 89. doi: https://doi.org/10.1016/j.tics.2013.12.001.

Staeren, N., Renvall, H., Martino, F. D., Goebel, R., Formisano, E., 2009. Sound categories are represented as distributed patterns in the human auditory cortex. Current Biology 19 (6), 498 – 502. doi: https://doi.org/10.1016/j.cub.2009.01.066.

Stanford Encyclopedia of Philosophy, 2013. Descartes and the pineal gland. https://plato.stanford.edu/entries/pineal-gland.

Stelzer, J., Chen, Y., Turner, R., 2013. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. NeuroImage 65, 69–82. doi: https://doi.org/10.1016/j.NeuroImage.2012.09.063.

Stokes, M., Spaal, E., 2016. The importance of single-trial analyses in cognitive neuroscience. Trends in Cognitive Sciences 20 (7), 483–6. doi: https://doi.org/10.1016/j.tics.2016.05.008.

Talairach, J., Szikla, G., 1980. Application of stereotactic concepts to the surgery of epilepsy. Acta Neurochirurgica 30, 35–54. doi: https://doi.org/10.1007/978-3-7091-8592-6_5.

Talairach, J., Tournoux, P., 1988. Co-planar stereotaxic atlas of the human brain: 3-Dimensional proportional system: An approach to cerebral imaging. Thieme Medical Publishers. ISBN = 978-086-57-7293-9.

Tan, H., Meyer, C. H., 2009. Estimation of k-space trajectories in spiral MRI. Magnetic Resonance in Medicine 61 (6), 1396–1404. doi: https://doi.org/10.1002/mrm.21813.

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society: Series A 58 (1), 267–288.

Tie, Y., Suarez, R. O., Whalen, S., Radmanesh, A., Norton, I. H., Golby, A. J., 2009. Comparison of blocked and event-related fMRI designs for pre-surgical language mapping. NeuroImage 47 (Suppl 2), T107–15. doi: https://doi.org/10.1016/j.NeuroImage.2008.11.020.

Triantafyllou, C., Hoge, R. D., Wald, L. L., 2006. Effect of spatial smoothing on physiological noise in high-resolution fMRI. NeuroImage 32, 551–577. doi: https://doi.org/10.1016/j.NeuroImage.2006.04.182.

Turkeltaub, P. E., Eden, G. F., Jones, K. M., Zeffiro, T. A., 2002. Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. NeuroImage 16 (3 Pt 1), 765–80. doi: https://doi.org/10.1006/nimg.2002.1131.

Turner, B., 2010. Comparison of methods for the use of pattern classification on rapid event-related fMRI data. In: Annual Meeting of the Society for Neuroscience. San Diego, CA.

Turner, B. O., Mumford, J. A., Poldrack, R. A., Ashby, F. G., 2012. Spatiotemporal activity estimation for multivoxel pattern analysis with rapid event-related designs. NeuroImage 62(3), 1429–1438. doi: https://doi.org/10.1016/j.NeuroImage.2012.05.057.

Van Dijk, K. R., Sabuncu, M. R., Buckner, R. L., 2012. The influence of head motion on intrinsic functional connectivity MRI. NeuroImage 59 (1), 431–8. doi: https://doi.org/10.1016/j.NeuroImage.2011.07.044.

van Zijl, P. C., Hua, J., Lu, H., 2012. The BOLD post-stimulus undershoot, one of the most debated issues in fMRI. NeuroImage 62 (2), 1092–102. doi: https://doi.org/10.1016/j.NeuroImage.2012.01.029.

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. NeuroImage 145, 166–179. doi: https://doi.org/10.1016/j.NeuroImage.2016.10.038.

Visconti di Oleggio Castello, M., Halchenko, Y. O., Guntupalli, J. S., Gors, J. D., Gobbini, M. I., 2017. The neural representation of personally familiar and unfamiliar faces in the distributed system for face perception. Scientific Reports 7, 12237. doi: https://doi.org/10.1038/s41598-017-12559-1.

Viswanathan, S., Cieslak, M., Grafton, S. T., 2012. On the geometric structure of fMRI searchlight-based information maps. Vol. 1. Available online at: http://arxiv.org/abs/1210.6317.

Wager, T. D., Lindquist, M., Kaplan, L., 2007. Meta-analysis of functional neuroimaging data: current and future directions. Society Cognitive and Affective Neuroscience 2 (2), 150–158. doi: https://doi.org/10.1093/scan/nsm015.

Wager, T. D., Smith, E. E., 2003. Neuroimaging studies of working memory: a meta-analysis. Cognitive, Affective& Behavioral Neuroscience 3 (4), 225–74. doi: https://doi.org/10.3758/CABN.3.4.255.

Waldorp, L., 2009. Robust and unbiased variance of GLM coefficients for misspecified autocorrelation and hemodynamic response models in fMRI. International Journal of Biomedical Imaging 2009, 723912. doi: http://dx.doi.org/10.1155/2009/723912.

