A multi-source perspective on inter-subject learning. Contributions to neuroimaging.

Sylvain Takerkart

▶ **To cite this version:**

Sylvain Takerkart. A multi-source perspective on inter-subject learning. Contributions to neuroimaging.. Machine Learning [cs.LG]. Aix-Marseille Universite, 2015. English. NNT : . tel-01251384v3

HAL Id: tel-01251384
https://hal.science/tel-01251384v3

Submitted on 27 Jan 2016

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

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

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Aix-Marseille Université

Ecole Doctorale 184 Mathématiques et Informatique
Laboratoire d'Informatique Fondamentale, UMR 7279

Thèse présentée pour obtenir le grade universitaire de docteur
Discipline : informatique

Sylvain TAKERKART

A multi-source perspective on inter-subject learning Contributions to neuroimaging

Soutenue le 24/09/2015.

Rainer Goebel	Maastricht University	Rapporteur
Patrick Gallinari	Université Pierre et Marie Curie, Paris	Rapporteur
Jean-François Mangin	CEA	Examinateur
Bertrand Thirion	INRIA	Examinateur
Olivier Coulon	CNRS	Co-directeur de thèse
Liva Ralaivola	Aix-Marseille Université	Directeur de thèse

Cette oeuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 3.0 France.

Résumé

L'apprentissage inter-sujet intervient dans l'analyse des données enregistrées chez des sujets humains, lorsque le sujet chez lequel on doit faire une prédiction ne faisait pas partie de la base d'apprentissage. Le plus typique de ces problèmes est l'aide au diagnostic, lorsque on demande à un outil informatique si un sujet, inconnu jusque là, est sain ou malade. Dans cette thèse, nous défendons le point de vue que le problème d'apprentissage inter-sujet doit être formalisé comme un problème multi-source dans lequel chaque sujet de la base d'apprentissage fournit une source de données enregistrées dans un espace d'entrée potentiellement différent et qui sont des réalisations de distributions différentes. Le cadre multi-source est ainsi une généralisation du problème d'adaptation de domaine, dans lequel une seule source de données est disponible. Nous présentons ensuite trois contributions motivées par des problèmes d'apprentissage inter-sujet en neuroimagerie.

Le résultat de notre première contribution est une méthode qui permet de produire des prédictions inter-sujet sur des données d'IRM fonctionnelle en utilisant les patrons d'activation disponibles à des échelles spatiales relativement fines disponibles dans une région d'intérêt du cortex. Du à la forte variabilité fonctionnelle inter-sujet, les espaces d'entrée dans lesquels vivent ces patrons sont différents au travers des sujets. Notre contribution consiste à construire un espace commun pour tous les sujets en utilisant une représentation graphique des patrons d'activation ainsi qu'un noyau de graphe qui projette implicitement ces représentations dans un espace de hilbert à noyau reproduisant. Nous avons démontré l'efficacité de cette approche grâce à l'amélioration de la performance de classification dans un tâche de prédiction inter-sujet construite pour étudier l'organisation fonctionnelle du cortex auditif.

La deuxième contribution présentée dans cette thèse est une nouvelle méthode qui permet l'identification de différences de formes locales du cortex entre plusieurs groupes d'observations. Les objets utilisés sont, une fois de plus, des représentations graphiques, cette fois construites à partir des points correspondant à des extrema de profondeur des sillons corticaux. L'utilisation d'un noyau de graphe adapté à ces objets permet, dans l' espace de hilbert à noyau reproduisant correspondant, de quantifier les différences entre groupes d'observations par la performance d'un classifieur entraîné à reconnaître ces groupes. Une méthode d'inférence spatial non paramétrique permet ensuite la détection, c'est

à dire l'identification des zones du cortex qui présentent des différences significatives. Nous validons cette méthode en démontrant qu'elle permet d'identifier, sur une large population de sujets sains, des asymétries corticales ainsi que des différences inter-sexe.

La troisième contribution est une méthode d'adaptation de domaine pour le cas multi-source. Notre méthode se base sur le kernel mean matching, une procédure d'appariement de distributions qui adapte la distribution de l'ensemble d'entrainement à celle de l'ensemble de test par une pondération des exemples d'apprentissage. Nous décrivons une extension du kernel mean matching au cas où l'ensemble d'apprentissage se compose de plusieurs sources de données. Nous présentons des résultats préliminaires sur une tâche de classification inter-sujet dans une expérience de magnéto-encéphalographie.

Mots clés : apprentissage multi-source, méthodes à noyau, classification, neuroimagerie.

Abstract

Inter-subject learning is a family of learning problems encountered in the analysis of data recorded in human subjects where we need to perform predictions on data recorded from a subject that was not available at training time. The most usual problem that uses inter-subject learning is to ask whether an unknown individual is healthy or sick, i.e to design a computer-aided diagnosis tool. In this thesis, we argue that such inter-subject learning questions should be addressed within the *multi-source learning* framework, and we formalize it as such in the context of neuroimaging studies. Indeed, each subject is a different source of data, with data samples that potentially live in different feature spaces and that are drawn from different probability distributions. The *multi-source* setting therefore constitutes an extenstion of the domain adaptation problem where a single source of training data is available. We then introduce three original contributions motivated by inter-subject learning questions in neuroimaging.

The result of our first contribution is a method that is able to perform reliable inter-subject predictions from fMRI data using fine-scale spatial patterns defined within a region of interest. Because of the strong inter-subject variability present at such fine scale, the original feature spaces are different across subjects. Our contribution consists in designing a common space for the patterns of all subjects using graphical representations of the patterns together with a graph kernel that implicitly projects the samples into a reproducing kernel hilbert space. We show that this approach is effective through the increased accuracy achieved on an inter-subject prediction task designed to study the functional organization of the human auditory cortex.

Our second contribution is a new method that enables to detect local differences in cortical shape across groups of anatomical MRI scans. The objects used to detect such differences are, yet again, graphical reprentations, this time designed from the spatial organization of the sulcal pits – the deepest points of cortical sulci. Using a graph kernel designed for these objects allows to project them into a reproducing kernel hilbert space and to quantify the differences between groups through the performances of a classifier trained to recognize these groups. A non-parametric spatial inference method is then proposed to perform the detection of cortical zones where the differences are statistically significant. We validate this method by showing that it detects cortical asymmetries and gender differences using a large database of healthy subjects.

The third contribution of this thesis is a multi-source domain adaptation technique. Our method builds upon the kernel mean matching, a distribution matching procedure that estimates importance weights for the training samples so that the weighted source distribution matches more closely the target distribution than the unweighted one. We introduce an extension of the kernel mean matching for the multi-source case, i.e when the training samples are drawn from several sources of data. We present preliminary results of this framework on a inter-subject prediction task used to analyse data from a magneto-encephalography experiment.

