Cyclotron Research Centre

University of Liège

Pattern Recognition in NeuroImaging

What can machine learning based models bring to the analysis of functional brain imaging?

by

Jessica Schrouff

A Thesis submitted to the University of Liège in partial fulfillment of the requirements for the degree of

Doctor in Applied Sciences

January 2013

Liège Belgium

Abstract

The study of the brain development and functioning raises many question that are tracked using neuroimaging techniques such as positron emission tomography or (functional) magnetic resonance imaging. During the last decades, various techniques have been developed to analyse neuroimaging data. These techniques brought valuable insight on neuroscientific questions, but encounter limitations which make them unsuitable to tackle more complex problems. More recently, machine learning based models, coming from the field of pattern recognition, have been promisingly applied to neuroimaging data.

In this work, the assets and limitations of machine learning based models were investigated and compared to previously developed techniques. To this end, two applications involving challenging datasets were defined and the results from widespread methods were compared to the results obtained using machine learning based modelling.

More specifically, the first application addressed a neuroscience question: ***Is it possible to detect and characterize mnemonic traces?*** The fMRI experiment comprised a learning and a control tasks, both flanked by rest sessions. From previous studies, patterns of brain activity generated during the learning task should be spontaneously repeated during the following rest session, while no difference should be observed between the pre- and post-task rest session in the control condition. Using univariate and multivariate feature selection steps before a Gaussian Processes classification, mnemonic traces could be detected and their spatio-temporal evolution characterized. On the contrary, an analysis of the rest sessions based on the detection of independent networks did not provide any results supporting the theory of memory consolidation.

The second application tackled a clinical issue: ***Can a pattern of brain activation characteristic to idiopathic Parkinson's disease be detected and localized?*** The dataset considered to address this question comprised the fMRI images of aged healthy subjects and Parkinsonian patients while they were performing a task of mental imagery of gait at three different paces. The signal comprised in a priori selected regions of interest allowed for the support vector machines classification of healthy and diseased volunteers with an accuracy of 86%. To localize the discriminating pattern, a methodology based on the weight in labelled regions (e.g. from the anatomical automatic labelling or Brodmann atlases) was developed, which enabled the comparison between univariate and multivariate results and showed a nice overlap between them. Furthermore, models could then be compared quantitatively in terms of pattern localization, using a specifically defined measure of distance. This measure could then be used to compare the patterns generated from different folds of a same model, from different feature sets, or from different modelling techniques.

The present study concluded that machine learning models can clearly and fruitfully complement other analysis techniques to tackle challenging questions in neuroscience. On the other hand, more work is needed in order to render the methodology fully accessible to the neuroscientific community.

Acknowledgements

I would like to thank my supervisor, Christophe Phillips, for his trust and support during the whole period of this work. I wish to thank him for his availability, his patience and for the various opportunities he provided me with during these four years.

I am also grateful to Professor Pierre Maquet, who gave me the opportunity to start this thesis and provided constructive advices all along this work.

Thank you to the PRoNTo team, and especially to Janaina Mourão-Miranda who allowed me to take part in an exciting Pascal Harvest project which produced PRoNTo. Working with this team was a particularly valuable experience in many professional and personal terms. Special thanks to Maria João Rosa for her warm welcome during my stay in London.

A large part of this work would not have been possible without Caroline Kussé, who paired with me for the neuroscience application and endured my Matlab teaching/speaking without complaints. I would also like to thank Gaëtan Garraux and Julien Cremers for providing the data and advices for the second part of this work.

Thanks to all the members of the Cyclotron Research Centre for their help and availability, but above all for their friendship and the great moments spent together during these four years.

Furthermore, I want to thank my family and my friends for their support and encouragements throughout my studies. A special wink to the "Biomeds", and in particular to Laura Symul, who helped me with the figures of this work.

Finally, special thanks to Sylvain Quoilin for his support and insightful advices during these past six years, as well as for the proofreading of this text.

Table of Contents

List of Tables

List of Figures

List of Acronyms

(N)REM	(Non-)Rapid Eye Movement
AA	Animals Area
AAL	Automated Anatomical Labelling
AD	Alzheimer's Disease
ARD	Automatic Relevance Distribution
BOLD	Blood Oxygen Level Dependent
CBS	CorticoBasal Syndrome
CSF	Cerebro-Spinal Fluid
CV	Cross-Validation
EEG	Electro-EncephaloGraphy
EP	Expectation Propagation
FFA	Fusiform Face Area
fMRI	functional Magnetic Resonance Imaging
FN	False Negative
FP	False Positive
FWHM	Full Width at Half Maximum
GLM	General Linear Model
GP	Gaussian Processes
HR(F)	Haemodynamic Response (Function)
IC(A)	Independent Component (Analysis)
IDPC	Integrated Difference in Partial Correlation
IPD	Idiopathic Parkinson's Disease
LOO	Leave-One-Out

MNI	Montreal Neurological Institute, standard stereotactic space
MSA	Multiple Systems Atrophy
MV	Majority Vote
MVPA	Multi-Variate Pattern Analysis
NOI(s)	Network(s) of Interest
PD	Parkinson's Disease
PDF	Probability Density Function
PET	Positron Emission Tomography
PPA	Parahippocampal Place Area
PPS	Parkinson Plus Syndromes
PPV	Positive Predictive Value
PRoNTo	Pattern Recognition for Neuroimaging Toolbox
PSP	Progressive Supranuclear Palsy
RFA	Recursive Feature Addition
RFT	Random Field Theory
ROI(s)	Region(s) of Interest
RVM	Relevance Vector Machines
sMRI	structural Magnetic Resonance Imaging
SPM	Statistical Parametric Mapping
SV(M)	Support Vector (Machines)
TN	True Negative
TP	True Positive
TPM	Tissue Probability Map
TR	Repetition Time
VBM	Voxel-Based Morphometry
voxel	Volume element in an image

Notation

β_j	Parameters of the general linear model
$\hat{\beta}$	Estimated parameters β
ε	Noise model
γ	Margin of the SVM classifier
$\lambda(a)$	Sigmoid function squashing variable a into the $[0, 1]$ interval
Φ	Proportion of transitions according to the forward sequence
Φ_{inv}	Proportion of transitions according to the inverse sequence
σ^2	Variance of a normal distribution
$\mathcal{N}$	Gaussian distribution

$B0$	External magnetic field in MRI
c	Index representing confounds, i.e. effects of no interest
C	Soft-margin parameter of the SVM classifier
f	Function mapping inputs to outputs
F	F-distribution
I	Index representing effects of interest
K	Covariance or similarity matrix
$L(k, k')$	Loss function between classes k and k'
m	Number of features/voxels in dataset
$\boldsymbol{M}$	Design matrix
n	Number of samples in dataset
p	Probability value

$p(a\|b)$	Conditional probability of a, knowing b
t	Student's t-distribution
T	Tesla
$\boldsymbol{w}$	Parameters or weights of the machine learning based model
$\boldsymbol{x_i}$	Vector of features/voxels for the i^{th} sample
$\boldsymbol{x_j}$	Series of samples for the j^{th} feature/voxel
$\boldsymbol{x_\star}$	New/unseen sample, i.e. vector of features
X	Data matrix comprising $\boldsymbol{x_j}$ for $j = 1 \cdots m$
$\boldsymbol{y_i}$	Model output. Label or target associated to sample i

acc_b	Balanced accuracy
C_x	Correlation between $Pr(x)$ and behavioural performance
d'	D-prime, measure of behavioural performance
dr	Distance between Rankings
m_{ROI}	Number of regions of interest ROIs
NW_{ROI}	Normalized Weight for one ROI
$Pr(x)$	Proportions of scans significantly linked to the task x
$R1_x$	Pre-task rest session, for condition x
$R2_x$	Post-task rest session, for condition x
SF	Scaling Factor of memory traces
W_{ROI}	Absolute weights for one ROI

Chapter 1

Introduction

Neuroscience is a challenging field of study, trying to unravel the mysteries of brain development and functioning, in health but also in disease. Two of the most fundamental questions in the field of neuroscience are how information is represented in different brain structures, and how this information evolves over time. To investigate these questions, different powerful tools have been developed to record brain activity, within which functional Magnetic Resonance Imaging (fMRI). fMRI can map brain activity with a spatial resolution of a few cubic millimeters and a typical temporal resolution in the order of 1 or 2 seconds.

During the last decades, various methodologies have been developed to analyse such data. A well-known technique is Statistical Parametric Mapping (SPM, Friston et al., 2007), detecting which volume elements (a.k.a. “voxels”) show a statistically significant response to the experimental conditions. However, there are limitations on what can be learned about the representation of information by examining voxels in a univariate fashion, i.e. independently from one another. For instance, sets of voxels considered as non-significant by the SPM analysis of one experimental condition might still carry information about the presence or absence of that condition. Furthermore, univariate analytic techniques are agnostic of any a priori information, for example disease-specific information. They are also mainly designed to perform group-wise comparisons and would therefore be unsuitable to evaluate the state of the disease of each individual.

To overcome these issues, multivariate pattern analyses (see Haynes and Rees, 2006; Pereira et al., 2009 for reviews) have been successfully applied to neuroimaging data. Multivariate pattern analysis derives from the field of pattern recognition, which is concerned with the automatic discovery of regularities in data. Those regularities then serve as the basis for the classification of new data, which means that those models allow for predictions. When applied to neuroimaging data, multivariate methods, and more specifically machine learning based modelling, aim at associating a particular cognitive, behavioural or perceptual state to specific patterns of brain functional activity. Application of these methods enabled to decode the category of a seen object [Spiridon and Kanwisher, 2002; Cox and Savoy, 2003; Shinkareva et al., 2008] or the orientation of a stripped pattern seen by the subject [Kamitani and Tong, 2005; Haynes and Rees, 2005] from the brain activation of the imaged subject. Advances in pattern-classification algorithms also allowed the decoding of less-controlled conditions such as memory retrieval tasks [Polyn et al., 2005; Chadwick et al., 2010].

When applied to neuroimaging data (which have multivariate properties), machine learning methods are able to detect subtle, spatially distributed activations and patterns of brain anatomy and should therefore achieve greater sensitivity than univariate techniques. Furthermore, those models generate predictions, which allows their use as diagnostic tools (see Klöppel et al., 2008; Vemuri et al., 2008; Phillips et al., 2011; Orrù et al., 2012 for examples). The predictive aspect of machine learning methods is an important asset since they enable the characterization of new/unseen patterns that were not associated to any perceptual or behavioural state (i.e. labelled). In univariate analyses, the labels are inputs of the models, while in machine learning analyses the labels are inputs, but also outputs of the model. These techniques therefore allow for missing data (in terms of labelling). In regards of those advantages, machine learning based models seem to enable further investigations of the two fundamental questions in neuroscience.

In this work, we investigated the assets of machine learning models applied to neuroimaging data. More specifically, two applications were examined: the first defined a neuroscience question and the second was based on a clinical question. These two applications involved complex datasets, that could not be successfully analysed using classical techniques. The aim of this work was therefore to prove the advantages of multivariate modelling over previous methods, but also to establish and discuss the limits of those recent analysis techniques.

The present work was therefore divided into two parts, one for each application. After an introduction to the techniques of acquisition and analysis common to both parts (chapter 2), the neuroscience application (chapters 3 to 6) studied the formation of memory traces after a learning task. This study involved the modelling of complex data, based on constrained, semi-constrained and spontaneous brain activity. The second part (chapters 7 to 10) investigated the discrimination between healthy subjects and Parkinson's disease patients. For this application, particular attention was paid to the localization of the pattern (section 8.2.2), which is an important aspect of the analysis, especially for neuroscientists. In both applications, the results of the machine learning modelling were compared to the results obtained from other, widely used, analysis techniques. After a separate discussion (chapters 6 and 10), the results were discussed in chapter 11, as well as future work.

Aim: Investigate the assets and the limits of machine learning based models applied to neuroimaging data via a neuroscience and a clinical application, each involving complex datasets.

Contributions of this thesis: The modelling of semi-constrained brain activity required the selection of discriminative features. A forward wrapper selection method was hence developed, based on binary support vector machine classifiers [Burges, 1998]. Decoding spontaneous brain activity represented the main challenge of this work. To this end, a specific methodology was developed, based on an error-correcting output code scheme [Dietterich and Bakiri, 1995]. This approach was further refined to characterize the spatio-temporal evolution of memory traces. To localize the pattern discriminating between Parkinson's disease patients and healthy subjects, local averages of the pattern were computed, according to anatomical or functional atlases. Based on these averages, different models could then be compared in terms of their localization using a specific distance measure.

Chapter 2

Material and Methods

Contents

In this chapter, the type of brain images used throughout this thesis is introduced, as well as the methods considered to analyse them. Please note that this chapter only provides a general overview of the material and methods. Further details on the conducted experiments and analyses can be found in chapters 5 and 8, for each application respectively.

2.1 Acquisition techniques

The present work focused on neuroimaging data, i.e. on images of the brain, acquired using a non-invasive technique. The images consist in volume elements, called voxels in which an aspect of brain activity or structure is recorded (see Figure 2.1). The acquired aspect depends on the considered technique/hardware.

The type of images used in this work was “functional Magnetic Resonance Imaging” (fMRI). A brief explanation of Positron Emission Tomography (PET) is also provided in this introductory section, since it is further mentioned (see chapter 7).

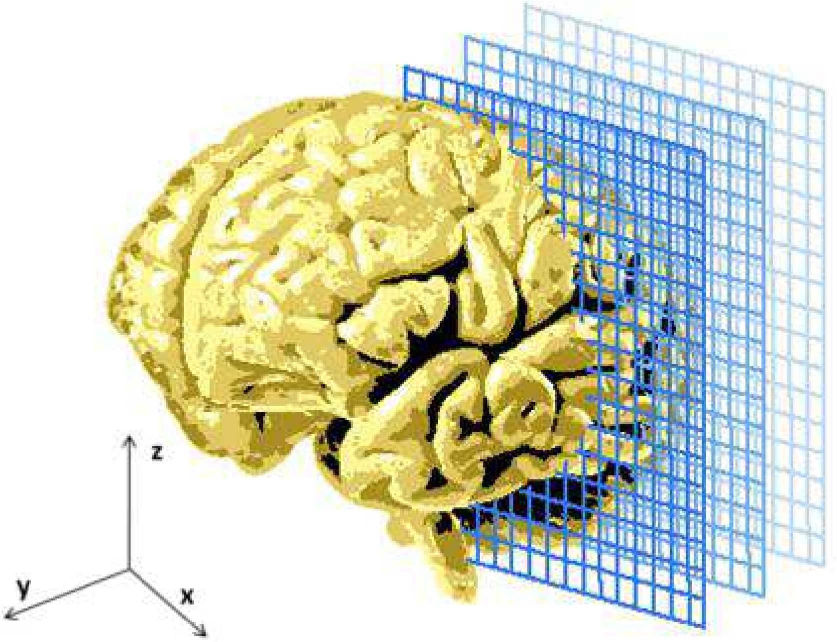

Figure 2.1: **Illustration of neuroimaging data.** The brain is divided into cubic volume elements called voxels. Each voxel contains one value representing the level of the functional/anatomical measure considered. Illustration by Laura Symul.

2.1.1 Magnetic Resonance Imaging and fMRI

Structural Magnetic Resonance Imaging (sMRI) is a medical imaging technique used to visualize internal body structures in detail. As indicated by its name, it makes use of the magnetic properties of the different tissues, and specifically of the hydrogen nuclei (protons). These can be viewed as small rotating magnets, due to their spin. In the presence of a large external magnetic field ($B0$ >1 Tesla, a.k.a. static field), the protons, at first randomly distributed, align in a (anti-)parallel way with the field. They keep rotating in a precessing movement, at a frequency called the Larmor frequency and which is proportional to the static field $B0$. If excited by a radiofrequency pulse at their particular precessing frequency, the protons absorb electromagnetic energy, which is then emitted when the protons return to the equilibrium (relaxation). From a macroscopic point of view, the relaxation process follows a time constant which is different across tissues and directions. These differences are measured and allow the construction of contrast images, enhancing one tissue property or another based on various parameters, such as the time between two excitation pulses and the time elapsed before acquiring the signal. Typically, a T1 contrast is used to obtain structural images (relaxation along the z-axis, differentiation between grey and white matter) while a $T2^{\star}$ contrast defines functional images (relaxation along the xy-plane, increased contrast for venous blood). The acquisition of a whole brain fMRI volume proceeds in the successive acquisition of signal from different slices along the z-axis.

In functional MRI (fMRI), it is not the contrast between tissues which is investigated but rather the contrast between ratios of oxygenated versus de-oxygenated blood: when neurons in one brain region are firing, their metabolism is increased, which means that they need more energy, delivered under the form of glucose and oxygen (Figure 2.2). This results in an increase in oxygenated blood flow (following the Haemodynamic Response, HR) to this region, which can be measured by the scanner due to its magnetic properties (diamagnetic in oxygenated form and paramagnetic when de-oxygenated). As illustrated in Figure 2.2, the Haemodynamic Response takes several seconds to peak and then come back to its baseline level. The time course of the neuronal activity therefore passes through a filter that has an assumed HR function.

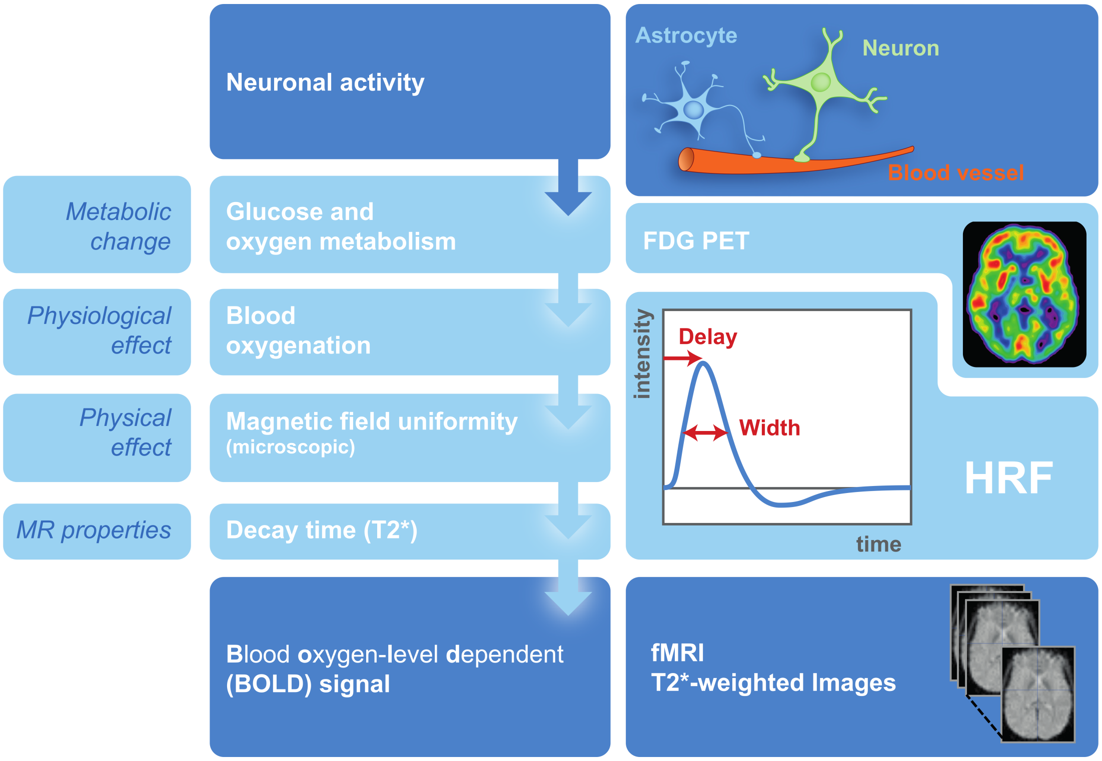

Figure 2.2: **From neuronal activity to BOLD signal.** This illustration displays the cascade of effects to generate a BOLD signal from neuronal activity: neuronal activity increases the glucose and oxygen intake, resulting in a metabolic change (signal acquired in FDG-PET). This metabolic change affects the blood oxygenation, according to the Haemodynamic Response Function (HRF, illustrated on the right side of the figure), which in turn affects the magnetic field uniformity that can be detected by the MR scanner to finally build T2*-weighted images. See [Hashemi and Bradley, 1997] for more details. Source: Doug Noll's primer, modified by Laura Symul.

The signal acquired in each voxel is called the Blood Oxygen Level Dependent signal (BOLD, Ogawa et al., 1990). Functional MRI thus allows the investigation of brain activity voxel per voxel. The Repetition Time (TR), i.e. the time between two successive acquisitions of a whole brain volume, is nowadays around 2 seconds which therefore enables to track changes in haemodynamic brain activity. However, the shape of the Haemodynamic Response Function (HRF) limits this resolution and has thus to be taken into account in further analyses.

2.1.2 Positron Emission Tomography

Positron Emission Tomography (PET) is a nuclear medical imaging technique, mainly used for the diagnosis of cancer or dementias. It implies the injection in the blood stream of a radiotracer that emits a positron when its radioactive isotope decays (see Figure 2.3 for an illustration of the PET principle). The positron then annihilates when encountering an electron, which results in the emission of two gamma photons in opposite directions. These gamma photons are then detected and when sufficient data is available, the distribution of radiotracers in the body can be reconstructed, i.e. an image of the spatial distribution of tracer uptake can be built. A commonly used tracer is fluorodeoxyglucose (^{18}FDG), allowing the mapping of metabolic activity of the tissues, in terms of regional glucose uptake. For practical details and a comparison with fMRI, please refer to appendix A.

2.2 Experimental design

fMRI investigations often imply the design of an experiment involving controlled stimulation in terms of content, timing and duration of the events. The brain response to these stimuli are then modelled in order to find differences between conditions (e.g. viewing a face versus viewing a building) or between (groups of) subjects (e.g. healthy controls versus diseased). Experiments can be designed in "block", i.e. periods of time during which the subject performs the task, separated by rest periods (usually lasting between 10 to 15 seconds). This allows for the BOLD signal to reach a plateau, providing a high signal-to-noise ratio, especially if the images are averaged over one block. On the other hand, experiments can consists of temporally isolated events. This is referred to as "event-related" designs and enables the investigation of brain responses to transient events (at the price of a lower signal-to-noise ratio). In the present work, particular experiments have been designed for both the neuroscience and the clinical applications. In the neuroscience application, transient events were investigated in each subject, while for the clinical application, a block-design experiment studied the (correct) discrimination between groups of subjects.

2.3 Preprocessing

For further analysis, the data from a specific voxel is assumed to correspond to the same region in the brain across volumes (time points or subjects). Movements or different brain shapes may lead to violation of this assumption. Furthermore,

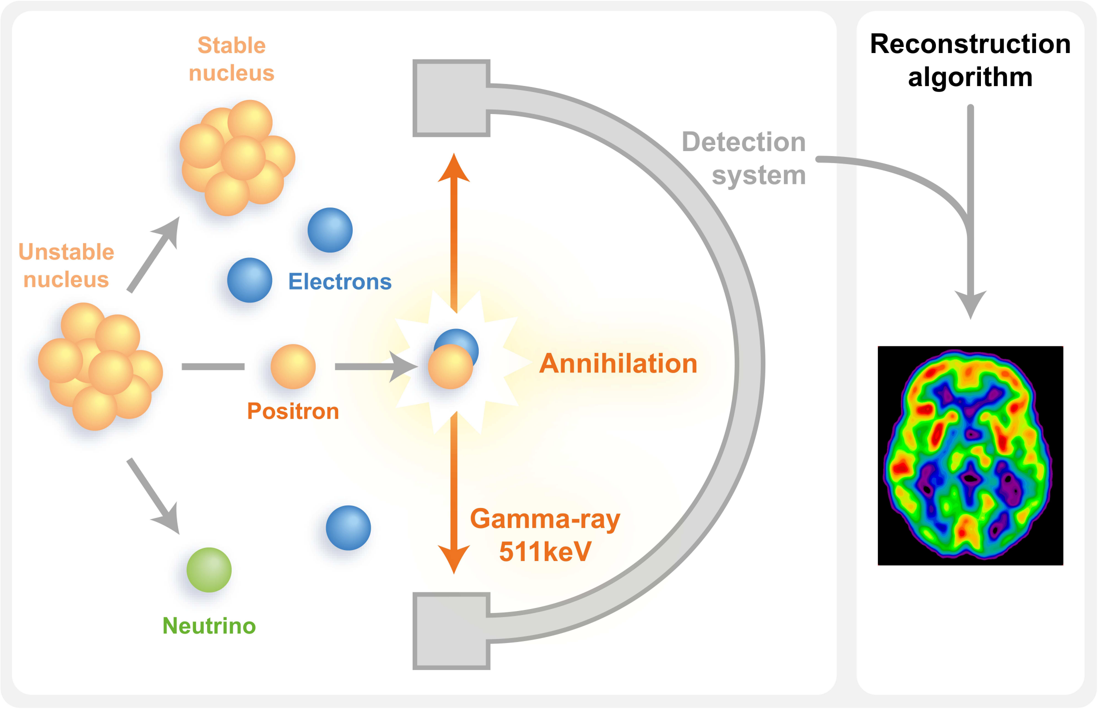

Figure 2.3: **Principle of the PET scanner.** The radiotracer contains unstable nuclei which decay into stable nuclei by emitting a positron. This positron randomly travels into the surrounding tissues (a few millimetres depending on the emitting atom) until it meets an electron, which results in the annihilation of both particles and in the emission of two (opposite) gamma rays. The gamma rays are detected by the scanner which then builds a PET image using a reconstruction algorithm. For more details, see [Valk et al., 2003]. Source [Lancelot and Zimmer, 2010], modified by Laura Symul.

structured or random noise can be added to the signal of interest (e.g. physiological noise due to heartbeat or breathing, intensity spikes). Corrections for these potential variabilities should therefore be performed before any further analysis, using temporal and spatial transformations.

Preprocessing usually consists in multiple steps, successively applying corrections for different sources of potential noise. These steps depend on the type of data (PET or fMRI) and level of analysis (within or between subjects). In this work, within-subject analysis will be conducted for the fMRI neuroscience application only. Therefore, the preprocessing steps described in section 2.3.1 are specific for fMRI.

2.3.1 Within-subject preprocessing

An example of classical within-subject preprocessing is presented in Figure 2.4, the different steps are described hereunder:

- **Slice time correction**: In fMRI, the slices (along the z-axis) are acquired successively, leading to differences in their sampling time. To correct this discrepancy, each volume was temporally realigned by interpolating the signal over time across volumes.

- **Realign and unwrap**: The two main sources of noise (i.e. variability unrelated to neuronal activity) for fMRI acquisition are movement of the subject and field inhomogeneities. While $B0$ (the static magnetic field) is spatially homogeneous, introducing the subject in the scanner disrupts it, leading to field inhomogeneities. Since the reconstruction of the contrast image is exact only under the assumption of a homogeneous static field, these field inhomogeneities can lead to distortions in the acquired images and must be accounted for. In this work, spatial deformations induced by the field inhomogeneities were estimated using the FieldMap toolbox [Hutton et al., 2002]. To correct for possible movement artefacts, rigid-body transformations were applied to realign the volumes on their mean (by estimating movement parameters at each TR). Simultaneously, the images were unwrapped, i.e. corrected for the elastic/non-linear deformations induced by the static field inhomogeneities and for the interaction between the subject's movements and the spatial deformations [Andersson et al., 2001].

- **Coregister**: The functional images were co-registered with the structural image, such that the anatomical localisation of single subject activations would be more accurate.

- **Smooth**: Finally, to increase the signal to noise ratio, the images are smoothed using a spatial low-pass Gaussian filter, characterized by its Full Width at Half Maximum (FWHM).

2.3.2 Between-subject preprocessing

The main problem when dealing with multi-subject analysis is that brains of different subjects vary in size and shape. To examine homologous brain regions across subjects, all images thus need to be brought into a common reference space, by normalizing them in the MNI space (Montreal Neurological Institute,Mazziotta et al., 2001). The "unified segmentation" [Ashburner and Friston, 2005] was used here. This approach relies on the optimization of the parameters of a generative model, including tissue segmentation, intensity non-uniformity correction and non-linear image registration. These deformations are then applied to all the images of a same subject (i.e. to the structural and the functional images in the case of fMRI, to the FDG uptake image in the case of PET). The normalisation does not perfectly match the images from different subjects, leaving some residual inter-subject anatomical variability. To dampen this variability, smoothing is usually applied, the size of the FWHM depending on the modality. For fMRI, the images entering the normalisation step have first undergone the within-subject preprocessing, except for the smoothing (performed at a later stage).

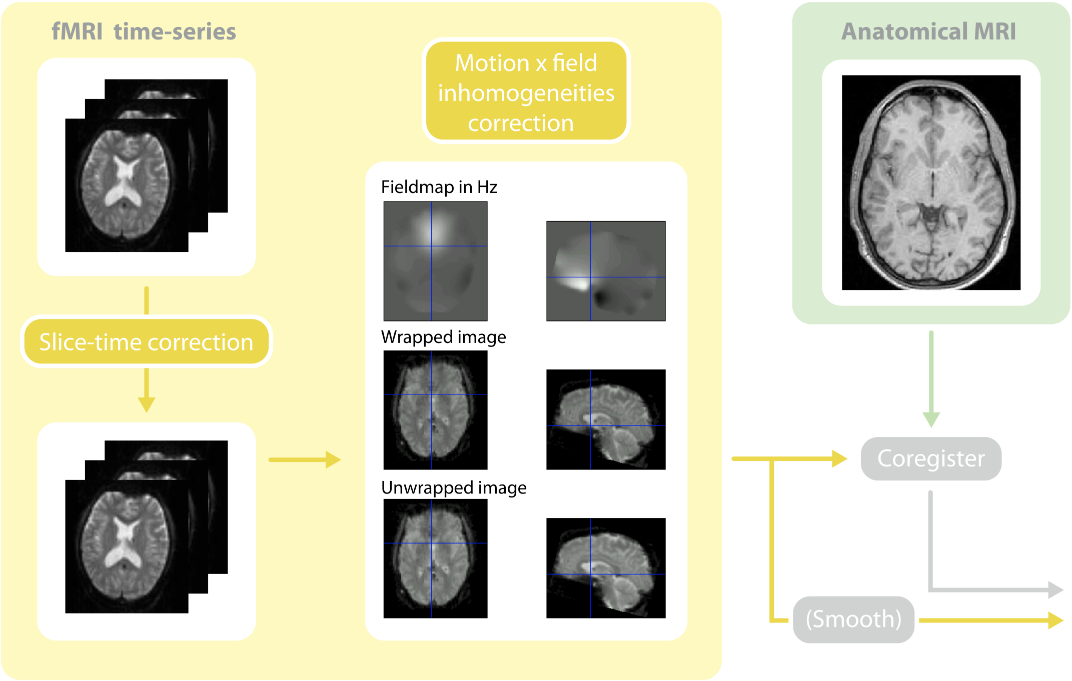

Figure 2.4: **fMRI within-subject preprocessing.** A typical fMRI within-subject preprocessing involves 4 main steps: the slice-time correction, correction for the field inhomogeneities, for motion and for their interaction, coregistering with an anatomical image and a smoothing (optional) using a Gaussian filter. See text for further details. Source: SPM course slides, modified by Laura Symul.

2.4 Univariate methods

This section briefly presents statistical parametric mapping, a widely used univariate technique to model neuroimaging data. For more details, see [Friston et al., 2007].

2.4.1 Principles

An fMRI time series can either be seen as a succession of 3D volumes, or as the collection of many voxels, each with its own temporal evolution. In univariate analysis, the second version is used: the time series of each voxel are modelled independently, the results being assembled into parametric images, from which statistical maps can be derived. When dealing with multiple subjects having only one image, each "time point" corresponds to a subject. Therefore, we'll refer to each scan (time point of fMRI time series or subject) as a *sample*.

Statistical Parametric Mapping (SPM, Friston et al., 2007) identifies regionally specific effects in neuroimaging data. These effects may be due to structural differences (e.g. in voxel-based morphometry, VBM, Ashburner and Friston, 2000) or to changes over a sequence of observations (as in fMRI). SPMs are images whose voxels are, under the null hypothesis, distributed according to a specific probability density function, such as the Student's t or the F-distribution. Observing "unlikely" large values or topological features is then interpreted as a regionally

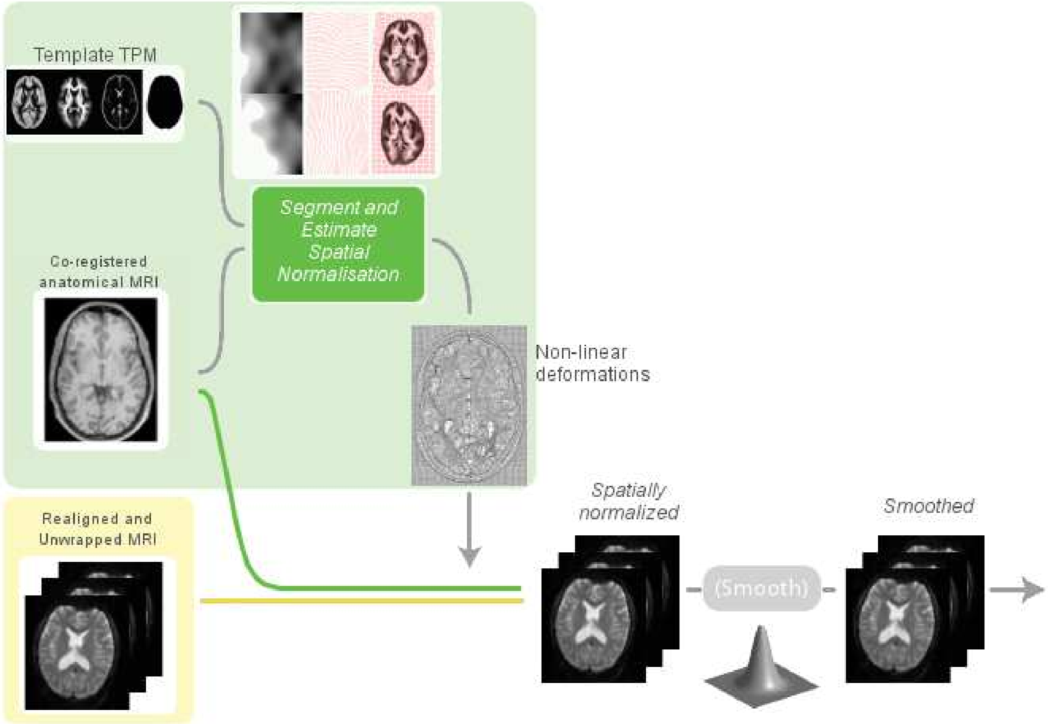
Template TPM
Segment and Estimate Spatial Normalisation
Co-registered anatomical MRI
Non-linear deformations
Realigned and Unwrapped MRI
Spatially normalized
(Smooth)
Smoothed

Figure 2.5: **Between subject preprocessing.** The images of each individual has to be transformed into a common, standard space. This operation is performed by first segmenting the co-registered anatomical image. Grey matter, white matter and cerebro-spinal fluid tissue probability maps (TPM) are included in the segmentation and define the reference space. The parameters of the spatial normalisation are then estimated, resulting in non-linear deformations, which can be applied on the (time-corrected and realigned) functional images. Finally, the images are smoothed using a spatial Gaussian filter. Source: SPM course slides, modified by Laura Symul.

specific effect, caused by the experimental design. In order to control for the risk of false positives across the whole image, it is then necessary to account for the problem of "multiple comparison". Statistical parametric mapping therefore consists in two steps: the modelling, using a *General Linear Model* (GLM) and the inference via random field theory (RFT).

2.4.2 General Linear Model

The general linear model, GLM, assumes that the responses to experimental conditions can be partitioned into three components: components of interest (such as the different conditions of stimulation), confounds (i.e. components affecting the signal but of no experimental interest, such as the movement parameters computed during the realignment) and the error, i.e. unexplained variance. It is expressed as:

$$\mathbf{x}_j = M\beta_j + \varepsilon_j \tag{2.1}$$

Where $\mathbf{x}_j$ is a vector representing the series of samples of the j^{th} voxel and is viewed as a random variable, M is the *design matrix*, containing both the components of interest and the confounds, β_j are the (unknown) parameters and ε_j is the error term. ε_j follows a probability density function (PDF) whose shape is fixed a priori. A GLM therefore consists in two parts: the model of the design, M, comprising components of interest and of no interest, and the model of the error ε_j.

One column in the design matrix corresponds to a regressor, modelling an effect of interest or confound. Typically, effects of interest are represented by the onsets and the durations of the corresponding stimuli, convolved with a canonical haemodynamic response function and its derivatives (to account for possible voxel-specific variability of the HR in terms of delay and amplitude). Confounds are usually noise effects which can be accounted for, such as events of no interest (like button presses), movement parameters or low frequency drift that can bias the results. Equation 2.1 can hence be written as:

$$\mathbf{x}_j = \begin{bmatrix} M_I M_c \end{bmatrix} \begin{bmatrix} \beta_{jI} \\ \beta_{jc} \end{bmatrix} + \varepsilon_j \tag{2.2}$$

With I corresponding to the components of interest and c to the confounds. Figure 2.6 shows an example of design matrix with components of interest and of no interest.

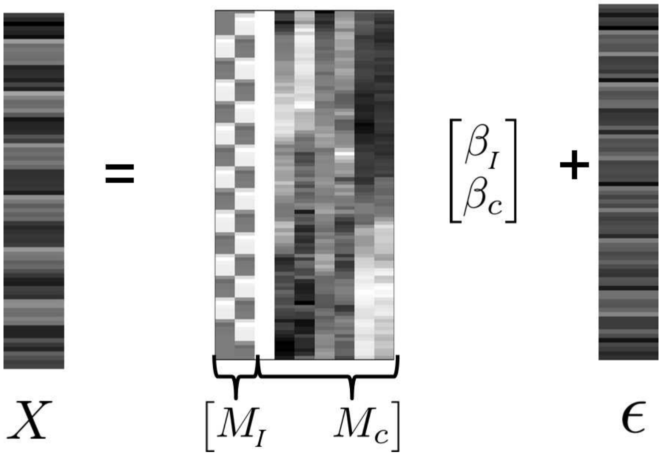

Figure 2.6: **General Linear Model.** The General Linear Model assumes that the acquired neuroimaging signal (data matrix X) is the combination of three components: the components of interest reflecting the design of the experiment (onsets of the stimuli convolved with the HRF, M_I), confounds (M_c) containing the sources of noise which can be modelled, such as the previously modelled movement parameters, and an error term ε. The weight assigned to each component of interest (β_I) or confound (β_c) are the parameters of the models and need to be estimated. Source: SPM8 software.