Wang, D., Buckner, R. L., Fox, M. D., Holt, D. J., Holmes, A. J., Stoecklein, S., Langs, G., Pan, R., Qian, T., Kuncheng, L., Baker, J. T., Stufflebeam, S. M., Wang, K., Wang, X., Hong, B., Liu, H., 2015. Parcellating cortical functional networks in individuals. Nature Neuroscience 18, 1853–1860. doi: https://doi.org/10.1038/nn.4164.

Wang, S., Summers, R. M., 2012. Machine learning and radiology. Medical Image Analysis 16 (5), 933–51. doi: https://doi.org/10.1016/j.media.2012.02.005.

Wansapura, J. P., Holland, S. K., Dunn, R. S., Ball Jr., W. S., 1999. NMR relaxation times in the human brain at 3.0 tesla. Journal of Magnetic Resonance Imaging 9 (4), 531–538. doi: https://doi.org/10.1002/(SICI)1522-2586(199904)9:4<531::AID-JMRI4>3.0.CO;2-L.

Wellcome Centre for Human Neuroimaging, 2018. Statistical Parametrical Mapping. https://www.fil.ion.ucl.ac.uk/spm/software/spm12.

Welvaert, M., Rosseel, Y., 2013. On the definition of signal-to-noise ratio and contrast-to-noise ratio for fMRI data. PLoS One 8 (11), e77089. doi: https://doi.org/10.1371/journal.pone.0077089.

Winkler, A., Ridgway, G., Webster, M., Smith, S., Nichols, T., 2014. Permutation inference for the general linear model. NeuroImage 92, 381–397. doi: http://dx.doi.org/10.1016/j.NeuroImage.2014.01.060.

Wolbers, T., Zahorik, P., Giudice, N. A., 2011. Decoding the direction of auditory motion in blind humans. NeuroImage 56 (2), 681 – 687. doi: https://doi.org/10.1016/j.NeuroImage.2010.04.266.

Woo, C. W., Krishnan, A., Wager, T. D., 2014. Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. NeuroImage 91, 412–419. doi: https://doi.org/10.1191/0962280203sm341ra.

Woolgar, A., Thompson, R., Bor, D., Duncan, J., 2011. Multi-voxel coding of stimuli, rules, and responses in human frontoparietal cortex. NeuroImage 56 (2), 744–52. doi: https://doi.org/10.1016/j.NeuroImage.2010.04.035.

Worsley, K. J., Andermann, M., Koulis, T., MacDonald, D., Evans, A. C., 1998. Detecting changes in nonisotropic images. Human Brain Mapping 8 (2).

Worsley, K. J., Evans, A. C., Marrett, S., Neelin, P., 1992. A three-dimensional statistical analysis for CBF activation studies in human brain. Journal of Cerebral Blood Flow & Metabolism 12 (6), 900–18. doi: https://doi.org/10.1038/jcbfm.1992.127.

Worsley, K. J., Friston, K. J., 1995. Analysis of fMRI time-series revisited–again. NeuroImage 2 (3), 173–81. doi: https://doi.org/10.1006/nimg.1995.1023.

Yarach, U., Luengviriya, C., Stucht, D., Godenschwenger, F., Schulze, P., Speck, O., 2016. Correction of B0-induced geometric distortion variations in prospective motion correction for 7T MRI. MAGMA 29 (3), 319–332. doi: https://doi.org/10.1007/s10334-015-0515-2.

Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zö, L., Polimeni, J. R., Fischl, B., Liu, H., Buckner, R. L., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. Journal of Neurophysiology 106 (3), 1125–65. doi: https://doi.org/10.1152/jn.00338.2011.

Zafar, R., Kamel, N., Naufal, M., Malik, A. S., Dass, S. C., Ahmad, R. F., Abdullah, J. M., Reza, F., 2018. A study of decoding human brain activities from simultaneous data of EEG and fMRI using MVPA. Australasian Physical & Engineering Sciences in Medicine 41 (3), 633–645. doi: https://doi.org/10.1007/s13246-018-0656-5.

Zaidi, A. D., Birbaumer, N., Fetz, E., Logothetis, E., Sitaram, R., 2018. The hemodynamic initial-dip consists of both volumetric and oxymetric changes correlated to localized spiking activity. bioRxivDoi: https://doi.org/10.1101/259895.

Zaitsev, M., Akin, B., LeVan, P., Knowles, B. R., 2017. Prospective motion correction in functional fMRI. NeuroImage 154, 37–42. doi: https://doi.org/10.1016/j.NeuroImage.2016.11.014.

Zhang, X., Wu, G., Dong, Z., Crawford, C., 2015. Embedded feature-selection support vector machine for driving pattern recognition. Journal of the Franklin Institute 352 (2), 669 – 685. doi: https://doi.org/10.1016/j.jfranklin.2014.04.021.

Zhang, Y., Tian, J., Yuan, K., Liu, P., Zhuo, L., Qin, W., Zhao, L., Liu, J., von Deneen, K. M., Klahr, N. J., Gold, M. S., Liu, Y., 2011. Distinct resting-state brain activities in heroin-dependent individuals. Brain Research 1402, 46 – 53. doi: https://doi.org/10.1016/j.brainres.2011.05.054.