Keywords : multi-source learning, kernel methods, classification, neuroimaging.

Remerciements

I want to start by thanking Liva Ralaivola for participating into my adventurous exploration of real neuroimaging data before accepting to guide me through my – at least as much – adventurous desire to complete this doctorate. Thanks also to Olivier Coulon for sharing this supervision duty. You two were my friends before being my advisors and I hope that now you're not my advisors anymore, you'll remain my friends for a looooong time! Thanks a lot as well to my two reviewers, Patrick Gallinari, who had to endure all the neuroimaging jargon during this read, and Rainer Goebel who has trusted me in a lot of different ways since we first met in 2000! Thanks also to the last two members of my defense committee, Bertrand Thirion, who, beyond being a member of this committe, has regularly helped me with practical advice, and Jean-François Mangin for sharing numerous fun conversations in various circumstances.

A special thank you goes to Guillaume Auzias! This enterprise would probably have never reached its end without you, your continuous ad hoc criticism of my work and your helpful advice. In this critical time, I sincerely hope you will still be on the first floor of the INT next year!

Thanks also to the other contributors and co-authors of the work described in this document: Lucile Brun for a lot of pits fun, Hachem Kadri for a lot of kernel fun, Romain Trachel and Daniele Schön for lots of other types of fun.

Because, as I used to say, during this whole PhD endeavour, I also had a *real job*, I want to thank all the people at INCM and then INT for making this possible, and especially the INCM and INT directors Driss Boussaoud and Guillaume Masson.

I also want to add that my adoptive research team, Qarma, was very fun to be part of. Almost all of you brought some contributions to this work, whether practical or philosophical. I hope I convinced you that neuroscience is fun and that it needs machine learners to progress. Yes, neuroscience needs you!

Finally, last but not least, thank you Sandrine for being with me all these years and for bringing us all our extra S-esses.

Contents

1 Neuroimaging: a primer

Contents

1.1 Neuroimaging acquisition techniques

Neuroimaging is a subfield of medical imaging which focuses on producing and making use of images of the central nervous system (CNS). The objectives of neuroimaging comprise the diagnosis of pathologies related to the CNS as well as its monitoring and understanding through the interpretation of visual representations of the structure and function of the brain and the spinal cord. Several technologies are available to produce images of the CNS, each providing different types of information. In this introductory section, we provide an overview of the most common acquisition techniques used in neuroimaging. They comprise standard technologies used in medical imaging as well as specific techniques that have been developed to better characterize the soft tissues of the brain or their functional properties.

1.1.1 Computed Axial Tomography

Computed Tomography (CT scan) is one the the oldest techniques available, directly derived from traditional radiography. It uses a series of X-ray scans in order to produce a three-dimensional image of the head through a computational reconstruction process that solves the inverse Radon transform. A CT scan therefore estimates the amount of X-rays absorbed in a given location, which is related to the tissue density. Even if this technique does not provide state-of-the-art quality, it remains widely used in clinical setting because a scan can be performed in less than a minute. The main indications of CT scan include the preparation of surgeries and the diagnosis of brain injuries thanks to its ability to accurately detect and localize tissue swelling and bleeding.

1.1.2 Positron emission tomography

Positron emission tomography (PET) is a technique that uses an array of sensors to measure the emissions of positrons from a radioactive tracer that is injected into the body prior to the scan. A computational reconstruction allows to image the concentration of the tracer as a three dimensional volume, hence to detect the locations where the chemical compound has accumulated. Depending on the chosen tracer, PET scanning can therefore highlight different properties of the body parts to be examined. When it comes to studying the CNS, the most commonly used tracer is the Fludeoxyglucose (FDG); indeed, this radioactive form of glucose makes it possible to directly study the metabolism of the brain, i.e to quantitatively measure brain activity. As a functional imaging technique, its advantages include the image quality offered and the short time necessary for acquisition, but its main disadvantage lies in the fast speed of decay of the radioactive compound concentration which limits the field of application of PET to studying tasks that are accordingly short.

1.1.3 Magnetic resonance imaging

Magnetic resonance imaging (MRI) is a technique that is based on the use of a high and homogeneous magnetic field in the imaging device. The energy of temporary pulses of radio waves sent by an emitting coil to the patient excite the hydrogen atoms (protons) in the target tissue, which emit themselves radio frequencies recorded during relaxation by a receiver coil. A modulation of the main magnetic field by gradient coils on the emission side allows to encode the position of the target tissue, which makes it possible to reconstruct an image. Because the water content of different types of tissues varies, the recorded magnetic resonance signal produced by the water protons will also change, which makes it possible to produce images with strong constrasts between different tissues. MRI has the nice advantages of being non-invasive (i.e it is possible to

do MRI without injecting any tracer) and avoiding exposition to X-rays. The design of pulse sequences, which define the successive changes on the operation of the gradient coils, leads to different image properties. We will now describe the main types of MR images used in neuroscience.

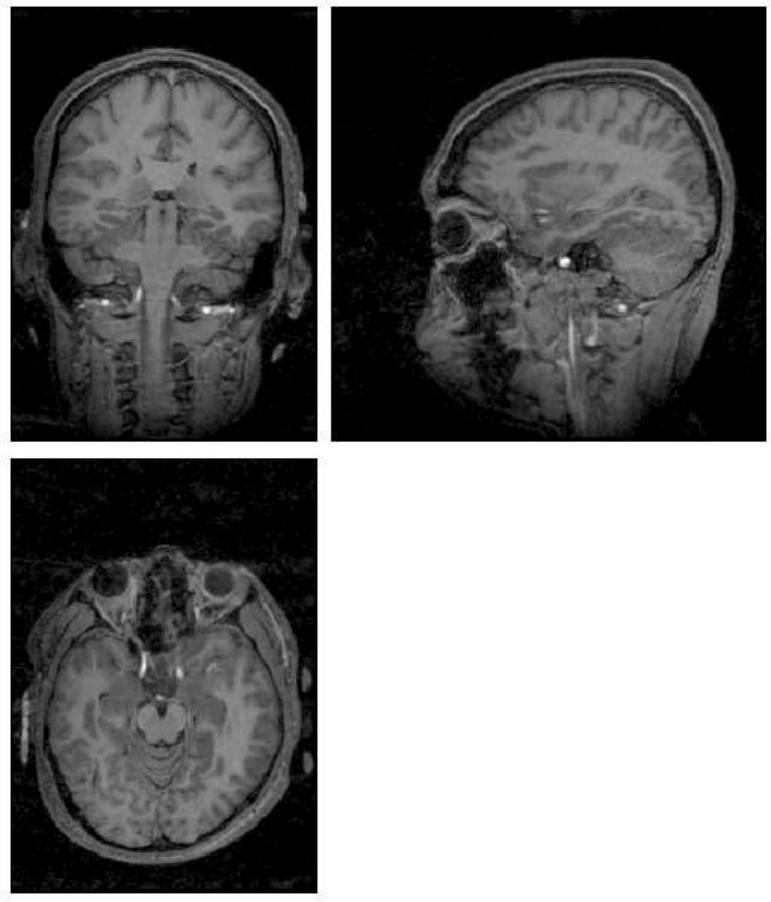

Figure 1.1: A 3D volume of anatomical MRI, represented by three slices cut through three orthogonal planes. One can directly observe the anatomy of the brain. The circumvolutions of the grey matter appear in grey, and the white matter, surrounded by grey matter, in white.