The parameters β_j are then estimated using ordinary or weighted least squares, giving $\hat{\beta}_j$. Once the parameters have been estimated, inference in terms of t or F statistics can be performed.

2.4.3 Inference

t- or F-scores are constructed from “contrasts”, i.e. linear combinations of the parameters β, and an estimation of the residual variance. A contrast defines the neuroscientific question driving the univariate analysis. For example, when considering two conditions of interest, a t-contrast of $[1 - 1]$ will investigate the differential regional effect of condition 1 compared to condition 2. On the other hand, a F-contrast with equation $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ will sum the effect of the two conditions. The statistical scores are then compared to the expected distribution under the null hypothesis, allowing the computation of a p-value for each voxel. However, p-values should be corrected for multiple comparisons (the number of comparisons being the number of voxels), especially in the case of an anatomically open hypothesis. Bonferroni correction [Dunn, 1961] could be used to correct for multiple comparisons but, as it assumes that all tests are independent (which is not true in neuroimaging), the adjustment becomes very severe when dealing with many voxels, as is usual in fMRI or PET data ($>$ 100,000). Random field theory, on the contrary, takes into account the fact that neighbouring voxels are not independent and provides a more parsimonious approach: With the estimated smoothness of the residuals of the GLM, the number of expected false positives in a statistical map can be controlled. A statistical map can then be thresholded, using some height and spatial extent thresholds of the clusters that are user-specified (Figure 2.7).

The results of a statistical parametric analysis are 3D maps, one per specified t or F-contrast. An example is shown in Figure 2.8, the design includes 3 stimulation conditions, all are considered as active in the corresponding F-contrast, while the movement parameters and low frequency drift are modelled as confounds. The map displayed is thresholded at $p < 0.05$, corrected for multiple comparisons using RFT.

2.5 Multivariate methods

Although mass univariate analysis brought significant insight on regionally specific inferences on brain function and structure, there are limitations on what can be learned by examining voxels in a univariate fashion. For instance, spatially distributed sets of voxels considered as non-significant by a SPM analysis of one experimental condition might still carry information about the presence or absence of that condition. Furthermore, classic univariate analytic techniques are also mainly designed to perform group-wise comparisons and would therefore be unsuitable to evaluate the state of a disease in individuals.

On the other hand, machine learning based Multi-Variate Pattern Analyses (MVPA, see Norman et al., 2008, Friston et al., 2008 and Haynes and Rees, 2006 for a review) allow an increased sensitivity to detect the presence of a particular mental representation. These multivariate methods, also known as *brain decoding* or *mind reading*, attempt to link a particular cognitive, behavioural, perceptual or medical state to specific patterns of voxels’ activity. Application of these methods made it possible to decode the category of a seen object [Spiridon and Kanwisher, 2002;

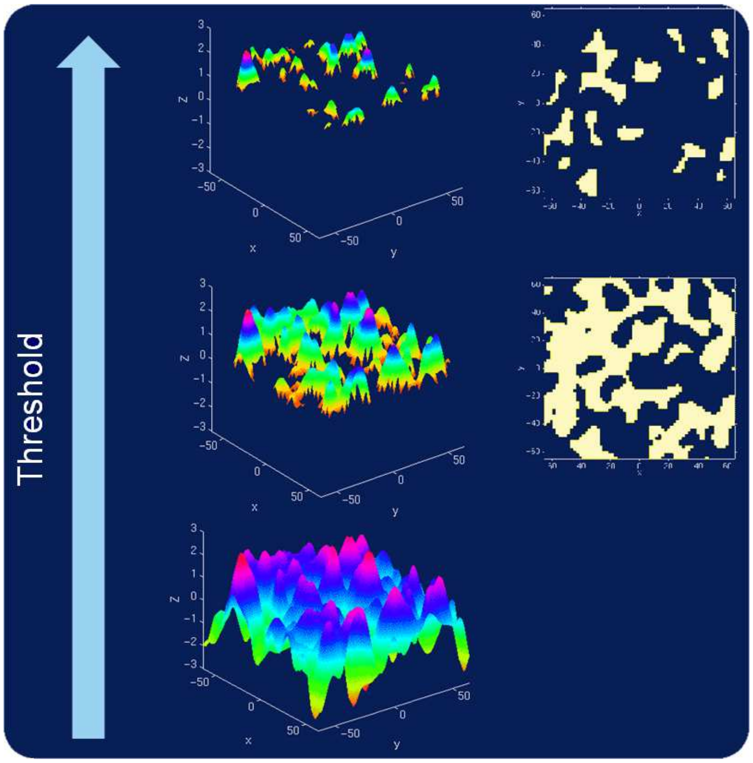

Figure 2.7: **Random field theory.** The theory assumes that the statistical parametric map is a discrete approximation of a smooth and continuous random field. The expected properties of a thresholded field can then be estimated and significance can be assessed in terms of voxel amplitude and cluster extent. In this figure showing a synthetic Gaussian random field, the height threshold is varied, displaying its impact on cluster number and extent. Source: SPM course slides.

Cox and Savoy, 2003; Shinkareva et al., 2008] or the orientation of a stripped pattern seen by the subject [Kamitani and Tong, 2005; Haynes and Rees, 2005] from the brain activation of the imaged subject. Advances in pattern-classification algorithms also allowed the decoding of less-controlled conditions such as memory retrieval tasks [Polyn et al., 2005; Chadwick et al., 2010].

2.5.1 Principles

Brain decoding derives from the fields of pattern recognition and machine learning, which are concerned with the automatic discovery of regularities in data. Those regularities then serve as basis for the classification of new data [Bishop, 2006]. A classical example of pattern recognition is the automatic classification of handwritten digits (illustrated in Figure 2.9 from Bishop, 2006): each digit is represented by a grey scale image of 256×256 pixels and the goal is to build an algorithm capable of classifying each image into the correct category (i.e. 0, 1,. . ., 9).

We therefore need to build a function (or machine), f, which will take images as inputs, $\mathbf{x}_i$, $i = 1 \ldots n$, with n the number of image samples, and produce their corresponding digit as outputs, y_i. Due to the high variability in handwritings, this operation is not trivial and the use of machine learning is necessary. This means

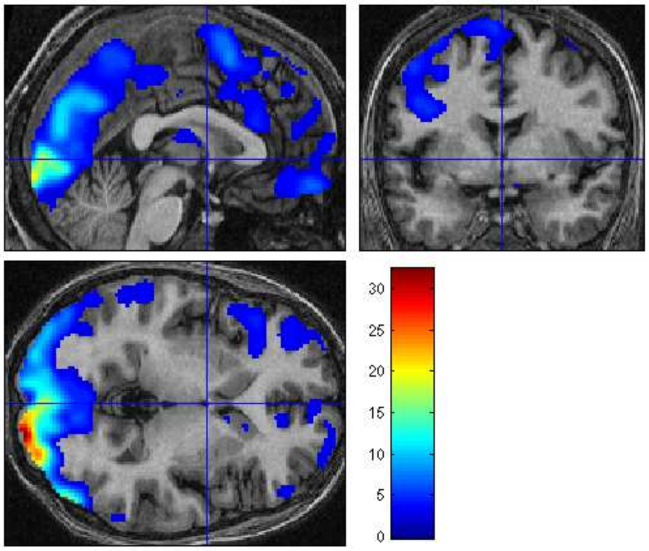

Figure 2.8: **Example of thresholded F-map.** Axial, coronal and sagittal views of a thresholded F-map (colour-coded F-values, $p < 0.05$, corrected for multiple comparisons), overlaid over a single subject anatomical image.

that the computer has to learn which pattern in the images corresponds to which digit (learning phase). This learning is achieved providing a learning set, which is a set comprising both images (inputs) and corresponding digit (outputs), which are known a priori and often hand-labelled. This is called *supervised learning*. The machine can then build the required function using this learning set and finally predict outputs, $y_\star$ when given new/unseen inputs, $\mathbf{x}_\star$ (test phase):

$$f : X \rightarrow \boldsymbol{y} \tag{2.3}$$
$$f : \mathbf{x}_\star \rightarrow y_\star \tag{2.4}$$

More specifically, the function f represents the "true" underlying function of the data from which only noisy samples can be observed:

$$\boldsymbol{y} = f(X) + \varepsilon \tag{2.5}$$

With ε, the noise, being distributed according to a Gaussian with zero mean and variance σ^2.

There are three main approaches to determine f: the discriminant function, the discriminative and the generative approaches. A discriminant function directly assigns a label $y_\star = +1$ or $y_\star = -1$ via $sign(f(X))$ (in the binary case). The predictions are referred to as "hard" predictions, because they cannot be associated to any confidence measure such as a probability. On the other hand, the discriminative and generative approaches model the conditional probability distribution $p(\boldsymbol{y}|X)$ in an inference stage, which is then used to make optimal decisions. This distribution can either be modelled directly (discriminative) or using Bayes' theorem (generative):

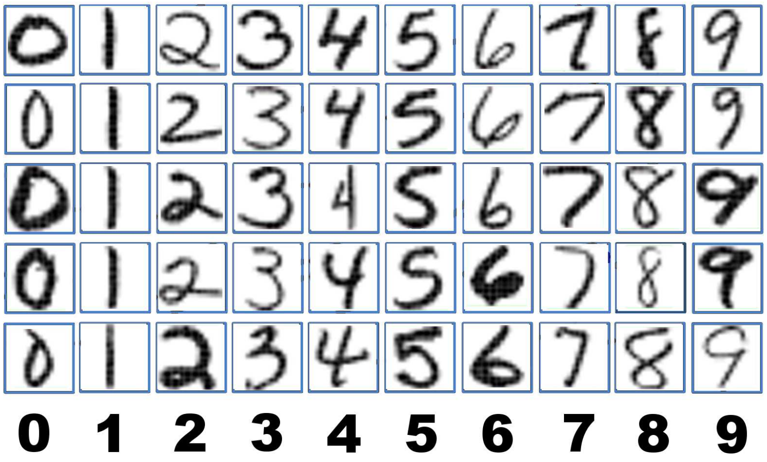

Figure 2.9: **Handwritten digits example.** The machine has to learn a model able to correctly assign a 256×256 pixels image of a handwritten digit to the corresponding number (i.e. 0 to 9). This is not a trivial task due to the multiplicity of the handwritings.

$$p(\boldsymbol{y}|X) = \frac{p(\boldsymbol{y}) \cdot p(X|\boldsymbol{y})}{p(X)} \tag{2.6}$$

In the latter, the likelihood $p(X|\boldsymbol{y})$ is the core of the model since it is possible to sample from this distribution afterwards. However, the aim here is to distinguish between different categories (of conditions or groups of subjects) such that only discriminant functions and discriminative models will be considered.

The built model is then evaluated in terms of generalization ability, i.e. its ability to (correctly) classify new samples (see section 2.5.4).

2.5.2 Inputs of the model

The learning set is generally represented under the form of a matrix $X \in \mathbb{R}^{n \times m}$: each sample $\boldsymbol{x}_i$ is represented by a feature vector, which is the collection of the m variables to feed in the machine, and a label, the output of the function. $\boldsymbol{x}_i$ corresponds to a point in the input space and the dimensionality of the space corresponds to the number of features in the samples. In the case of neuroimaging data, the variables are the values of the signal in each considered voxel and the samples are either the points of the time-series corresponding to a condition (after correction for HRF delay) or the images of a multi-subject analysis. Compared to the GLM approach, the inputs are:

$$X = \left[\mathbf{x}_j\right], \qquad j = 1...m \tag{2.7}$$

With m being the number of voxels, i.e. features. However, since X contains noise that can be modelled (i.e. the confounds of the design matrix), it is usually an adjusted version of X which is used as input of the machine learning based model:

$$X_a = X - M \times \begin{bmatrix} 0 \\ \hat{\beta}_C \end{bmatrix} \tag{2.8}$$

In general for fMRI, only the samples corresponding to a condition are considered for further modelling. The selection of the samples has to take the HRF shape into account, in terms of its delay to peak and its width (see Figure 2.10). In the present work, the onset of an event was computed as:

$$onset_a = round\left(\frac{onset + HRF_{delay}}{TR}\right) \tag{2.9}$$

Where $onset_a$ is the adjusted onset in TRs, $onset$ is the onset as in the design, in seconds and HRF_{delay} is a principled value, usually comprised between 3 and 6 seconds [Frackowiak et al., 2004].

The duration of each event was computed as:

$$duration_a = \lfloor\frac{duration}{TR}\rfloor \tag{2.10}$$

Where $duration_a$ is the adjusted duration in TRs, $duration$ is the duration as in the design, in seconds.

To ensure limited overlap of BOLD signal between different conditions, following samples of different categories had to be separated by at least $\lceil HRF_{width}/TR \rceil$ scans/TRs. The HRF_{width} can be set between 0 and 10 seconds, depending on the experimental design and application, since it influences the number of selected samples (as illustrated in Figure 2.10). A trade-off therefore exists between the number of samples n and their signal-to-noise ratio, especially in the case of fast event-related designs.

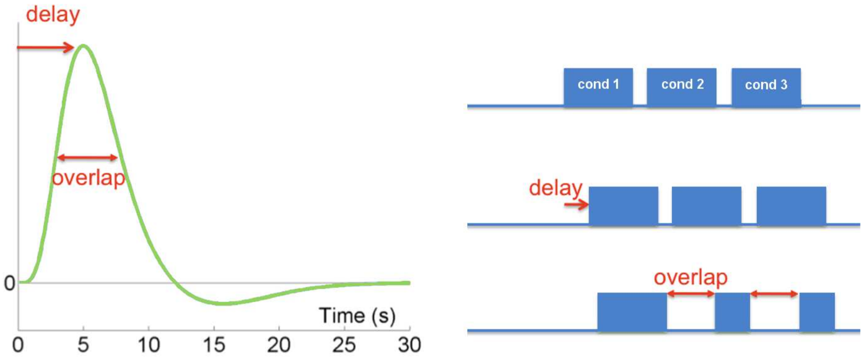

Figure 2.10: **Effect of the HRF delay and overlap.** On the left is the standard HRF response. On the right is the effect of the delay and overlap on the number of independent scans (cond 1, 2 and 3 correspond to three different experimental conditions and the blue boxes correspond to various scans acquired during each condition). In fMRI datasets, the nature of the HRF (i.e. being a delayed and dispersed version of the neuronal response to an experimental event) might lead to less independent scans/events than the ones originally acquired. Here, this issue is accounted for by discarding overlapping scans in terms of BOLD signal.

The outputs of the model, $\boldsymbol{y}$ (called labels or targets) are usually coded depending on the algorithm. In a binary case (i.e. one class versus another), the first category is usually attributed a +1 label, while the second category has a −1 or 0 label. In a multiclass classification, the labels are usually represented by the index of the class, i.e. $y_i = 1 \dots K$, with K being the number of disjoint classes/conditions. They are provided along their corresponding feature vector $\mathbf{x}_i$, $i = 1 \dots n$, during the learning phase (i.e. as inputs) and are predicted for new samples during the test phase (i.e. as outputs).

2.5.3 Classification algorithms

In this section, the linear model is presented, as well as three different methods to evaluate it. Please note that although most formulations hold for the multiclass case, this section focuses on binary classification.

The function f divides the input space into decision regions whose boundaries are defined by parameters learned during the learning phase. For example, a linear function defining $(m-1)$-dimensional hyperplanes has the form:

$$f(X) = \mathbf{w}^T X + w_0 \tag{2.11}$$

Where $\mathbf{w} \in \mathbb{R}^m$ and w_0 are the parameters of the classifier, also called weights and bias, respectively. They represent the relative contribution of each feature to the predictive task. For simplicity, w_0 is usually comprised in $\mathbf{w}$ by augmenting X.

According to equation 2.5, the outputs y are continuous and comprised in the $[-\infty, +\infty]$ interval. However, for classification, the outputs have to be “coded”, either as $+1/-1$ predictions or as probabilities lying in the [0 1] interval. When building a binary discriminant function, the sign of the output determines its class (see 2.5.3.1), such that the coding to the $+1/-1$ labels is direct. On the other hand, probabilities cannot be obtained directly. The values of $f(X)$ have thus to be “squashed” using a logistic function $\lambda(f(X))$ (eq. 2.12).

$$\lambda(f(X)) = \frac{1}{1 + exp(-f(X))} \tag{2.12}$$

To make predictions, the class leading to the highest probability is naturally chosen. This actually corresponds to the optimal decision rule under a zero-one *loss function*, for which any misclassification is penalized by one unit and correct classification by 0. Although the space of possible loss functions is infinite (with $L(k, k') \neq L(k', k)$), the zero-one loss function is a common choice. It is also the one made here.

In the present work, three algorithms were considered to estimate f: the Support Vector Machines (SVM, Burges, 1998), the Relevance Vector Machines (RVM, Tipping, 2001) and Gaussian Processes (GP, Rasmussen and Williams, 2006).

2.5.3.1 SVM

Support Vector Machines (SVM, Burges [1998]) are binary discriminant functions, which define decision boundaries to assign a label to each input. In the case of a linear SVM (eq. 2.11), this decision boundary can be drawn in many ways. This is shown in Figure 2.11 (left panel) for the case of two-dimensional samples. To find the solution leading to the best generalization ability, SVM makes use of the concept of *functional margin*, which is defined as the smallest distance between the decision hyperplane and any of the samples. For each sample $\mathbf{x}_i$, its geometric margin is defined as:

$$\gamma_i = \frac{y_i(\mathbf{w}^T\mathbf{x}_i)}{||\mathbf{w}||_2} \tag{2.13}$$

The geometric margin γ over the training set is then the minimum of γ_i, $i = 1 \ldots n$. The selected decision hyperplane is then the one leading to the largest margin (Figure 2.11, right panel):

$$\text{minimize } \tfrac{1}{2}||\mathbf{w}||_2^2 \qquad \text{over } \mathbf{w}, w_0$$
$$\text{Subject to } y_i(\mathbf{w}^T\mathbf{x}_i) \geq 1$$

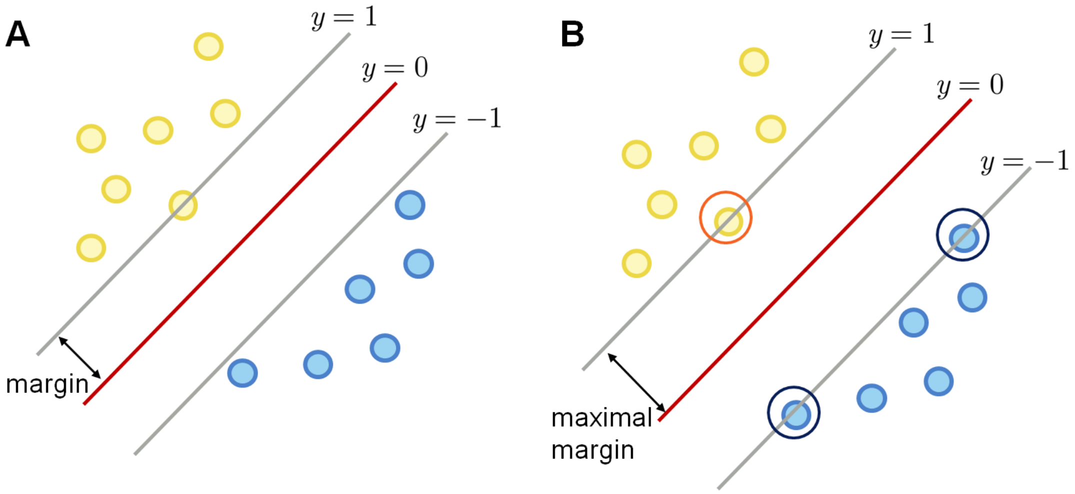

Figure 2.11: **SVM principle. A** When distinguishing between two classes (e.g. yellow versus blue), there are multiple ways of defining the decision boundary. The margin is defined as the distance from the closest point from one class to the decision boundary (in red). **B** Intuitively, the selected hyperplane would lead to the maximum margin. This results in the definition of Support Vectors (SV, circled in purple) which characterize the decision boundary. Thereby, SVM is a sparse technique.

The location of this boundary is determined by a subset of data points, the *support vectors* (circled in purple on Figure 2.11), satisfying the constraints exactly. The number of support vectors being smaller than the number of samples makes SVM a *sparse* technique.

In the case of non-separable classes, violations of the constraint are allowed but penalized:

$$\text{Minimize } \tfrac{1}{2}||w||_2^2 + C\textstyle\sum_{i=1}^{n}(1 - y_i f(\mathbf{x}_i))_+$$
$$\text{Subject to } y_i(\mathbf{w}^T\mathbf{x}_i) \geq 1$$

with $(z)_+ = z$ if $z > 0$, 0 otherwise and C being a *hyperparameter* whose value has to be optimized. This variation of SVM is called a "soft-margin" SVM classifier and might be more robust to outliers than the classical SVM. In neuroimaging, $m > n$, such that linear separability can usually be assumed and C is set to 1 [Mourão-Miranda et al., 2006; Hassabis et al., 2009].

While SVM is widely used to classify neuroimaging data (Laconte et al., 2005; Mourão-Miranda et al., 2006; Vemuri et al., 2008 for example), it has the disadvantage of not providing posterior probabilities, which might bring important information, especially in the case of clinical applications: a subject classified as healthy with an associated probability of 99% does not have the same implication as a subject classified as healthy with a probability of 51% and for which more testing might be needed.

2.5.3.2 RVM

Relevance Vector Machines (RVM, Tipping, 2001) follow the discriminative approach, such that this technique provides probabilistic predictions. Using eq. 2.5 and 2.11, the posterior in the Bayesian formulation (eq. 2.6) can be written as:

$$p(y_i|X, \mathbf{w}) = \mathcal{N}(y_i|\lambda(f(\mathbf{x}_i)), \sigma^2) \tag{2.14}$$

With $\lambda(z)$ representing any sigmoid function (e.g. eq2.12).

To limit overfitting, additional constraints have to be imposed on the parameters $\mathbf{w}$. This is usually done by adding a complexity term (regularization) and is performed implicitly in SVM using the notion of margin. Here however, the constraints are "preference bias", i.e. the constraint is in the form of a prior distribution over the parameters. To increase smoothness, these are chosen as:

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=0}^{n} \mathcal{N}(w_i|0, \alpha_i^{-1}) \tag{2.15}$$

With $\boldsymbol{\alpha}$, the $n + 1$ hyperparameters, having a Gamma distribution.

It is important to note that there is one hyperparameter per weight, which consists in an Automatic Relevance Distribution (ARD, MacKay, 1994) prior. During evaluation, this prior makes the probability mass concentrate at very high values of $\boldsymbol{\alpha}$, such that the distribution over $\mathbf{w}$ peaks at 0 mean with a variance around 0 and the corresponding weights are then *pruned*. As SVM, RVM is thus also a sparse method.

To estimate the model, Bayesian inference proceeds by:

- Computing, using Bayes' rule, the posterior over all unknowns given the data:

$$p(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\mathbf{w}, \boldsymbol{\alpha}, \sigma^2) \cdot p(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2)}{p(\boldsymbol{y})} \tag{2.16}$$

- Making predictions for a new test point in terms of the predictive distribution:

$$p(y_\star|\boldsymbol{y}) = \int p(y_\star|\mathbf{w}, \boldsymbol{\alpha}, \sigma^2) \cdot p(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2|\boldsymbol{y}) d\mathbf{w} d\boldsymbol{\alpha} d\sigma^2 \tag{2.17}$$

Both steps include normalising integrals, whose expression cannot be derived analytically because of the non-Gaussianity induced by the sigmoid function. Those integrals are thus approximated using Laplace's method as implemented in [MacKay, 1992]. It should be noted that the analytical and computational aspects of inference are out of the scope of this work and will therefore be only briefly mentioned. The interested reader can refer to [MacKay, 1992; Tipping, 2001; Rasmussen and Williams, 2006].

2.5.3.3 GP

A more general approach to the machine learning problem is the Gaussian Processes [Rasmussen and Williams, 2006], which assumes a Gaussian distribution over the latent function $f(X)$. In this model, the prior is placed on the function values $f_i = \mathbf{w}^T\mathbf{x}_i$, and the posterior can be written:

$$p(\mathbf{f}|X, \boldsymbol{y}) = \frac{p(\boldsymbol{y}|\mathbf{f}, X) \cdot p(\mathbf{f}|X)}{p(\boldsymbol{y}|X)} \tag{2.18}$$

Contrarily to the SVM and RVM techniques, the "function space" now replaces the weight space. The linear function $f(X)$ is replaced by a Gaussian Process, $\mathbf{f}|X \sim \mathcal{N}(0, K)$, with K, the $n \times n$ covariance matrix with entries depending on the samples (see 2.5.3.4). The log of the GP prior on f has the form:

$$log\, p(\mathbf{f}|X) = -\frac{1}{2}\mathbf{f}^T K^{-1}\mathbf{f} - \frac{1}{2}log|K| - \frac{n}{2}log2\pi \tag{2.19}$$

The prior on f also places a prior on the posterior $\pi(X) = p(\boldsymbol{y} = +1|X) = \lambda(f(X))$. In this model, we are not interested in the values of the latent function f but rather in $\pi(X)$, such that f, although allowing a convenient formulation of the model, will be integrated out.

Inference is then divided in two steps:

- Compute the distribution of f corresponding to a test case $\mathbf{x}_\star$:

$$p(f_\star|X, \boldsymbol{y}, \mathbf{x}_\star) = \int p(f_\star|X, \mathbf{x}_\star, \mathbf{f})\, p(\mathbf{f}|X, \boldsymbol{y}) d\mathbf{f} \tag{2.20}$$

- Use this distribution to produce a probabilistic prediction $\pi_\star$:

$$\pi_\star = \int \lambda(f_\star)\, p(f_\star|X, \boldsymbol{y}, \mathbf{x}_\star) df_\star \tag{2.21}$$

As for RVM, exact inference of the GP model isn't possible. Two approximation techniques have been used in this work: the Laplace method, which approximates the non-Gaussian posterior $p(\boldsymbol{y} = +1|X, \mathbf{w})$ by the Gaussian $q(\boldsymbol{y} = +1|X, \mathbf{w})$, using a second-order Taylor expansion, and the Expectation Propagation (EP) method, which uses local approximations of the likelihood $p(y_i|f_i)$ and optimizes the local parameters via constraints on the moments of the distribution. Both give similar accuracies in a binary context, but it has been shown in [Rasmussen and Williams, 2006] that the EP approximation is closer to the "true" distribution than Laplace's approximation. For binary problems, the EP inference method was therefore used. Both methods are fully described in [Rasmussen and Williams, 2006].

2.5.3.4 Kernel trick

In neuroimaging, the number of variables or features m is usually large, such that a linear separability of the classes can reasonably be assumed. However, m is also larger than the number of samples n, in such a way that solving equation 2.11 is an ill-posed problem. One possible solution to overcome this issue is to use regularization, which constrains the number of solutions. To perform regularization efficiently, *kernels* are usually computed, consisting in pair-wise similarity measures between all samples or patterns, summarized in a kernel matrix ($n \times n$ dimensions, instead of $n \times m$). An example of a feature space mapping $\phi(X)$ is the linear kernel:

$$\begin{aligned} K(\mathbf{x}_p, \mathbf{x}_q) &= \phi(\mathbf{x}_p)^T \phi(\mathbf{x}_q) && (2.22) \\ &= \mathbf{x}_p^T \mathbf{x}_q && (2.23) \end{aligned}$$

with $\phi(\mathbf{x}_p) = \mathbf{x}_p$, the identity mapping and $\mathbf{x}_p$, $\mathbf{x}_q$, two feature vectors.

In the algorithms in which the feature vectors $\mathbf{x}_i$, $i = 1 \ldots n$, enter only in the form of dot products, these can be substituted by the kernel. This is called the *kernel trick* and leads to dual formulation of the algorithms, which become kernel methods [Laconte et al., 2005]. Kernel methods are extremely useful, and allow to perform the learning using the kernel matrix instead of the data matrix, which is computationally more efficient. In addition to the computational advantages, using the kernel formulation together with proper regularization (i.e. restricting the choice of functions to favour those having a small norm) enables the solution of ill-conditioned problems and therefore avoids over-fitting [Shawe-Taylor and Cristianini, 2004]. SVM, RVM and GP can all be expressed as kernel methods [Bishop, 2006; Rasmussen and Williams, 2006] and have been implemented using the kernel trick.

2.5.4 Evaluation of the model accuracy

Once the model has been built, its performance has to be assessed. The quality of a model is evaluated as its ability to predict the labels for unseen/new samples, which is also defined as the *generalization ability* of the model. An error rate is then computed using a loss function (see section 2.5.3), which assigns a "penalty" to any misclassification. Computing the error rate on the dataset used to build the model would lead to overoptimistic results, since a function complex enough would lead to the perfect modelling of the training set [Hastie et al., 2003]. To this end, a test set, completely separated from the training set, must be provided to the classifier. The predicted labels for the test samples are then compared to the “true” labels and accuracy measures (1-error rate) can be derived.

In the present work, the zero-one loss function was used to present accuracies in terms of total, balanced and class accuracies, as well as under the form of a confusion matrix. This corresponds to a penalty of 1 for any misclassification and no penalty for a correct classification.

Table 2.1: Confusion matrix: example for the binary problem. TP stands for True Positives, the number of correctly classified positives (label: +1), FP for False Positives, the number of incorrectly classified positives, FN for False Negatives, the number of incorrectly classified negatives, and TN for True Negatives, the number of correctly classified negatives (label: -1). P and N represent the total numbers of +1 and -1 labelled samples, respectively.

		Truth	
Predicted		**+1**	**-1**
	+1	TP	FP
	-1	FN	TN
	Total	P	N

From the confusion matrix, the total accuracy is defined as:

$$Acc_t = \frac{TP + TN}{P + N} \tag{2.24}$$

However, when P and N are not balanced, the total accuracy is over-optimistic. The total accuracy is then replaced by the balanced accuracy:

$$Acc_b = \frac{1}{2}\{\frac{TP}{P} + \frac{TN}{N}\} \tag{2.25}$$

which actually corresponds to the mean of the class accuracies. For diagnosis purposes, when making errors in one class or the other do not have the same implications, results are usually reported in terms of sensitivity (if a disease is present, true label: +1, it is indeed detected, predicted label: +1), specificity (the disease is absent and is not reported) and positive predictive value (PPV, equivalent to a false discovery rate), defined as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.26}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.27}$$

$$PPV = \frac{TP}{TP + FP} \tag{2.28}$$

There are many ways to split the dataset into train and test sets. However, in neuroimaging, the data are scarce and this split is usually performed in a *cross-validation* (CV) scheme, which rotates the partition. A model is then estimated and tested for each partition, called *fold*, leading to measures of accuracy for each fold. The model quality is finally presented in terms of averaged accuracies across folds. In this work, a leave-one-out (LOO) CV approach was used, either to leave one subject out (all the samples related to one subject) or one block out (one block of contiguous samples, in a within-subject fMRI analysis).

More than just assessing the performance of the model, it is necessary to know if the observed result could be obtained by chance. In the present case, the null hypothesis stating that the labels do not bring any information was tested for rejection at a certain significance threshold (typically, $p < 0.05$). Assessment of the significance can be performed using either parametric or non-parametric statistic tests. The first category leads to confidence intervals at the specified threshold (e.g. 95%) but present the disadvantages of making assumptions about the data distribution. More specifically, statistical parametric tests assume that the test samples are identically and independently distributed. This condition is usually not fulfilled, especially when dealing with fMRI which induces a correlation between successive scans due to the shape of the HRF. Therefore, non-parametric testing was used in this work, the data distribution being built (as a histogram) instead of assumed. Under the null hypothesis, the labels would not bring any information and thus any random permutation of the labels should lead to the same level of performance (Figure 2.12).

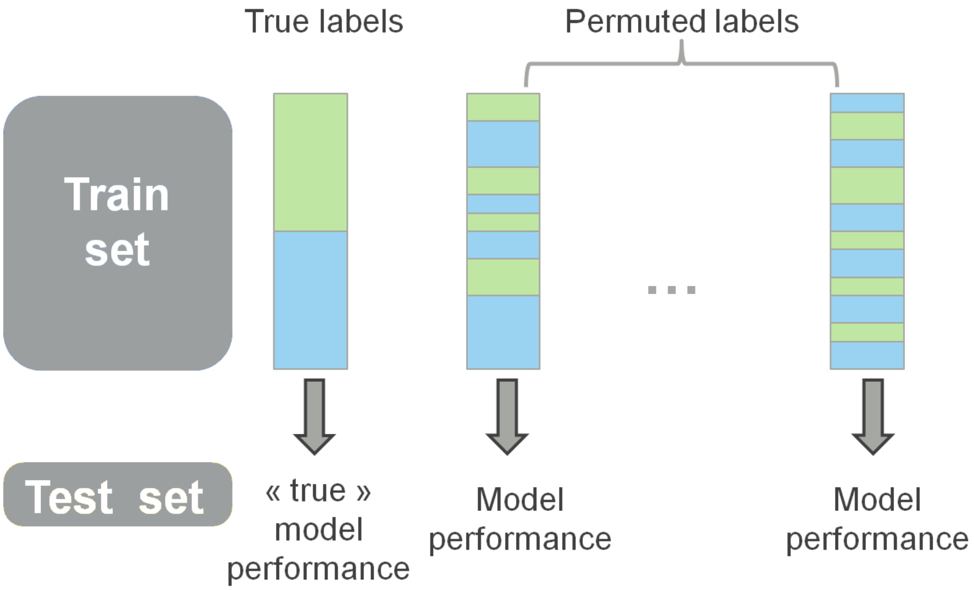

Figure 2.12: **Non-parametric testing using permutations.** To assess the significance of the performance of a machine learning based model, permutations are used: the labels of the training set are randomly permuted to obtain "baseline" model performances. The "true" model performance is then compared to the baseline level obtained from the permutations. Typically, a model performance is significant if the performance obtained by chance does not exceed or equal the true model performance more than 5% of the time ($p < 0.05$).

It is thus possible to associate a p-value to the model performance (the balanced accuracy in this case), by computing p as the number of times that a random permutation of the labels led to equal or higher performance than the performance obtained with the original labels. If the performance with the original labels significantly exceeds the level that would be expected by randomly attributing the labels, the researcher can conclude that the algorithm has truly learned some property of the data, and can therefore reject the null hypothesis that there is no information in the data about the label being predicted.

2.5.5 Interpretation of weights

As mentioned earlier, the weights represent the relative contribution of each of the features to the classification. In neuroimaging, the features correspond to the signal in each voxel, and a relative contribution of each voxel to the linear decision function is therefore obtained (see Figure 2.13 for an example of weight map). Unlike in statistical parametric mapping, it is not possible to threshold this map because the weights at each voxel are dependent on one another and no direct localization inferences or voxel-wise statistical test assuming independence can be performed on them. Although no regionally specific effects can be determined from the weight map, intuitions on which regions participated in the classification into one class or another can still be obtained. Furthermore, feature selection strategies can make use of the weights as the selection criterion to determine which voxels contribute most to the considered discrimination (see section 5.3).

2.5.6 Multiclass classification

Up to now, only binary classification was considered. However, distinguishing between more than two categories is often desirable (e.g. in the case of multiple forms of a disease). Different strategies allow performing multiclass classification: using K binary one-versus-all classifiers, using $K(K-1)/2$ binary one-versus-one classifiers [Fürnkranz, 2002] or using a multiclass model.

In the binary one-versus-all approach, the samples of one class have a +1 label while all the others a -1 or 0 label. This problem can be solved via fast and extensively tested algorithms, such as RVM or GP. For predictions, the class leading to the largest value of the decision function is selected. The main disadvantage of this technique is that it implies imbalances across categories. To avoid this issue, binary classifiers were used in a one-versus-one fashion, in which the classifiers between all possible pairs of classes are estimated, leading to $K(K-1)/2$ predictions and/or probabilities. These outputs then need to be recombined in order to obtain a unique multiclass prediction. Different techniques exist to perform the recombination, starting by a simple vote. However, an Error-Correcting Output Code (ECOC, Dietterich and Bakiri, 1995), based on [Hassabis et al., 2009] was used here: each class was represented by a codeword of length $K(K-1)/2$, the number of binary classifications. Each classifier votes for the two classes it was built for, and for each class the votes of all the classifiers are assembled to constitute a "codeword", which is further used for comparison with test points. For each test instance, the distance between the vector computed from the predictions of the set

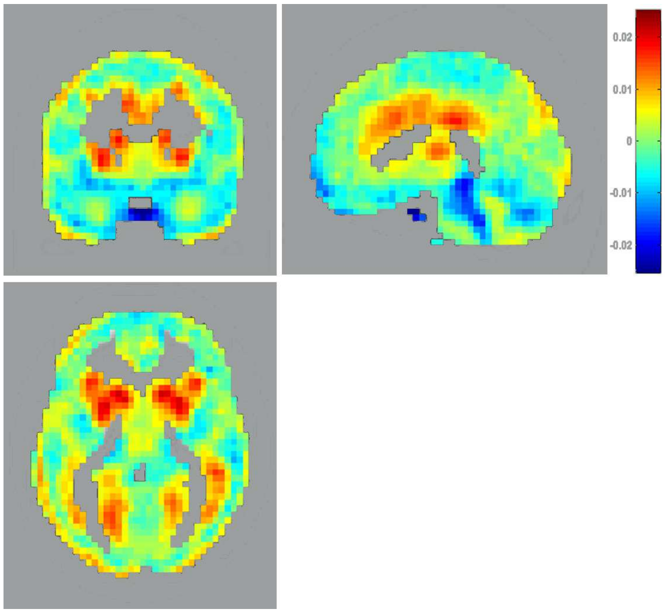

Figure 2.13: **Voxels' weights.** (Adapted from [Phillips et al., 2011]). Distribution over the brain volume of the voxel relevance for a machine trained to discriminate between PET images of vegetative state patients and healthy subjects. A positive value (yellow-red) indicates that relatively large metabolic activity in those voxels will drive the classification towards the control group. On the contrary, a negative value (blue) indicates that relatively large metabolic activity in those voxels will drive the classification towards the patients' group. The voxels with little relevance (green) hardly contribute to the classification of data. Image from [Schrouff and Phillips, 2012].

of classifiers and the correct codewords associated to each possible class can then be computed and the class characterized by the smallest distance from the predicted vector is selected. With SVM binary classifiers, the outputs are defined by +1/-1 labels (see Table 2.2, left) and the final class of a test point was attributed according to the smallest Hamming distance between this vector and all the candidate class codewords [Hassabis et al., 2009]. With GP classifiers, the codewords were defined in terms of probabilities obtained from each binary classification (see Table 2.2, right, Schrouff et al., 2012), and the distance was computed as the sum of the differences between the table and the probabilities obtained from each binary classifier (L_1 distance).