Anatomical MRI (also called structural MRI) uses pulse sequences that produce images where the structures of the brain and spinal cord are highly contrasted. This is the tool of choice to study the morphology of the brain in a quantitative manner, i.e to perform morphometry studies. The most usual pulse sequences aim at measuring the difference in T_1 relaxation time (spin-lattice relaxation time) between tissues, and in particular between the grey and white matter. See Fig. 1.1 for an example.

Functional MRI (fMRI) measures the so-called Blood-Oxygen-Level Dependent

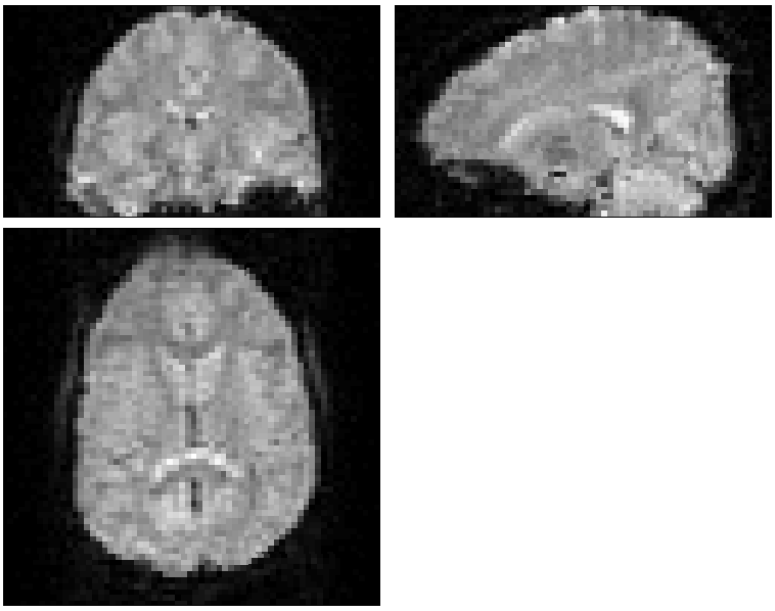

Figure 1.2: A 3D volume of fMRI data. A full fMRI dataset is composed of a timeseries of such volumes.

(BOLD) effect. A local increase in neural activity demands energy consumption which requires oxygen; this oxygen demand is actually over-compensated, which result in a increased concentration of oxy-hemoglobin compared to deoxy-hemoglobin, which results in an increased MR signal measured with the T_2^* relaxation time contrast. Functional MRI therefore measures an indirect signature of neural activity – the precise links between neural activity and the fMRI signal remaining to be elucidated. Since its discovery in the 1990s, fMRI has become widely used to study brain function, mostly because it provides very good spatial resolution over the whole brain in a non-invasive manner, which made it contribute significantely to the *brain mapping* field. See an example of an fMRI volume on Fig. 1.2.

1.1.4 Electroencephalography and magnetoencephalography

Electroencephalography (EEG) and magnetoencephalography (MEG) respectively measure the electrical and magnetic field over a set of electrodes positioned on or over the scalp. The signals recorded are induced by synchronized neuronal electrical currents over large populations of neurons which share the same orientation, thus creating local modulations of the electrical and magnetic fields, large enough to be detected by scalp electrodes. While EEG is a very old tool with origins dating from the XIX-th century, MEG is more recent since it was

developed in the 1960s. Both EEG and MEG provide a very high temporal resolution (on the order of a few milliseconds), which make them very useful in research settings, in particular to study the oscillatory behaviors of neuronal activity. Moreover, EEG is a standard tool in clinical settings, where it can help characterize epileptic seizures, diagnose psychiatric disorders or prognosticate the evolution of comatose patients. Although some research results are encouraging for a future implantation of MEG in hospitals, it is not, as of today, an approved tool for clinical applications.

1.2 Inference in neuroimaging

The main objectives of neuroimaging as a field of medical imaging can be categorized as follows:

- in clinical settings, the goal of neuroimaging is to help to decide whether a patient carries a neurological or psychiatric disease, or to predict his/her evolution with regard to such pathologies;
- in research settings, functional neuroimaging attempts to understand brain function in its normal healthy state;
- in pharmacology, the effectiveness of a drug can be quantified by examining its spread throughout the brain or its effect on the modulation of brain processes thanks to neuroimaging;
- finally, another objective is to build databases to describe what is the normal CNS.

Overall, reaching any of these four objectives requires to examine groups of individuals and solve two main questions that consist in:

- finding commonalities across subjects within a population;
- finding differences between subjects belonging to different populations.

These two main questions can be addressed using univariate or multivariate statistical tools in different ways that we describe in the following section. In short, univariate methods examine a single voxel[a] of the images at a time, before applying the same analysis model repeatedly and independently at each pixel. In contrast, multivariate methods consider groups of pixels – or even all the pixels – in a single model.

[a] a *voxel* – a volume element – is the equivalent of a pixel in volumetric imaging

1.3 Univariate techniques

1.3.1 The General Linear Model

The tool of choice for building univariate methods in neuroimaging is the so-called General Linear Model (GLM). In neuroimaging, it is used as follows:

$$Y = X\beta + \epsilon, \tag{1.1}$$

where

- Y is a vector of length n which contains n data points recorded at a given location, being a pixel, voxel or single electrode;
- X is the so-called design matrix, of size $n \times m$, which is composed of m regressors that each contains a variable that we believe should contribute to explain Y; it is to be specified by the experimenter;
- β is a weight vector of size m, the i-th value weighting the i-th regressor of X; it is the vector that needs to be estimated;
- ϵ is the residual vector of size n, which contains everything in Y that cannot be explained by X.

An example of application of the GLM is shown on Fig. 1.3. This model is tagged as *general* because it comprises several classical statistical models such as the simple linear regression, the multiple linear regression and the analysis of variance (ANOVA). It has been massively used in neuroimaging, mostly because of the success of the SPM software[b] (see [Ashburner 2012] for a historical perspective on SPM). SPM, which stands for *Statistical Parametric Mapping* had first implemented the GLM for PET data, before making it available for fMRI and aMRI. It falls into the realm of *massively univariate* methods. Indeed, because it works on data from a single voxel, the model attempts to explain the behavior of a single variable, hence the use of the *univariate* term. The *massive* term follows the repeated use of the same model (i.e the same design matrix X) on the very large number of voxels available in neuroimaging datasets.

In practice, the following steps are used to perform a GLM analysis:

- at each voxel v, fit the model in order to obtain an estimate of the β vector, denoted $\hat{\beta}^v$;
- interrogate the model at each voxel v using contrasts: define the null hypothesis $c^T\hat{\beta}^v = 0$ for a contrast c (see below for details); perform hypothesis testing using t or F tests;

[b] http://www.fil.ion.ucl.ac.uk/spm/

- obtain a statistical parametric map that covers the brain with t or F values, with their associated map of p-values;
- perform inference on this statistical map to detect *where* the null hypothesis can be rejected, including corrections for the multiple comparison problem (see below).

A contrast c is a vector or a matrix of weights that are applied to the parameters β of the model. The most simple contrast is a vector $c = [0 \cdots 1 \cdots 0]$, where only the i-th weight is non zero and equal to one. In this case, $c^T\beta = \beta_i$, and the null hypothesis is simply $\beta_i = 0$. When this null hypothesis is rejected, it means that the i-th regressor of the design matrix X actually contributes to explaining the data Y. For instance, it Y contains one data point per subject and X_i is the age of each subject, the rejection of this null hypothesis gets interpreted as the fact that *age has a significant effect on explaining* Y. In general, when c is a vector, the null hypothesis is a linear combination of the different β_i-s and the associated statistical test is a Student t-test; when c is a matrix, thus testing for different linear combinations of β_i-s at the same time, the associated statistics if Fischer's F.

Another important point lies in the fact that the application of the same model at all voxels implies performing a number of tests equal to the number of voxels, which can be on the order of 10^5 to 10^6 depending on the modality. Even in the scenario that the null hypothesis is true everywhere, this will produce a large number of voxels that will pass the test defined by $p < 0.05$. We therefore need to correct for this effect, which is called the multiple comparison problem. The most simple technique for voxel-wise inference is the Bonferroni procedure to control the family-wise error rate: the critical p-value is simply divided by the number of tests, which increases the threshold on the statistic. But it is known to lack power. A standard strategy consists in examining clusters of suprathreshold voxels (for a given fixed threshold that can be informed by uncorrected pointwise p-values), and performing statistical assessment on the clusters, which vastly reduced the number of tests. This can be done using the Random Field Theory [Worsley et al. 1992] which parametrically linked point-wise statistics with the expected size of suprathreshold clusters using smoothness assumptions on the statistical map. Besides these parametric approaches, one can also resort to non-parametric strategies for either voxel-wise or cluster-wise inference, for instance using permutation-based approaches [Bullmore et al. 1999; Nichols et al. 2002a].

We will now describe the implementations of this General Linear Model to address the two most common examples of the two types of questions that were described in Section 1.2.

1.3.2 Group analysis in fMRI

Functional MRI experiments consist in having the subject sequentially perform several repetitions of one or several tasks while lying in the scanner, in a predetermined manner that defines the *experimental paradigm*. During the several minutes of the experiment, fMRI volumes are acquired continuously, with typically one volume every 2 to 3 seconds, to form a 3D+time dataset. Most often, the same experiment is performed on several subjects, and the objective of a *group* analysis is to find significant effects, i.e locations for which we can reject the null hypothesis for a contrast of interest, that are common across the population. In this case, we use a two-level GLM.

The first level is the *subject* level. The fMRI data of each subject s is composed of timeseries available for each voxel v of the brain, that we define as $Y^{v,s}(t)$. Some parts of the brain will be *activated* by the experimental paradigm and the BOLD response should then *correlate* with the paradigm, which we therefore use to define the design matrix X. If the subject is asked to alternatively performs several tasks – or several variants of the same task, the timeseries of each task/variant will be included as a regressor in the design matrix and the GLM will implement a multiple regression: $Y^{v,s}(t) = X\beta^{v,s} + \epsilon^{v,s}$ is estimated at each voxel v and for each subject s. The *subject-level* contrast maps $c^{v,s} = c^T\hat{\beta^{v,s}}$ are then computed for a given contrast c (where c can for example implement the null hypothesis that two of the tasks produce the same BOLD response). This first level GLM is illustrated in details on Figs 1.3 and 1.4.

Then, a second level GLM is estimated at the *group* level, where the data Y_g^v at voxel v contains s data points which are the contrast values $c^{v,s}$ estimated on each subject with the first level GLM [c]. The most simple question that can be asked at the group level is to determine where in the brain the contrast values have a non-null value over the population. This can be done by including a constant regressor with a one value in the group-level design matrix X_g. The model that we estimate at each voxel v is then $Y_g^v = X_g\beta^v + \epsilon$, which in fact implements a one-sample t-test. The resulting t map can then be processed by the spatial inference described previously in order to deal with the multiple comparisons problem. The clusters that will survive are locations where the contrast c is significantely non-null over the population, thus answering our initial problem.

1.3.3 Voxel-based morphometry

When processing anatomical MR data, traditional morphometry consists in measuring – often manually – the volume of a given brain structure and performing statistical analysis to estimate the potential differences between subjects belonging to different populations: a simple two-sample t-test can then be used to

[c] Note that this requires that the voxel numbered v designates the same brain location for all subjects, which is achieved by a processed called *spatial normalization*