The difference between the two tables, binary and probabilistic, lies in the precision of the distance between the vector of predictions associated to a test instance and the different codewords. This is illustrated in Table 2.3.

When using the SVM predictions, the distance measure used to assign the final class is the Hamming distance (i.e. the number of differing bits), which would give

Table 2.2: Example of codewords for a 3 classes problem using predictions (left part) and probabilities (right part) of the binary classifier. The lines correspond to the considered classes while the columns represent the different binary comparisons (A, B, C, any three classes).

	Prediction codewords			Probabilistic codewords		
	A-B	A-C	B-C	A-B	A-C	B-C
class A	1	1	0	1	1	0.5
class B	-1	0	1	0	0.5	1
class C	0	-1	-1	0.5	0	0

Table 2.3: Outputs of SVM and GP classifiers (second and third rows respectively) applied on one example data point (true class B) for a 3 classes problem.

Test point (B)	A-B	A-C	B-C
SVM predictions	-1	-1	-1
GP probabilities	0.2	0.3	0.5

for the three classes:

$$\begin{aligned}
L_A &= 1+1+1=3 \\
L_B &= 0+1+1=2 \\
L_C &= 1+0+0=1
\end{aligned}$$

where L represents the final score of each class. In the present case, the class C is assigned to the test point since it shows the smallest final score L, which leads to a misclassification (true class: B, see Table 2.3). It is interesting to note that a simple vote would lead to the same result, i.e. a misclassification. On the other hand, the probability based codewords lead to the following scores:

$$\begin{aligned}
L_A &= |1-0.2|+|1-0.3|+|0.5-0.5| = 1.5 \\
L_B &= |0-0.2|+|0.5-0.3|+|1-0.5| = 0.9 \\
L_C &= |0.5-0.2|+|0-0.3|+|0-0.5| = 1.1
\end{aligned}$$

Where class B was correctly assigned to the test point. Therefore, whenever possible (i.e. when the classifier returned probabilities), the probability based codewords were used to perform the ECOC scheme.

Finally, multiclass classification can also be performed by replacing the logistic function by a softmax function $\lambda(f_c(X))$ for each class:

$$\lambda(f_c(X)) = \frac{exp(f_c(X))}{\sum_{c'=1}^{K} exp(f_{c'}(X))} \tag{2.29}$$

In a GP formulation, generalisation to a multi-class case can be performed quite directly in the Laplace method, but not in the EP method, such that Laplace approximation was used in the case of multi-class problems [Rasmussen and Williams, 2006]. While many other multiclass models exist (e.g. DAGSVM, Platt et al., 2000, . . .), it was shown that their use doesn't bring any improvement in terms of generalization ability when compared to combinations of binary classifiers [Hsu and Lin, 2002; Rifkin and Klautau, 2004]. Therefore, we limited our choice of multiclass models to a multiclass GP, with Laplace inference.

2.5.7 Softwares

All methods described here and in following chapters were implemented in Matlab (Mathworks). Preprocessing and univariate analysis of the images was performed using SPM8 (`www.fil.ion.ucl.ac.uk/spm`). The SVM implementation used is the LIBSVM toolbox (Chang C. C. and Lin, C. J., `www.csie.ntu.edu.tw/~cjlin/libsvm`) with a PROBID interface (A. Marquand and J. Mourão-Miranda, `www.brainmap.co.uk`), which is a standard implementation of a classical SVM as is commonly employed in decoding neuroimaging data. RVM was implemented by M. Tipping (`www.miketipping.com`, Tipping [2001]). The GP implementation used is the compiled version coded by C. E. Rasmussen and C. K. I. Williams (Rasmussen and Williams [2006], `www.gaussianprocess.org/gpml`) and also interfaced in PROBID.

The codes written during the first part of this work (i.e. for the neuroscience application) helped in the implementation of a new software, in collaboration with J. Mourão-Miranda, J. Richiardi, J. Ashburner, A. Marquand, C. Chu, C. Phillips, J. Rondina and M.J. Rosa. This software, called PRoNTo (Pattern Recognition for Neuroimaging Toolbox, Schrouff et al., 2013, `www.mlnl.cs.ucl.ac.uk/pronto`), was then used in the second part of this work (clinical application).

Part I

neuroscience Application

Chapter 3

Introduction

Contents

Classically, brain operations are considered as essentially reflexive and mainly driven by external stimuli. In this perspective, brain function is predominantly geared to interpreting incoming stimuli and programming motor output. Another view posits that the bulk of brain's activity is intrinsic, spontaneous (i.e., it emerges in the absence of any identified external stimulus), and essentially aims at maintaining and processing information [Raichle, 2006]. Consistent with this view, the energy required for the brain to respond to external stimuli is extremely small compared to the ongoing amount of energy that the brain normally and continuously expends [Raichle and Mintun, 2006]. It is assumed that perception, memory, and even the stream of consciousness result from this spontaneous activity. In consequence, the characterization of spontaneous brain activity now stands as a central issue in understanding how human brain processes information conveyed by external stimuli or endogenous processes, including those related to past experience or current stream of thoughts.

The application of machine learning based models on functional neuroimaging data has recently made it possible to decode mental states, based on objective measurements of regional brain activity. Although decoding spontaneous brain activity stands as a fascinating challenge, it faces a number of technical, methodological and ethical difficulties. First, the absence of objective control of mental representations associated with spontaneous brain activity complicates the signal extraction and feature selection steps. Second, to be able to decode, a model needs to be trained on a specific problem. This is in contradiction with the multiple possible states thought to take place during spontaneous activity and would suppose a model with innumerable categories to distinguish. Finally, decoding spontaneous brain activity brings up ethical questions: how much would you like your personal thoughts to be revealed? Although mind reading is still far away, applications of machine learning based models to lie detection is now investigated, showing that ethical issues already arise and should be carefully taken into account.

A first step in the decoding of spontaneous brain activity would be to consider experimental conditions in which the nature of spontaneously active mental representations is constrained by the experimental protocol. Such a situation can arise in the framework of a study of human declarative memory, and more specifically in the context of memory consolidation.

In this chapter, we briefly introduce the theories on memory, and memory consolidation, as well as previous attempts to characterize its functioning in animals. State-of-the art methods to investigate this phenomenon in healthy humans are then presented in section 3.2, along with their promising results and limitations. Finally, section 3.3 presents how machine learning based models might bring new insight on the characterization of mnemonic traces and summarizes the aim of the present study.

3.1 Memory consolidation

Although being able to memorize information is necessary for living, little is known about how new memories are formed and can then be accessed as long as years after. According to [Gazzangina et al., 2002], the theory of memory actually assumes three steps:

- the *encoding*, which is the processing of incoming information.
- the *mnemonic consolidation*, during which a permanent record of that information is created and maintained.
- the *retrieval*, which consists in retrieving the information on purpose.

Memory consolidation is a necessary step to retain information in the long-term (from days to years). Current research suggests that the brain activity patterns generated during the encoding phase are spontaneously repeated and that this repetition arises predominantly when the cortex is “off-line”, i.e. not engaged in the processing of external stimulation, which is referred to as *spontaneous brain activity*. Mnemonic traces have hence been detected in animals during different vigilance states such as Non-Rapid Eye Movement (NREM) sleep [Ji and Wilson, 2007], Rapid Eye Movement (REM) sleep [Louie and Wilson, 2001] or resting-state wakefulness [Hoffman and McNaughton, 2002]. Their exact timing and duration is still poorly characterized but previous works suggest that the time frame of these patterns is compressed or expanded according to an unknown scaling factor, SF [Louie and Wilson, 2001; Lee and Wilson, 2002]. [Ji and Wilson, 2007] also showed that the number of mnemonic traces decreases in time, starting around 30 minutes after the encoding. These studies therefore support the theory of memory consolidation and give hints on its temporal aspect.

Regarding its spatial location, the theory suggests that episodic memory consolidation takes place in the hippocampus [Gazzangina et al., 2002; Peigneux et al., 2006]. However, case studies have shown that impairing the medial temporal lobe (containing the hippocampus) did not impair remote memories, suggesting that consolidated memories are not located in the hippocampus, but rather in the neocortex [Gazzangina et al., 2002]. This further implies that memory consolidation moves the burden of retention from hippocampo-noecortical circuits to purely cortical long-term stores, as suggested by [Ji and Wilson, 2007].

Detailed characterization of firing replays during memory consolidation revealed their specific temporal structure: [Louie and Wilson, 2001] and [Lee and Wilson, 2002] showed a reactivation of temporally sequenced information in sleeping rats and [Foster and Wilson, 2006] discovered a reverse replay of behavioural sequences in the awake state in rats. Although the direction of the replay is different, it seems that the structure of the learning material (referred to as the "phase" information) has to be maintained during consolidation of memories.

Finally, [Girardeau et al., 2009] showed that disrupting memory consolidation by hippocampal stimulation during post-training sleep resulted in impairments in behavioural performance to a spatial memory task, suggesting that the “strength” of reactivation could be linked to behavioural performance.

While these studies provide some evidence supporting the theory of memory consolidation, they were performed on animals, using intra-cranial recordings. Apart from the obvious fact that intra-cranial recording cannot be done on healthy humans, the conclusions hold for individual neurons only and need to be verified for large neuronal populations. This is particularly the case for the hypothesis regarding the phase information: showing that discharges of individual neurons follow the sequence imposed by previous waking activity does not show whether this firing in sequences involves large neuronal populations. Furthermore, in most experiments, the animals had to be trained before the experiment, such that the actual learning achieved could not be estimated.

3.2 State-of-the art

In humans, non-invasive neuroimaging techniques such as PET and fMRI have been used to investigate memory consolidation. Spontaneous activity was acquired in “resting-state” sessions, during which the subject lies (in the scanner), eyes closed, and is not submitted to any external stimulation. Analyzing rest sessions represents a big challenge, due to their dimensionality (number of scans $\times m$, the number of voxels) and to the absence of “ground truth”. Up to now, researchers aimed at reducing the dimensionality of the data [Margulies et al., 2010], mostly by computing activation maps (GLM analysis) or by selecting Regions of Interest (ROIs). This allowed indirect characterization of mnemonic traces, through its effect on other tasks [Peigneux et al., 2006] or via seed-correlations [Tambini et al., 2010]. Although these studies only brought indirect support of the theory of memory consolidation, they are in line with what was shown in animals using intra-cranial recordings (see further).

Previous work on memory consolidation can be classified in two groups: the "activation" studies and the analyses of ROI interactions.

3.2.1 Activation studies

Activation studies usually consist in searching brain areas responding to an external stimulation. In the case of memory consolidation, GLM analyses are usually performed on both the learning task and the rest sessions and the obtained maps are then compared in search for (statistical) overlap.

Experience-dependent reactivations in the processing of mnemonic traces could be observed during human REM sleep following a sequence learning task [Maquet et al., 2000]. Moreover, [Peigneux et al., 2003] showed that this reactivation reflected the reprocessing of high-order components of sequence learning, i.e. the sequential contingencies contained in the learning material. They further showed that the strengthening of memories only occurs when the learning material is structured, which is in line with the work of [Louie and Wilson, 2001] on rats.

When investigating declarative memories (in contrast with motor sequence learning which involves implicit, non-declarative memories), [Peigneux et al., 2004] showed that the amount of hippocampal activity during NREM sleep (and more particularly in the deep sleep stage) was related to the subject's overnight improvement in behavioural performance. This confirms the hypothesis linking the "strength" of the reactivation to the behavioural performance.

As shown in animals, evidence about memory consolidation was also found during active wakefulness following a learning task [Peigneux et al., 2006]. The authors indeed showed that the brain responses to an unrelated task were modulated by a previous learning task and that this post-training activity correlates with behavioural performance.

While these studies support the theory of memory consolidation, they do not allow a direct characterization of mnemonic traces during resting-state wakefulness. Furthermore, some activation studies are not immune to confounds such as order effects not controlled for or, more importantly, concurrent practice of the learned material.

3.2.2 Interactions between ROIs

When computing interactions between regions of interest in the search of mnemonic traces, the functional connectivity between specific ROIs thought to be activated during the learning process is usually computed.

Different techniques exist to select and compute functional interactions between ROIs. This work focuses on two main procedures: the seed-correlation and spatial network analyses, which both average the signal within each ROI, giving m_{ROI} (the number of ROIs) time-courses serving as the basis to compute interactions between ROIs.

A seed-correlation analysis is simply the manual selection of ROIs. This selection can be based on an a priori hypothesis about the problem or on (previous) activation studies. After signal averaging within each ROI, comparison of different resting conditions is performed via the correlation coefficients computed between all possible pairs of ROIs. Regarding the detection of mnemonic traces, this technique has proven useful to detect changes in ROI interactions before and after a memory task. More specifically, Tambini et al. [Tambini et al., 2010] investigated the off-line transfer of information between the hippocampus and the neocortex. They showed an enhanced marginal correlation between the hippocampus and neocortex (lateral occipital cortex) during post-task rest compared to baseline rest, which predicted individual differences in later associative memory. While their results are promising, they do not directly characterize mnemonic consolidation. Furthermore, this technique presents the disadvantage of an arbitrary selection of ROIs as well as becoming quickly intractable since the number of pairwise comparisons increases according to $\mathcal{O}(\frac{m_{ROI}^2}{2})$, with m_{ROI} the number of regions.

A more advanced technique relies on the selection of extended large-scale functional brain networks, which consist of segregated regions (potentially spread over the brain) that interact in order to perform a functional task [Marrelec et al., 2008]. When using networks of regions, integration measures can be computed hierarchically, i.e. at the whole brain, network and ROI levels. This hierarchical decomposition of the problem enables a deeper insight on the results and reduces the tractability issue since interactions are not computed pairwise but on a within- and between-networks basis (usually no more than 10 networks, [Damoiseaux et al., 2006]), based on the entropy of each ROI ($\mathcal{O}(m_{ROI})$). The analysis of interactions within- and between-networks allowed investigating the differences in integration between various states of consciousness, such as between wakefulness and sleep [Boly et al., 2012] or wakefulness and anaesthesia [Schrouff et al., 2011].

To compare the results from machine learning based models with state-of-the art methods, integration within- and between-networks will be computed to try to detect mnemonic traces. More specifically, an increase in integration between ROIs in the hippocampus and ROIs in the neocortex would be expected. Due to the visual and auditory character of the considered experiment (see section 4.2), changes in integration within the early visual and/or auditory areas would also be expected.

3.3 Aim of this study

When looking for mnemonic traces, spatially distributed and, more importantly, transient events are expected. In the analysis of ROI interactions, the whole time series is reduced to only one measure per ROI (its entropy or its pair-wise correlation coefficient). This temporal reduction is also performed in the activation studies since they rely on the β parameters of a GLM analysis (one parameter per condition and per voxel). We can therefore suppose that these techniques are not the most appropriate to directly highlight memory consolidation or characterize its temporal evolution. Furthermore, they usually rely on a (manual) selection of ROIs, which can lead to biases in the results as explained in [Schrouff et al., 2011].

Due to the limitations of both the activation studies and ROI interaction analyses, models based on machine learning were considered to help supporting the theory of mnemonic consolidation. The first argument in favour of machine learning based models is that they detect *patterns* and therefore do not assume the voxels to be independent or reduce the spatial dimensionality of the data. Mnemonic traces can thus be assimilated to the reactivation of patterns of brain activity generated during encoding. Furthermore, there is no need to reduce the temporal dimension since the scans can be treated as a succession of samples. This enables the detection of mnemonic traces in each scan and thereby allows investigating the temporal evolution of memory consolidation.

In this work, we aim at finding evidence supporting the theory of memory consolidation by characterizing its different aspects as suggested by previous studies. More specifically, a specific experiment was designed, comprising a control and a memory conditions, both consisting of a control or memory task, respectively, flanked by two rest sessions (see section 4.2). If the theory is valid, we expect to find:

1. Scans in the resting-state sessions which can be linked to the task [Hoffman and McNaughton, 2002], named here reactivation patterns. Note that the detection of false positives is expected, such that reactivation patterns should be found in all the considered rest sessions.

2. The proportion of these reactivation patterns should be higher in the post-task than in the pre-task rest session in the memory condition, their difference being (significantly) larger than in the control condition [Tambini et al., 2010].

3. The increase in proportion of scans linked to the task should be related to the subject's behavioural performance, and this for the memory condition only [Peigneux et al., 2006].

4. If the learning material contains temporally structured spatial locations, the strengthening of the material should follow this structure (or its reverse) [Louie and Wilson, 2001; Lee and Wilson, 2002; Foster and Wilson, 2006].

5. The proportion of scans linked to the task in the post-task rest session (for the memory condition) should decrease along time [Ji and Wilson, 2007].

Finally, a prospective scaling factor, SF [Louie and Wilson, 2001], will be investigated since no previous work could study this parameter of memory consolidation in resting-state healthy humans.

Aim: Apply machine learning based models to detect and characterize patterns of brain activity generated during a learning task unconsciously rehearsed during following spontaneous activity.

Chapter 4

Material and Methods

Contents

In this chapter, the data considered to investigate mnemonic consolidation is presented. Particular attention was paid to the design of the experiment, especially in terms of means of control since this was a limitation in previous studies. The techniques envisaged to model both the constrained and spontaneous brain activity are further exposed. Finally, section 4.7.2 details the computation of functional interactions between resting-state networks, as detected on the spontaneous brain activity sessions.

4.1 Population

A group of 14 volunteers (7 females), aged between 19 and 29 years (mean 24.44), participated in the study. This study was approved by the Ethical Committee of the Faculty of Medicine of the University of Liège. All participants were fully informed, gave their written informed consent and were paid for their participation. All included participants were non-smoking, healthy right-handed students. The volunteers were screened for anxiety (Beck anxiety inventory,[Beck et al., 1988]), depression (Beck depression inventory II, [Steer et al., 1997]), sleep quality (Pittsburgh sleep quality index, [Buysse et al., 1989]), chronotype (Horne and Ostberg morningness - eveningness questionnaire, [Horne and Ostberg, 1976]), excessive daytime sleepiness (Epworth sleepiness scale, [Johns, 1991]), laterality (Edinburgh Inventory, [Oldfield, 1971]), amount and content of daydreams (Imaginal Process Inventory- `www.themeasurementgroup.com/evaluationtools/ipi`). The participants presented no medical, traumatic, psychiatric or sleep disorders. During the 7 days preceding the experiment, volunteers followed a regular sleep schedule, verified by wrist actigraphy and sleep diaries.

4.2 Experimental Design

fMRI acquisition for all the volunteers was split into two main activation conditions: a *control* condition and a *memory* condition, their order being randomized (Figure 4.1). The control condition (represented by "o") consisted in an auditory discrimination task based on the oddball paradigm [Squires et al., 1975], flanked by two rest sessions lasting 10 minutes each and further referred to as $R1_o$ and $R2_o$, respectively.

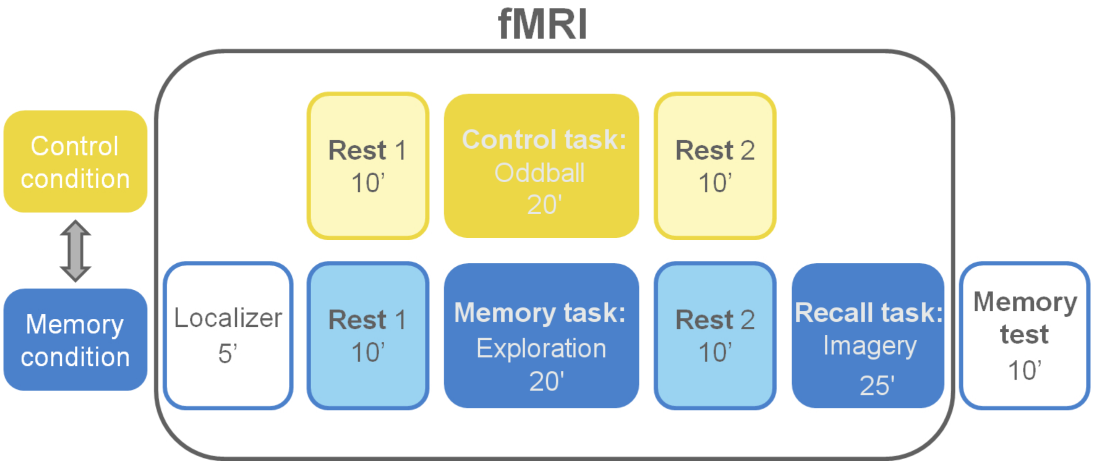

Figure 4.1: **Experimental design.** Subjects underwent a control task flanked by two rest sessions and a memory task, flanked by two rest sessions and followed by a recall or mental imagery session. A functional localizer preceded the memory condition to avoid novelty effects. Finally, the subjects were tested on their learning of the memory task outside the scanner.

The memory condition (represented by "m") consisted in five successive sessions:

- **Localizer.** During the first session, images of faces, buildings and animals were presented in random order at the centre of the screen during 500 ms with an inter-stimulus interval of 1500ms (Figure 4.2, **A**). The purpose of this session was both to identify brain areas responding to the three image types and to eschew novelty effects during subsequent sessions.
- $\boldsymbol{R1_m}$. Pre-task rest session, eyes closed during 10 minutes. No instruction.
- **Exploration.** During the memory task, the images shown during the localizer session were displayed one at a time for 3 seconds, each image being assigned a specific location on the screen. The order of presentation followed a predefined sequence of contiguous screen positions in such a way that volunteers had the impression of following a path throughout a bidimensional maze (Figure 4.2, **B**). The complete maze consisted of three blocks of 27

consecutive images within which the 3 categories of images were always presented in the same order (i.e. 9 faces, 9 buildings and 9 animals). Between blocks, a fixation cross was displayed for 15 to 18 seconds. To ensure optimal encoding, the whole path was repeated five times during the scanning session. Volunteers were instructed to pay attention to each image, to their location on the screen and to their succession.

- ***$R2_m$*.** Post-task rest session, eyes closed during 10 minutes. No instruction.
- **Mental imagery.** During the last session, volunteers were presented with 54 memory tests. During each test, two images, simultaneously displayed on the screen for 4 seconds, represented the starting and target positions of a trajectory that the volunteers would have to follow mentally (Figure 4.2, **C**). The mental trajectories included 3 to 6 images (average 4.5) of a same category. For each image that they could conjure up during this mental travel, volunteers had to signal by a key press whether it was a face, a building or an animal (one finger and key per condition). However, participants had the possibility to skip a path if they could not remember any part of it. The expected number of images of each type was perfectly balanced between categories.

A memory test was finally performed outside the scanner, in order to behaviourally assess the accuracy of the spatial knowledge acquired by the volunteers. They were presented with the previously seen pictures and 48 novel images in random order. For each trial, an image was displayed on the screen at a specific location and participants had to specify whether this image was part of the maze and, if they believed it was, if it was displayed at its correct location. The behavioural performance was then computed using a d' [Green and Swets, 1966] measure, which takes into account the percentage of hits (i.e. recognition when the image has been previously displayed) and of false alarms (i.e. recognition although the image was not included in the memory task).

The classification procedures were first applied on the exploration and mental imagery sessions. These were designed such that the participant performed a totally controlled task during the exploration session, whereas during mental imagery, the pace and succession of mental representations were not constrained by external stimuli but only by the volunteer's capacity to retrieve the learned stimuli and their location. The latter led to possibly unbalanced data across categories if one type of images was better remembered than the others. A further characteristic of the exploration session was that within a block, no rest period was introduced at the transition between images of different categories. As a consequence, fMRI signals of different classes of events were expected to overlap, making correct classification more complex. Furthermore, during the mental imagery session, the event duration was not fixed and depended entirely on the speed at which each participant recalled the requested images. This resulted in event durations varying between 200 ms and 4000 ms, with most events during less than 2000 ms.

Finally, the previously built models were applied to all rest sessions (both from the control and memory conditions) to try to highlight mnemonic traces. The pro-

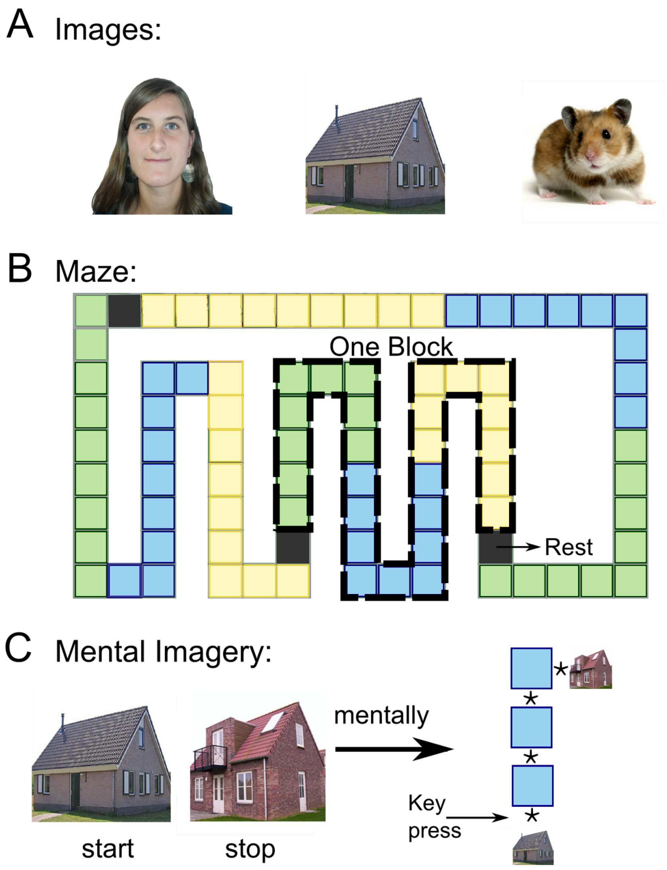

Figure 4.2: **Illustration of the experimental setup for the memory condition.** **A** Example of images of faces, buildings and animals presented to the subject. In total, 81 different images were used. **B** Synoptic view of the maze, green areas stand for images of faces, blue areas, buildings and yellow areas, animals. The succession of three areas of each color is called a block. **C** A mental trajectory begins with the start and stop points being displayed on the screen. The subject then travels mentally in the maze, mentally visualizing all the images comprised in the path and pressing a key every time he visualizes the required image.

portions of detected patterns were then compared across conditions and correlated with the participant's behavioural performance.

4.3 Data acquisition

Functional MRI time series were acquired on a 3T head-only scanner (Magnetom Allegra, Siemens Medical Solutions, Erlangen, Germany) operated with the standard transmit-receive quadrature head coil. Multislice T2*-weighted functional images were acquired with a gradient-echo echo-planar imaging sequence using axial slice orientation and covering the whole brain (34 slices, FoV = 192×192 mm^2, voxel size= 3×3×3 mm^3, 25% interslice gap, matrix size 64×64×34, TR = 2040 ms, TE = 30 ms, FA = 90°). The three initial volumes were discarded to avoid

T1 saturation effect. The static field inhomogeneities were measured using a field mapping sequence (32 slices, FoV =220×220 mm^2, voxel size = 3.4×3.4×3 mm^3, 30% interslice gap, TR of one slice=517 ms, TE= 4.92 and 7.38 ms, FA= 90°), using the same brain coverage and slice orientation as for the EPI sequence. Finally, a high-resolution T1-weighted image was acquired for each participant (3D MDEFT Deichmann et al., 2004; TR of one slice = 7.92 ms, TE = 2.4 ms, TI = 910 ms, FA = 15°, FoV = 256 ×224×176 mm^3, 1 mm isotropic spatial resolution).

4.4 Signal extraction

For the sessions considered for further modelling, the whole time series of all voxels were extracted. The data matrix X was then adjusted for movement effects (estimated by realignment parameters) and low frequency drifts (cutoff: 1/128 Hz) using a GLM, as described in section 2.5.2. For exploration and mental imagery, the signal corresponding to stimulus onsets was then extracted, considering a hemodynamic response function (HRF) delay of 6 seconds (according to Frackowiak et al., 2004) and an HRF width of 0 seconds to keep as many samples as possible. To avoid decoding the signal linked to motor activity in the mental imagery session, the scans selected for further classification were the ones preceding the key presses (after correction for HRF delay). The signal was finally averaged over specific time-windows to increase the Signal-to-Noise Ratio (SNR; Kamitani and Tong, 2005; Mourão-Miranda et al., 2006). For the exploration session, the average was performed over the time the stimulus was presented (i.e. 3 seconds). For the mental imagery session, this average was performed over the interval between two key presses, with a maximum of 2 scans (i.e. 4.080 seconds) to avoid the inclusion of episodes of task-unrelated thoughts. For rest sessions, each scan was considered as a sample and the adjusted data matrix directly entered the test phase of the classification, without any further treatment.

4.5 Feature selection

Feature selection consists in selecting a subset of features which contains as much information as the whole set with the advantages of reducing memory requirement [Formisano et al., 2008] and increasing the signal-to-noise ratio, thereby improving overall performance [Guyon and Elisseeff, 2003]. There are three families of methods to select features (please see [Guyon and Elisseeff, 2003] for a complete description of each method):

- *filters*, which rank the variables according to a predefined criterion. A filter is a preprocessing step, independent of the choice of the predictor.

- *wrappers*, which build subsets of features according to their relevance to a given classifier.

- *embedded methods*, which penalize large number of features during the estimation of model parameters.

In this work, a univariate filter based on the results of a GLM analysis and a multivariate wrapper were considered, selecting features according to the accuracy of SVM classifiers, both separately and then joined as suggested in [Guyon and Elisseeff, 2003].

The univariate filter determined (from a GLM analysis) the subset of "active" voxels (i.e. whose activity was statistically significantly correlated with the three conditions, Mitchell et al., 2004). From the resulting F-maps, two voxel sets were selected: (1) all voxels above an F-threshold of 0.5 (referred to as "global GLM feature selection") and (2) the 1000 most significant voxels [Shinkareva et al., 2008] (referred to as "specific GLM feature selection"). This number of 1000 was chosen as the trade-off between an under-constrained space (dimensionality much larger than 1000), which might lead to overfitting, and an over-constrained space (dimensionality much smaller than 1000), which would make the second feature selection step useless. The GLM feature selection was performed on the training set only, to ensure unbiased estimations of the accuracy.

As mentioned earlier, the weights of a machine learning model represent the relative contribution of each of the variables to the classification. These weights can then be used as a criterion to select features: a binary SVM using linear kernels was used to rank the voxels according to their "discriminating power", which was computed from their specific weights [Mitchell et al., 2004]. The voxels with the largest absolute weights were selected for further modelling. The number of selected voxels systematically varied from 5 (per condition and binary comparison) to 150 at most, by increments of 25 (respectively m_{min}, m_{max} and Δm in Figure 4.3). These parameters were fixed arbitrarily. At each iteration, the sum of the accuracy of the three binary models on a left-out block was taken as a global accuracy measurement. In general, the addition of relevant features increases the accuracy of the classification while adding irrelevant features leads to a decrease in accuracy [Bishop, 2006]. When recursively adding features, the global accuracy is therefore expected to increase and then decrease (when irrelevant features are being added to the relevant ones). The set of voxels leading to the highest global accuracy (i.e. when the global accuracy starts decreasing compared to the 2 previous iterations) was then selected for the classification analysis (see Figure 4.3 for an illustration of the process). Features are thus added recursively following a "Recursive Feature Addition" (RFA) procedure, in contrast to "Recursive Feature Elimination" in which features are recursively discarded [De Martino et al., 2008]. RFA can then be assimilated to a forward wrapper feature selection, with a cost function based on the global accuracy as objective.

4.6 Modelling constrained brain activity

Classification was performed using binary SVM and GP classifiers with the ECOC approach to obtain multiclass predictions. While exploration and mental imagery can be treated the same way, rest sessions present a different issue which has to be solved separately. Therefore, the (semi-) constrained brain activity was first modelled by testing different combinations of feature selection and classifiers. Then,

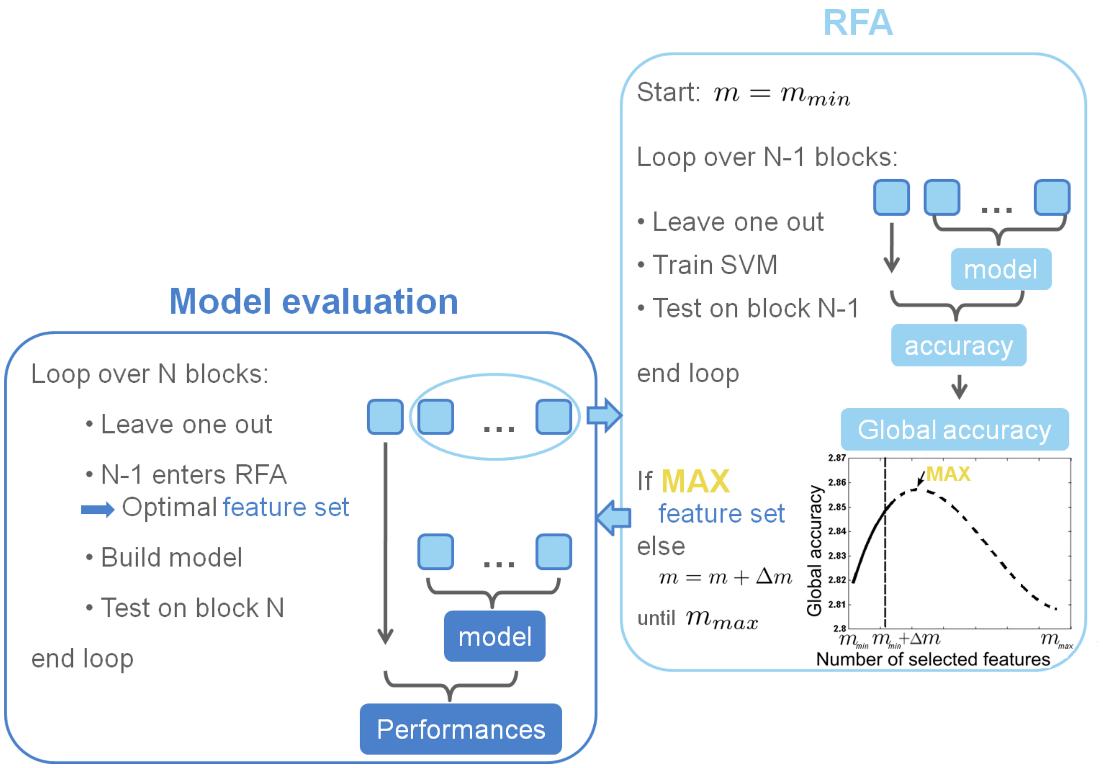

Figure 4.3: **Recursive Feature Addition (RFA) process and Cross-Validations (CV)** used in procedures 3 to 5. The accuracy measure is presented in the left box: from the outer CV, one block (containing n_{block} events) is left out which will be used later to test the final classification accuracy and does not enter the feature selection process. The $N-1$ other blocks enter the RFA process (right box) to define the optimal set of features, which will be used to build the model: the inner CV tests an SVM model built on $N-2$ blocks, leading to a value for the global accuracy (sum of the accuracies obtained for each binary comparison). This inner CV loop is repeated until the accuracy curve starts to decrease and hence a maximum value of global accuracy is reached, corresponding to an optimal subset of variables. N represents the number of blocks, m_{min} (respectively m_{max}) represents the minimum (respectively maximum) number of selected features and Δm, the step size.

the procedure leading to the best results was selected for modelling the rest sessions and to detect mnemonic traces.

The different feature extraction, GLM and Recursive Feature Addition (RFA), and classification (SVM and GP) methods were combined in five distinct "procedures" (Table 4.1), which were conducted as follows:

- Procedure 1: The specific GLM feature selection method identifies the 1000 most *active* voxels in the considered experimental design. SVM binary classification is then performed for each pair of image types.

- Procedure 2: The specific GLM feature selection method identifies the 1000 most *active* voxels in the considered experimental design. GP binary classification is then performed for each pair of image types.

- Procedure 3: The global GLM feature selection method identifies all *active* voxels above the F-threshold of 0.5. Then, RFA is performed with a range of selected features from 5 to 150, selecting the number of features leading to the best generalization accuracy. GP classification is then performed using this number of selected features.
- Procedure 4: The specific GLM feature selection method identifies the 1000 most *active* voxels in the considered experimental design. RFA is then performed with a range of selected features from 5 to 150 [Shinkareva et al., 2008], selecting the number of features leading to the best generalization accuracy. GP classification is then performed using this number of selected features.
- Procedure 5: The specific GLM feature selection method identifies the 1000 most *active* voxels in the considered experimental design. RFA is then performed with a range of selected features from 5 to 150 [Shinkareva et al., 2008], selecting the number of features leading to the best generalization accuracy. SVM classification is then performed using this number of selected features.

Table 4.1: Outline of the different combinations of features extraction and classification methods used in the present study.