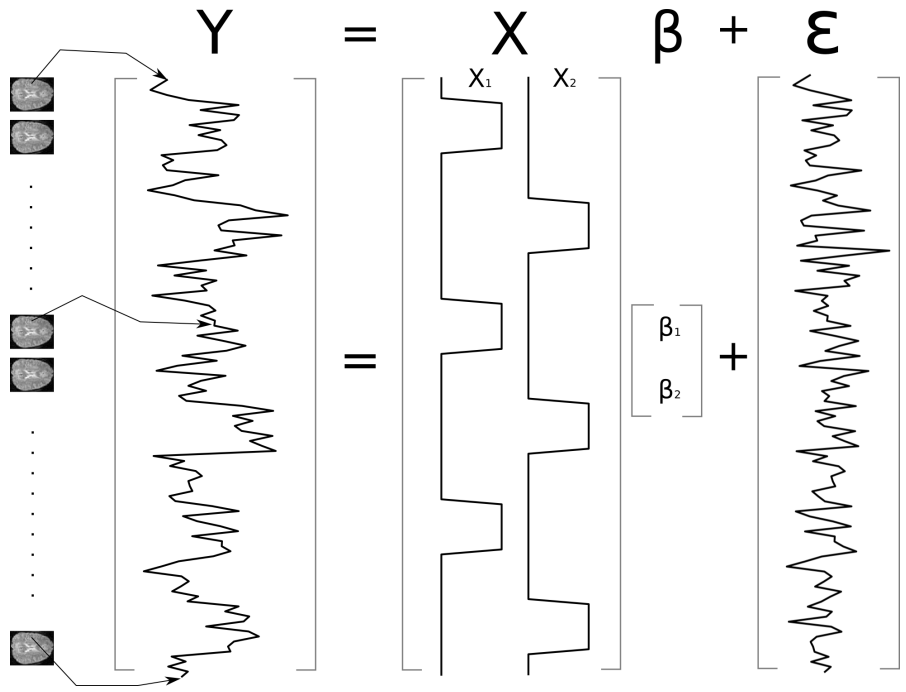

Figure 1.3: Illustration of the subject level GLM applied on one voxel of an fMRI 3D+time dataset. In this case, the experimental paradigm – encoded in the design matrix X – simply includes two types of experimental trials. For instance, when $X_1 = 1$, the subject is passively looking at a *face* presented on the screen and when $X_2 = 1$, he/she is looking at a *scrambled image*; when $X_1 = 0$ and $X_2 = 0$, the screen is empty. This paradigm is similar to the one of the MEG dataset used in Chapter 6

assess the difference in volume between healthy controls and patients. For instance, the volume of the hippocampus is known to be smaller in patients with Alzheimer's disease than in healthy subjects [Schott et al. 2003].

In order to detect morphological differences smaller than with region-based approaches, the *voxel-based morphometry* framework (see [Ashburner 2009] for a review) starts by estimating the density of gray matter $Y^v(s)$ at each voxel v of the brain of each subject s, after spatial normalization. By defining the vector Y^v containing the density values at voxel v for all subjects , the GLM $Y^v = X\beta^v + \epsilon^v$ can be used to assess differences between populations. In order to do so, X needs to encode the fact that our set of observations comprises two populations (for instance, with one regressor that takes the 1 value for subjects belonging to the first population and -1 for subjects of the other), and a contrast c should be defined to test the null hypothesis that there are no differences between the two populations. In this case, we will obtain a map of t values that will then be processed as previously to determine in which locations of the brain the den-

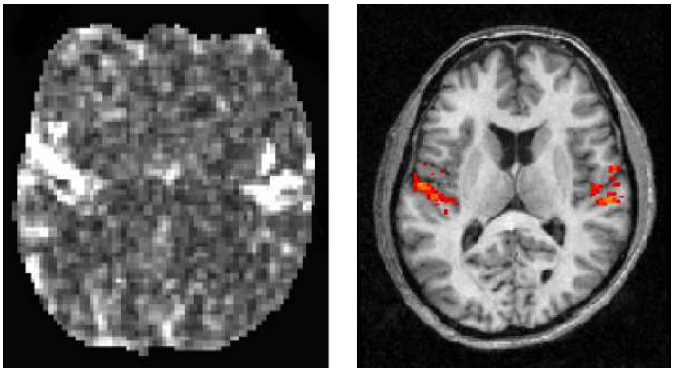

Figure 1.4: Once the first level GLM is estimated on all voxels of the fMRI 3D+time data, one can compute a contrast map; the raw contrast map presented on the left corresponds to a F-test showing activation in any of the conditions of the paradigm used in Chapter 4. After thresholding and estimation of the corrected p-value of each cluster, the leftover significant clusters are resampled to match the resolution of the anatomical MRI, resulting in the overlay presented on the right. Activation is present bilaterally in the auditory cortex, as sought after with this paradigm.

sity of gray matter does not verify the null hypothesis of no differences across populations.

1.4 Multivariate machine learning techniques

1.4.1 General setting: supervised learning

The goal of *supervised learning* is to learn a function that expresses as explicitely as possible the relationships between two spaces, an input space $\mathcal{X}$ and a target space $\mathcal{Y}$. When $\mathcal{Y}$ is a discrete set such as $\{1, \ldots, C\}$, this problem is known as *classification*; when $\mathcal{Y}$ is continuous, for instance when $\mathcal{Y} = \mathbb{R}$, it is known as *regression*.

A pair (X, Y) is a random variable of $\mathcal{X} \times \mathcal{Y}$ that follows an unknown, but fixed, joint probability distribution $\mathcal{P}$. We dispose of a finite set $\mathcal{D}$ of size N, composed of labeled examples $\{(x_n, y_n)\}_{n=1}^{N}$ independently drawn from $\mathcal{P}$ and the objective is to *learn* a function $f : \mathcal{X} \to \mathcal{Y}$ that allows to make accurate *predictions* of the y value associated with an example x.

In order to quantify the quality of the predictions, we use a *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ to compute the discrepancy between the true label y and the predicted label $f(x)$. Given such function, *learning* the optimal prediction

function f consists in minimizing the true risk

$$R(f) = \mathbb{E}_{(x,y)\sim P}(\ell(f(x), y)).$$

In practice, given the labeled dataset $\mathcal{D}$, we will attempt to minimize the *empirical risk*

$$R_S(f) = \frac{1}{N}\sum_{n=1}^{N}(\ell(f(x_n), y_n))$$

to find a function f within a restricted class of functions F.

1.4.2 Multi-Voxel Pattern Analysis of functional MRI data

Multi-Voxel Pattern Analysis (MVPA) is a recently introduced analysis technique for fMRI data, which is based on supervised learning. Let us consider an fMRI experiment where several tasks were performed (each task being labeled with a discrete value of Y), a classifier is learned from a set of labeled examples $\{(x_n, y_n)\}_{n=1}^{n=N}$, where the examples x_n are elements of $\mathbb{R}^d$ representing a $pattern$ of activation recorded over a set of d voxels, usually located within a given region of the brain. Besides the original MVPA name that was introduced with the release of the Princeton MVPA Toolbox [Polyn et al. 2005], this framework also takes different appellations: i) because the role of the paradigm (X in the GLM and Y for MVPA) and the data (Y in the GLM and X with MVPA) are inverted, it is sometimes refered to as *reverse inference*; ii) because the classifier attempts to guess what the subject was doing (which task $y \in \{1, \ldots, C\}$ he/she was performing) from a brain recording x_n, it is also called *brain decoding* analysis.

The basic type of inference performed with this model is then of the following type: if the classifier learned in such manner is able to provide predictions with an accuracy significantely above chance level, it means that the set of d voxels used to construct the input patterns carries information about the way the brain differentially performs the tasks $\{1, \ldots, C\}$. Provided with a group of subjects that performed the same experiment, this is often assessed in several steps.

- Perform cross-validation to assess the model accuracy individually on each subject. This consists in splitting a fully labeled dataset in a number of subsets, and repeat the following operations on each subset: i) data from this subset form the *test* set; ii) data from all other subsets form the *training* set; iii) learn the classifier on data from the training set; iv) compute its empirical accuracy on the test set.

- Use a statistical model to assess the null hypothesis of no differences between tasks, for each subject. A well suited method for this is to used permutation-based non parametric statistics [Nichols et al. 2002b]. Under this null hypothesis, the labels of the examples carry no information and

we could shuffle them without any effect on the classification accuracy. We can therefore estimate the distribution of the classification accuracy under this null hypothesis by performing a – large – number of cross-validations using each time a different set of randomly permuted labels. This enables to compute the p-value that the classification accuracy obtained with the true labels actually follows this null distribution, and to reject the null hypothesis if it is too weak (typically if $p < 0.05$).

- The group level analysis then looks at the consistency of within-subject results across the population.

In the univariate GLM framework, this question of whether the brain activation is different across tasks would have been asked independently at each voxel using a contrast on the β estimates. Because with MVPA, the same question is asked using a set of d voxels – with $d > 1$, it is often thought that the multivariate nature of MVPA provides additional statistical power compared to the GLM. However, the interpretation of MVPA results may reveal to be more challenging than with the GLM because it is not straightforward to know which of the d voxels are directly involved in the processing of each of the tasks. MVPA therefore does not offer the direct possibility to perform *brain mapping*, in the sense that it does not provide a statistical map of the whole brain that makes it possible to visualize results in a glimpse, a process that neuroscientist are used to since the GLM has become a *de facto* standard.

In order to overcome this limitation, it suffices to implement a sliding window strategy: the MVPA analysis and inference described above are performed in a small contiguous region centered around a given point of the brain, and repeated many times by changing the center point so that it browses the full brain. This is the so-called *searchlight* strategy , which has received a lot of attention since its introduction by [Kriegeskorte et al. 2006], because it can be thought as offering the best of both worlds: the easy mapping capability of the GLM and the statistical power of MVPA.

1.4.3 Computer-aided diagnosis tools for aMRI

The most classical question that can be addressed in a machine learning setting using anatomical MRI is to design a computerized tool to help clinicians refine their diagosis, i.e to develop a *computer-aided diagnosis* (CAD) tool. This problem is naturally addressed as a supervised learning question and requires to go through the following steps.

- Build a database gathering a large number of labeled T1 MR images belonging to two populations: i) healthy control subjects and ii) diseased patients suffering from a given neurological disorder.

- Extract a set of relevant characteristics from the images that can be used to differentiate the patients from the healthy subjects
- Learn a classifier using the labeled database as the training dataset and the extracted feature set.

Then, the clinician can use this classifier on a previously unseen incoming patient to help with the diagnosis process. Note that each of the three aforementionned steps is critical when it comes to providing a tool that will be accurate, hence useful to the clinicians.

- The database construction needs some special care in order to be useful for the considered diagnosis task. First, the sample size needs to be large enough to represent both populations and their statistical distributions. Second, the acquisition scheme should be standardized as much as possible. Ideally, all images should be acquired with the same device, because the effect of the scanning device can be far larger than the effect of the disease itself (as we have shown in [Auzias, Breuil, et al. 2014]). However, single-scanner studies are often limited in terms of sample size; in the case of multi-site studies (i.e when the database is composed of images acquired at different scanning site), special care should be taken to ensure that the same pulse sequence is used, as done for example in the Alzheimer's Disease Neuroimaging Initiative (ADNI, see [Mueller et al. 2005]).
- The feature extraction, which often relies on automatic image processing tools, should be thoroughly validated. Indeed, it can be an important source of noise if the automatic algorithms fail or yield erroneous measurements. In this case, a manual intervention can be benefitial (as it was performed in our study of sulcal anatomy in autism: [Auzias, Viellard, et al. 2014]), but it remains costly in time and not tractable for very large sample sizes.
- Although the classification method seems to have less importance for a given feature set [Sabuncu, Konukoglu, et al. 2015], it remains that the use of complex, potentiall more inforative, feature sets can ask for the design of a dedicated classifier, as for instance with structured objects.

To conclude, the most prominent applications of CAD tools comprise the study of white matter lesions in multiple sclerosis [Bilello et al. 2013], as well as the early diagnosis of Alzheimer's disease, for which it has been shown that such tool can perform reliable predictions before a standard clinical diagnosis is possible [Frisoni et al. 2010].

2 Inter-subject learning as a multi-source problem

Contents

When dealing with both of the problems described above (finding invariants across a population of subjects, or finding differences across several groups of subjects), and regardless of the type of neuroimaging modalities studied, we face a major challenge in the inter-subject variability. Our proposal in this thesis is to use the *machine learning* setting known as *multi-source learning* in order to handle this variability, by stating that each subject is a different *source* of data.

In this chapter, we first describe the standard *multi-source* setting (Section 2.1). We then frame the problem of learning with data from multiple subjects in neuroimaging as a multi-source learning question (Section 2.2).

2.1 Multi-source learning

2.1.1 Multi-source setting

Let us first define the multi-source setting encountered in learning problems. Such problem arises when one has data available that come from different *sources*, the word source being here employed in its most ordinary signification. Most often, it is the process that generated the data that differs across sources. A typical example occurs in image categorization, where one might have different types of images available, for instance drawings and photographs. It is intuitive that all the drawings share some common characteristics that are different from the

photographs and that both types of images are informative for the common task of image categorization. Those can therefore be modeled as different *sources* of data. In a more technical sense, this means that the data points – here the images – are drawn from different distributions, and potentially live in different feature spaces. The goal of *multi-source learning* is to agglomerate the information provided by each source for the task to be solved in order to be able to generalize to yet another source of data that was not available during training, for instance paintings in our image categorization example.

We can then formalize the definition of a multi-source dataset by first denoting $\mathcal{S} = \{1, \cdots, S\}$ the set of sources. Then, for each source $s \in \mathcal{S}$, we have some data available $\{x_n^s\}_{n=1}^{n=N^s}$, where x_n^s are elements of a feature space $\mathcal{X}^s$ and N^s is the number of examples for this s-th source. These examples are associated with a variable $y_n^s \in \mathbb{R}$, thus forming labeled examples (x_n^s, y_n^s). We define $\mathcal{D}^s = \{(x_n^s, y_n^s)\}_{n=1}^{n=N^s}$ the training data for source s.

The goal of multi-source learning is to estimate a function that is able to perform accurate prediction of the value of the variable y^t from an observation x^t drawn in a space $\mathcal{X}^t$ with a different distribution than the S sources previously described, i.e a function that can generalize to a target domain $\mathcal{D}^t$.