	Feature selection		**Classification technique**	
	GLM	**SVM**	**SVM**	**GP**
Procedure 1	specific		x	
Procedure 2	specific			x
Procedure 3	global	x		x
Procedure 4	specific	x		x
Procedure 5	specific	x	x	

With procedures 1 and 2, features were only selected by a GLM analysis and accuracies were computed in terms of leave-one-block-out cross-validations: at each step, one block containing n_{block} data (i.e. 27 consecutive images of faces, buildings and animals for the maze exploration session, and all the mentally represented images of 6 consecutive paths for the mental imagery session) was left out as a test set while the others were used as a training set to build the SVM or GP model. With procedures 3, 4 and 5, a nested leave-one-block-out cross-validation was needed (Figure 4.3) to ensure the independence of feature extraction and classification [Mitchell et al., 2004; Guyon and Elisseeff, 2003]. The inner cross-validation was used to determine the number of features to be selected by RFA and therefore obtain an optimal SVM model on N-1 blocks while the outer leave-one-block-out cross-validation tested the built SVM or GP model. The outer cross-validation was performed on the same folds for all procedures, which therefore allowed their comparison.

Due to class imbalance in the mental imagery session, balanced accuracies were computed (as the mean of the class accuracies) to take the different frequencies of the classes into account and therefore replaced the total accuracy. To assess the significance of the classification of each procedure on each participant, permutations of the training set labels were performed (labels were permuted within each block to preserve class frequencies in each temporally correlated block). P-values were then associated to the balanced accuracy value of each participant, by comparing it to the balanced accuracy obtained when shuffling the labels 100 times per cross-validation step (i.e. 1500 times for exploration and 900 times for mental imagery in total).

Finally, the different procedures were compared using Friedman tests based on the balanced accuracy across participants but also on the accuracies for each class, which allowed a better insight on the particularities of each modelling technique. In particular, the proportions of Support Vectors (SV) for each class was computed for the three SVM binary classifiers of procedure 5 to investigate the effect of an unbalanced data set on the SVM technique and link them to class accuracies.

4.7 Modelling spontaneous brain activity

As mentioned in chapter 3, analysing spontaneous brain activity is usually performed by reducing the temporal and spatial spaces (spatial network analysis or analysis of Regions of Interest, ROIs). However, when looking for mnemonic traces, spatially distributed and transient events are expected. Therefore, it seems that neither of those methods would be able to directly highlight memory consolidation. Machine learning based models were hence considered, which do not (directly) reduce the temporal or spatial space.

4.7.1 Machine learning based models

Assessing the modelling of spontaneous brain activity (as in rest sessions) is a bigger challenge due to the absence of ground “truth”: there are no means of checking the accuracy of the model or comparing different rest sessions. To solve this issue, the “confidence” of the classifier was used, rather than its predictions. It was indeed assumed that if a scan was really linked to the task, the classifier would be more confident about its prediction than for a random prediction. This confidence measure allowed assessing scans which were *significantly* linked to the task. Finally, to compare the memory and control conditions, each rest session was summarized by only one value.

4.7.1.1 Confidence of the classifier

More specifically, the model built on the mental imagery session was applied to each scan of the rest sessions. In this case, the ECOC approach was not used to associate a label to each scan but to assess the confidence of the prediction by computing the distance between the two most probable classes (in terms of distances to the codewords in Table 4.1, [Dietterich and Bakiri, 1995]). A unique measure, referred to as L, is therefore attributed to each scan and is computed as:

$$L = L_{k1} - L_{k2} \tag{4.1}$$

with $L_{k1} = \min\limits_{k=1...K} L_k$ and $L_{k2} = \min\limits_{k=1...K\setminus k1} L_k$, the difference between the two most likely classes.

The baseline-level of L was then computed using permutations of the training labels (Figure 4.4, $P = 1000$). For each permutation, the maximum of L was considered (as in the correction for multiple comparisons), which allowed comparing the L value of each scan (L_i) to the 1000 L values of the permutations and thereby associating a probability to each scan. The proportion of scans linked to the memory task was then computed as the percentage of scans for which the associated p-value was smaller than 0.05.

$$\begin{aligned} L_{perm} &= L_p, & p = 1 \dots P \qquad (4.2)\\ p(i) &= \frac{L_{perm} > L_i}{L_{perm}} & (4.3)\\ Pr(Rest) &= \frac{1}{n}\sum_{i=1}^{n} p(i) < 0.05 & (4.4) \end{aligned}$$

For each participant, four values were obtained (one per rest session), referred to as Pr.

To support the theory of mnemonic consolidation, an increase in Pr from pre-task to post-task rest in the memory condition should be observed, and this increase should be significantly larger than that observed in control condition. This is the first hypothesis and can be translated as:

$$Pr(m) = Pr(R2_m) - Pr(R1_m) > Pr(o) = Pr(R2_o) - Pr(R1_o) \tag{4.5}$$

The significance of the result being assessed by a Friedman statistical test.

The second hypothesis relates to the link between the proportion of mnemonic traces and the participant's behavioural performance. According to the theory, the higher $Pr(m)$, the higher the participant's performance, and there should be a positive correlation between $Pr(m)$ and d', denoted by C_m. This correlation coefficient should be larger than the one obtained for the control condition (denoted by C_o). To assess the significance of both C_m and C_o, the values of d' were randomly permuted 1000 times, which allowed computing the baseline level of both correlation coefficients and thereby associating a p-value to C_m, C_o and also to their difference $C_m - C_o$.

4.7.1.2 Temporal structure of memory consolidation

The temporal structure of the mnemonic consolidation was then investigated by assuming that the sequence followed in the exploration session (i.e. faces-buildings-animals) would be replayed during the rest session. When considering a scan significantly linked to the task (i.e. $p(i) < 0.05$), its transition should hence follow

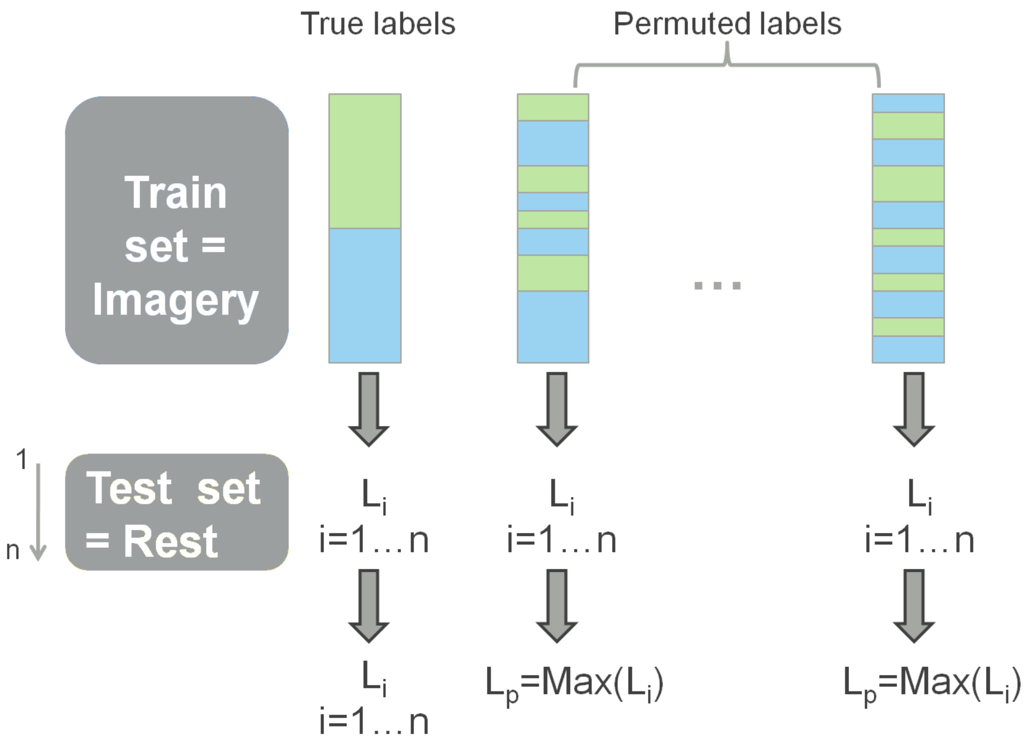

Figure 4.4: **Baseline level of the confidence measure.** The baseline level of the confidence measure, L, is computed by permuting the labels of the training set (mental imagery) and applying the built model on the test set (rest sessions). For each scan and permutation, one measure of L is obtained (L_i). The maximum of L_i per permutation is retained (L_p) to be compared with each of the L_i computed using the true labels.

Table 4.2: Possible transitions for the considered experimental design. The diagonal terms represent identical transitions. The blue off-diagonal terms represent transitions according to the forward sequence, i.e. the sequence designed in the exploration session. The red off-diagonal terms represent the inverse sequence, which will be used as control. F:faces, B:buildings, A:animals, i:i^{th} scan, x:number of transitions.

	$\boldsymbol{F_i}$	$\boldsymbol{B_i}$	$\boldsymbol{A_i}$
$\boldsymbol{F_{i-1}}$	×	×	×
$\boldsymbol{B_{i-1}}$	×	×	×
$\boldsymbol{A_{i-1}}$	×	×	×

the sequence (Figure 4.5). In this experiment, there are 3 classes, and 9 possible transitions which can be represented by the Table 4.2.

In this table, the selected transitions are the ones from $i-1$ to i (Figure 4.5). It is hence possible that the $i-1^{th}$ scan is not significantly linked to the task. This should especially arise in the case of false positives or during the start of a succession of scans with $p < 0.05$ (further referred to as an “episode”). Although this might

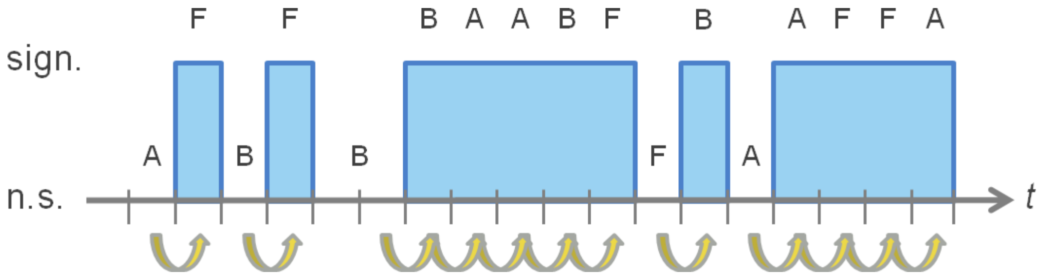

Figure 4.5: **Investigating the phase information.** The p-value of each scan can be binarized into "significant" (sign.) or "not significant" (n.s.) to form a vector of 0s and 1s of the size of the time-series (t). To investigate the phase information, a transition matrix was built, counting the number of transitions from one class to another, considering the transition between the scan at (t-1) and the scan significantly linked to the task at (t). The transitions taken into account are represented by arrows.

result in a bias, the bias should be equally present in all rest sessions and both conditions. Increases in proportions of transitions corresponding to the sequence from the pre-task to the post-task rest session will therefore be investigated and compared between the memory and the control conditions.

$$\Phi(Rest) = \frac{\sum_{p(i)<0.05} \times}{n} \tag{4.6}$$

with $\times$ representing any transition following the sequence imposed during learning and n, the total number of scans in the considered rest session.

To ensure that this increase is not only reflecting an increase in Pr, the inverted sequence will be used as control, with the hypothesis that there is no increase in Φ from $R1_m$ to $R2_m$ when considering the inverted sequence. Regarding the diagonal terms corresponding to identical transitions, they could pertain to either the forward or inverted sequence and will therefore not be taken into account.

The fourth and last hypothesis suggests a decrease in $Pr(m)$ with time. This is investigated by dividing the time-series in two parts, early and late, and computing Pr_1 and Pr_2 for each rest session. If the hypothesis is verified, a decrease in Pr should be observed in time, i.e. $Pr_1(R2_m) > Pr_2(R2_m)$. For the other sessions, the temporal evolution of Pr should be random, but not necessarily stable or increasing. Performing statistical tests on the decreases in Pr would therefore be difficult and will not be conducted in this work. Only a qualitative impression can thus be given.

A possible scaling factor (i.e. factor of compression or expansion of the pattern time frame) was also investigated by computing the number of successive scans significantly linked to the task. By assuming that most false positives will be isolated (i.e. with a duration of 1 scan), the proportions of episodes with a duration larger than a certain threshold (ranging from 1 to 10 scans), as well as their duration, were computed for each rest session and then compared. This is illustrated by a toy example in Figure 4.6.

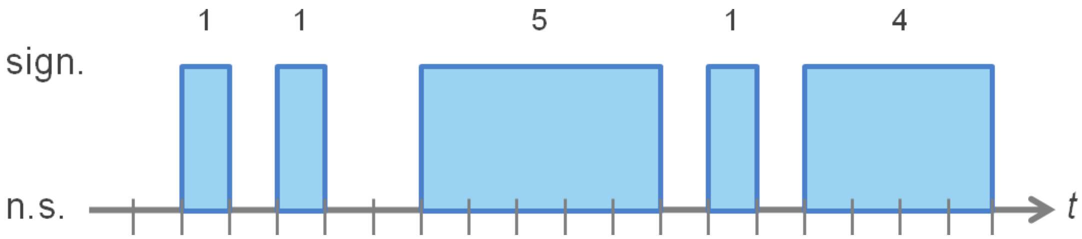

Figure 4.6: **Investigating the Scaling Factor.** The p-value of each scan can be binarized into "significant" (sign.) or "not significant" (n.s.) to form a vector of 0s and 1s of the size of the time-series (t). An episode is the succession of one or more scans with $p < 0.05$. To investigate the Scaling Factor, SF, we computed the proportion and duration of episodes lasting a defined threshold or more. If the threshold is set to 2 scans, the duration is 4.5 scans ($= \frac{5+4}{2}$), and the proportion of episodes lasting at least two scans is 0.4 ($\frac{2}{5}$).

Finally, the effect of the threshold at $p < 0.05$ to define a scan as significantly linked to the task was investigated. More specifically, this threshold was varied from $p < 0.01$ to $p < 0.1$ with a step of 0.01 and the proportions Pr_o and Pr_m as well as the significance of the correlations C_o and C_m were computed.

4.7.2 Spatial Networks

To compare the results from machine learning based models with a state-of-the-art analysis of resting-state fMRI data, a three-step procedure was used to compute functional interactions between networks (Figure 4.7), as implemented in the NetBrainWork toolbox (`sites.google.com/site/netbrainwork`).

First, the detection of functional networks at the group level was achieved using NEDICA (NEtwork Detection using ICA, [Perlbarg et al., 2008]), which detects networks at the individual level using spatial independent component analysis (ICA). After registration into the MNI standardized space (using the SPM2 software), a hierarchical clustering was performed on the independent components (IC) from all participants, yielding a similarity tree. The partitioning of the similarity tree into classes relied on the idea that each class should ideally be composed of one and only one IC from each participant. Two parameters were computed to quantify this idea, which allowed selecting the consistent classes across participants. A group t-map was associated with each selected class [Perlbarg et al., 2008]. The group representative classes, the spatial structure of which was characteristic of known functional networks according to the literature [Damoiseaux et al., 2006; Smith et al., 2009], were used for subsequent analysis as the main networks of interest (NOIs). Maps corresponding to noise processes or not characteristic of any previously identified functional network were discarded.

Second, 20-voxels regions of interest were selected around the peaks of each group t-map and corresponded to the main nodes of functional networks. At this stage, none of these nodes correspond to the main activated areas during the learning task since they were detected on all rest sessions concatenated. We therefore manually added the regions which were significantly activated during the exploration session,

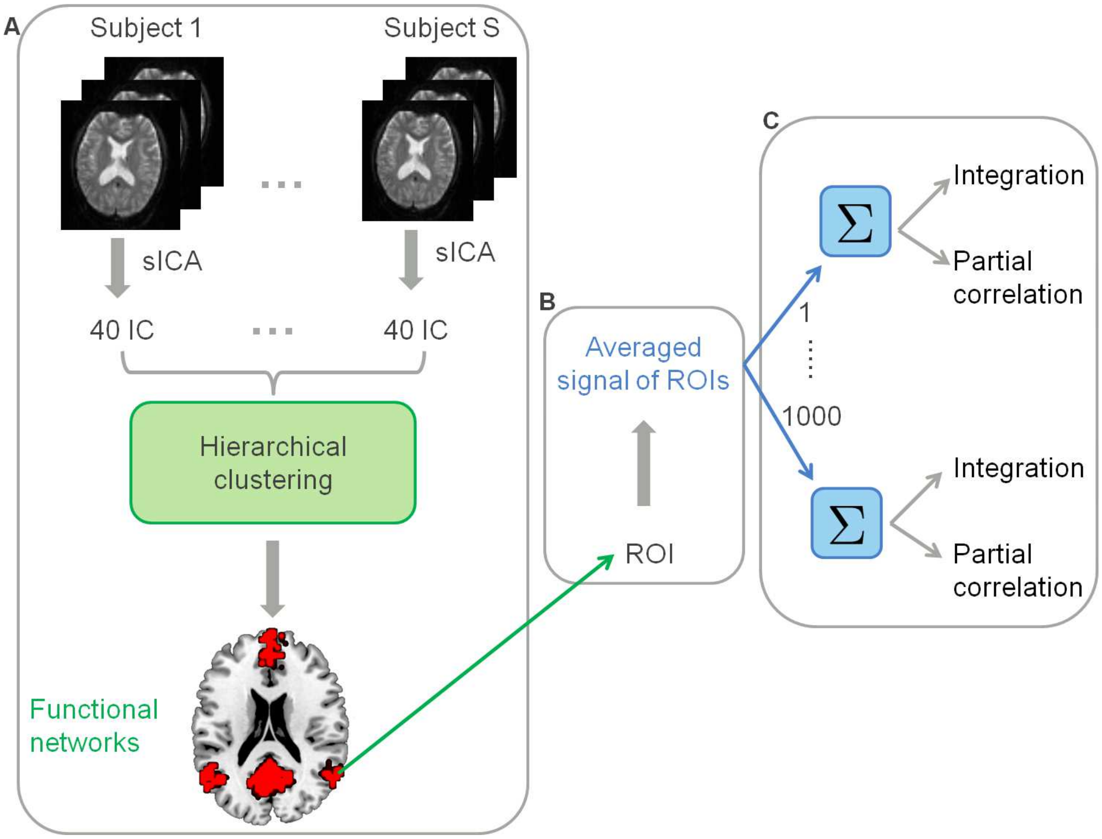

Figure 4.7: **Three-step procedure used to compute integration and partial correlations at the group level. A** Spatial ICA was applied on each subject, leading to 40 IC (default value). These $S \times 40$ IC were then hierarchically clustered and IC corresponding to cardiorespiratory artefacts were discarded. Group t-maps were then computed revealing functional networks at the group level. **B** From these maps, ROIs were automatically selected and used for the computation **C** of integration and partial correlations via a 1000 samples Bayesian numerical sampling scheme of the posterior distribution.

as assessed by a GLM and defined two NOIs: one “hippocampal” comprising three regions in the hippocampus and one “maze” comprising the Fusiform Face Area (FFA), the Parahippocampal Place Area (PPA) and two regions activated during the display of images of animals (Animals Area, AA).

Third, the functional interactions within and between the NOIs were quantified using two types of measures, hierarchical integration and partial correlation. As a preprocessing step before the computation of functional interactions, the CORSICA method (CORrection of Structured noise using Spatial Independent Component Analysis, [Perlbarg et al., 2007]) was used to take physiological noise into account. This technique takes advantage of the fact that the spatial distribution of physiological noise or head motion signals is independent of the TR of the acquisitions. In particular, CSF pools such as the ventricles appear to act as detectors of head motion and physiology-related movements, and the major blood vessels as detectors of cardiac activity. CORSICA includes three successive steps: spatial ICA decomposition, selection of noise-related components using specific masks of interest comprising the ventricles, the brainstem and the basilar arteries, and removal of those components.

Hierarchical integration corresponds to the mutual information between time courses of BOLD signal recorded in the various ROIs [Marrelec et al., 2005, 2009] . It provides a global measure of functional information exchanges within and/or between brain systems. Furthermore, if a system is divided into subsystems, the total integration of this system can be decomposed into within-subsystem and between-subsystem integration (Marrelec et al., 2008, illustrated in Figure 4.8). In particular, the total integration of the brain is equal to the sum of within-NOIs integration and between-SOIs integration. To infer the integration measures, a Bayesian numerical sampling scheme approximating the posterior distribution of the parameters of interest in a group analysis is necessary [Marrelec et al., 2006]. In the present work, 1000 samples were used to perform this approximation, therefore leading to a thousand estimations of integration measures. The results are presented in terms of the mean and standard deviation of the 1000 estimates. The interested reader can find further information about the concept and the computation of hierarchical integration in [Marrelec et al., 2008].

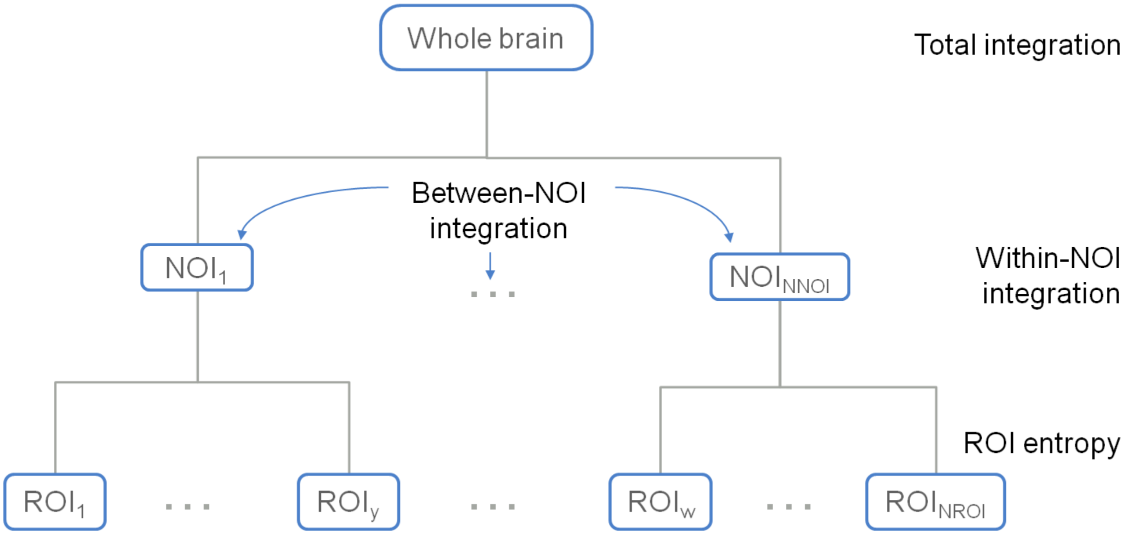

Figure 4.8: **Illustration of the hierarchical computing of integration.** The hierarchical tree comprises the whole brain, NNOI networks of interest (NOIs), and NROI regions of interest (ROIs). The top level is the level denoted “whole brain” and is associated with total integration. Total integration is computed as the sum of NNOI terms of within-NOI integration and one term of between-NOI integration. Both within- and between-NOI integrations are computed at the NOI level using the entropy of the ROIs, which is calculated at the bottom level of the hierarchy.

Since hierarchical integration provides a global measure of interaction, it is unable to quantify pairwise functional connectivity between ROIs. To do so in a given network, we resorted to partial correlation. Partial correlation is a measure of functional connectivity that is more closely related to effective connectivity than simple correlation [Marrelec et al., 2007, 2009; Smith et al., 2011]. It was computed in the same manner as integration (by using the 1000 samples of the Bayesian numerical sampling scheme, [Marrelec et al., 2005, 2009]). The density of connections at a given threshold was computed as the number of ROI pairs for which partial correlation was above the threshold. A "connectivity" curve was derived by computing the density of connections across a range of thresholds. Finally, the integral of the difference between two curves obtained under different conditions, hereafter referred to as Integrated Difference in Partial Correlations (IDPC) was computed. IDPC is independent from the partial correlation threshold and quantitatively estimates the differences in connectivity between conditions.

According to [Smith et al., 2009], networks can be detected both at rest and when the brain performs a task. The authors also showed that functional networks at rest corresponded to the brain maps of activation under certain tasks. In particular, a cognition and memory task would mainly activate regions related to the executive and ventral attentional networks. However, such relationship between the behavioural domain and the detected networks cannot be thresholded and all networks should therefore be considered when performing integration, correlation or partial correlation analyses.

In the present case, it is tempting to use only the regions which were detected as the main peaks during the exploration session in the GLM analysis. While this can be justified in view of the considered hypotheses, this however does not represent the reality, since the signal from each supplementary region is taken into account in integration and partial correlation measures. Therefore, the results depend heavily on the selected number of components and ROIs. To avoid the (possibly biased) selection of components, all detected networks were considered for further analysis, as well as the manually detected ROIs.

Chapter 5

Results

Contents

In this chapter, we present the results corresponding to the different steps of the analysis and for all sessions. The first results (section 5.1) are used to ensure the absence of any outlier subject in terms of behavioural parameters (e.g. anxiety, depression and sleep quality), as well as the participants' performance to the learning task. Secondly, the data and feature sets were built in sections 5.2 and 5.3, respectively. The results of the different modelling procedures are then reported and compared in section 5.4 for constrained (exploration session) and semi-constrained (mental imagery) brain activity. Application of the best procedure to spontaneous brain activity is exposed in section 5.5.1. The rest sessions were also analysed in terms of interactions within- and between-networks (section 5.5.2). In both cases and when possible, each aspect of the memory consolidation theory was investigated.

5.1 Behavioural data

The behavioural results are presented in terms of scores to the different questionnaires and behavioural performance d' computed from the test led outside the scanner.

5.1.1 Scores

During screening, two participants were identified as outliers in terms of anxiety and depression. Since the potential effects of these parameters are still poorly understood, results for these participants will be displayed in red.

Another participant admitted to have used a strategy based on the alphabet to mentally represent the maze (i.e. as an alphabetical explicit list not relying on mental image representation). This might affect our ability to decode the rest sessions since our model is based on mental imaging only. Results for this participant will be displayed in blue.

Furthermore, these three participants admitted having consciously rehearsed the bi-dimensional maze during the post-task rest session. This might lead to either a facilitated detection of mnemonic traces or, on the opposite, to increased noise in the considered session.

The different hypotheses supporting the theory of memory consolidation will therefore be investigated considering (1) all participants and (2) without the behavioural outliers.

5.1.2 Participants' performance

The participants' performance to the learning task was computed by taking into account the hits, misses, false alarms and correct rejection rates to the memory test conducted outside the scanner, as assessed by d' in terms of content of the images (Table 5.1).

A Friedman test showed an effect of category on the participants' performance (p=0.0042). Post hoc paired Wilcoxon signed rank tests showed that d' in the faces and animals categories was significantly larger than in the buildings category (F-B: p = 0.0017, A-B: p=$6 * 10^{-4}$, Bonferroni corrected for multiples comparison) whereas no significant difference was detected between d' corresponding to faces and animals (p = 0.8508).

5.2 Signal extraction

The results of the signal extraction step are presented in terms of number of events extracted for each session. These numbers rely on the considered formulation for event onset and duration (see equations 2.9, 2.10) but also on the participant's ability to mentally retrieve the images in the mental imagery session.

During exploration, 135 events were extracted for each category, each one lasting 3 seconds. During mental imagery, the number of extracted events and their corresponding duration were variable depending on the volunteer's ability to retrieve the different images forming the requested mental path (Table 5.2).

These findings were consistent with the participants' performance d': the number of events is significantly lower (Friedman test: $p = 0.0054$, Post hoc Wilcoxon tests,

Table 5.1: Participants' behavioural performances for each class as well as for all categories confounded (d' on the content). Statistical outliers in terms of behaviour are highlighted with a different colour.

Participant	d'(Faces)	d' (Buildings)	d' (Animals)	d' (Total)
S1	2.73	2.40	3.56	3.61
S2	4.74	1.99	3.87	3.17
S3	4.74	3.87	3.56	4.69
S4	4.74	2.08	3.87	3.23
S5	3.87	2.93	4.74	4.29
S6	4.74	2.44	4.74	3.57
S7	4.74	1.65	3.87	2.96
S8	3.56	3.19	3.87	4.29
S9	4.74	0.72	3.19	2.17
S10	2.33	3.05	2.60	2.54
S11	3.19	1.68	2.47	3.18
S12	2.44	0.94	3.13	1.94
S13	2.35	1.99	3.87	2.56
S14	2.60	2.64	4.74	3.01

$p < 0.05$, Bonferroni corrected for multiple comparisons) for the buildings category than for the other 2 categories whereas no significant difference was detected between the number of events in the faces and animals categories ($p = 0.0347$, does not survive the Bonferroni correction).

This result potentially affects the classification based on fRMI data as it relies on binary comparisons.

It should be noted that no direct interaction could be detected between the performances of the participant and the number of events in each category (correlations: $p > 0.05$).

5.3 Feature selection

Figure 5.1 displays the 1000 voxels selected from the "specific GLM" feature selection, for both the exploration and mental imagery session of participant S1.

The "global GLM" feature selection option considered in procedure 3 led to about 35,000 selected voxels for both sessions (range: 30,372-40,527, mean: 36,203 for the exploration session; range: 29,996-38,541, mean: 34,172 for the mental imagery session).

Table 5.2: Number of extracted events for the mental imagery session. Percentages are in brackets and show the possible imbalances between categories. Statistical outliers in terms of behaviour are highlighted with a different colour.

Subject	Faces	Buildings	Animals	Total
S1	69 (44.23%)	30 (19.23%)	57 (36.54%)	156
S2	76 (55.07%)	11 (07.97%)	51 (36.96%)	138
S3	47 (30.32%)	58 (37.42%)	50 (32.29%)	155
S4	63 (37.72%)	50 (29.94%)	54 (32.34%)	167
S5	74 (39.36%)	42 (22.34%)	72 (38.30%)	188
S6	65 (39.39%)	43 (26.06%)	57 (34.55%)	165
S7	70 (41.42%)	36 (21.30%)	63 (37.28%)	169
S8	43 (40.19%)	31 (28.97%)	33 (30.84%)	107
S9	67 (38.29%)	44 (25.14%)	64 (36.57%)	175
S10	18 (21.69%)	32 (38.55%)	33 (39.76%)	83
S11	37 (45.68%)	20 (24.69%)	24 (29.63%)	81
S12	69 (41.07%)	31 (18.45%)	68 (40.48%)	168
S13	55 (32.54%)	58 (34.32%)	56 (33.14%)	169
S14	77 (38.89%)	53 (26.77%)	68 (34.34%)	198
Mean	59.29 (39.17%)	38.50 (25.44%)	53.57 (35.39%)	151.36

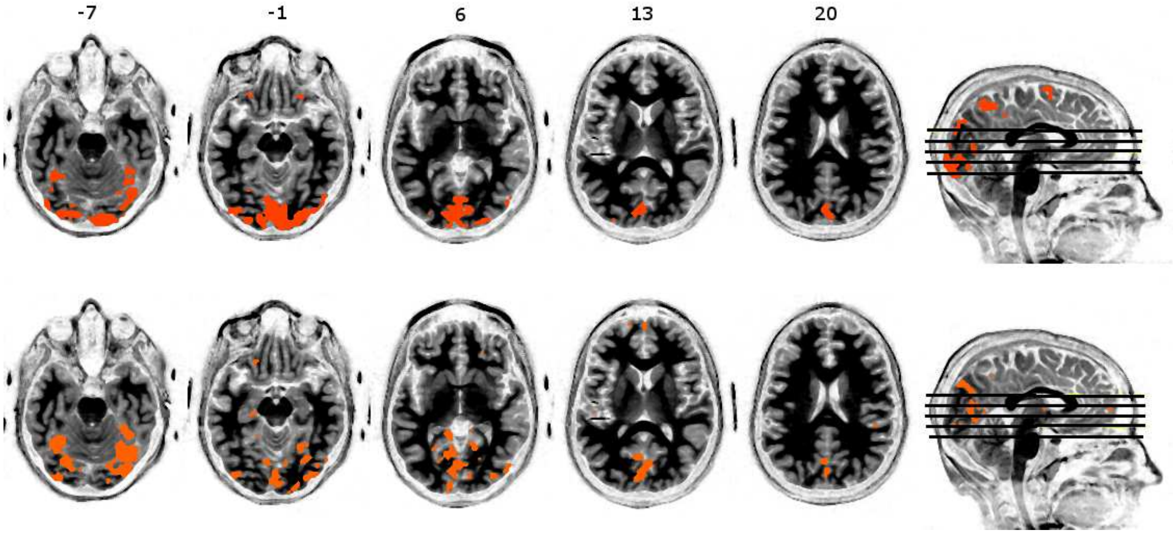

Figure 5.1: **Specific GLM feature selection.** Voxels selected after the "specific" GLM feature selection in the exploration (top) and mental imagery (bottom) sessions, for subject 1, when leaving the first block out. The z-coordinate of the slices are displayed on top.

The number of features extracted by RFA in procedures 3, 4 and 5 are summarized in Table 5.3.

Table 5.3: Number of RFA selected features for procedures 3, 4 and 5. The optimal subset of variables is represented for each participant by its average size (second and fourth columns) and standard deviation (third and fifth columns) across blocks for the exploration (second and third column) and mental imagery (fourth and fifth) sessions. The last line gives the mean and standard deviation across participants. Results are presented in terms of mean and standard deviation across the number of features obtained after each cross-validation step. Statistical outliers in terms of behaviour are highlighted with a different colour.

	Procedure 3				**Procedures 4 and 5**			
	Exploration		**Imagery**		**Exploration**		**Imagery**	
Subject	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
S1	369	5	220	142	337	65	254	173
S2	360	66	385	149	277	88	258	150
S3	369	3	191	134	274	76	272	185
S4	324	88	351	245	310	78	314	164
S5	341	89	357	274	326	67	297	163
S6	363	55	266	249	338	66	290	172
S7	375	4	349	262	365	29	278	204
S8	234	100	225	256	216	88	217	134
S9	354	89	346	153	314	103	306	190
S10	364	54	258	168	313	65	148	185
S11	352	58	509	201	366	28	304	222
S12	344	85	313	108	333	40	294	183
S13	391	45	272	139	291	109	326	227
S14	370	30	241	201	327	67	205	165
All	350.71	55.07	305.93	182.36	313.36	69.21	268.79	179.79

Procedure 3 identified 350.71 optimal features (305.93 for mental imagery, mean across blocks and across participants) while procedures 4 and 5 selected 313.36 features (268.79 for mental imagery), the difference between procedures being significant (Friedman test, $p < 10^{-4}$) for exploration. Standard deviations in the number of voxels selected indicate a high variability across blocks for mental imagery, independently of the procedure. This high variability across blocks, precluding from any conclusion at the procedure level, is directly linked to the design of the session. For exploration, the variability across blocks is small for both procedures, suggesting that the computation of a GLM for each LOO-CV does not induce much variability in the subset of voxels selected (Friedman test on the residuals, $p = 0.7276$). Procedure 5 being identical to procedure 4 in terms of feature selection, Table 5.3

displays the sizes of the selected subsets of features for both procedures.

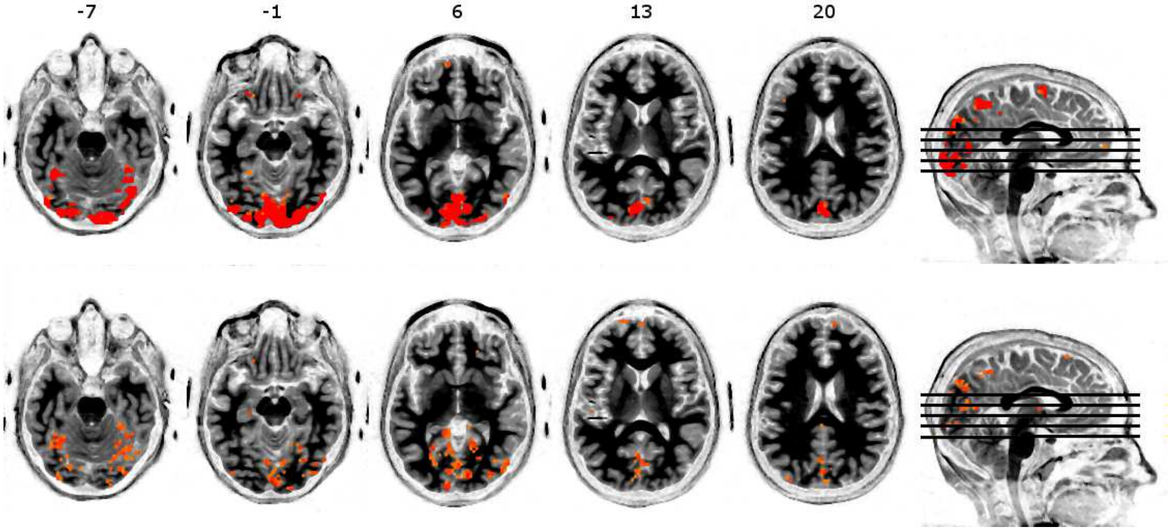

Figure 5.2: **Specific GLM and RFA feature selection.** Voxels selected after the "specific" GLM feature selection followed by RFA (procedures 4 and 5) in the exploration (top) and mental imagery (bottom) sessions, for subject 1. The images across blocks have been averaged, such that darker areas correspond to voxels commonly selected for different LOO-CV folds, while lighter areas correspond to "outliers". We can see that the selection is more consistent across blocks for the exploration than for the mental imagery session. The z-coordinate of the slices are displayed on top.

For both sessions and all procedures, the selected voxels were mostly comprised in the ventral visual path (primary areas, Fusiform Face Area), parietal regions linked to spatial features and hippocampus related to navigation (see Figure 5.2 for an average across blocks for participant 1, procedures 4 and 5). Activation in these areas represented properly the different aspects of both tasks.

5.4 Modelling (semi-)constrained brain activity

5.4.1 Classification accuracy

The exploration session was first modelled, before considering the mental imagery session. In the following sections, the results for both sessions and the five procedures are expressed for each category in terms of balanced accuracy (mean across blocks and significance for each participant, Figures 5.3 and 5.4).

Procedure 1. For exploration, the mean balanced accuracies were all above chance level, ranging from 54.57 to 89.88 % ($p < 0.05$). For mental imagery, mean balanced accuracies ranged from 26.59 to 67.01 %. Low accuracy measures led to non-significant results for participants S5, S10 and S13 ($p > 0.05$).

Procedure 2. For exploration, GP classification provided mean balanced accuracies ranging from 56.05 to 90.12% ($p < 0.05$). For mental imagery, mean balanced

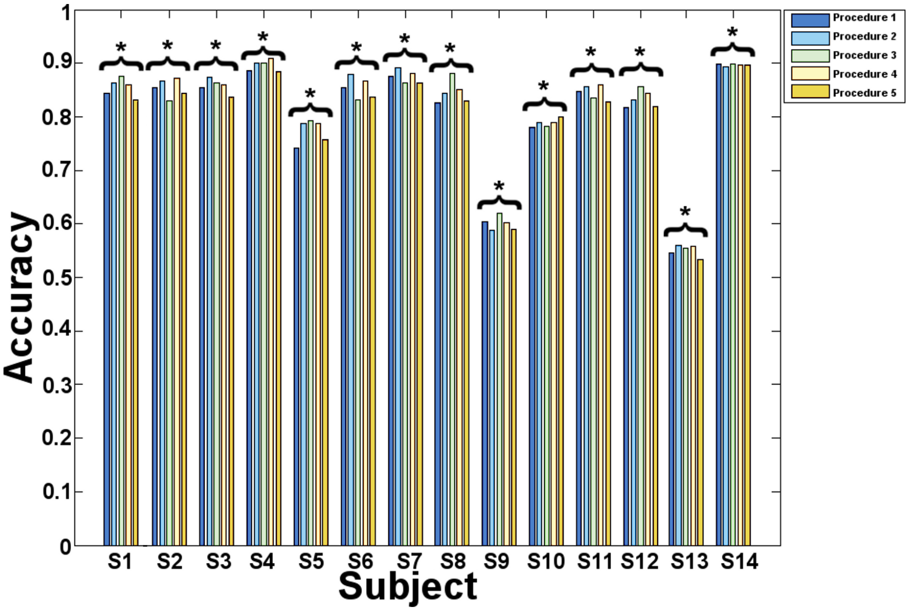

Figure 5.3: **Exploration: Mean balanced accuracies obtained in the different procedures for all subjects.** Procedure 1: specific GLM feature selection and SVM classification. Procedure 2: specific GLM feature selection and GP classification. Procedure 3: global GLM and RFA feature selections and GP classification. Procedure 4: specific GLM and RFA feature selections with GP classification. Procedure 5: specific GLM and RFA feature selections with SVM classification. All results are significant.

accuracies were comprised between 24.09 and 63.48%. The classification was not significant for participants S2, S10, S11 and S13 ($p > 0.05$).

Procedure 3. For exploration, mean balanced accuracies obtained using the RFA feature selection ranged from 55.56 to 90.12 % ($p < 0.05$). For mental imagery, mean balanced accuracies ranged from 27.82 to 61.65%. These results were not significant for participants S5, S10, S11 and S13 ($p > 0.05$).

Procedure 4. For exploration, the optimal subsets of features defined by GLM and RFA were associated with mean balanced accuracies ranging from 55.80 to 90.86 % ($p < 0.05$). For mental imagery, mean balanced ranged from 32.55 to 69.78 %. However, non-significant results were found for participants S10, S11 and S13 ($p > 0.05$).

Procedure 5. For exploration, the optimal subsets of features defined by GLM and RFA were associated with mean balanced accuracies ranging from 53.33 to 89.63 % ($p < 0.05$). For mental imagery, mean balanced accuracies ranged from 33.04 to 67.50 %, leading to non significant results for participants S5, S10, S11 and S13.

Overall mean balanced accuracies for the exploration session were significantly above chance for all the participants and all procedures. For the mental imagery sessions, mean balanced accuracies were not significant for some participants and some procedures: S2 (procedure 2), S5 (procedures 1, 3, 5), S10 (all procedures),

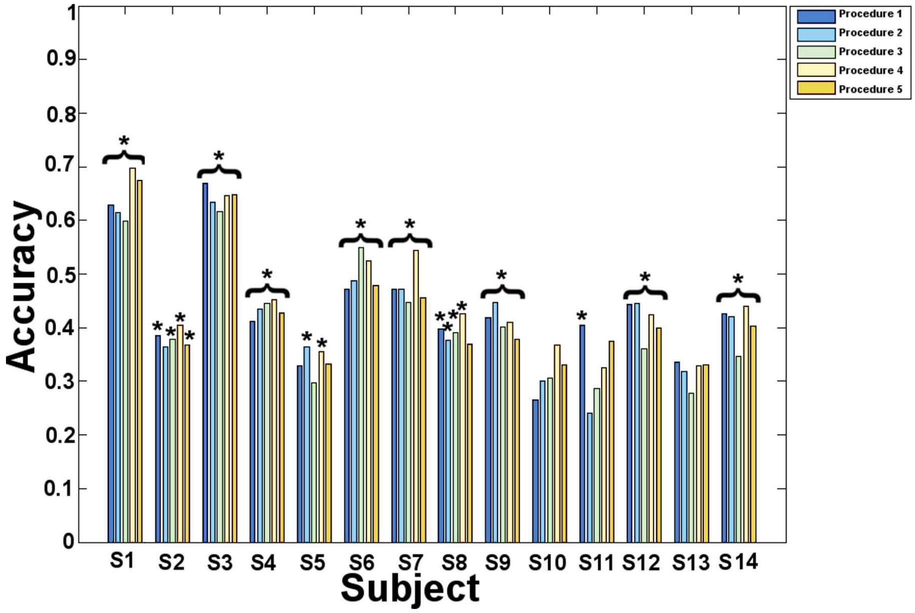

Figure 5.4: **Mental imagery: Mean balanced accuracies obtained in the different procedures for all subjects.** Procedure 1: specific GLM feature selection and SVM classification. Procedure 2: specific GLM feature selection and GP classification. Procedure 3: global GLM and RFA feature selections and GP classification. Procedure 4: specific GLM and RFA feature selections with GP classification. E Procedure 5: specific GLM and RFA feature selections with SVM classification. Significant classification accuracies are marked by stars $\star$.

S11 (all except procedure 1) and S13 (all procedures).

5.4.2 Comparison of procedures

The different procedures were first compared in terms of balanced accuracy. To obtain more insight on the results, they were also compared based on the class accuracies. Finally, since SVM seems to behave differently with unbalanced data sets, the number of Support Vectors, SV for each binary comparison were reported and correlated with the model accuracy.

5.4.2.1 Balanced accuracy

For exploration, the Friedman test on the over categories accuracy measures revealed significant differences ($p < 10^{-4}$) between procedures. Paired Wilcoxon signed rank tests showed that procedures 1 and 5 (SVM classification) performed significantly worse than procedures 2, 3 and 4 ($p < 0.05$, Bonferroni corrected for multiple comparisons, Figure 5.5.A.I).

Similarly, for mental imagery, a Friedman test on the balanced accuracies showed a significant effect of procedure ($p = 0.0094$). The paired Wilcoxon signed rank tests showed that procedure 4 performed significantly better than procedure 3 ($p = 6.1 * 10^{-4}$) and better than all other procedures ($p < 0.05$, but does not survive Bonferroni correction for multiple comparisons). Procedure 3 also tended

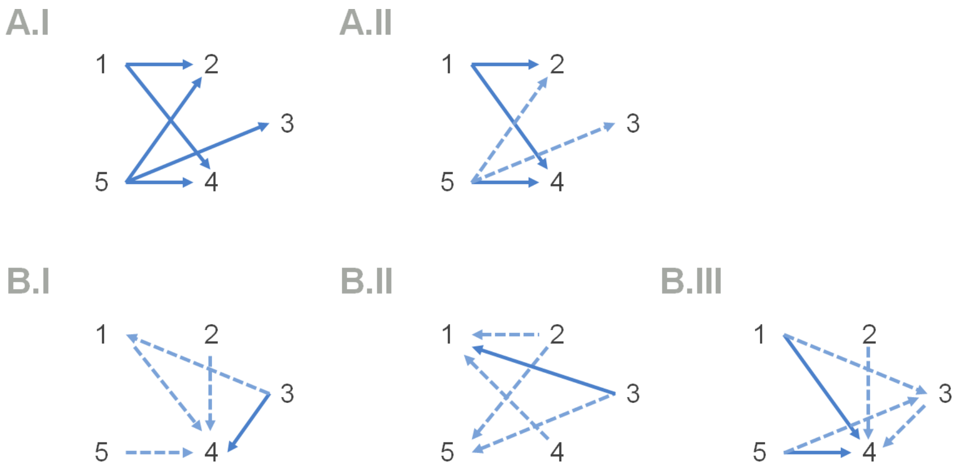

Figure 5.5: **Schematic comparisons between procedures.** The full arrows represent a significant difference ($p < 0.05$, survives Bonferroni correction for multiple comparisons) in performance between the two procedures linked, the arrow pointing to the best. The dashed arrows represent trends ($p < 0.05$, but does not survive Bonferroni correction). **A.I** Exploration: differences in balanced accuracy. Procedures 2, 3 and 4 (GP) performed best. **A.II** Exploration: differences in animals class accuracy. **B.I** Mental imagery: differences in balanced accuracy. Procedure 4 tended to perform best. **B.II** Mental imagery: differences in faces class accuracy. Procedures 1 and 5 (SVM) tended to perform best. **B.III** Mental imagery: differences in buildings class accuracy. Procedures 3 and 4 (tended to) perform best.

to perform worse than procedure 1 ($p = 0.0437$, not significant after Bonferroni correction, Figure 5.5.B.I).

5.4.2.2 Class accuracy

For exploration, there was an effect of procedure on the class accuracy measures only for the animal category (F: $p = 0.0672$, B:$p = 0.1594$ and A: $p = 0.0017$). Paired Wilcoxon signed rank tests showed that procedures 1 and 5 tended to perform worse than procedures 2, 3 and 4 for the animal category ($p < 0.05$, corrected for multiple comparisons using Bonferroni correction, Figure 5.5.A.II).

For mental imagery, Friedman tests showed a significant effect of procedure on the classification of faces and buildings ($p < 10^{-3}$). Paired Wilcoxon signed rank tests on the class accuracy for faces showed that procedure 1 performed significantly better than procedures 2, 3 and 4 ($p < 0.05$, Bonferroni correction, Figure 5.5.B.II). Trends also indicated that procedure 5 led to higher accuracies than procedures 2 and 3 (2-5: $p = 0.0068$, 3-5: $p = 0.0269$, do not survive Bonferroni correction). The paired Wilcoxon signed rank tests on the class accuracy for buildings showed that procedure 4 performed significantly better than procedures 1 and 5 ($p < 0.05$, Bonferroni corrected for multiple comparisons) and tended to perform better than procedures 2 and 3 (2-4: $p = 0.0353$, 3-4:$p = 0.0081$). Trends showing better performance of procedure 3 over procedures 1 and 5 were also noticed but not significant (1-3: $p = 0.0327$, 3-5: $p = 0.0327$, do not survive the Bonferroni correction, Figure 5.5.B.III). No other significant differences in class accuracies were noted.

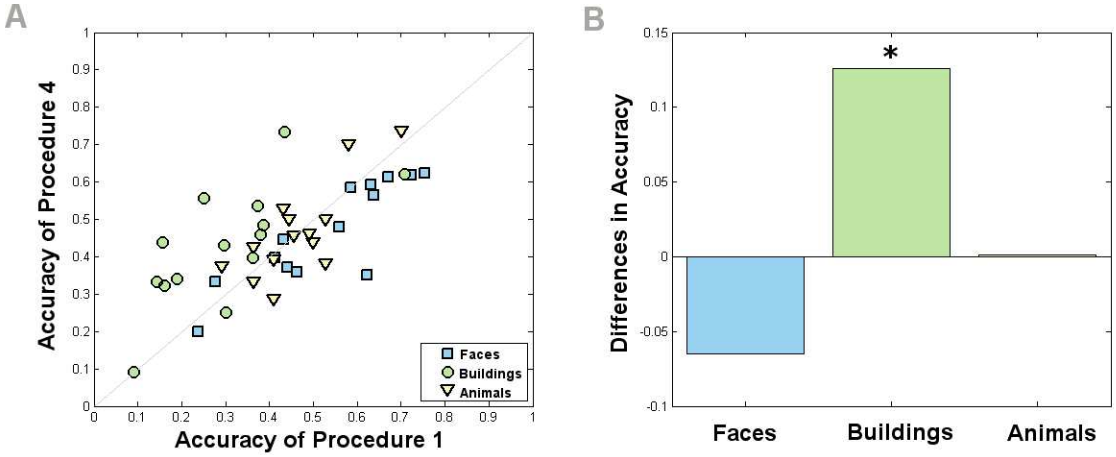

Figure 5.6: **Class accuracies obtained by the procedures 1 and 4 for all subjects.** The faces category is represented in blue, the buildings category, in green, and the animals category, in yellow. **A** Comparison of the accuracy values. x-axis: class accuracies obtained using procedure 1. y-axis: class accuracies obtained using procedure 4. Most points corresponding to class accuracies of buildings (represented by green circles) are above the 45° line (in light grey), meaning that the buildings were better classified using procedure 4. **B** Average difference in class accuracy between procedures 1 and 4. This figure shows that the difference in buildings classification is significant across subjects, while this is not the case for the faces and animals categories.

This result is illustrated in Figure 5.6, comparing the class accuracies obtained for each participant with procedures 1 and 4 (Figure 5.6.A). It was observed that procedure 4 performed always better than procedure 1 to classify buildings. Figure 5.6.B assessed the significance of this difference in performance between procedure 1 and procedure 4. Only the building classification was significantly different, i.e. worse for procedure 1 compared to procedure 4. Similar results were obtained when comparing procedures 4 and 5 in terms of buildings accuracy (not shown).

5.4.2.3 Support Vector proportions

Procedures 1 and 5 showing no significant difference in balanced or class accuracies, support vectors (SV) proportions were computed from each SVM binary classifier of procedure 5 (percentage of faces SV for the F-B and F-A comparisons and percentage of animals SV for the B-A comparison).

For exploration, a significant effect of the binary classifier on the proportions of SV was assessed ($p < 0.05$): post hoc Wilcoxon tests revealed that the proportion of SV in the faces category (F-B classifier) was significantly higher than in the faces category for the F-A classifier and in the animals category for the B-A classifier ($p < 0.05$, Bonferroni corrected). Whilst the proportions of faces SV in the F-B and F-A classifiers differed significantly from 50% (F-B>50%: $p = 0.0084$, F-A<50%: $p = 0.0045$), no significant correlation could be found between the SV proportions and the class accuracies ($0.1350 < p < 0.6479$). Moreover, the sign of some correlation coefficients were not consistent with the expected effect on class accuracy (for example, a positive correlation was found between the proportion of faces SV and the class accuracy of buildings, while one would expect the class accuracy of buildings to decrease when increasing the proportion of faces SV).

For mental imagery, a Friedman test also showed an effect of the binary classifier on the proportions of SV ($p = 1.25 * 10^{-5}$). Post hoc Wilcoxon signed rank tests revealed that the proportions of SV in the faces (for the F-B classifier) and in the animals (for the B-A classifier) categories were significantly higher than in the faces category for the F-A classifier ($p < 0.05$, Bonferroni corrected). SV proportions in the faces (F-B) and animals (B-A) categories were significantly higher than 50% (faces in F-B: $p = 0.0151$ and animals in B-A: $p = 0.0013$). Significant anti-correlations were found between the class accuracy of buildings and the proportion of faces SV in the F-B classifier ($p = 0.0166$), and the proportion of animals SV in the B-A classifier ($p = 0.0166$). Although no other significant correlation was assessed between class accuracy and SV proportions, the signs of all correlation coefficients were consistent with the expected effects.

5.4.3 Effect of behavioural data

In this section, we investigated whether participant's behaviour impacts the performance of the machine learning based model, using two parameters: the participants' behavioural performance as computed by d' and the number of extracted events.

5.4.3.1 Behavioural performances

No significant correlations were found between the accuracy of all classifiers and the performance of the participants at the test session led outside the scanner. The same result was obtained when taking into account the procedures individually.

It should be noted that when considering only the proportions of hits (i.e. by computing the percentage of correct answers), trends indicated an effect of the total number of correct answers on the balanced accuracy ($p = 0.0867$) as well as a correlation between the classification of buildings and the number of correct answers in the buildings category ($p = 0.0575$). In particular, procedures 2 and 3 (resp. 3, 4 and 5) showed significant correlation between the participants' performances and the balanced accuracy over the three categories (resp. for the buildings category). The other procedures still showed trends, but the correlations were not significant (over categories: $p(P1) = 0.1367$, $p(P4) = 0.0560$, $p(P5) = 0.092$, buildings: $p(P1) = 0.0509$, $p(P2) = 0.0686$).

5.4.3.2 Number of events

When considering the classification of mental imagery using all procedures (i.e. accuracies have been averaged across procedures), trends indicating an effect of the number of events could be detected for the faces and buildings category (correlations, faces: $\rho = 0.5189$, $p = 0.0573$, buildings: $\rho = 0.4804$,$p = 0.0821$). When investigating the procedures individually, significant correlations were found between the number of faces events and the classification of faces in procedure 5. Furthermore, procedures 1 and 5 (SVM classifier) showed a significant correlation between the number of buildings events and the classification of images of that category ($p < 0.05$).

5.5 Modelling spontaneous brain activity

In this section, we present the results from the modelling of spontaneous brain activity using (1) a new methodological approach based on machine learning models (section 5.5.1) and (2) a state-of-the art technique to analyse interactions between spatial networks (section 5.5.2, Margulies et al., 2010).

5.5.1 Machine learning based models

From the results of section 5.4, procedure 4 was selected as the best procedure to model mental imagery. In this case, only one LOO-CV was needed to select voxels via RFA and build the model on the whole session (no test set apart).

5.5.1.1 Proportions

The proportions Pr (see section 4.7.1) can be found for each participant and rest session in Table 5.4. This table shows that the proportions Pr are non-null for all rest sessions, suggesting either that spontaneous brain activity is associated with recurrent activation of these areas or that the considered methodology might detect false positives. Results are also highly variable across participants, which may be due to different levels of noise and/or the quality of the model of mental imagery.

To support the second hypothesis (formulated in section 3.3), the increases in Pr from pre-task to post-task rest in both conditions were compared. These results are reported in Table 5.4, under “Pr(m/o)”, as computed in equation 4.5.

Surprisingly, Table 5.4 shows that the statistical outliers in terms of behaviour display a large decrease in Pr from pre-task to post-task rest in the memory condition. Furthermore, their results prevent any significant difference between $Pr(m)$ and $Pr(o)$, as reported in Table 5.5.

The Friedman test on the differences in Pr between the memory and control conditions shows a clear trend in support of the second hypothesis formulated in section 3.3, when the statistical outliers are discarded. Furthermore, a one-tailed Wilcoxon signed rank test showed that the increase in Pr is significantly larger in the memory condition than in the control condition ($p = 0.0293$). This suggests that the

Table 5.4: : Proportions Pr of scans significantly linked to the task for each participant and rest session. Increases from the pre-task to the post-task rest sessions are reported under *Pr(m)* and *Pr(o)*, as computed in equation 4.5. Behavioural outliers are highlighted in colour.

	Memory			**Control**		
Subject	$R1_m$	$R2_m$	$Pr(m)$	$R1_o$	$R2_o$	$Pr(o)$
S1	38.11	38.44	0.33	40.07	38.76	-1.30
S2	19.54	17.27	-2.28	14.01	13.03	-0.98
S3	50.49	57.66	7.17	50.81	47.23	-3.58
S4	33.55	22.15	-11.40	33.55	33.22	-0.33
S5	62.54	64.50	1.95	60.59	61.24	0.65
S6	34.20	28.01	-6.19	27.69	28.34	0.65
S7	51.14	53.09	1.95	48.86	50.49	1.63
S8	24.76	14.01	-10.75	20.52	20.52	0.00
S9	50.98	51.79	0.81	44.44	42.81	-1.63
S10	2.32	5.96	3.64	4.30	3.64	-0.66
S11	25.83	26.82	0.99	27.81	28.81	0.99
S12	38.74	39.40	0.66	49.34	49.34	0.00
S13	43.05	39.07	-3.97	41.72	33.77	-7.95
S14	51.32	52.32	0.99	38.08	48.01	9.93

Table 5.5: : Increases in proportions Pr from pre-task to post-task rest sessions for both the memory and the control conditions (with standard deviations). The last column represents the p-value obtained from a Friedman test comparing $Pr(m)$ and $Pr(o)$. “No outliers” means that behavioural outliers have been discarded.

Selection	$Pr(m)$	$Pr(o)$	**p**
All subjects	-1.15 (5.28)	-0.18 (3.77)	0.4054
No outliers	1.11 (2.88)	-0.26 (4.29)	0.0578

proportions of scans significantly linked to the learning task is larger in the post-task than in the pre-task rest session, and this effect is significant in comparison to a control condition.

Since each scan significantly linked to the task also has a categorical prediction, we can derive the increases in Pr for each class (Table 5.6).

No significant differences could be found between the memory and control conditions (although p≃0.1 for buildings). One can however observe that the increase is the largest for the buildings category.

Table 5.6: : Increases in proportions Pr from pre-task to post-task rest sessions for both the memory and the control conditions for each category (with standard deviation). "No outliers" means that behavioural outliers have been discarded.

	$\boldsymbol{Pr(m)}$			$\boldsymbol{Pr(o)}$		
Selection	Faces	Buildings	Animals	Faces	Buildings	Animals
All subjects	-1.06 (3.38)	0.94 (2.40)	-1.03 (2.37)	0.04 (3.00)	0.27 (2.62)	-0.49 (2.23)
No outliers	-0.14 (2.83)	1.52 (2.09)	-0.27 (1.81)	-0.28 (3.26)	0.49 (2.82)	-0.48 (2.53)

5.5.1.2 Correlations with behavioural performance

The values of $Pr(m)$ and $Pr(o)$ were then correlated with the participants' behavioural performance d' to investigate the third hypothesis of section 3.3. Permutations of d' allowed assigning p-values to the correlation coefficient as well as to their difference (see Table 5.7 and Figure 5.7).

Table 5.7: : Correlation coefficients between Pr and d', the participants' behavioural performance, as well as their attributed p-value. $p(difC)$ represents the p-value assigned to the difference between the two correlations C_m and C_o. "No outliers" means that behavioural outliers have been discarded.

Selection	$\boldsymbol{C_m}$	$\boldsymbol{p(C_m)}$	$\boldsymbol{C_o}$	$\boldsymbol{p(C_o)}$	$\boldsymbol{p(difC)}$
All subjects	-0.06	0.5880	0.00	0.5060	0.3970
No outliers	0.50	0.0580	-0.01	0.5040	0.040

As shown in Table 5.7 and Figure 5.7, a clear trend suggests that there is a link between the increase in Pr from the pre-task to the post-task rest session in the memory condition and the participant's behavioural performance d'. Furthermore, this link is absent in the control condition and the difference in correlation coefficients between both conditions is significant. These results bring evidence regarding the validation of the third hypothesis of section 3.3.

No significant correlation could be found between the participants' behavioural performance in each category and the increase in Pr for the respective class. This might be due to the decrease in statistical power.

5.5.1.3 Phase information

During encoding, the learning material is temporally structured according to the cycle "F-B-A", i.e. images of faces, buildings and animals are always presented in this order (see section 4.2). To control for both the condition and the order of the cycle, the proportions of transitions according to the forward cycle ("F-B-A") and to the inverted cycle ("A-B-F") were computed for the memory and

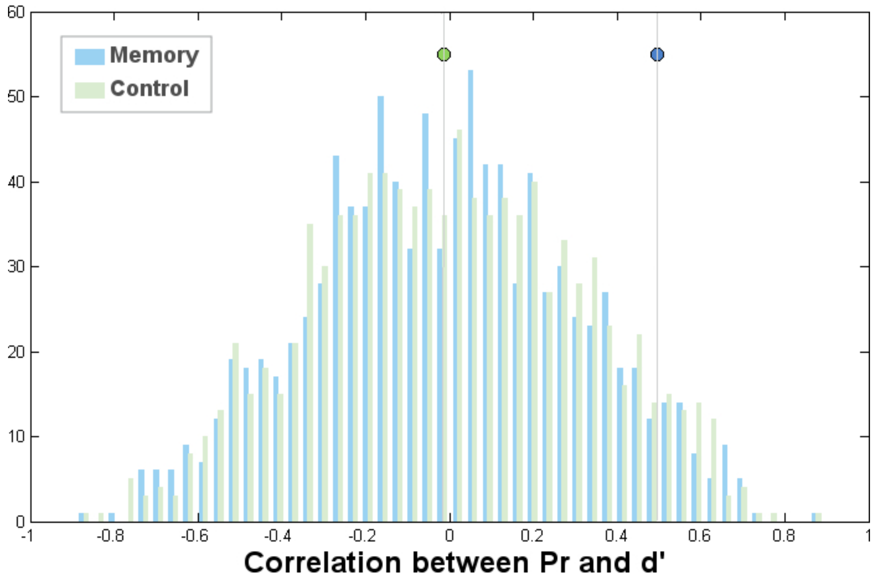

Figure 5.7: **Correlation between the differences in proportion Pr and the subjects' behavioural performances.** The correlation coefficients between the increases in Pr from the pre-task to the post-task rest sessions and the subjects' behavioural performance d' are represented by the blue and green circles on the top of the plot, for the memory and control conditions, respectively. The distribution of the correlation coefficient obtained from permutations of d' are displayed as a histogram, for the memory (light blue) and control (light green) conditions.

control conditions. Results are presented in terms of increases in the proportions of transitions from the pre-task to the post-task rest session 5.8.

When considering all participants, no significant difference could be found between the two conditions or the order of the cycle. However, discarding the behavioural outliers led to a significant difference between conditions when considering the forward cycle (F-B-A, $p = 0.0305$). Furthermore, this difference was absent when considering the inverted cycle (A-B-F, $p = 0.2179$). Although these results support the fourth hypothesis regarding the temporal structure of the replay, no statistically significant difference was found between the proportion of forward and reverse cycles, suggesting that the increase in Pr follows both the forward and the inverted cycle, with a slight preference for the forward cycle, as shown by the median (No outliers) in Table 5.8.

5.5.1.4 Temporal evolution

Separating the time-series in two equal parts led to the computation of Pr_1 and Pr_2 for each rest session and participant. As represented in Figure 5.8, no clear decrease in Pr was observed during the post-task rest session in the memory condition: it rather seems that the level of Pr is maintained across time. In the other rest sessions, Pr increased from the first to the last 5 minutes.

Table 5.8: : Increases in the proportions of transitions following the forward, F-B-A cycle ("forward") or the inverted, A-B-F cycle ("reverse") from pre-task to post-task rest sessions for both the memory and the control conditions. The median of the presented values are also shown for all participants (All subjects) and when discarding the behavioural outliers (No outliers).

	Memory		**Control**	
Subject	**Forward**	**Reverse**	**Forward**	**Reverse**
S1	0.98	-0.98	-2.93	-2.61
S2	-1.30	0.00	-0.65	-2.28
S3	-1.63	2.61	0.65	-0.33
S4	-4.89	-3.26	1.30	1.95
S5	0.33	1.95	-2.61	-1.30
S6	0.33	-0.33	-0.98	0.00
S7	-2.9316	-2.2801	-0.6515	-0.0000
S8	-0.33	-0.33	-0.65	-0.65
S9	-0.03	1.94	-0.65	0.00
S10	0.66	-0.33	0.00	0.33
S11	0.33	-2.32	-2.32	2.32
S12	0.99	0.99	-2.65	2.98
S13	1.66	-2.32	-3.64	-3.64
S14	0.99	3.64	0.00	-1.32
All subjects	**0.33**	**-0.33**	**-0.65**	**-0.16**
No outliers	**0.33**	**0.00**	**-0.65**	**-0.33**

5.5.1.5 Scaling factor

A possible scaling factor was investigated by computing the proportions and average duration of episodes lasting a fixed duration. This duration was varied from 1 to 10 scans. The median of these values across participants are represented in Figures 5.9 and 5.10, and in Table 5.9 for thresholds $th = 1$ and $th = 2$.

When considering all participants, no significant difference could be found, for any of the computed measures. Discarding the behavioural outliers does not lead to significant differences for the proportion of long episodes or for the duration of all episodes ($th = 1$). However, it leads to a significant effect of the session on the duration of longer episodes ($th = 2$, Friedman test, $p = 0.0493$), and more specifically with the episodes in $R2_m$ lasting longer than in $R1_o$ and $R2_o$ ($R2_m$-$R1_o$: $p = 0.0420$, $R2_m$-$R2_o$: $p = 0.0137$). Although this result is not represented by the median across participants, it is better highlighted using the mean across participants ($R1_m$:2.79, $R2_m$:3.14, $R1_o$:2.82, $R2_o$:2.79).

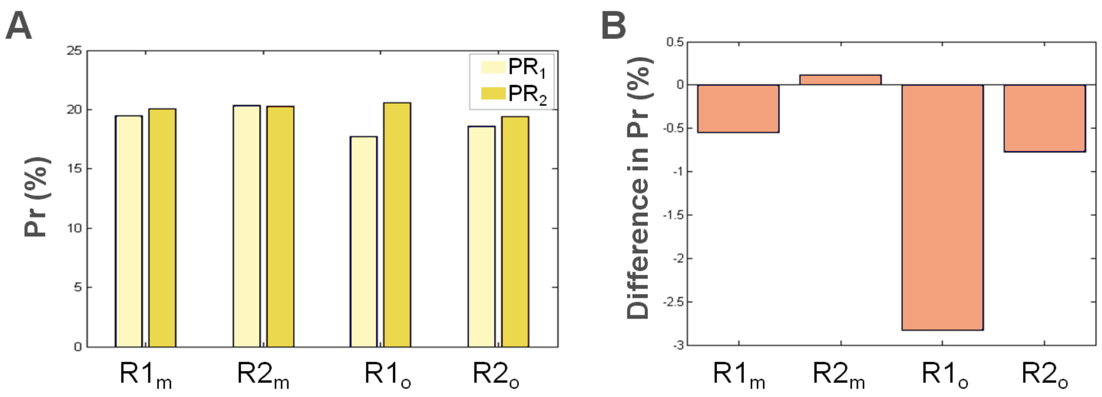

Figure 5.8: **Evolution of Pr along time. A** Average across subjects of Pr_1 (0 to 5 minutes) and Pr_2 (5 to 10 minutes) for each rest session. **B** Decrease of Pr along time for each session (i.e. $Pr_1 - Pr_2$). Except for $R2_m$, all show an increase in Pr along time.

Table 5.9: : Proportion (in %) of episodes of scans significantly linked to the task having a duration larger than one scan, average duration (in scans) of these episodes and average duration of all episodes (i.e. those having a duration of one scan included). These values are summarized by the median across participants for each rest session. "No outliers" means that behavioural outliers have been discarded.

Median (All subjects)	$\boldsymbol{R1_m}$	$\boldsymbol{R2_m}$	$\boldsymbol{R1_o}$	$\boldsymbol{R2_o}$
Proportion $(th = 2)$	40.42	37.20	41.56	40.34
Duration $(th = 2)$	2.93	2.89	2.71	2.74
Duration $(th = 1)$	1.82	1.68	1.68	1.68

Median (No outliers)	$\boldsymbol{R1_m}$	$\boldsymbol{R2_m}$	$\boldsymbol{R1_o}$	$\boldsymbol{R2_o}$
Proportion $(th = 2)$	41.10	46.97	43.10	41.79
Duration $(th = 2)$	3.24	2.91	2.89	2.74
Duration $(th = 1)$	1.89	1.79	1.73	1.89

These results show that the average duration of an episode is around 2 scans (when taking all episodes into account, $th = 1$). However, this duration increases when discarding the episodes lasting only one scan, i.e. those that can be suspected to be mostly false positives. This increase is larger in the post-task rest session of the memory condition than in the other rest sessions, as shown by Figure 5.10 for thresholds of 2 to 4 scans. This result suggests an average duration of episode comprising 2 to 4 scans, i.e. between 4 and 8 seconds. It should be noted that this duration was measured from the BOLD signal and does not represent the neuronal activity, which supposedly comprises more fast and transient events.

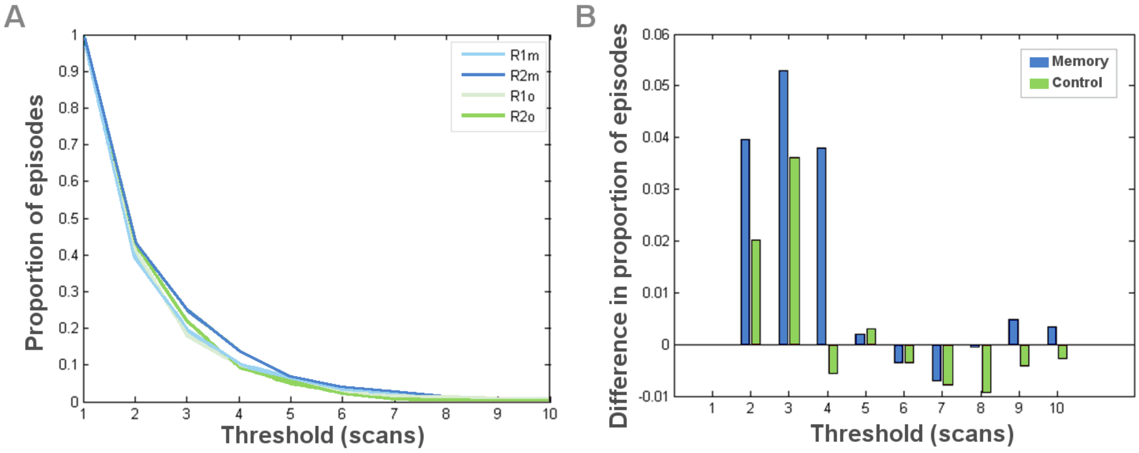

Figure 5.9: **Proportions of episodes.** When increasing the threshold on the duration of the episodes from 1 to 10 scans (x-axis), the average proportions of episodes lasting at least as long as the threshold decrease (**A**). However, it seems that the proportions of episodes lasting from 2 to 4 scans are larger in the post-task rest session in the memory condition, than in any other rest session. This is further confirmed by the differences between pre-task and post-task rest sessions (**B**), displayed for the memory (blue) and control (green) conditions (without the behavioural outliers).

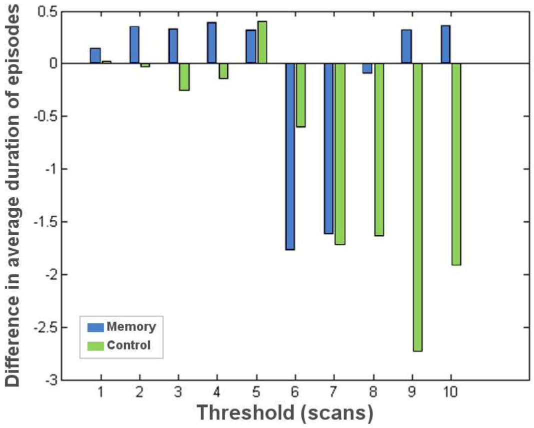

Figure 5.10: **Duration of episodes.** X-axis: threshold on the minimal duration of one episode. Y-axis: Average across subjects (without the behavioural outliers) of the difference between the duration of episodes in the pre- and post-task rest sessions, for the memory (blue) and control (green) conditions. As for the proportions of episodes, the duration of the episodes in the post-task rest session showed an increase compared to the pre-task rest session for episodes lasting 2 to 4 scans.

5.5.1.6 Selection of scans significantly linked to the task

During the definition of the methodology, certain choices were made that might have affected the results presented above. To ascertain that some of these choices did not randomly lead to the observed results, we investigated the effect of the threshold deciding whether a scan is significantly linked to the task or not.

First, Pr was computed for each threshold and rest session, discarding the behavioural outliers. This allowed plotting the increases in Pr from the pre-task to the post-task rest sessions in both the memory and control conditions (Figure 5.11) as a function of this threshold. Although the p-value threshold clearly affects the increases in Pr, $Pr(m)$ is always larger than $Pr(o)$. Furthermore, the p-values assigned to the correlations between the increase in Pr and the participant's behavioural performance seem to stabilize around the reported values for $0.05 < p < 0.1$ (Figure 5.11).

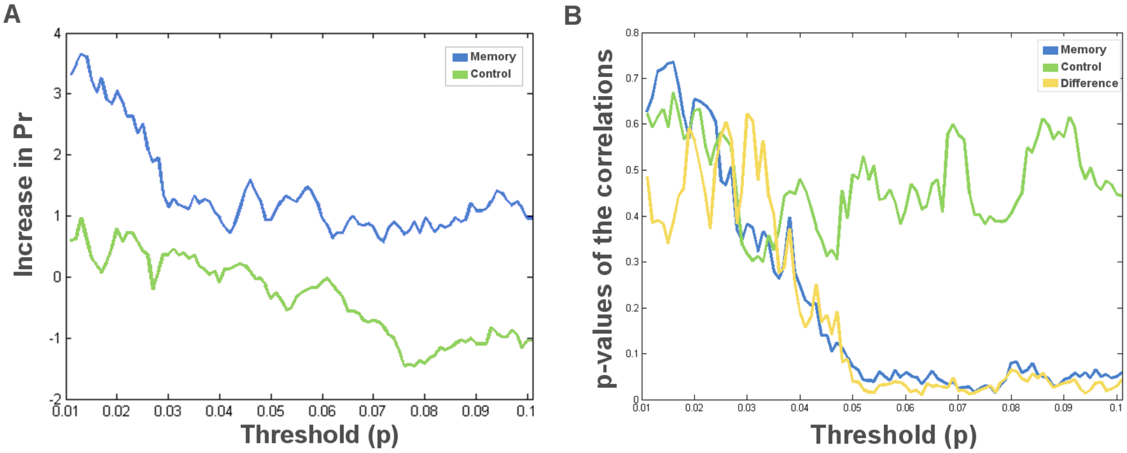

Figure 5.11: **Varying the p-value assessing scans as significantly linked to the task.** X-axis: p-value varied from 0.01 to 0.1. **A** Increases in Pr (in %) from the pre-task to the post-task rest session in the memory condition (blue) and in the control condition (green). **B** P-value assigned to C_m (memory condition, in blue) and C_o, (control condition, in green).