In most cases, one suppose that all the spaces are $\mathcal{X}^1, \cdots, \mathcal{X}^S, \mathcal{X}^t$ are identical, and the problem is then known as multi-source domain adaptation. If the input spaces are actually different, one usually attempts to find a set of transformations $\{\mathcal{R}^i : \mathcal{X}^i \to \mathcal{X}\}_{i \in \{1, \cdots, S, t\}}$ that bring all observations into a common representation space $\mathcal{X}$. We are then brought back to a multi-source domain adaptation question.

2.1.2 Link with multi-view and multi-task learning

In multi-view learning, the training dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{n=N}$ is characterized by the fact that each example x_n is composed of several *views*, i.e $x_n = (x_n^1, \cdots, x_n^S)$, where each of the x_n^s lives in a different feature space $\mathcal{X}^s$. A survey on multi-view learning can be found in [S. Sun 2013]. In some cases, not all views are available for all examples. We then talk about *missing views*. A multi-source problem can then be framed as a multi-view question where for all the examples, only one view is available and all the other ones are missing. The goal is then to be able to learn a predictor for examples described by yet another view. In practice, this model is too complex and it has not been implemented in the literature.

In multi-task learning, the dataset available at training time is $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{n=N}$, and this time, it is the output variable that does not take the usual form of a single scalar: in fact, each y_n is a multi-dimensional vector $\{y_n^1, \cdots, y_n^S\}$, each coding for a given task. A multi-source learning question can be modeled as a multi-task problem for which only one task variable y_n^t is available for each of the samples x_n: the s-th source is then composed of the samples for which the s-th y value is available. The different tasks are then the same, but because the distribution of

each data source is different, the conditional distributions, of x given y^s will also be different, which makes the multi-task model relevant. We will see later in the literature review that this model has been used in several methods.

2.2 A multi-source setting for inter-subject prediction

Here, we describe in details a generic model that frames the different inference problems encountered in neuroimaging within a common machine learning setting. In short, when dealing with data from multiple subjects, we argue that each subject provides a *source* of data, and that the different inter-subject learning questions – described in Section 1.2, which all aim at performing predictions on data from new subjects, can be embedded within a multi-source learning setting

2.2.1 Dataset and probabilistic model

First, we denote $\mathcal{S} = \{1, \cdots, S\}$ the set of subjects, where S is the number of subjects available at training time;

Then, for each subject $s \in \mathcal{S}$

- a variable $z^s \in \mathbb{R}$ is available that characterizes subject s; for instance, if several populations of subjects are present in $\mathcal{S}$ – for instance some patients that we would like to differentiate from healthy subjects, we might encode this information into the z^s variable by giving it a specific categorical value for each population, i.e $z^s \in \mathcal{G} \doteq \{1, \ldots, G\}$
- we have some data samples at hand $\{x_n^s\}_{n=1}^{n=N^s}$, where x_n^s are elements of a feature space $\mathcal{X}^s$ and N^s is the number of examples available for subject s
- these examples might be associated with a variable $y_n^s \in \mathbb{R}$, thus forming a labeled example (x_n^s, y_n^s);
- we note $\mathcal{D}^s = \{z^s, (x_n^s, y_n^s)\}_{n=1}^{n=N^s}$ the training data for subject s.

The full training set is then defined as

$$\mathcal{D} \doteq \cup_{s=1}^{S} \mathcal{D}^s.$$

In addition to the training data, there exist a dataset $\mathcal{D}^t$ of the same nature for a test subject t not in $\mathcal{S}$

$$\mathcal{D}^t = \{z^t, (x_n^t, y_n^t)\}_{n=1}^{n=N^t},$$

except the labels z^t and $\{y_n^t\}_{n=1}^{n=N^t}$ are not observed.

We then define a *hierarchical* probabilistic setting that may be associated with the generation of this data. There is an unknown and fixed distribution $\mathscr{L}$ that governs the distribution of subjects within the global population. Each realisation of the law is a subject s, i.e a set $\mathcal{D}^s = \{z^s, (x_n^s, y_n^s)\}_{n=1}^{n=N^s}$. The pairs (x_n^s, y_n^s) are themselves realisations of a law $\mathcal{L}^s$, which is different for each subject. Also note that the conditional law $\mathscr{L}_{|z}$ describes the distribution of the subjects of the population within several sub-groups, if these sub-groups are labeled by the z variable. This model is illustrated on Fig. 2.1.

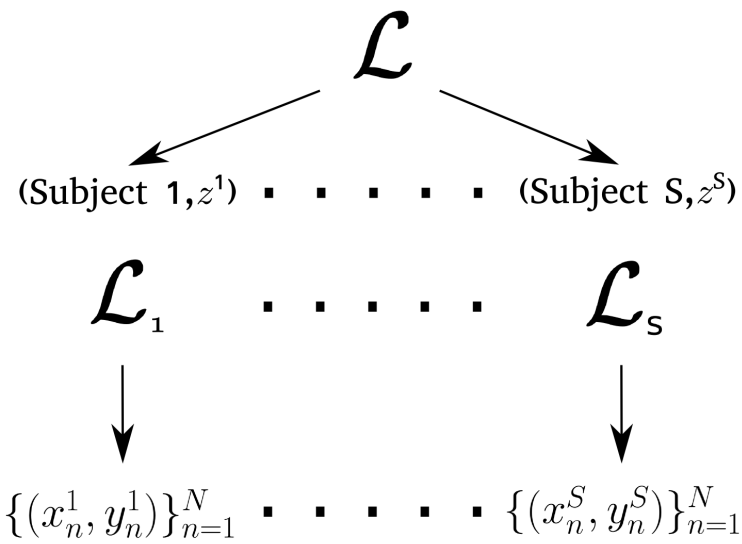

Figure 2.1: Illustration of the generative multi-source model for neuroimaging datasets. The subjects are random realizations of a fixed law. The data points are random realizations of subject-specific laws.

Within this framework, an *inter-subject prediction* problem can be formalized as a *multi-source learning* question for which each subject provides a *source* of data and we aim at performing predictions on data from a source, i.e a subject, not available during training. It is essential to understand that the *multi-source* nature of the probelm stems from i) s being an independent random variable of law $\mathscr{L}$, ii) the data from subject s living an input space $\mathcal{X}^s$ and being drawn from a distribution $\mathcal{L}^s$ which can be different for each subject.

2.2.2 Addressed problems

With such a setting, we can address two problems that each belongs to the two main categories of group analyses performed in neuroimaging (see Section 1.2).