Therefore, although the choice of the threshold affects the previously presented results, the conclusions drawn from these results still hold when varying the threshold from 0.05 to 0.1. Moreover, while local minima and maxima can be spotted on both graphs, our choice of $p < 0.05$ does not correspond to any extreme local variation of the computed measures and thus seems appropriate.

5.5.2 Analysis of network interaction

5.5.2.1 Manual selection of ROIs

The manually selected regions correspond to the main peaks of the statistical parametric map built from the GLM analysis of the exploration session. These regions were defined as spheres centred around the coordinates provided in Table 5.10, with two contingency layers (i.e. 25 voxels in total, Figure 5.12).

The selected ROIs define two subsets of interest: a "hippocampal" NOI and a "maze" NOI (i.e. FFA+PPA+AA).

Table 5.10: Coordinates of the center of the manually selected regions in the MNI space as well as their attributed names. The first three regions form the hippocampal NOI, while the six others correspond to regions activated during the display of the images (“maze” NOI). “sym” refers to the symmetry in the definition of the ROIs.

Region	Name	x	y	z
Left Hippocampal 1	L-Hipp1	-31.68	-6.82	-19.84
Left Hippocampal 2	L-Hipp2	-17.82	-14.32	-14.42
Right Hippocampal	R-Hipp	29.70	-12.89	-24.59
Left FFA	L-FFA	-37.62	-52.27	-15.89
Right FFA (sym)	R-FFA	39.60	-49.36	-16.03
Left PPA	L-PPA	-21.78	-43.21	-9.61
Right PPA (sym)	R-PPA	19.80	-43.21	-9.61
Left Animals	L-AA	-43.56	-49.11	-11.00
Right Animals	R-AA	49.50	-65.79	5.13

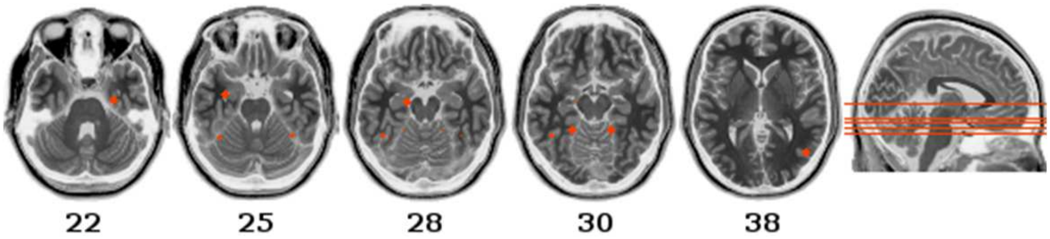

Figure 5.12: **Manually selected ROIs.** Regions of interest manually selected from the statistical peaks computed by a GLM on the exploration session. 3 regions were selected in the hippocampus and 6 in the neocortex.

5.5.2.2 Automatic ROI selection

When manually identifying the components according to [Smith et al., 2009], six main networks of interest were detected (Figure 5.13):

1. Visual network (VIS): a first medial and lateral component was detected and added to the occipital component.
2. Ventral attentional network (vATT): this network was found in its lateralized components (left and right), which were then merged.
3. Dorsal attentional network (dATT)
4. Auditory network (AUD)
5. Sensori-motor network (MOT)
6. Default mode network (DM)

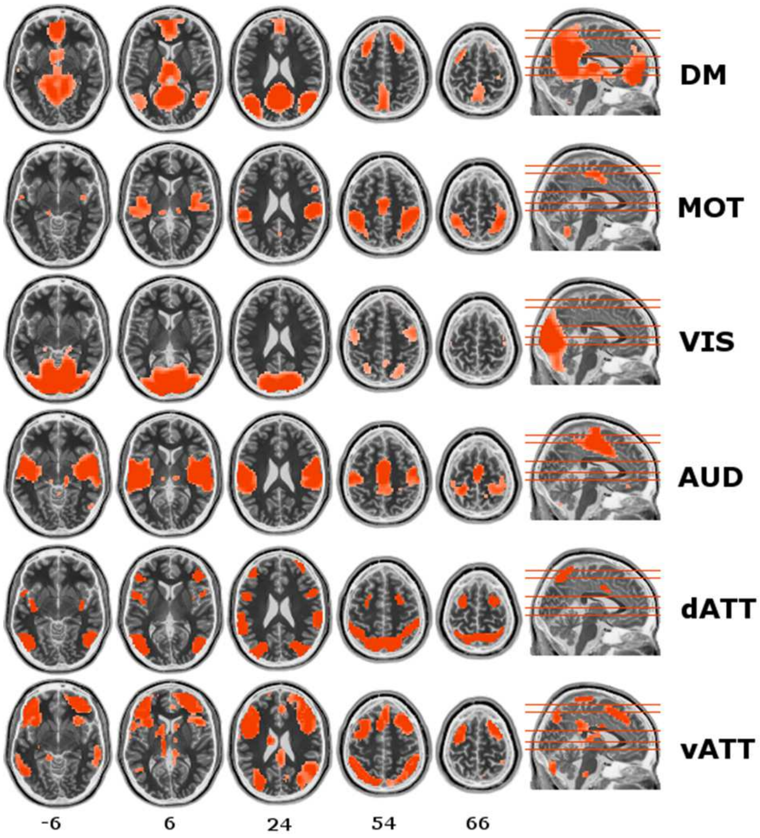

Figure 5.13: **Networks of interest manually selected.** The t-maps display the 6 detected networks computed with $p < 0.05$ with correction for multiple comparisons, at different locations on the z-axis.

The ROIs were then automatically selected from the thresholded t-maps. They counted a maximum of 20 voxels in extent and were at least 4 cm apart. In total, 48 ROIs were considered for further analysis (57 with the manually selected ROIs).

5.5.2.3 Integration

Total integration. Both conditions (i.e. $R1_m - R2_m$ and $R1_o - R2_o$) show a significant increase in total integration. This increase is larger for the control condition (increase=1.25) than for the memory condition (increase=0.44), at $p < 0.05$ ($p = 0.035$).

Within-NOI integration. For the memory condition, significant increases in within-network integration were observed for the VIS and AUD NOIs. For the control condition, significant increases were found for the DM, dATT and MOT NOIs. Please note that the only significant differences between conditions were found for the DM and MOT NOIs (marked by grey $\star$ in Figure 5.14).

A first interesting result is that no increase in within-NOI integration could be found for the Hipp and Maze NOIs.

Between-NOI integration. All pairwise between-NOI integrations are reported in Figure 5.15 (significant differences between pre-task to post-task rest sessions

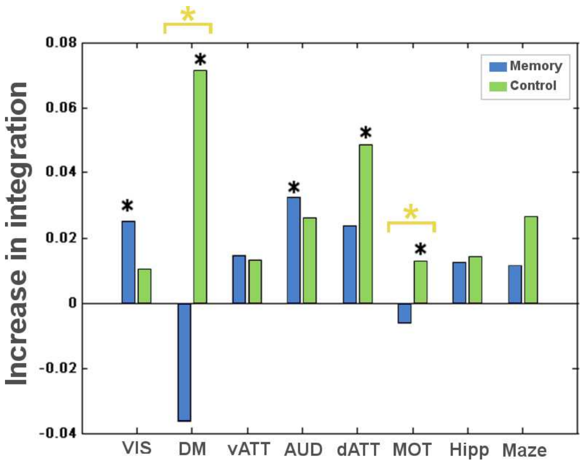

Figure 5.14: **Within-NOI integration.** Increases in within-NOI integration from the pre-task to the post-task rest session for the memory (blue) and control (green) conditions. Significant increases are marked by black stars ⋆, while yellow stars represent a significant difference between both conditions.

being marked by ⋆).

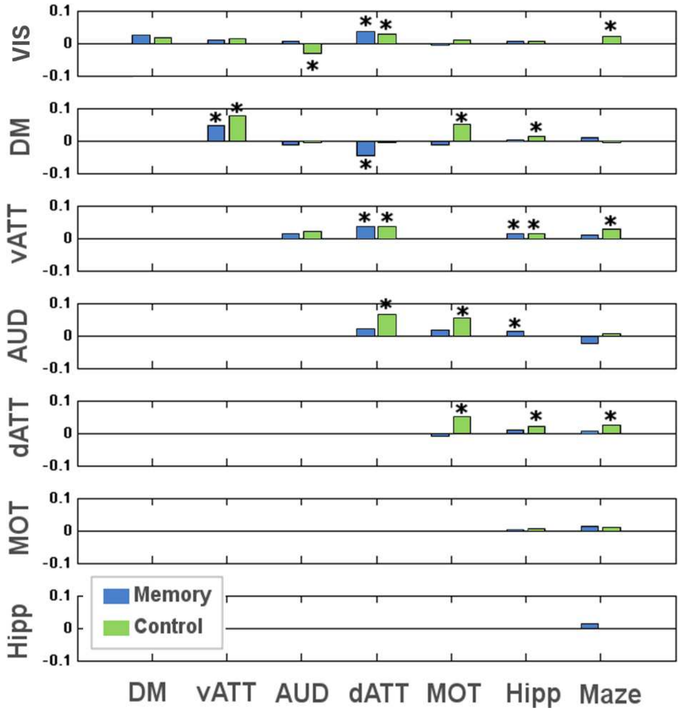

Figure 5.15: **Between-NOI integration.** Increases in between-NOI integration from the pre-task to the post-task rest session for the memory (blue) and control (green) conditions. Significant increases are marked by black stars ⋆. It should be noted that the results are symmetric.

The results show that the vATT and dATT NOIs were the most affected by the tasks. Although this could have been expected [Smith et al., 2009], increases in integration between the Hipp and Maze NOIs were also expected, or between the Hipp and VIS NOIs in the memory condition, which are not present. Furthermore, some results are surprising, such as the significant increases in integration between the DM and vATT NOIs, and between the DM and Hipp NOIs for the memory condition. According to [Smith et al., 2009], the DM network shouldn't be affected by any of the considered tasks.

More than the differences between the pre-task and post-task rest sessions, we were interested in the differences between conditions, i.e. are the increases in between-NOIs integration larger for the memory or the control conditions? The p-values of such comparisons are represented in Table 5.11, a p-value > 0.9 meaning that the increase is larger for the memory condition while $p < 0.1$ reveals a higher increase for the control condition.

Table 5.11: P-value assigned to the differences between conditions for each of the pairwise between-NOI integration. Significant results are highlighted in bold. It should be noted that results are symmetric.

NOI	**DM**	**vATT**	**AUD**	**dATT**	**MOT**	**Hipp**	**Maze**
VIS	0.65	0.43	**0.97**	0.63	0.14	0.46	0.13
DM		0.27	0.41	**0.07**	**0.08**	0.15	0.80
vATT			0.36	0.51	0.49	0.51	0.20
AUD				0.12	0.10	**0.91**	0.19
dATT					**0.02**	0.22	0.14
MOT						0.31	0.61
Hipp							0.40

5.5.2.4 Partial correlation

The partial correlation measure was used to assess any change in the "connectivity" between the Hipp NOI and all other NOIs (i.e. modification in partial correlation between the 3 hippocampal ROIs and any other ROI). As shown in Figure 5.16, the effect of the task on the hippocampal connectivity is small. Furthermore, this effect is the same for both the memory and the control conditions, leading to equal values of the IDPC (section 4.7.2): IDPC(m)=0.0084 and IDPC(o)=0.0090.

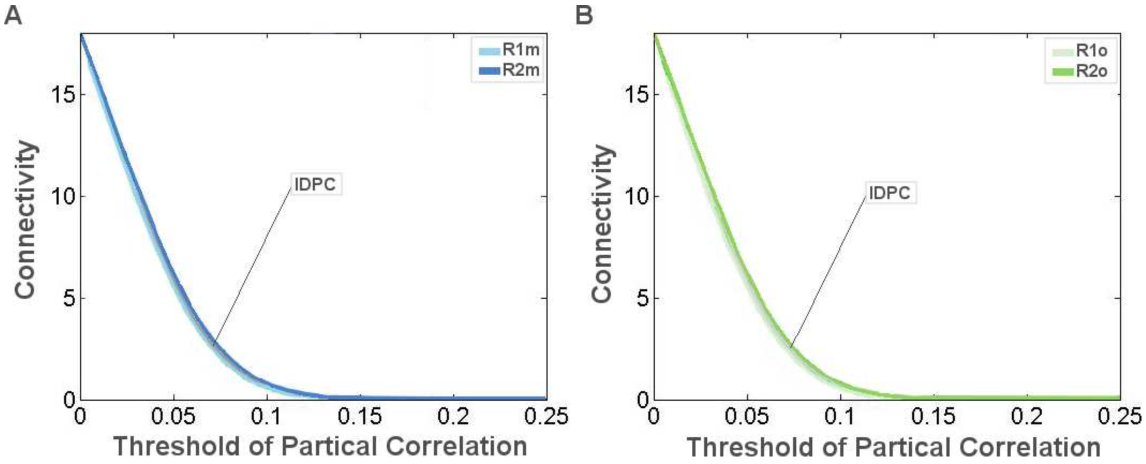

Figure 5.16: **Connectivity of the hippocampal ROIs.** Connectivity of the 3 hippocampal ROIs for the memory (**A**) and control (**B**) conditions. The pre-task connectivity is plotted in light blue or green while the post-task connectivity is displayed in dark blue or green, for the memory and control conditions respectively. A small difference can be observed for both conditions, leading to small and almost equal values of the IDPC (grey area between the two curves).

Chapter 6

Discussion

Contents

6.1 (Semi-) constrained brain activity

We tested the performance of different classification procedures on two separate fMRI time series. Whereas the experimental design of the exploration session imposed a paced and regular succession of stimulus categories, the mental imagery session was characterized by imbalanced numbers of trials between categories, a self-paced succession of individual trials and subject-related task performance. The uneven number of events across categories was related to the disparity in individual memory performance, pictures of buildings being significantly less well remembered than the two other classes of stimuli. In addition, the succession of events of variable durations, sometimes beyond the temporal resolution of fMRI, put a further strain on classification procedures. The best combinations of techniques (namely procedures 1 and 4) were able to classify accurately, i.e. significantly above chance level, the mental images from 11 out of the 14 subjects.

When classifying the controlled session, results indicated that SVM performed significantly worse than GP. No effect of the feature selection (either specific GLM, global GLM and RFA or specific GLM and RFA) could be detected, which does not correspond to what was reported in the literature [Mitchell et al., 2004; Mourão-Miranda et al., 2006; Formisano et al., 2008; De Martino et al., 2008]. This result indicates that for this well-controlled experiment, using a GLM filter or a RFA embedded wrapper leads to the same performance.

However, when considering the mental imagery session, performance of the considered procedures indicates that GP might be more sensitive to the addition of irrelevant features than SVM. This hypothesis is supported by the fact that a univariate feature extraction by a specific GLM substantially improved classification accuracy. Indeed, procedure 4 (specific GLM-RFA-GP) achieved better accuracies than procedure 3 (global GLM-RFA-GP), which needed more computational time for significantly poorer results. While the authors of [Mourão-Miranda et al., 2006] suggested that such a univariate feature selection step "may improve" the accuracy of intrasubject classification, we showed that this improvement is significant for the considered GLM contrasts, GP classifier and data sets. In addition, combining a specific GLM with a second multivariate step further improved feature selection, as indicated by the higher accuracy achieved by procedure 4 relative to 2 (specific GLM-GP). This result is in agreement with [Mitchell et al., 2004; Formisano et al., 2008] and [De Martino et al., 2008], which stated that the combination of a univariate selection of "active" voxels combined to a multivariate selection of "discriminant" voxels led to the best performance of classifiers. However, procedures 1 and 5 (specific GLM-RFA-SVM) showed similar performance, suggesting that the RFA step did not bring further relevant information to the SVM classifier.

Once the optimal subset of features was defined, the performance of GP and SVM classifiers showed only slight differences (trend that procedure 4 performs better but not significantly). However, GP seemed more robust than SVM for classifying imbalanced data sets, as the former achieved a significantly better accuracy for the least represented class (i.e., buildings in the current study). This result might be explained by the sparseness of SVM since significantly different proportions of support vectors between the binary classifiers were revealed. Furthermore, the proportions of support vectors correlated with the obtained class accuracies.

Regarding the effect of behavioural measures on the results of the classification, it seems that the number of events in each category has an impact on the accuracy measures, especially for procedures 1 and 5 (SVM classifier). These significant correlations are likely to be directly due the poor ability of SVM to deal with imbalanced datasets. On the other hand, no relationship could be drawn between the subjects' performance d' and the performance of the procedures. However, trends indicated an effect of the percentage of correct answers on the obtained accuracy, especially for the least represented class. This might be explained by the fact that d' is a logarithmic measure, while the percentage of correct answers or of correct predictions is linear. Therefore, although the small number of observations precludes any definitive conclusion, these findings suggest that the ability to reinstate category-specific activity patterns within specific occipito-temporal areas supports memory retrieval.

Conclusion: The results show that for fMRI time series which include complex, unbalanced self-generated mental states, best accuracies are obtained by a feature selection combining a specific GLM and a recursive feature addition. Whilst the advantage of GP over SVM to classifying this type of data is small (in terms of balanced accuracy), the former seems more appropriate for markedly unbalanced data sets, and thus preferable for more realistic experimental setups.

6.2 Spontaneous brain activity

The procedure leading to the best performance was then applied to the different rest sessions which resulted in the computation of the proportion of scans of spontaneous brain activity significantly linked to the memory task (Pr). This proportion Pr was non-null for all subjects and all rest sessions, implying that the method detected false positives. Furthermore, Pr showed a high variability across subjects, suggesting that the differences from pre-task to post-task rest sessions should be investigated, instead of considering the absolute values of Pr.

The results showed an increase in Pr from the pre-task to the post-task rest session, in the memory condition. When discarding the statistical outliers in terms of behaviour (i.e. S4, S6 and S8), the increase in Pr tended to be larger for the memory condition than for the control condition. This result supports the first hypothesis of the theory of memory consolidation, which assumes that patterns of brain activity generated during encoding are unconsciously rehearsed during post-task rest [Hoffman and McNaughton, 2002; Tambini et al., 2010].

Furthermore, for the same selection of subjects, the behavioural performance of the subjects, d', correlated with the increase in proportion Pr. This correlation could not be achieved when considering permutations of the behavioural measure and was significantly higher than the correlation value obtained from a control task (C_o). This result suggests that the larger the increase in proportion Pr from pre-task to post-task rest, the better the memorization of task features by the subject. This is in agreement with [Peigneux et al., 2006] and [Tambini et al., 2010] who linked the subject's performance to the amount of hippocampal activity [Peigneux et al., 2006] or correlation with the neocortex [Tambini et al., 2010].

Another hypothesis regarded the temporal structure of the replays [Louie and Wilson, 2001; Lee and Wilson, 2002; Foster and Wilson, 2006]. To investigate whether the increase in Pr followed the phase information contained in the design (i.e. the succession of images of faces, buildings and then animals), we computed the proportions of transitions according to the forward cycle, as well as to the inverted cycle (used as a control cycle). The results showed that the increase in the proportion of transitions according to the forward cycle from pre-task to post-task rest session was significantly higher for the memory condition than for the control condition. Furthermore, this difference was not present for the inverted cycle. However, no significant difference could be found when comparing the increases in the transitions according to the forward or inverted cycle from pre-task to post-task rest session in the memory condition. The results therefore suggest that the increase in Pr from pre-task to post-task rest session is the largest effect, affecting both the transitions according to the forward and inverted cycles, with a preference for the forward cycle (as shown by the median across subjects). It is the first time that evidence regarding the temporal structure of the replays at the region level is found.

According to [Wilson and McNaughton, 1994], the strength of the correlations decreased along time, with a time constant around 12 minutes. [Tambini et al., 2010] tried to reproduce this result but could not show an effect of time on the seed correlations. They explain this by a too short time range (10 minutes), and therefore too close to the time constant described by [Wilson and McNaughton, 1994]. Our results show that $R2_m$ is the only session during which Pr decreases, but this decrease is not significant compared to the large increases observed during the other rest sessions, and particularly $R1_o$. We therefore conclude that the considered time-scale is also too small to show that this effect is significant, as proposed in [Tambini et al., 2010]. The comprehensive characterization of activity-induced neural reactivations will require the assessment of these processes over a longer time period including during sleep.

When investigating a possible scaling factor (SF), the average duration of a detected episode was computed as around 2 scans, i.e. 4 seconds. This average duration increased to around 3 scans when considering episodes lasting more than one scan. Furthermore, the duration of these episodes was significantly longer in the post-task rest session for the memory condition than in the pre-task and post-task rest session of the control condition. This suggests that the mnemonic traces (i.e. the scans significantly linked to the task) detected in $R2_m$ might comprise less false positives than in the other rest sessions. Regarding the scaling factor, the average duration of 6 seconds leads to either a reactivation of only a few patterns at a time (i.e. in average only 2 images), or to a temporal compression of these reactivations. However, the absence of independent measure of neural activity prevents any definitive conclusion, since it is not possible to infer the true number of rehearsed patterns during an episode.

Finally we studied the effect of the arbitrary choice made when defining the threshold to which a scan would be assessed as significantly linked to the task ($p<0.05$). In regard of the results (see section 5.5.1.6), we can reasonably conclude that although our choice affected the results, these were not obtained merely by chance.

Although all the results taken together provide evidence supporting the theory of memory consolidation, it is useful to stress the inter-subject variability in the sensitivity of our decoding scheme. Behaviour was variable in the first place : two subjects were discarded due to their anxiety or depression score whilst another used a strategy based on the alphabet to remember the memory task. As shown in Table 5.5, these indeed showed large decreases in Pr from pre-task to post-task rest sessions in the memory condition. This was associated with a lower sensitivity in detecting mnemonic traces, such as the correlation with the subjects' behavioural performance or the temporal structure of the replays. This however does not mean that subjects considered as statistical outliers did not learn the task (as shown by their performance) but rather that they might have used different learning strategy, decreasing our ability to model the corresponding rest sessions. Different factors might lead to these poor predictions, such as the anxiety or depression scores (S4 and S8) or the strategy used to memorize the images (S6). Another important parameter is the fact that these subjects particularly rehearsed the bi-dimensional maze during the post-task rest session in the memory condition. We can only assume that this conscious repetition led to a higher level of noise in the data (compared to our effect of interest), maybe due to the selection of irrelevant features. However, the effects of the subject's behaviour on memory consolidation remain poorly understood and would need deeper investigation.

When comparing the obtained results to state-of-the art methods to analyse ROI interactions, it appears that some hypotheses cannot be tested using such an approach due to the reduction of the time-series into a single measure (i.e. the ROI entropy used to compute integration and partial correlation values). Studying the temporal evolution of the ROI/NOI interactions is therefore impossible using the considered technique, as well as investigating the temporal structure of the replays or a prospective scaling factor. Regarding increases in integration within- or between-NOIs, only an effect of the task on the attentional streams was found, for both conditions. Although this result is in agreement with [Smith et al., 2009], other modifications in ROI interactions are quite surprising, such as the significant increases in within- and between-DM integration. Finally, no significant change in hippocampal connectivity could be observed from pre-task to post-task rest sessions, in neither condition. The absence of results supporting the theory of memory consolidation leads to the conclusion that analysing ROI interactions might not be suited to detect and characterize mnemonic traces.

Conclusions: The classification of rest sessions could be performed by applying previously built models on a mental imagery session. While the results should be more deeply investigated, some evidence was found supporting the theory of memory consolidation. The proposed methodology also allowed to directly investigate the reactivations during post-experience rest. Although there is room for improvement, machine learning modelling therefore seems a promising technique to study memory consolidation and tackle the complex issue of decoding spontaneous brain activity.

6.3 Future work

Although this work presents promising results, improvements would be welcome in terms of methodology and acquisitions. First, the proposed models do not allow the study of causality, and thereby cannot verify the theory of a transfer of information from the hippocampus to the neocortex [Ji and Wilson, 2007]. Causal machine learning models are appearing [Peters et al., 2011] and could be used in this application in a near future. Regarding the causality in ROI interactions, one could consider "Dynamic Causal Modelling" [Friston et al., 2003]. This technique computes effective connectivity between ROIs, thereby inferring causality in the interactions[1].

Second, the successive scans of the rest sessions were considered as independent from one another. However, they contain auto-correlation structures, as explained in [Friston et al., 2007]. These might contain information that was not extracted here: if a scan is defined as significantly linked to the task, would it be more likely that the following scan is also linked to task or less likely? This question cannot be answered in part because the scaling factor is unknown. Turning to EEG (Electro-

[1] Dynamic Causal Modelling will be applied to the considered dataset in a further work.

EncephaloGraphy) would therefore bring further insight on the temporal evolution of mnemonic traces. More specifically, epochs of signal could be more reasonably treated as independent due to the high temporal resolution of this acquisition technique. Furthermore, the outcome of an increased temporal resolution would lead to an easier computation of the scaling factor for example, by identifying the N170, a marker of visual stimulation. While modelling spontaneous activity using EEG brings considerable assets, these are balanced by a lower spatial resolution and a decreased signal-to-noise ratio, which makes decoding EEG signals still a challenge.

Part II

Clinical Application

Chapter 7

Introduction

Contents

7.1 Clinical challenges

In the application presented in part I, we focused on investigating the brain functioning in healthy subjects. However, studying the brain dysfunctions and trying to isolate the affected structures or the possible causes of a degenerative disease is another very interesting challenge. Degenerative syndromes have indeed become a large burden in today's society: dementia affects 1 in 20 people over the age of 65 and 1 in 5 over the age of 80 [Ferri et al., 2006]. Worldwide, there are an estimated 35.6 million people with dementia and this number is not likely to decrease due to the ageing of the populations, especially in developing countries (Figure 7.1).

Beyond the number of people affected, the most common dementias are often misdiagnosed using classical clinical exams. As an example, a definitive diagnosis of Alzheimer's disease (AD) can only be obtained using post-mortem histopathological analysis. Currently, AD is diagnosed using clinical exams, neuropsychological testing and manual measurements on brain images (MRI or PET), leading to time-consuming criteria and accuracies of the diagnosis around 80% at best [Knopman et al., 2001]. AD is therefore often misdiagnosed, although an early treatment would be more effective. This example illustrates the need for automated and objective diagnostic procedures.

During the past decades, advances in neuroimaging techniques enabled the identification of biomarkers in dementias, such as in Alzheimer's disease [Zakzanis et al., 2003] or in different states of consciousness [Monti et al., 2010]. Statistical parametric mapping (for functional images) and voxel-based morphometry, ([Ashburner and Friston, 2000] for structural images) helped to infer group differences, e.g. between

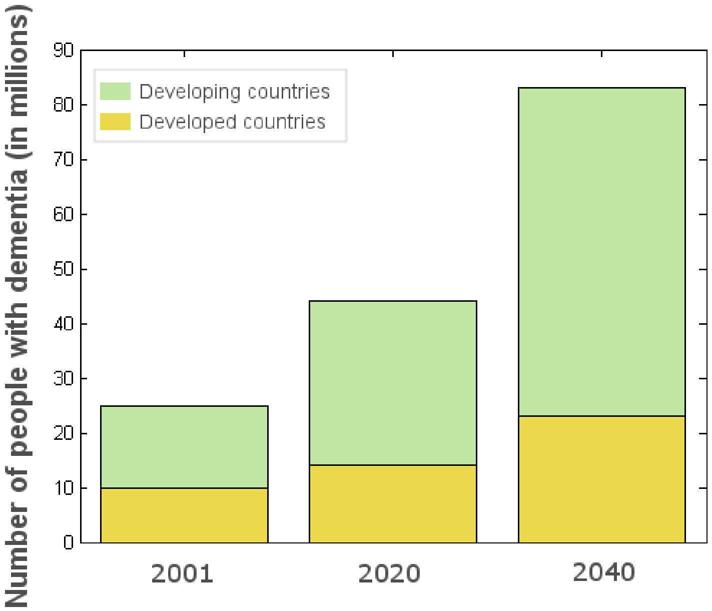

Figure 7.1: **Number of people with dementia.** Identified and projected number of people with dementia in developed and developing countries. [Ferri et al., 2006] estimates 4.6 million new cases of dementia each year, with most diseased people living in developing countries.

healthy subjects and patients. However, conclusions at the subject's level would also be desirable for diagnosis, prognosis, treatment planning or the monitoring of disease progression [Orrù et al., 2012].

7.2 Machine learning models

In order to build automated and objective diagnostic aids, multivariate analysis could become particularly useful. Multivariate decoding of neuroimaging data can be used to achieve two different objectives: firstly and obviously predict the perceptual, cognitive or medical state of one or many subjects, referred to as *pattern discrimination* [Pereira et al., 2009]. Therefore, once the machine learning based model has been trained, it can then be used as a "black box" that predicts the category of any new data fed in. This can be viewed as a diagnostic tool in the case of a disease-versus-healthy classification (or any variation). Beyond the final diagnosis/prognosis, clinicians are also interested in where the information about the variable of interest is coded in the brain. In this case, machine learning based modelling can be used to reveal the pattern of voxels leading to the discrimination of different states, referred to as *pattern localization*. With linear kernel machines (such as the ones used in this work), these two goals can be reached simultaneously: the estimated weight associated to each voxel reveals the patterns of voxels considered as important by the model to perform the classification. However, as already mentioned in chapter 2.5.5, the pattern has to be considered as a whole, which leads to a difficult interpretation.

In the following subsections, the latest methods and results are reviewed for pattern discrimination and localization. To illustrate the considered issues, we focused on the case of Parkinson's Disease (PD), which presents challenges for both applications.

7.2.1 Diagnostic tool

The use of machine learning models as diagnostic tools has already demonstrated its promises. For example, [Vemuri et al., 2008; Klöppel et al., 2008] performed the distinction between healthy subjects and AD patients with an accuracy higher than 86%. These methods also allowed to assess the level of consciousness of vegetative states patients [Phillips et al., 2011], which can lead to a dramatic increase in their quality of life.

When considering Parkinson's disease, two types of problems have been envisaged: discriminating Idiopathic Parkinson's Disease (IPD) from healthy controls, and distinguishing between IPD and Parkinson Plus Syndromes (PPS), which represent atypical forms of Parkinson's disease. The latter issue can either be treated as a binary problem (IPD vs PPS), as in [Duchesne et al., 2009] or divided into its sub-classes, when considering the atypical syndromes separately. [Focke et al., 2011] therefore successfully discriminated between IPD and Progressive Supranuclear Palsy (PSP) and between IPD and Multiple Systems Atrophy (MSA), using the grey matter extracted from structural MRI. In a more recent work, [Garraux et al., submitted] derived a multiclass classifier to directly discriminate between the ^{18}FDG PET images of IPD, PSP, MSA and Cortico-Basal Syndrome (CBS) patients. In contrast with previous works, they were able to associate a "confidence" measure to each prediction, which is particularly useful in a clinical context. Therefore, although there is room for improvement, especially in the multiclass case, machine learning methods already proved to be useful when discriminating IPD from PSP [Duchesne et al., 2009; Focke et al., 2011; Garraux et al., submitted].

Regarding the discrimination between IPD patients and healthy controls, various parameters such as movement parameters [Aubin et al., 2012], voice measurements [Geetha Ramani and Sivagami, 2011] or eye movements [Tseng et al., 2012] provided significant results. In contrast, when considering neuroimaging data, it seems that only PET images allowed the significant classification of IPD patients and healthy controls: [Jokinen et al., 2009] and [Acton and Newberg, 2006] achieved accuracies higher than 90%, while structural MRI [Focke et al., 2011], gave no significant results [Orrù et al., 2012]. However, both works extracted specific features in the images (voxels in the striatum for Acton and Newberg, 2006 and dopamine uptake striatal to occipital cortices ratio in Jokinen et al., 2009) and [Acton and Newberg, 2006] used non-linear classifiers, thereby precluding the building and localization of the pattern. Furthermore, in view of the advantages of (f)MRI compared to PET (see Table A.1), finding MRI biomarkers of idiopathic Pakinson's disease would be desirable.

In order to identify MRI biomarkers of Parkinson's disease when compared to healthy subjects, we referred to previous works investigating gait disturbances due to IPD in fMRI [Snijders et al., 2011; Maillet et al., 2012; Cremers et al., 2012b].

More specifically, the univariate results of [Cremers et al., 2012b] showed different activation patterns in controls and patients during the mental imagery of gait in fMRI (see Table 7.1). In the present work, mental imagery of gait in fMRI was therefore investigated as a possible informative feature set to discriminate between IPD patients and healthy controls.

Table 7.1: Mental gait activation patterns in controls (Ctrl) and patients. SMA stands for supplementary motor area, DLPFC for dorsolateral prefrontal cortex, PPN for pedunculopontine nucleus and MLR for mesencephaliclocomotor region. Laterality is indicated by L (left), R (right) or bi (bilateral). For coordinates, please see [Cremers et al., 2012b]

.

Area	**Ctrl**	**Patients**	**Ctrl>IPD**
Lateral premotor cortex	bi	-	-
Pre-SMA	bi	R	-
Anterior cingulate cortex	bi	R	-
Middle frontal gyrus (DLPFC)	R	-	-
Inferior frontal gyrus	bi	-	-
Anterior insula	bi	-	-
Intraparietalsulcus	bi	-	bi
Precuneus	bi	-	R
Parieto-occipital sulcus	bi	-	bi
Posterior hippocampus	bi	-	L
Parahippocampalgyrus	bi	-	-
Lingual gyrus	bi	-	R
Caudate nucleus (head)	R	-	-
Anterior putamen	bi	-	-
Anterior pallidum	L	-	-
PPN/MLR area	L	-	L
Lateral pons	L	-	-
Cerebellar vermis	Midline	-	Midline
Cerebellar hemisphere	bi	-	bi

7.2.2 Pattern localization

There are different approaches to investigate pattern localization. Most studies report the “peaks”, i.e. the voxels corresponding to the highest weights for each category. However, without the ability to threshold the weight maps, such interpretation can be complex and not easily illustrated. Another approach is to compute one machine learning model for each voxel, or for each voxel and its neighbourhood (referred to as the searchlight scheme, [Kriegeskorte et al., 2006]), thereby constructing a map containing the accuracies for each voxel. The obtained accuracies can then be tested for significance, as in second-level univariate analyses, and hence thresholded [Pereira et al., 2009]. This technique, whilst giving thresholded maps of accuracies, is only locally multivariate and therefore does not take full advantage of the multivariate nature of the data. Furthermore, it does not provide a diagnostic tool.

In this study, techniques were developed to ease the interpretation of the weights associated with a diagnostic tool. To illustrate the proposed methodology, the results were compared with the univariate results from [Snijders et al., 2011; Maillet et al., 2012] and more particularly to [Cremers et al., 2012b] (Table 7.1).

7.3 Aim of this work

In the present work, both the issue of fMRI pattern classification and localization were tackled by considering the discrimination between aged healthy subjects and idiopathic Parkinson's disease patients. Based on previous works [Cremers et al., 2012a,b], we investigated whether the mental imagery of gait could predict the presence or absence of IPD. Furthermore, a methodology to help interpreting the model weights was developed and allowed comparing models in terms of pattern localization.

Aim: Apply machine learning based models in a clinical context to classify and localize the fMRI patterns of idiopathic Parkinson's disease.

Chapter 8

Material and Methods

Contents

In this chapter, we present the material and methods used to investigate the patterns of IPD. After the description of the data and design, the use of pattern classification techniques as a diagnostic tool is exposed in section 8.2. Finally, the methods developed to localize the patterns of IPD are presented in section 8.2.2.1.

8.1 Data and design

The material considered in this work being the same as in [Cremers et al., 2012b], only a brief description of the population and experimental design will be provided. For more details, please refer to [Cremers et al., 2012a,b].

8.1.1 Population

In total, 29 subjects participated in the study: 14 patients (7 males; mean age: 65.1 ± 9.5 years) diagnosed with IPD with different degrees of severity of gait disturbances and 15 controls matched for age (63.8 ± 8.1 years) and gender (7 males). The volunteers did not have any history of intracranial lesion, neuroleptic agents exposure or excessive alcohol consumption. Written informed consents for this research protocol approved by the local ethics committee were obtained from all participants.

8.1.2 Experimental design

Before fMRI, the subjects were asked to walk comfortably and then briskly on a 25m path. After gait evaluation, they were trained to mentally rehearse themselves walking on the path without making any voluntary movement.

All subjects then underwent a block-design fMRI session comprising three tasks: mental imagery of standing on the path (STAND), walking at a comfortable pace along the path (COMF) and walking briskly along the path (BRISK). The COMF and BRISK conditions were self-paced, subjects indicating when they had completed each trial by a key press, while each trial of the STAND condition was constrained by the duration of the previous COMF trial. Eight trials of each condition (12 for BRISK to account for a shorter duration of the trials) were randomly presented to each subject (Figure 8.1). Mental imagery was performed in a visuokinesthetic first-person perspective.

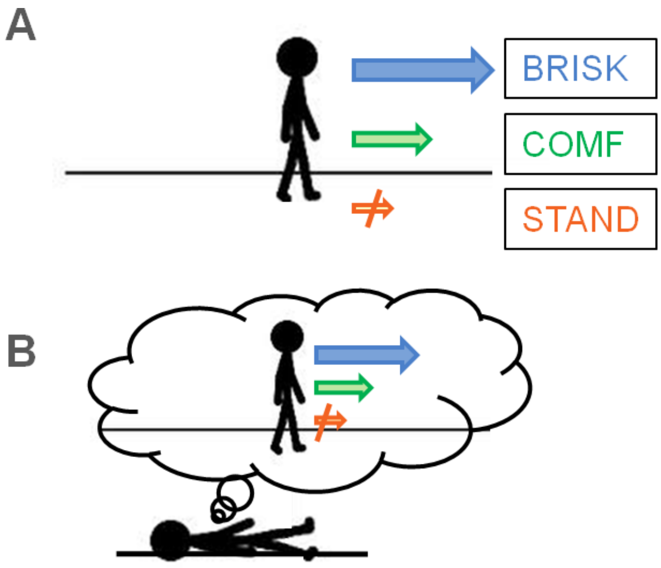

Figure 8.1: **Design of the mental imagery of gait experiment. A** Before scanning, subjects are trained to walk at a comfortable and then brisk pace on a 25m path. **B** In the scanner, subjects mentally rehearse standing (8 blocks) or walking comfortably (8 self-paced blocks) or briskly (12 self-paced blocks) on the 25m path.

8.1.3 Data acquisition

BOLD fMRI data were obtained on a 3T Magnetom Allegra MR Head scanner (Siemens AG Medical Solutions, Erlangen Germany) using a single-shot 2D gradient-echo echo-planar imaging (GRE-EPI) sequence (32 axial slices, slice thickness = 3 mm, slice gap = 30%, TR = 2,130 ms, TE = 40 ms, flip angle = 90°; bandwidth = 3,552 Hz; matrix size = 64 × 64, yielding an in-plane resolution of 3.44 mm × 3.44 mm). The first three images of the BOLD time series were discarded to allow for T1 saturation effects. Head movement was minimized by restraining the subject's head using a vacuum cushion.

In addition to BOLD fMRI, all participants underwent a high-resolution volumetric anatomical MRI of thebrain using a T1-weighted MDEFT sequence (TR = 7.92 ms TE = 2.4 ms, TI = 910 ms, flip angle = 15°; matrix size 240 × 256, yielding 176 contiguous sagittal slices with a isotropic voxel size of 1 mm^3).

8.1.4 Preprocessing

fMRI data preprocessing and univariate analysis were performed using SPM8[1]. Functional images were realigned and co-registered to the structural image before normalisation using DARTEL [Ashburner, 2007]. Finally, smoothing was applied using a 8mm FWHM Gaussian kernel.

A general linear model then summarized the time series from each subject by modelling each condition by a boxcar function convoluted with a canonical haemodynamic response function. In the end, three images per subject were considered for further analysis: the parametric maps of STAND, COMF and BRISK representing the BOLD signal activity associated with each condition.

8.2 Multivariate analysis

The multivariate analysis was performed using PRoNTo[2]. This Matlab-based software (MathWorks, Natick, MA) provides a flexible framework to perform pattern recognition based on machine learning models [Schrouff et al., 2013].

8.2.1 Pattern discrimination

Pattern discrimination was performed using binary SVM, as in [Focke et al., 2011] and [Orrù et al., 2012], with a linear kernel and the soft-margin hyperparameter C set to 1 [Mourão-Miranda et al., 2006]. In part I, feature selection led to an improvement in model performance when dealing with within-subject classification. This effect of feature selection on between-subject classification has however been questioned (e.g. by [Mourão-Miranda et al., 2006]). In the present work, both whole brain and space selection analyses were therefore conducted.

In its current version, PRoNTo does not provide wrapper or embedded feature selection. Three masks (Figure 8.2) were therefore used as filters before building the linear kernel, based on [Cremers et al., 2012b]:

- A "whole brain" mask, selecting all voxels within the brain.
- A "motor mask", built with a digital neuro-anatomical atlas (using the WFU-PickAtlas, Maldjian et al., 2003) and comprising the areas involved in gait

[1] www.fil.ion.ucl.ac.uk/spm

[2] www.mlnl.cs.ucl.ac.uk/pronto

(both in healthy subjects and patients), as described in Table 1 of [Maillet et al., 2012].

- A mask comprising the Mesencephalic Locomotor Region (MLR) and pedunculopontine nucleus, further referred to as "MLR mask". These areas were previously and consistently reported in univariate analyses comparing healthy subjects and IPD patients [Cremers et al., 2012b; Karachi et al., 2010; Snijders et al., 2011; Alam et al., 2011]. The mask consists in a box of $1.7cm^3$ and was built using anatomical markers (coordinates in MNI space: x=[8.5,-8.5], y=[-26,-36], z=[-12,-22] mm).

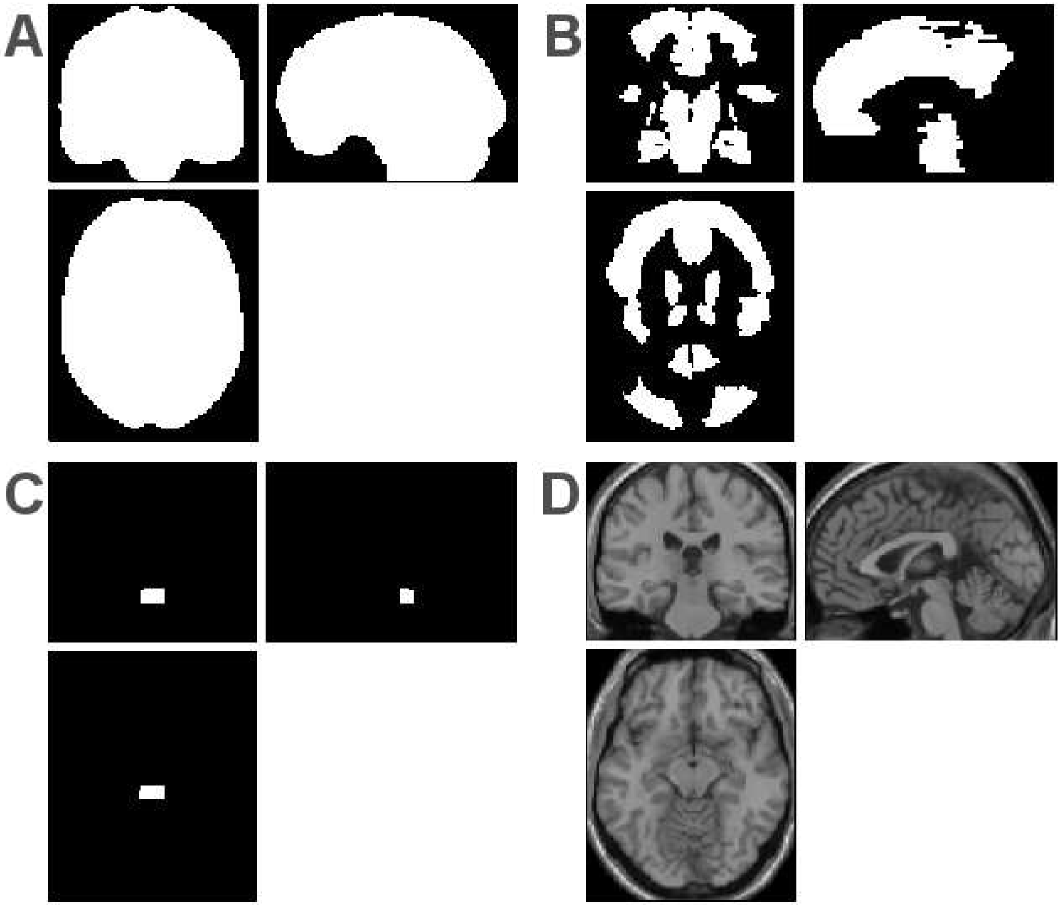

Figure 8.2: **Masks considered for the discrimination between healthy controls and IPD patients**. **A** Whole brain (517845 voxels), **B**, motor (217814 voxels) and **C** MLR masks (225 voxels, note that the cross-hair position has been centred on the MLR for this panel). **D** displays the SPM single subject canonical structural image for better representation.

To distinguish between IPD and healthy controls (i.e. between groups classification), the three tasks will be combined in every possible way (e.g. BRISK, BRISK+COMF, . . . , see Table 8.1) and tested for each mask, leading to 7 combinations times 3 masks, resulting in 21 models. Using more than one condition simply means that the corresponding images were added as samples (e.g. the STAND

and COMF images of control 1 were labelled as “control” and tested independently when left out as test set).

Table 8.1: Combinations of the three conditions (STAND, COMF and BRISK) used to discriminate between IPD and CTRL. The combination of all conditions is further referred to as “All”(last column).

	Combination						
Condition	1	2	3	4	5	6	All
STAND	x			x	x		x
COMF		x		x		x	x
BRISK			x		x	x	x

To benefit from the different models built from each mask, their predictions were taken together in a majority vote (MV). This voting operation is based on the idea that combining weak models (i.e. models leading to significant but low accuracies) could lead to a strong model (i.e. a model leading to significant and high accuracies). MV has been increasingly used in the field of machine learning and pattern recognition, as well as for clinical applications [Garraux et al., submitted]. In the present case, a majority vote was computed from the STAND, BRISK and COMF models (first three columns in Table 8.1) for each test data (which are the same across models, i.e. one subject per fold). Since three binary models are involved in the voting operation, no ties are possible.

In addition to the IPD vs. CTRL comparison, the discrimination between the three tasks was also assessed by pooling together the data of both groups. This classification required the use of a multiclass model, implemented in PRoNTo in the form of multiclass Gaussian Processes.

For both classification problems, leave-one-subject-out cross-validation was performed to compute the balanced and class accuracies, as well as positive predictive values. The significance of accuracy measures was assessed by non-parametric testing using 1000 random permutations of the training labels (100 permutations for Gaussian Processes due to its higher computational expenses). Particular care was taken during the permutation testing of the MV model: the labels have to be the same for the three models considered for voting, such that the permutation testing has to be performed simultaneously for the different models.

8.2.2 Pattern localization

To localize the pattern discriminating between IPD and healthy controls, the first requirement is to find an intuitive way to display the weights (section 8.2.2.1). Beyond the representation of the weights, it is asked how similar two patterns are, i.e. compare models in terms of pattern localization. This leads to the question of quantifying the similarity between two patterns, which is addressed in section 8.2.2.2.

8.2.2.1 Representing the weights

In univariate analyses, significant voxels are grouped into clusters that can then be compared to known anatomically or functionally labelled regions. Neuroscientists are therefore used to refer to these regions, which define atlases, such as the Automated Anatomical Labelling atlas (AAL, Tzourio-Mazoyer et al., 2002) or the Talairach atlas based on Brodmann areas [Talairach and Tournoux, 1988].

In order to facilitate the interpretation of the weights associated to each voxel, the weight maps were "smoothed" according to the regions labelled by these atlases. One measure of weight per labelled region was defined, further referred to as the Normalized Weights, NW_{ROI}. This measure is computed from the absolute weight in one region W_{ROI}:

$$W_{ROI} = \sum_{v \in ROI} |W_v| \tag{8.1}$$

with v representing the index of a voxel in the weight image and W_v its weight.

Since W$_{ROI}$ consists in a sum of absolute values, it (partly) reflects the size of the regions. To account for the region size, the normalized weight of one region NW_{ROI} was defined as its weight W_{ROI} divided by the volume of the region (number of voxels). From this measure, it is then possible to rank the regions according to the percentage of the normalized weight that they explain. This is then similar to a Principal Component Analysis, in which the components are ranked on the proportion of the signal variance they explain.

To illustrate this approach, the univariate results of [Cremers et al., 2012b] were reproduced. They investigated the patterns of activity generated by the COMF>STAND contrast in each group separately and then in their comparison. SVM models were therefore built on the discrimination between the COMF and STAND conditions in the control (CTRL) and in the patients (IPD) groups separately. The whole brain pattern discriminating best between the two groups (built in section 8.2.1) was then localized and compared to the univariate results (to a certain extent, see further).

The labelled regions used to localize the patterns were defined by the AAL atlas from the WFU-PickAtlas [Maldjian et al., 2003] toolbox in SPM. To the classic, lateralized 117 AAL regions, the pons, the midbrain and the medulla regions were added, since they were reported in the univariate results of [Cremers et al., 2012b; Maillet et al., 2012] and comprise the mesencephalic locomotor region and the pedunculopontine nucleus. The 120 regions from this manually generated atlas are illustrated in Figure 8.3.

The top ranked regions in terms of normalized weights, NW, were then compared to the univariate results of [Cremers et al., 2012b]. It should be noted that univariate results are directed (COMF>STAND), while multivariate results are not (COMF$\neq$STAND), which further precludes from any direct comparisons between univariate and multivariate results. For display purposes, an image of the normalized weights of the models leading to the best discrimination between IPD and

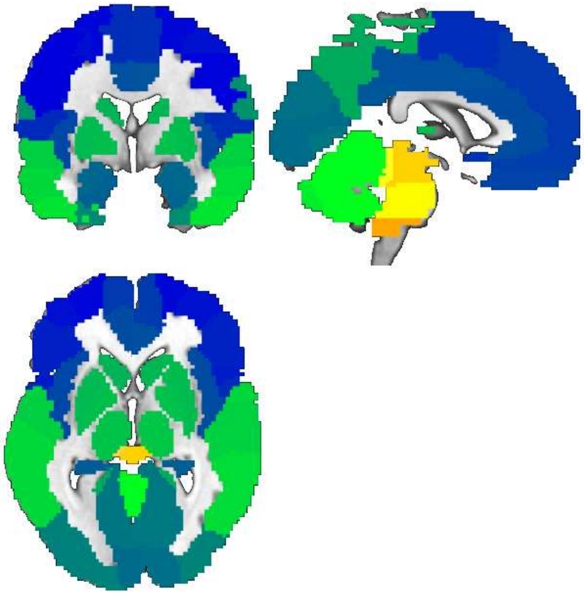

Figure 8.3: **AAL atlas with pons, medulla and midbrain regions.** Views of the 117 labelled regions defined in the AAL atlas (in green and blue), with the addition of medulla, pons and midbrain areas (in yellow). In total, the brain has been parcelled into 120 labelled regions of interest.

healthy controls was built.

This methodology, although simple, gives a ranking of the regions contributing to the pattern classification, without the need to threshold the weights. However, two aspects have to be accounted for when computing those measures:

1. The overlap between the mask used to perform the classification and the atlas defining the regions.

2. The variability of the ranking across folds.

The first issue regards the overlap between the mask and the atlas: when performing whole-brain pattern classification, the mask can comprise white matter or cerebrospinal fluid in addition to the grey matter. On the other hand, the atlases usually comprise grey matter only, such as those generated from AAL labels or Brodmann areas using the WFU-PickAtlas [Maldjian et al., 2003]. As a result, some voxels v are not associated to any region and their weight is not taken into account. To overcome this problem, a new region was created, called *others*, that pools all those voxels into a single region for which the (normalized) weight can be computed (Figure 8.4). The overlap between the mask and the atlas is hence reflected by their common volume and by the weight of the *others* region. These values were therefore reported for each pattern.

The second issue relates to the variability of the weights across folds: if the data contains outliers, then the results will be driven by only a few folds and the average of the folds will not correctly represent the pattern classification model . To identify

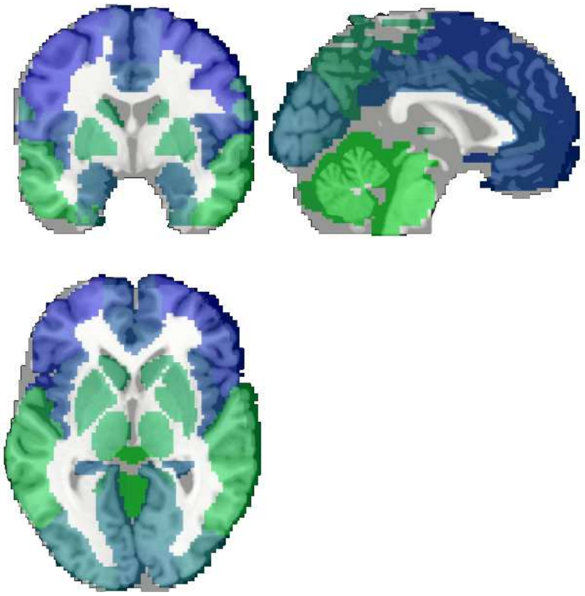

Figure 8.4: **Account for mask-atlas overlap.** The mask is represented as a white, transparent overlay on a structural brain image while the atlas is defined by the blue/green labelled regions. Any voxel v displayed in gray-scale, i.e. not covered by any blue/green region was pooled into the *others* region.

such situations, the rankings of the regions are presented in terms of expected values across folds:

$$E(Rank_{ROI}) = 1 \times f(1) + 2 \times f(2) + 3 \times f(3) + \ldots + NROI \times f(NROI) \quad (8.2)$$

With $f(x)$ being the frequency that the region ROI was ranked x^{th}, and $NROI$, the total number of labelled regions (i.e. 120 in the present case).

8.2.2.2 Compare patterns

Computing the expected value of the ranks across folds provides a qualitative idea on the stability of the ranking across folds. However, defining a quantitative measure to compute the differences in pattern localization across folds would bring more insight on the variability of the weights. From a more general point of view, being able to compare two models in terms of pattern localization is a parameter that is highly desirable: models are usually compared in terms of performance (accuracy, sensitivity, specificity, ...) or in terms of goodness of fit (e.g. marginal likelihood when using Gaussian Processes). However, neuroscientists are particularly interested in the localization of the obtained patterns and therefore, a quantitative assessment of the differences in the ranking of the labelled regions would provide them with an additional way to compare various models.

In order to assess the difference between patterns in terms of their localization, a measure of distance between rankings was defined, dr. This measure compares the (expected values of) the ranking of the labelled regions and is inspired from those used in the field of web search [Lempel and Moran, 2005]. It is computed as:

$$dr(f_1, f_2) = \frac{2}{N * (N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} I_{f_1,f_2}(i,j) \tag{8.3}$$

where

$$I_{f_1,f_2}(i,j) = \begin{cases} 1 & \text{if } f_1(i) < f_1(j) \text{ and } f_2(i) > f_2(j) \\ 0 & \text{otherwise} \end{cases}$$

with $dr(f_1, f_2)$, the ranking distance between the folds/models f_1 and f_2 and N, the number of elements in the ranking, which corresponds to the number of regions in the atlas in the present case (i.e. 120 labelled regions). The values of *dr* range from 0 (exactly the same rankings) to 1 (exactly opposite rankings). This distance measure is illustrated in Table 8.2 for different rankings.

Table 8.2: Illustration of the ranking distance measure, for 4 rankings of 5 labelled regions. The distance between rankings *dr* was computed between the first and the other rankings.

Region	R_1	R_2	R_3	R_4
Region A	1	1	5	5
Region B	2	3	1	4
Region C	3	2	4	3
Region D	4	4	2	2
Region E	5	5	3	1
$dr(R_1, -)$	0	0.1	0.6	1

In this clinical application, the ranking distance was computed between each fold and their average on the basis of the normalized weights for the COMF vs. STAND comparison in the control (CTRL) and patients (IPD) groups, separately. A Kruskal-Wallis non-parametric statistical test then enabled the identification of outliers and assessed the stability of the ranking between each fold and the average across folds.

Beyond the identification of potential outliers in terms of pattern localization, it is possible to compute the ranking distance between different models using the expected values of the ranks across folds. In the present work, we identified the (combinations of) conditions leading to the largest distance between the patterns of the IPD and control groups. The idea behind this test is the following: if the comparison of two (combinations of) conditions generate dissimilar patterns in the two groups, then these conditions are probably suited to discriminate between those two groups. On the contrary, if the obtained ranking distance between groups is small, the binary comparison generated similar patterns, such that those conditions contain few information about the groups.

To compare groups in terms of patterns, the SVM models of all binary combinations of conditions were built within each group independently. The distance (*dr*) between each pair of CTRL - IPD models was then computed. To illustrate this process, consider the COMF vs STAND comparison previously investigated for pattern localization. The COMF vs STAND model was computed for the control group, and thereby an expected ranking (across folds) was obtained for each region. The same comparison was performed when considering the IPD group, which also gave expected rankings per region. These two expected rankings were then compared using the ranking distance. The operation was then repeated for each binary modelling of the conditions (i.e. COMF vs BRISK, BRISK vs STAND, COMF+BRISK vs STAND, ...), allowing to compare the different binary combinations of conditions in terms of distance between groups. Please note that this operation was led on the whole brain and motor masks only, since the MLR mask already defined a specific ROI (more specifically, rankings would only involve three regions at most).

Finally, the SVM and GP techniques were compared in terms of pattern localization by computing the ranking distance between SVM and GP models on all binary combinations of the conditions in the control (CTRL) group. This comparison was performed in the same way as when comparing the two groups. According to [Pereira et al., 2009], binary SVM and GP models should generate similar patterns. The ranking distance was therefore expected to have values closer to those observed between folds than to those between groups.

Computing the ranking distance dr between groups or modelling techniques provides a first quantitative idea on how similar two patterns are. However, although dr varies from 0 to 1, no probability value is associated to the ranking distance. Drawing conclusions from these values can hence be complicated[3]. To obtain p-values associated with the ranking distance, we resorted to the permutation testing performed to assess the significance of the balanced/class accuracies (see section 2.5.4). For a given model (e.g. the COMF vs STAND comparison performed within the CTRL group, using the whole brain mask), a weight image was built for each random permutation of the labels. The normalized weights per region NW_{ROI} were then computed for each fold of the permutation (using the same atlas as previously). Thereby, the expected values of the ranking across folds could be calculated for each permutation. The ranking distance dr between two models (e.g. COMF vs STAND comparison in the CTRL and IPD groups) was then computed between each pair of permutations, providing a “null” distribution of dr between these two models. This distribution can then be compared to the “true” value of dr: if dr is significantly ($p < 0.05$) smaller than the dr values computed from the random permutations of the labels, the patterns can be considered as significantly similar.

In this work, 100 permutations were computed for each model comparing the CTRL and IPD groups (24 models in total, for the whole brain and motor masks), giving 100 vectors of expected ranking. The binary comparison of these vectors led to $\frac{100\times(100-1)}{2}$ values for dr. The same approach was applied to the comparison between SVM and GP models on the control group (whole brain mask).

[3] dr can actually be compared to a correlation coefficient without a p-value.

Chapter 9

Results

Contents

9.1 Pattern discrimination

9.1.1 Between groups comparison

Results are presented for each (combination of) condition(s) in Table 9.1 in terms of balanced accuracy while weights of the BRISK+COMF model are represented in Figure 9.1 for each mask. The results of the majority vote computed from the STAND, BRISK and COMF models are presented in the last line of Table 9.1.

Overall, the whole brain mask led to a poor discrimination of IPD vs. CTRL, with the accuracy reaching a maximum at 62.3% when considering both the COMF and BRISK conditions together. This is the only significant result with the whole brain mask.

Slightly better results were obtained from the features in the motor mask, as shown by a higher balanced accuracy for the BRISK-COMF combination, as well as for the BRISK condition (both significant at $p < 0.05$).

The signal comprised in the MLR mask led to the best results, the highest performance being reached when considering the COMF condition. For this model, the balanced accuracy had a value of 76% (p=0.01), with the class accuracies reaching 78.6% for IPD and 73.3% for CTRL (both significant at $p < 0.05$). PPV are in the same range, with a PPV of 78.6% for CTRL and 73.3% for IPD.

Regarding the majority vote, the results showed no improvement for the whole

Table 9.1: Balanced accuracy (in %) for the IPD vs. CTRL classification for each combination of the three tasks (rows) and for each mask (columns). "All" represents the combination of the three tasks, while "MV" refers to the majority vote computed from the first three models. Significant results are displayed in bold.

Conditions used	**Masks used**		
Condition	**Whole brain**	**Motor**	**MLR**
STAND	14.3	34.5	**72.6**
COMF	58.3	62.1	**76.0**
BRISK	59.0	**66.2**	62.1
STAND+COMF	36.3	36.2	**72.4**
STAND+BRISK	36.7	39.7	**65.4**
COMF+BRISK	**62.3**	**65.8**	**62.1**
All	42.9	48.3	56.4
MV	44.83	44.83	**86.19**

brain and motor masks, with balanced accuracies below the chance level of 50%. On the opposite, a clear increase in balanced accuracy was found for the MLR mask. For this mask, the balanced accuracy reached 86.19%, with 86.67% of correct classification and PPV for the CTRL group and 85.71% for the IPD group. This result suggests that the errors made individually by the three models are different and can be compensated by the voting operation.

9.1.2 Between tasks comparison

In terms of balanced accuracy, significant results could be obtained from the discrimination between the three tasks across the two groups of subjects when considering the whole brain and motor masks (Table 9.2). This result is confirmed by the PPV for each class, which are quite high. However, the signal in the MLR does not lead to a significant classification of the three tasks, which is further confirmed by the PPV (almost all samples are classified as STAND).

Table 9.2: Balanced acc_b and class accuracies (in %, PPV in brackets) of the multiclass GP model discriminating between the three tasks (STAND, COMF and BRISK) when considering both groups jointly. Significant results are displayed in bold.

Mask	$\boldsymbol{acc_b}$	**STAND**	**COMF**	**BRISK**
Whole brain	**65.5**	58.6 (65.4)	41.4 (75.0)	**96.6** (62.2)
Motor areas	**66.7**	62.1 (64.3)	41.4 (70.6)	**96.6** (66.7)
MLR area	32.2	**89.7** (32.1)	0.0 (0.0)	6.9 (33.3)

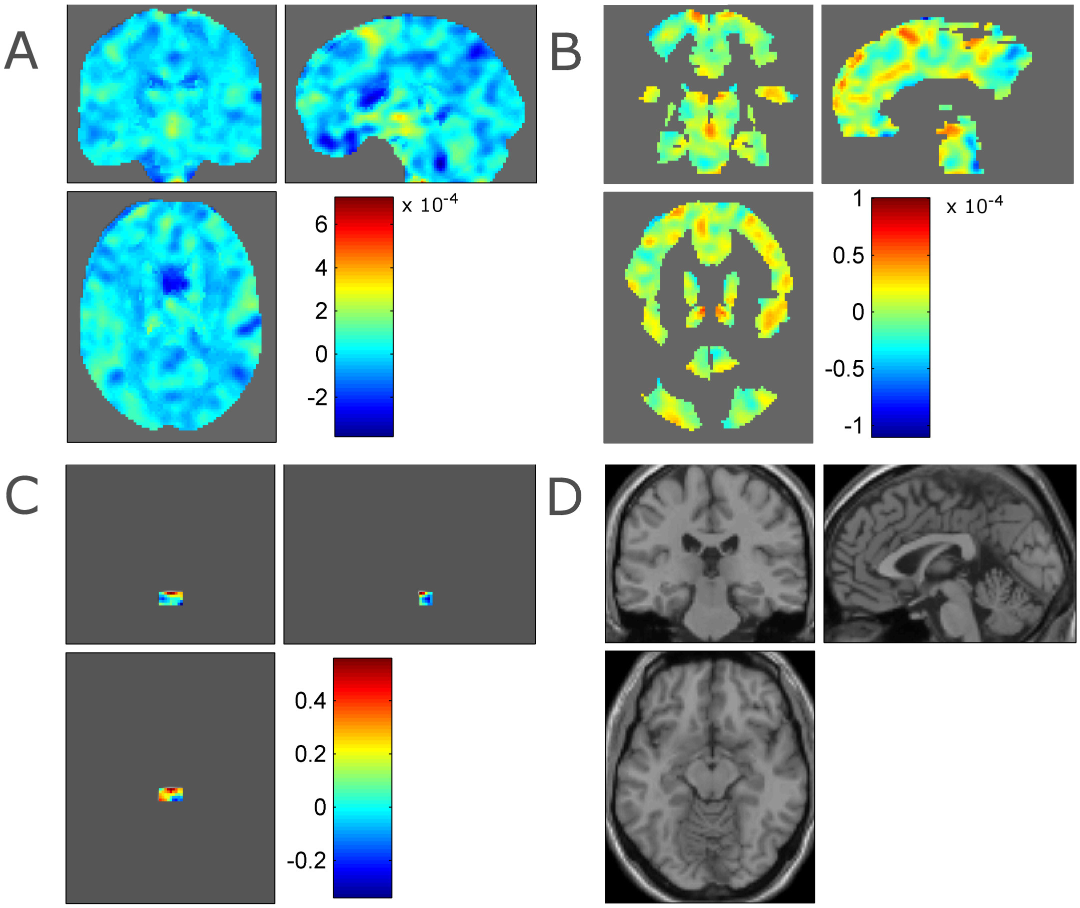

Figure 9.1: Weights of the SVM model discriminating between CTRL (label +1) and IPD (label -1) based on the combination of the BRISK and COMF conditions, for **A** the whole brain, **B**, motor and **C** MLR masks (Note that the cross-hair position has been centred on the MLR for this panel). **D** displays the SPM single subject canonical structural image for better representation.

9.2 Pattern localization

9.2.1 Representing the weights

The results of the COMF vs. STAND comparison are presented below for each group separately in terms of balanced and class accuracies. The overlap between the mask and the atlas in terms of absolute weights is also reported, as well as the ranking of the *others* region. Please note that the *others* region represent 16.39% of the total volume of the whole brain mask, and 7.47% of the motor mask. Furthermore, the top ten (arbitrarily fixed number) regions according to normalized weights, NW, are represented for each group in Table 9.3 and compared to the univariate results from Table 7.1.

In terms of comparison between the two groups (i.e. CTRL vs. IPD discrimination), section 9.1 has shown that using the combination of both the BRISK and COMF conditions led to the most stable results across feature sets, giving significant accuracies for each mask. Therefore, to illustrate the smoothed representation of the weights, the pattern of this model was localized and displayed in Figure 9.2 for the whole brain mask. As for the comparisons within each group, the top ten regions according to *NW* were reported in Table 9.3 and compared to the univariate results of [Cremers et al., 2012b].

Control group. The mental imagery of the COMF and STAND conditions could be discriminated with an accuracy of 86.7% (STAND: 100%, COMF: 73.3%). The weight of the region *others* represented 13.03% of the total weights, meaning that the atlas and the mask overlap at 86.87% when considering the sum of the absolute values of the weights. This region was ranked first in terms of absolute weights W_{ROI}, but 117/120 in terms of normalized weights *NW*, suggesting that if some voxels were important for the classification, this information was lost when building the *others* region. Regarding the comparison of the results with Table 7.1, the medulla, cerebellar vermis and hemisphere regions, the SMA, the caudate nuclei and middle and inferior frontal regions were ranked in the top 15.

IPD group. The accuracy of the classification between the COMF and STAND conditions reached 85.7%, with class accuracies of 100% for STAND and 71.4% for COMF. The *other* region represented 18% of the weights (rank 1), and 0.61% of the normalized weights (rank 87). Although some regions were not reported in previous univariate studies [Maillet et al., 2012; Cremers et al., 2012b], the SMA and the anterior cingulate cortex were ranked in the top 10.

IPD versus control. The *others* region was ranked first in terms of weights (21.0906%) and 120/120 in terms of normalized weights *NW*. The medulla, cerebellar vermis and hemisphere regions were ranked in the top 10, which is in agreement with the univariate results of [Cremers et al., 2012b]. However, the parametric maps of each group were compared in a directed way (controls > IPD) considering the COMF>STAND contrast, which precludes from any conclusion.

9.2.2 Comparing patterns

To quantify the difference between patterns in terms of localization, the ranking distance was computed in different situations:

Distance across folds. When considering the COMF-STAND comparison, the ranking distance between each fold and their average varied from 0.0108 to 0.1571 for the control group and from 0.0157 to 0.2076 for the IPD group. A Kruskal-Wallis statistical test revealed no significant difference in the distance distribution across groups ($p = 0.7766$). However, it identified patient 3 as an outlier in the IPD group. This suggests that the average pattern across folds might be importantly influenced by the data of this patient.

To investigate whether this subject was also an outlier in terms of ranking distance when it comes to the group comparison, a Kruskal-Wallis test was performed on the ranking distances across folds obtained from the CTRL-IPD comparison based on the COMF and BRISK conditions (whole brain). As shown in Figure 9.4, no statistically significant difference was observed between the two groups in terms

Table 9.3: Regions ranked according to their normalized weights in the control and IPD groups when classifying COMF and STAND, and for their comparison when considering the BRISK and COMF conditions jointly. The whole brain mask was used for all the cases. 'SMA' stands for 'Supplementary Motor' Area, 'Inf' for 'inferior', 'Mid' for 'middle', 'Med' for 'medial', 'Sup' for 'superior' and 'Ant' for 'anterior'. Lateralization is displayed using L (left) and R (right), when sound.

Rank	**COMF vs. STAND (CTRL)**	**COMF vs. STAND (IPD)**	**CTRL vs. IPD (COMF+BRISK)**
1	Medulla	Olfactory (L)	Caudate (L)
2	Vermis 3	SupraMarginal (L)	Vermis 10
3	Cerebellum 3 (L)	Cerebellum 10 (l)	Medulla
4	Vermis 4-5	Vermis 3	Mid Frontal (L)
5	SMA (R)	Rectus (L)	Inf Frontal (R)
6	Sup Temporal (L)	SMA (R)	Rectus (L)
7	Caudate (L)	Rectus (R)	Inf Frontal (L)
8	SMA (L)	Med Frontal (R)	Cerebellum 7 (R)
9	Inf Frontal (L)	Olfactory (R)	Cerebellum 10 (L)
10	Angular (L)	Ant Cingulate (R)	Sup Frontal (R)

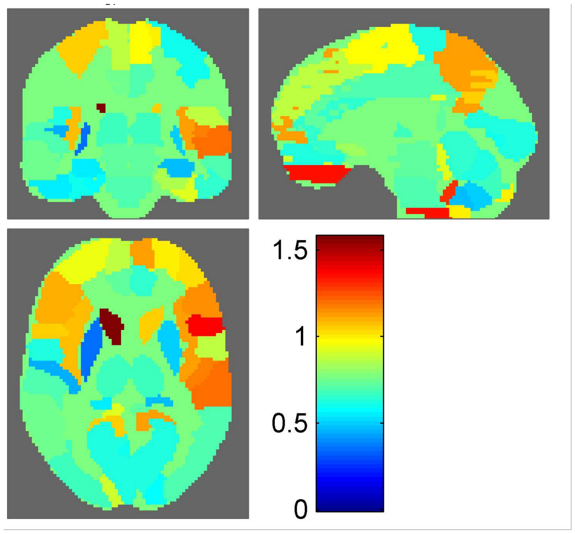

Figure 9.2: **Smoothed weights of the SVM model discriminating between CTRL and IPD** based on the combination of the BRISK and COMF conditions, for the whole brain mask. The proportions of *NW* are represented for each labelled region, the regions with the highest proportions of *NW* in red, the lowest (close to zero), in blue.

of ranking distance across folds. This result suggests that patient 3 displays a different pattern only for the COMF versus STAND comparison. This could be due to increased noise or movements during the acquisition of one of these two conditions or both. Since COMF is common to both the within and between groups model, the noise probably comes from the STAND condition.

Distance between groups. The ranking distance across groups using all possible

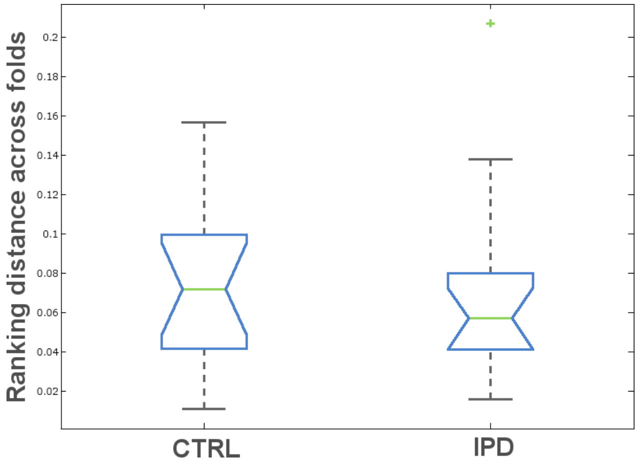

Figure 9.3: **Ranking distance across folds** for the COMF-STAND comparison, for each group separately. A Kruskal-Wallis statistical test revealed no significant difference between the two groups but displayed patient 3 as an outlier for the COMF vs. STAND comparison.

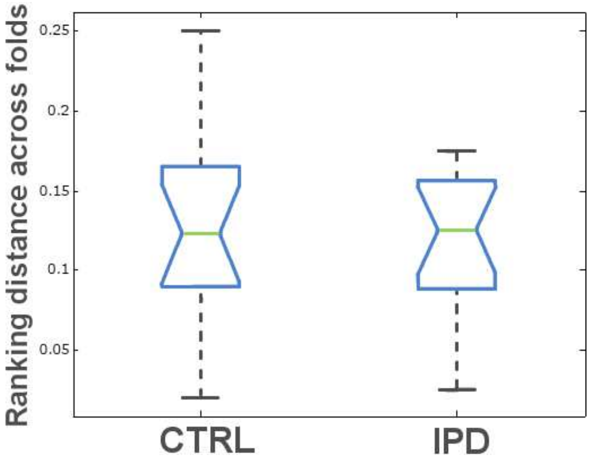

Figure 9.4: **Ranking distance across folds** for the CTRL-IPD comparison, when considering the whole brain mask and the COMF and BRISK conditions jointly. A Kruskal-Wallis statistical test revealed no significant difference between the two groups in terms of ranking distance.

binary combinations of conditions and for both the whole brain and motor masks are presented in Table 9.4, along with the balanced accuracy of the binary model in each group. The largest ranking distances between groups are observed for the STAND vs COMF (Figure 9.5) and BRISK+COMF vs STAND models, with a slight increase in distance from the whole brain to motor mask. This is in agreement with the combinations of conditions leading to the best discrimination between groups, although no direct comparison can be performed.

Although the COMF vs BRISK comparison also showed a large ranking distance between the two groups, the models within each group showed low balanced accuracy values, such that these models did not really learn from the labels. This

Table 9.4: The balanced accuracy of the model of each possible binary combination of the conditions was displayed for each group and mask, as well as the ranking distance between the control and IPD groups. CTRL represents the balanced accuracy for the control group, IPD, the balanced accuracy for the IPD group and *dr*, the ranking distance between the two groups. Significant balanced accuracies are marked in bold. Ranking distances marked in bold are significantly smaller than for permuted labels ($p < 0.05$), while ranking distances marked with a star* show a trend ($0.05 < p < 0.1$).

	Whole brain			Motor regions		
Model	**CTRL**	**IPD**	***dr***	**CTRL**	**IPD**	***dr***
BRISK vs COMF	**66.7**	50.0	0.3688	**63.3**	57.1	0.3757
BRISK vs STAND	**83.3**	**75.0**	0.3317*	**83.3**	**75.0**	0.3320*
COMF vs STAND	**86.7**	**85.7**	0.3990	**86.7**	**85.7**	0.4040
BRISK+STAND vs COMF	**61.7**	**62.5**	0.3387*	**63.3**	**58.9**	0.3340*
BRISK+COMF vs STAND	**85.0**	**85.7**	0.3721	**85.0**	**87.5**	0.3725
COMF+STAND vs BRISK	**75.0**	57.1	**0.3052**	**75.0**	57.1	**0.3108**

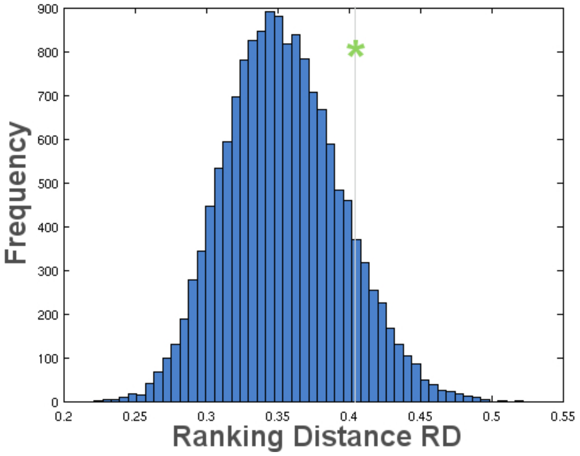

Figure 9.5: **Permuted ranking distance between CTRL and IPD** for the COMF-STAND comparison. The values of *dr* obtained from the random permutations of the labels are displayed in the blue histogram. The “true” value of *dr* is displayed as a green star. One can see that the “true” value of *dr* is not smaller than the *dr* obtained from the permuted labels.

suggests that the generated patterns could have been generated from any random permutation of the labels, as shown by the ranking distance (which is not significant).

It should be noted that the COMF+STAND vs BRISK model led to ranking distances significantly smaller for the true labels than for random permutations. Trends were also noted for the BRISK vs STAND and BRISK+STAND vs COMF models, suggesting that these three models generate (significantly) similar patterns for both groups.

Distance between techniques. Table 9.5 displays the ranking distance between the SVM and GP models built on the same comparisons between (combinations of) conditions. The ranking distances between techniques are much closer to the ranking distances between each fold and their average than to the ranking distances between the two groups (CTRL vs IPD). This is further confirmed by their associated p-value, which reveals that the ranking distances for the true labels are all significantly smaller than for permuted labels (see Figure 9.6 for the COMF-STAND comparison).

Table 9.5: The balanced accuracy of the model of each possible binary combination of the conditions was displayed for each technique on the control group (whole brain mask), as well as the ranking distance between the obtained patterns. SVM represents the balanced accuracy for the SVM models, GP, the balanced accuracy for the Gaussian processes models and dr, the ranking distance between the two techniques. Significant balanced accuracy are represented in bold. Ranking distances marked in bold are significantly smaller than for permuted labels ($p < 0.05$).

Model	SVM	GP	***dr***
BRISK vs COMF	**66.7**	**63.3**	**0.0847**
BRISK vs STAND	**83.3**	**80.0**	**0.1349**
COMF vs STAND	**86.7**	**86.7**	**0.1193**
BRISK+STAND vs COMF	**61.7**	**68.3**	**0.1056**
BRISK+COMF vs STAND	**85.0**	**85.0**	**0.1422**
COMF+STAND vs BRISK	**75.0**	**68.3**	**0.0996**

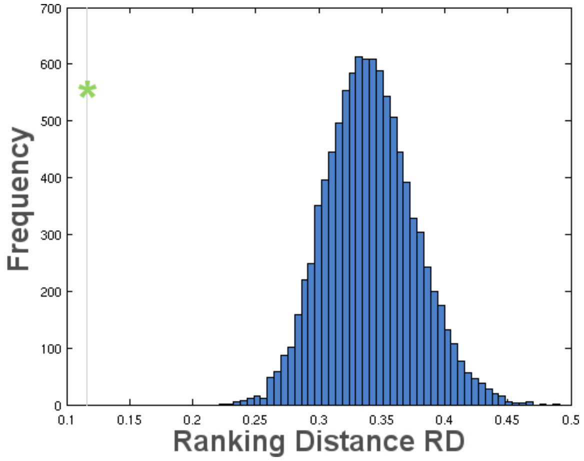

Figure 9.6: **Permuted ranking distance between SVM and GP** for the COMF-STAND comparison. The values of *dr* obtained from the random permutations of the labels are displayed in the blue histogram. The “true” value of *dr* is displayed as a green star. One can see that the “true” value of *dr* is smaller than any of the *dr* obtained from the permuted labels.

Chapter 10

Discussion

Contents

In this clinical application, two issues were tackled, namely the pattern discrimination and localization of idiopathic Parkinson's disease using the mental imagery of gait in fMRI.

10.1 Pattern discrimination

To identify idiopathic Parkinson's disease patients from healthy controls, the parametric maps of three mental imagery tasks were classified (STAND, COMF and BRISK). The best model discriminated significantly between IPD and controls with a balanced accuracy of 76%, when using the signal comprised in the mesencephalic locomotor region. The considered voting operation (i.e. majority vote from the STAND, COMF and BRISK classifiers) led to an increase in performance of 10.19%, yielding a balanced accuracy of 86.19%. This result suggests that the errors made by the three models are different and that combining their outputs through a simple majority vote can provide a stronger classifier.

As revealed by table 9.1, the ability of a model to discriminate between IPD and controls depends heavily on the selected voxels. However, combining the mental imagery of gait at comfortable and brisk paces led to significant results across feature sets, suggesting that the combination of these conditions could lead to a consistent model across feature sets.

Although these results are not overwhelming, it is (to the best of our knowledge) one of the first significant classification between IPD and controls [Focke et al., 2011; Orrù et al., 2012] using (f)MRI. Furthermore, the performance of the boosted model competes with the correct diagnostic rates obtained by clinicians (which is of 89% in average when considering dopamine uptake, Acton et al., 2006). In view of the simplicity of the model (SVM with a filter feature selection), this result is promising and suggests that mental imagery of gait could be a biomarker of Parkinson's disease although improvements need to be performed before being able to deal with more complex issues, such as diagnosing Parkinson's disease in its early stage.

However, distinguishing between the three tasks using both groups led to significant results when considering the whole brain and motor masks, suggesting that the between-subject variability within one group is large compared to the between-groups variability for those features. This result highlights the importance of feature selection in the present case and favours the use of wrapper or embedded feature selection techniques to increase the performance of the machine learning based models. Regarding the Parkinsonian group, the between-subjects variability might further be explained by the heterogeneity of the gait disorders in patients. Distinguishing between patients with light or severe gait disorders, for example by considering the Freezing of Gait (FoG, Karachi et al., 2010), might increase the ratio of between versus within group variability and thereby improve the classification. Another issue to consider for diagnostic purposes is disease duration; there was a large inter-individual variability in disease duration (and severity, see Cremers et al., 2012b for a table presenting different disease parameters for each patient). A possible improvement would hence be the inclusion of de novo patients (i.e. early stage patients). Finally, medication was another confounding factor since all patients were scanned on medication, with a variability in the equivalent doses of medicine across subjects. In conclusion to this comment, the inclusion of early stage de novo patients who are not yet treated should decrease the within-group variability and thereby might improve the performance of the classification.

Finally, although a large overlap has been observed between mental imagery of gait and actual gait in healthy subjects [Dobkin et al., 2004], our results question the overlap between mental imagery of disturbed gait and actual disturbed gait, especially in the STAND and BRISK conditions. Solving this issue is not straightforward but developments in ambulatory Electro-EncephaloGraphy (EEG) acquisition systems and in the decoding of this type of signal might bring a solution by directly acquiring the brain activity under actual gait.

10.2 Pattern localization

The patterns generated by machine learning based discriminative models can be difficult to interpret since no thresholding can be performed. Secondly, models are rarely compared in terms of weight patterns. To allow cognitive interpretations of the weights, a region specific scalar was defined: the normalized weights per region (*NW*). This value can be displayed over all the regions and/or ranked in descending order according to its proportion. We also aimed at quantifying the differences between patterns in terms of localization, which corresponded to the ranked list of regions in this work.

10.2.1 Interpreting the weight

To facilitate the interpretation of the weights, the normalized weights were computed within labelled regions as defined by an atlas. The atlas can be generated from classic atlases (Brodmann or AAL) or manually [Maldjian et al., 2003], on the brain structure or functioning. The normalized weights can then be ranked, hence offering a more intuitive way to visualize the weights and allowing for cognitive conclusions.

When comparing the ranked list of regions to univariate results [Cremers et al., 2012b], a nice overlap could be observed. Although the univariate and multivariate analyses do not represent exactly the same comparisons, they seem to provide similar lists of regions. It should be noted that this has been verified in the case of a sound model, with accuracy higher than 85%. Since weights can be generated from any model, they do not represent where in the brain is the information about the considered categories, but rather the ability of each region to discriminate between the categories, which could be linked to noise or confounds. Therefore, when univariate results are available for comparison, the list of regions ranked according to their normalized weights could provide information about the quality of the model. In case there is no functional/anatomical a priori on pattern localization, one could build accuracy maps, using the searchlight approach [Kriegeskorte et al., 2006] for example. Another way to localize the pattern would be to build one model per region, using multiple kernel learning algorithms [Gönen and Alpaydin, 2011]. This approach would further solve the issue of the overlap between the mask and the atlas.

However, when considering regions, the choice of the atlas is important, since the size and shape of the regions can be quite different from one atlas to another. Therefore, the atlas should be picked carefully, taking a priori information into account when available (e.g. lateralization or further dividing specific regions). Testing different atlases could also provide further information but remains an open question.

10.2.2 Comparing patterns

In this work, we provided the *ranking distance* [Lempel and Moran, 2005], which quantifies the difference between two ranking vectors. This distance was computed within and between models and proved useful for different aspects of multivariate analysis. First, it enabled assessing the homogeneity of the data in terms of pattern within one category. In the present case, the control group was more homogeneous than the IPD group for the COMF vs STAND comparison, which showed an outlier in terms of the distance between each fold and their average. The heterogeneity in the group of patients might be due to the state of disease, medication or movements in the scanner. However, this heterogeneity did not seem to affect the comparison between groups, since no significant difference was found between the ranking distances across folds for the CTRL and IPD group, when considering the BRISK and COMF conditions jointly.

When comparing similar models computed on different groups, the results suggested that the larger the distance between groups, the higher the performance when discriminating between groups (if the considered models are sound). In exploratory cases, the ranking distance could thereby help finding which features or conditions could become biomarkers of the variable of interest.

Finally, the ranking distance allowed the comparison of different modelling techniques (namely binary SVM and GP) in terms of pattern localization. This information could be added to comparisons in terms of accuracy (balanced accuracy, area under the curve) or model fitting (e.g. maximum a posteriori likelihood in Gaussian processes), and could become particularly helpful for the interpretation of the pattern when feature selection steps or sparse models are involved.

Conclusions: In this clinical setting, we tackled the issues of pattern discrimination and localization of Parkinson's disease, when compared to healthy subjects. Although there is room for improvement, the mental imagery of gait at both comfortable and brisk paces proved to be a promising fMRI biomarker of IPD. Techniques to display and compare patterns in terms of their localization were further developed and might help the cognitive interpretation of the multivariate results.

Chapter 11

Conclusions and final remarks

In this work, we investigated the assets and disadvantages of machine learning based modelling of neuroimaging data via two applications. The first application was designed to study mnemonic traces during conscious resting-state directly following a learning task, while the second aimed at discriminating and localising the patterns of idiopathic Parkinson's disease. These two applications involved complex datasets and presented challenges that could not be successfully solved using other techniques, such as univariate models or network analyses.

In both cases, machine learning based models enabled to overcome issues that other methods encountered: they allowed the modelling of spontaneous brain activity without the suppression of the temporal evolution and permitted the significant discrimination between healthy controls and Parkinson's diseased patients.

The main disadvantage with machine learning based models is that the voxels' weights are not easily interpretable since weight maps cannot be thresholded due to their multivariate nature. Neuroscientists thus prefer univariate (SPM, Friston et al., 2007) or locally multivariate (Searchlight, Kriegeskorte et al., 2006) techniques to infer cognitive conclusions on the location of the information discriminating between groups/conditions. In the present work, we proposed an approach to localize multivariate patterns, by parcelling the weight image into functionally or anatomically labelled regions that can then be ranked according to their normalized weight (NW_{ROI}). Although this procedure is recent and needs more testing, the results displayed in chapter 9 are promising, showing a good overlap with previously published univariate results [Cremers et al., 2012b].

It is interesting to note that feature selection approaches improved the model performance in both applications: the best procedure to model semi-constrained brain activity involved univariate (F-test filtering) and multivariate feature selections (RFA, section 4.5). In part II, the MLR mask (i.e. ROIs selected on prior knowledge) performed better at discriminating between Parkinson's diseased patients and controls than the whole brain mask. These results questions the conclusions from recent works stating that data-driven feature selection approaches (i.e. GLM or RFA) did not bring any increase in model accuracy [Chu et al., 2012] or that space compression (i.e. ROI selection) had no effect on the performance of multisubject classifiers [Mourão-Miranda et al., 2006]. While the debate on the usefulness of feature selection approaches when modelling high-dimensionality data is legitimate, the contradictions in the conclusions across studies suggests that the usefulness of feature selection steps is data-dependent. This might be due to the level of noise, or of brain activity unrelated to the variable of interest in the data, which would influence the relationship between the number of features and model performance (illustrated in figure 4.3).

For both applications, some of the limitations of the obtained results were caused by the limitations of the datasets: the scaling factor of mnemonic traces could not have been further investigated due to the temporal resolutions of the haemodynamic response and the fMRI. Similarly, having access to the pattern of brain activity generated during real gait (and not during mental imagery of gait) would be desirable, which is not possible in the context of fMRI or PET imaging. As proposed in chapters 6 and 10, EEG could provide answers to the aforementioned limitations of fMRI or PET. However, while EEG might provide a higher temporal resolution and freedom of movement (to a certain extent), the classification of such datasets remains an open issue, due to their low signal-to-noise ratio.

In view of the results, this work leads to the conclusion that machine learning models can indeed bring insights on complex problems that could hardly be addressed with other techniques. However, many challenges remain. A first example is that for the built models to be used by other neuroscientists, they have to be subject and centre independent. As shown by the three behavioural outliers in part I, obtaining models that can successfully be applied to any subject can be complex. This issue was further illustrated by the large variability in proportions Pr across subjects, especially when considering the pre-task rest session in the memory condition for which Pr was associated to the detection of false positives. In part II, the data was acquired from a unique centre. The built models hence depend on the acquisition machine, timing and environment and would therefore certainly not perform as well on datasets from other centres. Centre-independent models require the training set to comport data from many centres and thereby represents one of the biggest challenges for machine learning based models to be distributed. Therefore, although the models built on both applications gave satisfactory results, they would not perform similarly on datasets from other subjects and/or centres.

Among the other challenges still to overcome: the early diagnosis of a disease, predicting the evolution of a disease for a specific patient or the response to treatment.

Although much work remains, it seems likely that machine learning models will constitute a new way of analysing data, that can complement other techniques and bring new insights on the two fundamental questions in neuroscience.

Conclusion: In this work, the assets and the limits of machine learning based models applied to neuroimaging data were investigated via a neuroscience and a clinical application, each involving complex datasets. Although much work remains for the obtained models to be useful as neuroscience or clinical tools, our results showed that multivariate modelling could overcome some of the issues encountered with other techniques.

Bibliography

P. D. Acton and A. Newberg. Artificial neural network classifier for the diagnosis of Parkinson's disease using [99mtc]trodat-1 and spect. *Phys Med Biology*, 51: 3057–3066, 2006.

P. D. Acton, A. Newberg, K. Plössl, and P. D. Mozley. Comparison of region-of-interest analysis and human observers in the diagnosis of Parkinson's disease using [99mtc]trodat-1 and spect. *Phys Med Biology*, 51:575, 2006.

M. Alam, K. Schwabe, and J. Krauss. The pedunculopontine nucleus area: critical evaluation of interspecies differences relevant for its use as a target for deep brain stimulation. *Brain*, 134:11–23, 2011.

J. Andersson, C. Hutton, J. Ashburner, R. Turner, and K. Friston. Modelling geometric deformations in EPI time series. *NeuroImage*, 13(5):903–919, 2001.

J. Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38: 95 – 113, 2007. ISSN 1053-8119.

J. Ashburner and K. Friston. Voxel-based morphometry – the methods. *NeuroImage*, 11:805–821, 2000.

J. Ashburner and K. Friston. Unified segmentation. *NeuroImage*, 26:839–851, 2005.

P. Aubin, A. Serackis, and J. Griskevicius. Support vector machine classification of Parkinson's disease, essential tremor and healthy control subjects based on upper extremity motion. In *International Conference on Biomedical Engineering and Biotechnology*, pages 900–904, 2012.

A. Beck, N. Epstein, G. Brown, and R. Steer. An inventory for measuring clinical anxiety: Psychometric properties. *J. Consult. Clin. Psychol.*, 56:893–897, 1988.

C. M. Bishop. *Pattern Recognition and Machine learning*. Springer, 2006.

M. Boly, V. Perlbarg, M. G, M. Schabus, S. Laureys, J. Doyon, M. Pélégrini-Issac, P. Maquet, and H. Benali. Hierarchical clustering of brain activity during human non rapid eye movement sleep. *Proc Natl Acad Sci USA*, 109:5856–5861, 2012.

C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

D. Buysse, C. 3rd Reynolds, T. Monk, S. Berman, and D. Kupfer. The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research. *Psychiatry Res.*, 28:193–213, 1989.

M. Chadwick, D. Hassabis, N. Weiskopf, and E. Maguire. Decoding individual episodic memory traces in the human hippocampus. *Curr. Biol.*, 20:1–4, 2010.

C. Chu, A.-L. Hsu, K.-H. Chou, P. Bandettini, and C. Lin. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Neuroimage*, 60:59 – 70, 2012. ISSN 1053-8119.

D. D. Cox and R. L. Savoy. Functional magnetic resonance imaging (fMRI) 'brain reading': detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage*, 19:261–270, 2003.

J. Cremers, A. Dessoullieres, and G. Garraux. Hemispheric specialization during mental imagery of brisk walking. *Hum. Brain Mapp.*, 33:873–882, 2012a.

J. Cremers, K. D'Ostilio, J. Stamatakis, V. Delvaux, and G. Garraux. Brain activation pattern related to gait disturbances in Parkinson's disease. *Mov. Disord.*, 27:1498–1505, 2012b.

J. S. Damoiseaux, S. A. R. B. Rombouts, F. Barkhof, P. Scheltens, C. J. Stam, S. M. Smith, and C. F. Beckmann. Consistent resting-state networks across healthy subjects. *Proceedings of the National Academy of Sciences*, 103:13848–13853, 2006.

F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano. Combining multivariate voxel selection and support vector machines for mapping and classification of fmri spatial patterns. *NeuroImage*, 43:44–58, 2008.

R. Deichmann, C. Schwarzbauer, and R. Turner. Optimisation of the 3D MDEFT sequence for anatomical brain imaging: technical implications at 1.5 and 3 T. *Neuroimage*, 21:757 – 767, 2004. ISSN 1053-8119.

T. G. Dietterich and G. Bakiri. Solving multiclass learning problem via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

B. H. Dobkin, A. Firestine, M. West, K. Saremi, and W. R. Ankle dorsiflexion as an fMRI paradigm to assay motor control for walking during rehabilitation. *NeuroImage*, 23:370–381, 2004.

S. Duchesne, Y. Rolland, and M. Vérin. Automated computer differential classification in Parkinsonian syndromes via pattern analysis on MRI. *Acad. Radiol.*, 16:61–70, 2009.

O. Dunn. Multiple comparisons among means. *J. Am. Stat. Assoc.*, 56:52–64, 1961.

C. P. Ferri, M. Prince, C. Brayne, H. Brodaty, L. Fratiglioni, M. Ganguli, K. Hall, K. Hasegawa, H. Hendrie, Y. Huang, A. Jorm, C. Mathers, P. R. Menezes, E. Rimmer, and M. Scazufca. Global prevalence of dementia: a delphi consensus study. *The Lancet*, 366:2112 – 2117, 2006.

N. K. Focke, G. Helms, S. Scheewe, P. M. Pantel, C. G. Bachmann, P. Dechent, J. Ebentheuer, A. Mohr, W. Paulus, and C. Trenkwalder. Individual voxel-based subtype prediction can differentiate progressive supranuclear palsy from idiopathic Parkinson syndrome and healthy controls. *Hum. Brain Mapp.*, 32: 1905–1915, 2011. ISSN 1097-0193.

E. Formisano, F. de Martino, and G. Valente. Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. *Magn. Reson. Imaging*, 26:921–934, 2008.

D. J. Foster and M. A. Wilson. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, 440:680–683, 2006.

R. Frackowiak, K. Friston, C. Frith, C. Price, S. Zeki, J. Ashburner, and W. Penny. *Humain brain function*. Elsevier Academic Press, New York, 2004.

K. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *NeuroImage*, 19:1273–1302, 2003.

K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny. *Statistical Parametric Mapping: the analysis of functional brain images*. Elsevier Academic Press, London, 2007.

K. Friston, C. Chu, J. Mourão-Miranda, O. Hulme, G. Rees, W. Penny, and J. Ashburner. Bayesian decoding of brain images. *NeuroImage*, 39:181 – 205, 2008. ISSN 1053-8119.

J. Fürnkranz. Round robin classification. *Journal of Machine Learning Research*, 2:721–747, 2002.

G. Garraux, C. Phillips, J. Schrouff, and E. Salmon. Multiclass classification of FDG PET scans for the distinction between Parkinson's disease and atypical Parkinsonian syndromes. *NeuroImage Clinical*, submitted.

M. S. Gazzangina, R. B. Ivry, and G. R. Mangun. *Cognitive neuroscience*. Norton, 2002.

R. Geetha Ramani and G. Sivagami. Parkinson's disease classification using data mining algorithms. *International journal of computer applications*, 32:17–22, 2011.

G. Girardeau, K. Benchenane, S. I. Wiener, G. Buzsaki, and M. B. Zugaro. Selective suppression of hippocampal ripples impairs spatial memory. *Nat. Neurosci.*, 12: 1222–1231, 2009.

M. Gönen and E. Alpaydin. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.

D. Green and J. Swets. *Signal detection theory and psychophysics.* Wiley, New York, 1966.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

R. H. Hashemi and W. G. J. Bradley. *MRI: the basics.* Williams and Wilkins, 1997.

D. Hassabis, C. Chu, G. Rees, N. Weiskopf, P. Molyneux, and E. Maguire. Decoding neural ensembles in the human hippocampus. *Curr. Biol.*, 19:546–554, 2009.

T. Hastie, R. Tibshirani, and J. H. Friedman. *Elements of Statistical Learning.* Springer, 2003.

J. Haynes and G. Rees. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nat. Neurosci.*, 8:686–691, 2005.

J.-D. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.*, 7:523–534, 2006.

K. Hoffman and B. McNaughton. Coordinated reactivation of distributed memory traces in primate neocortex. *Science*, 297:2070–2073, 2002.

J. Horne and O. Ostberg. A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *Int. J. Chronobiol.*, 4: 97–110, 1976.

C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. 13:415–425, 2002.

C. Hutton, A. Bork, O. Josephs, R. Deichmann, J. Ashburner, and R. Turner. Image distortion correction in fmri: A quantitative evaluation. *Neuroimage*, 16: 217 – 240, 2002. ISSN 1053-8119.

D. Ji and M. Wilson. Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nat. Neurosci.*, 10:100–107, 2007.

M. Johns. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *Sleep*, 14:540–545, 1991.

P. Jokinen, H. Helenius, E. Rauhala, A. Brãck, O. Eskola, and J. O. Rinne. Simple ratio analysis of 18F-fluorodopa uptake in striatal subregions separates patients with early Parkinson disease from healthy controls. *J. Nucl. Med.*, 50:893–899, 2009.

Y. Kamitani and F. Tong. Decoding the visual and subjective contents of the human brain. *Nat. Neurosci.*, 8:679–685, 2005.

C. Karachi, D. Grabli, F. A. Bernad, D. Tandé, N. Wattiez, H. Belaid, E. Bardinet, A. Prigent, H.-P. Nothacker, S. Hunot, A. Hartmann, S. Lehéricy, E. Hirsch, and F. Chantal. Cholinergic mesencephalic neurons are involved in gait and postural disorders in Parkinson disease. *J. Clin. Invest.*, 120:2745–2754, 2010.

P. E. Kinahan and D. C. Noll. A direct comparison between whole-brain PET and BOLD fMRI measurements of single-subject activation response. *Neuroimage*, 9:430 – 438, 1999. ISSN 1053-8119.

S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, N. C. Fox, C. R. Jack, J. Ashburner, and R. S. J. Frackowiak. Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131:681–689, 2008.

D. S. Knopman, S. T. DeKosky, J. L. Cummings, H. Chui, J. Corey–Bloom, N. Relkin, G. W. Small, B. Miller, and J. C. Stevens. Practice parameter: Diagnosis of dementia (an evidence-based review): Report of the quality standards subcommittee of the american academy of neurology. *Neurology*, 56(9): 1143–1153, 2001.

N. Kriegeskorte, R. Goebel, and P. Bandettini. Information-based functional brain mapping. *PNAS*, 103:3863–3868, 2006.

S. Laconte, S. Strother, V. Cherkassky, J. Anderson, and X. Huber. Support vector machines for temporal classification of block design fMRI data. *Neuroimage*, 26: 317–329, 2005.

S. Lancelot and L. Zimmer. Small-animal positron emission tomography as a tool for neuropharmacology. *Trends in Pharmacological Sciences*, 31(9):411 – 417, 2010. ISSN 0165-6147.

A. K. Lee and M. A. Wilson. Memory of sequential experience in the hippocampus during Slow Wave Sleep. *Neuron*, 36:1183–1194, 2002.

R. Lempel and S. Moran. Rank-stability and rank-similarity of link-based web ranking algorithms in authority-connected graphs. *Information Retrieval*, 8:245–264, 2005.

K. Louie and M. A. Wilson. Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, 29:145–156, 2001.

D. J. C. MacKay. The evidence framework applied to classification networks. *Neural Comput.*, 4:720–736, 1992.

D. J. C. MacKay. *Models of Neural Networks III*, chapter Bayesian methods for backpropagation networks, pages 211–254. Springer, 1994.

A. Maillet, P. Pollak, and B. Debũ. Imaging gait disorders in parkinsonism: a review. *Journal of Neurology, Neurosurgery & Psychiatry*, 83(10):986–993, 2012.

J. A. Maldjian, P. J. Laurienti, J. B. Burdette, and R. A. Kraft. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *NeuroImage*, 19:1233–1239, 2003.

P. Maquet, S. Laureys, P. Peigneux, S. Fuchs, C. Petiau, C. Phillips, J. Aerts, G. Del Fiore, C. Degueldre, T. Meulemans, A. Luxen, G. Frank, M. Van der Linden, C. Smith, and A. Cleeremans. Experience-dependent changes in cerebral reactivation during human REM sleep. *Nature Neuroscience*, 3:831–836, 2000.

D. Margulies, J. Böttger, X. Long, Y. Lv, C. Kelly, A. Schäfer, D. Goldhahn, A. Abbushi, M. Milham, G. Lohmann, and A. Villringer. Resting developments: a review of fmri post-processing methodologies for spontaneous brain activity. *Magnetic Resonance Materials in Physics, Biology and Medecine*, 23:289–307, 2010.

G. Marrelec, J. Daunizeau, M. Pélégrini-Issac, J. Doyon, and H. Benali. Conditional correlation as a measure of mediated interactivity in fMRI and MEG/EEG. 53: 3501–3516, 2005.

G. Marrelec, A. Krainik, H. Duffau, M. Pélégrini-Issac, S. Lehéricy, J. Doyon, and H. Benali. Partial correlation for functional brain interactivity investigation in functional MRI. *Neuroimage*, 32:228–237, 2006.

G. Marrelec, B. Horwitz, J. Kim, M. Pélégrini-Issac, H. Benali, and J. Doyon. Using partial correlation to enhance structural equation modeling of functional mri data. *Magn. Reson. Imaging*, 25:1181 – 1189, 2007. ISSN 0730-725X.

G. Marrelec, P. Bellec, A. Krainik, H. Duffau, M. Pélégrini-Issac, S. Lehéricy, H. Benali, and J. Doyon. Regions, systems, and the brain: Hierarchical measures of functional integration in fMRI. *Med. Image Anal.*, 12:484–496, 2008.

G. Marrelec, J. Kim, J. Doyon, and B. Horwitz. Large-scale neural model validation of partial correlation analysis for effective connectivity investigation in functional mri. *Hum. Brain Mapp.*, 30:941–950, 2009. ISSN 1097-0193.

J. Mazziotta, A. Toga, A. C. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike, C. Holmes, D. L. Collins, P. Thompson, D. MacDonald, M. Iaconobi, T. Schormann, K. Amunts, N. Palomero-Gallagher, S. Geyer, L. Parsons, K. Narr, N. Kabani, G. Le Goualher, D. Boomsma, T. Cannon, R. Kawashima, and B. Mazoyer. A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM). *Philosophical Transactions of the Royal Society of London B Biological Sciences*, 356: 1293–1322, 2001.

T. Mitchell, R. Hutchinson, R. S. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman. Learning to decode cognitive states from brain images. *Machine Learning*, 57:145–175, 2004.

M. M. Monti, A. Vanhaudenhuyse, M. R. Coleman, M. Boly, J. D. Pickard, L. Tshibanda, A. M. Owen, and S. Laureys. Willful modulation of brain activity in disorders of consciousness. *N. Engl. J. Med.*, 362:579–589, 2010.

J. Mourão-Miranda, E. Reynaud, F. McGlone, G. Calvert, and M. Brammer. The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. *NeuroImage*, 33:1055–1065, 2006.

K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *TRENDS in Cognitive Sciences*, 10: 424–430, 2008.

S. Ogawa, T. Lee, A. Kay, and D. Tank. Brain magnetic resonance imaging with contrast dependent on bloodoxygenation. *Proc. NatI. Acad. Sci. USA*, 87:9868–9872, 1990.

R. Oldfield. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, 9:97–113, 1971.

G. Orrù, W. Pettersson-Yeo, A. F. Marquand, G. Sartori, and A. Mechelli. Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. *Neuroscience & Biobehavioral Reviews*, 36: 1140 – 1152, 2012. ISSN 0149-7634.

P. Peigneux, S. Laureys, S. Fuchs, A. Destrebecqz, F. Collette, X. Delbeuck, C. Phillips, J. Aerts, G. Del Fiore, C. Degueldre, A. Luxen, A. Cleeremans, and P. Maquet. Learned material content and acquisition level modulate cerebral reactivation during posttraining rapid-eye-movements sleep. *Neuroimage*, 20:125–134, 2003.

P. Peigneux, S. Laureys, S. Fuchs, F. Collette, F. Perrin, J. Reggers, C. Phillips, C. Degueldre, D. Del Fiore, J. Aerts, A. Luxen, and P. Maquet. Are spatial memories strengthened in the human hippocampus during slow wave sleep? *Neuron*, 44:535–545, 2004.

P. Peigneux, P. Orban, E. Balteau, C. Degueldre, A. Luxen, S. Laureys, and P. Maquet. Offline persistence of memory-related cerebral activity during active wakefulness. *PLoS Biol.*, 4:e100, 2006.

F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and fmri: a tutorial overview. *Neuroimage*, 45:S199–S209, 2009.

V. Perlbarg, P. Bellec, J.-L. Anton, M. Pélégrini-Issac, J. Doyon, and H. Benali. Corsica: correction of structured noise in fmri by automatic identification of ica components. *Magn. Reson. Imaging*, 25:35–46, 2007. ISSN 0730-725X.

V. Perlbarg, G. Marrelec, J. Doyon, M. Pélégrini-Issac, L. S, and H. Benali. NEDICA: Detection of group functional networks in fMRI using spatial independent component analysis. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI'08)*, 2008.

J. Peters, D. Janzing, and B. Schölkopf. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:2436–2450, 2011.

C. Phillips, M.-A. Bruno, P. Maquet, M. Boly, Q. Noirhomme, C. Schnakers, A. Vanhaudenhuyse, M. Bonjean, R. Hustinx, G. Moonen, A. Luxen, and S. Laureys. 'Relevance vector machine' consciousness classifier applied to cerebral metabolism of vegetative and locked-in patients. *Neuroimage*, 56:797–808, 2011.

J. C. Platt, N. Christiani, and J. Shawe-Taylor. Large margin DAGS for multiclass classification. In *Advances in Neural Information Processing Systems (NIPS 1999)*, volume 12, pages 547–553, 2000.

S. Polyn, V. Natu, J. Cohen, and K. Norman. Category-specific cortical activity precedes retrieval during memory search. *Science*, 310:1963–1966, 2005.

M. Raichle. Neuroscience. The brain's dark energy. *Science*, 314:1249–1250, 2006.

M. E. Raichle and M. A. Mintun. Brain work and brain imaging. *Annu. Rev. Neurosci.*, 29:449–476, 2006.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning.* the MIT Press, 2006.

R. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

J. Schrouff and C. Phillips. *Coma and altered states of conscious*, chapter Multivariate pattern recognition analysis: brain decoding, pages 41–51. Springer, 2012.

J. Schrouff, V. Perlbarg, M. Boly, G. Marrelec, P. Boveroux, A. Vanhaudenhuyse, M.-A. Bruno, S. Laureys, C. Phillips, M. Pélégrini-Issac, P. Maquet, and H. Benali. Brain functional integration decreases during propofol-induced loss of consciousness. *Neuroimage*, 57:198–205, 2011.

J. Schrouff, C. Kussé, L. Wehenkel, P. Maquet, and C. Phillips. Decoding semi-constrained brain activity from fMRI using Support Vector Machines and Gaussian Processes. *PLoS One*, 7:e35860, 2012.

J. Schrouff, M. J. Rosa, J. Rondina, A. Marquand, C. Chu, J. Ashburner, C. Phillips, J. Richiardi, and J. M. ao Miranda. PRoNTo: Pattern Recognition for Neuroimaging Toolbox. *Neuroinformatics*, pages 1–19, 2013. doi: 10.1007/s12021-013-9178-1. URL `http://dx.doi.org/10.1007/s12021-013-9178-1`.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.

S. Shinkareva, R. A. Mason, V. L. Malave, W. Wang, T. M. Mitchell, and M. A. Just. Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLoS one*, 3:3:e1394., 2008.

S. M. Smith, P. T. Fox, K. L. Miller, D. C. Glahn, P. M. Fox, C. E. Mackay, N. Filippini, K. E. Watkins, R. Toro, A. R. Laird, and C. F. Beckmann. Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, 106:13040–13045, 2009.

S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and W. Woolrich, Mark. Network modelling methods for fMRI. *Neuroimage*, 54:875–891, 2011.

A. H. Snijders, I. Leunissen, M. Bakker, S. Overeem, R. C. Helmich, B. R. Bloem, and I. Toni. Gait-related cerebral alterations in patients with Parkinson's disease with freezing of gait. *Brain*, 134(1):59–72, 2011.

M. Spiridon and N. Kanwisher. How distributed is visual category information in human occipito-temporal cortex? an fMRI study. *Neuron*, 35:1157–1165, 2002.

N. K. Squires, K. C. Squires, and S. A. Hillyard. Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalogr. Clin. Neurophysiol.*, 38:387–401, 1975.

R. Steer, R. Ball, W. Ranieri, and A. Beck. Further evidence for the construct validity of the Beck depression inventory-II with psychiatric outpatients. *Psychol. Rep.*, 80:443–446, 1997.

J. Talairach and N. Y. Tournoux, P. *Co-planar stereotaxic atlas of the human brain.* New York, 1988.

A. Tambini, N. Ketz, and L. Davachi. Enhanced brain correlations during rest are related to memory for recent experiences. *Neuron*, 65:280–290, 2010.

M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

P.-H. Tseng, I. Cameron, G. Pari, J. Reynolds, D. Munoz, and L. Itti. High-throughput classification of clinical populations from natural viewing eye movements. *J. Neurol.*, pages 1–10, 2012. ISSN 0340-5354.

N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 15:273–289, 2002.

P. E. Valk, D. L. Bailey, D. W. Townsend, and N. Maisey, Michael. *Positron Emission Tomography.* Springer, 2003.

P. Vemuri, J. L. Gunter, M. L. Senjem, J. L. Whitwell, K. Kantarci, D. S. Knopman, B. F. Boeve, R. C. Petersen, and C. R. Jack. Alzheimer's disease diagnosis in individual subjects using structural MR images: Validation studies. *NeuroImage*, 39:1186–1197, 2008.

M. Wilson and B. L. McNaughton. Reactivation of hippocampal ensemble memories during sleep. *Science*, 265:676–679, 1994.

K. K. Zakzanis, S. J. Graham, and Z. Campbell. A meta-analysis of structural and functional brain imaging in dementia of the Alzheimer's type: A neuroimaging profile. *Neuropsychology Review*, 13:1–18, 2003. ISSN 1040-7308.

Appendix A

Positron Emission Tomography

As mentioned in section 2.1.2, PET data images the metabolic activity of the tissues. In practice, the radiologist would search for hyper-metabolic regions if he suspects cancer, while he would look for hypo-metabolic regions when diagnosing dementias. In the present work, we would expect to find hypo-metabolic regions since we are dealing with Parkinson's Disease (PD).

A.1 Comparison with fMRI

While PET and fMRI both acquire metabolic changes induced by neuronal at the whole brain scale, they show differences in acquisition parameters, leading to differences in functional activation studies [Kinahan and Noll, 1999]. In Table A.1, we briefly compare the two modalities in terms of spatial and temporal resolution, duration, ease of use, invasivity, dimension and content of output images.

Table A.1: Brief comparison of the PET and fMRI acquisition technologies

	PET	**fMRI**
Spatial resolution	5-10 mm	2-4 mm
Temporal resolution	One image per injection	One image per TR (∼2s)
Duration of acquisition session	45 minutes or more (decay time)	usually 30 to 45 minutes
Constraints	need of cyclotron + production line for the radiotracer close by, or buying the radiotracer from a company (expensive!)	no metal in or around the subject
Invasivity	injection of tracer	-
Dimension	3D file per injection	3D file per TR
Content	regional glucose uptake	BOLD signal