2.2.2.1 Inter-subject decoding

The first problem consists in finding commonalities across subjects within a given population by performing multivoxel pattern analysis in a functional neuroimaging experiment: we want to be able to guess the value of the variable y (for instance the task that was performed by the subject) from its brain activation pattern, typically measured with fMRI, MEG or EEG. To find what is common across subjects using such decoding approach, the most common approach consists in independently performing decoding within each subject's dataset, and, in a later stage, examining the commonalities of the decoding performances across subjects (see Section 1.4.2). However, we claim that it is more informative to perform a single-stage analysis through *inter-subject decoding*, i.e performing predictions on a pattern recorded in a subject that was not part of the training dataset. Indeed, obtaining good *inter-subject decoding* performances implies that the patterns are both informative – for the considered learning task – and reproducible across subjects. This then shows that some subject-invariant neural coding principles have been identified, at least at the spatial scale provided by the acquisition device.

This *inter-subject decoding* problem can be defined as an instance of our *multi-source* setting. Before going into the details, note that we here assume that the population from which the subjects are drawn is homogeneous, i.e that we do not have any knowledge about the potential existence of different groups within this population. As a consequence, z^s takes a constant value for all available subjects s. We can therefore drop the z^s variable in this section.

We assume we have at hand several examples for each subject (i.e $\forall s, N^s > 1$, and hopefully $N^s \gg 1$) and all the examples x_n^s of the training subjects are labeled with a categorical variable, i.e $y_n^s \in \mathcal{C} \doteq \{1, \ldots, C\}$, where C is the number of classes. Our training set is therefore

$$D \doteq \cup_{s=1}^{S} \{(x_n^s, y_n^s)\}_{n=1}^{n=N^s}$$

and the goal of *inter-subject decoding* is to be able to perform predictions on the test dataset $\mathcal{D}^t = \{x_n^t\}_{n=1}^{n=N^t}$. This can be formulated as

1. computing the targets $\{y_n^t\}_{n=1}^{n=N^t}$ associated with the data from D^t;
2. and/or *learning* from D a predictor $f : \mathcal{X}^t \to \mathcal{C}$ with risk

$$R(f) \doteq P_{(x,y)\sim\mathcal{L}^t}(f(x) \neq y)$$

as small as possible.

From a machine learning point of view, the former problem is a problem of *transductive* learning, where the only concern is to compute the targets associated with the test data, without any consideration for the issue of predicting

the labels of new data coming from subject s. The latter problem is a supervised *inductive* learning problem, where the objective is to have at hand a predictor capable of reliably computing the targets associated to any data from any subject s, even if unseen at training time.

We will propose two solutions for this *inter-subject decoding* problem: an induction-based framework based on structural representations in Chapter 4 and a transductive domain adaptation approach in Chapter 6.

2.2.2.2 Group prediction

The second problem we can address consists in guessing within which population an unseen subject belongs to, i.e predicting the value of z^t for the test subject t. A typical example of such problem is solved when designing *computer-aided diagnostic* tools, which attempt to use imaging data to predict whether a patient is healthy or sick (in this simple case, *healthy* and *sick* define two populations of our set of subjects, and therefore correspond to two different values of z). In neuroimaging, this question is usually addressed using one observation per subject, most often anatomical MRI or PET, which means that $\forall s, N^s = 1$. In this case we drop the n index and denote x^s the observation for subject s; furthermore, this observation is of the same type for all subjects, i.e there is no y label further associated with x. Therefore, the training dataset is simply

$$\mathcal{D} \doteq \{(x^s, z^s)_{s=1}^{s=S}\}.$$

Our goal with performing *group prediction*, is to learn a predictor g of the class $z \in \{1, \ldots, G\}$ from the image x^t of a test subject, i.e $f : \mathcal{X}^t \to \mathcal{C}$ with risk

$$R(g) \doteq P_{(x,z))\sim\mathscr{L}}(f(x) \neq z)$$

as small as possible.

In Chapter 5, we will propose a framework that uses such *group prediction* task in order to detect local differences in cortical morphology between two populations of subjects.

Note that having only one observation per subject s prevents us from estimating the distribution $\mathcal{L}^s$. Data that could allow the use of probabilistic methods on anatomical MRI has only started to be acquired and made available very recently, as in [Maclaren et al. 2014] where each subject has been scanned a large number of times ($N^s = 40$).

3 State of the art

Contents

Performing *inter-subject* predictions is a *transfer learning* problem [Pan and Q. Yang 2010] where one attempts to transfer the information conveyed by the examples of each source of data available at training time (i.e each training subject) to the data of the target subject. One can consider two main cases in order to tackle this question. First, if the input spaces $\mathcal{X}^s$ are not identical across subjects, a necessary step consists in constructing *representations* to bring the observations of all sources into a common space. Second, if all the input spaces are identical, the major problem lies in handling the fact that the distributions $\mathcal{L}^s$ are likely to be different across sources, i.e from a given subject to another, which defines a *domain adaptation* problem. We will now examine the different approaches that exist in the litterature to deal with the construction of invariant representations (in Section 3.1), to perform adaptation, from one domain to another

(in Section 3.2) and using multiple input sources (in Section 3.3). Finally, we will examine a few other approaches that can deal with inter-subject variability (Section 3.4).

3.1 Constructing invariant representations

In this section, we review the various types of approaches aimed at dealing with the fact that the original feature spaces might not be identical across sources, and sometimes across examples. These methods attempt to construct representations of the original examples, in the sense defined in [Marr 1982]: *a representation is a formal system for making explicit certain entities or types of information and that can be operated on by an algorithm in order to perform a certain task*. Powerful representations are such that they are invariant with respect to the effect induced by some sources of variability. In the case where the examples are images, which is our focus in this thesis, the acquisition protocol is very often not very well controlled which causes the images not to have the same size and makes the raw pixel space unusable as a feature space. For natural images, other typical sources of variability include differences in point of view or in illumination.

We distinguish three types of approaches that attempt to deal with this variability:

- feature engineering, which looks for characteristics of the examples that are themselves invariant with respect to some transformations and then constructs a vector-like feature set from these characteristics;
- structural representation design, which aims at encoding the samples into structured objects such as strings, trees or graphs;
- representation learning, which uses *machine learning* methods to obtain such invariant representations.

.

The *feature engineering* approach falls into the so-called *statistical pattern recognition* framework, while the design of *structural representations* is the base point of *structural pattern recognition*. These two subfields have a longlasting history of opposition and attempts at reconciliation and unification [Goldfarb et al. 2004], while the *representation learning* field is a newly emerging area of *machine learning* which attracts a lot of attention these days, notably thanks to its successes in classical tasks such as image classification [Ciresan et al. 2012]. Each of these encompasses a large array of methods. Our aim here is not to exhaustively review them but to give typical examples of such methods applied to imaging data, that are somehow relevant to the problems we are going to address in this thesis with neuroimaging data.

3.1.1 Feature engineering

When dealing with examples which are images that do not have the same size, one has to construct a feature set that can be compared across examples in order to perform a learning task such as categorization, i.e classification into a pre-determined number of categories. One method can be to first locate salient points from the image, such as the Scale Invariant Feature Transformation (SIFT), described in [Lowe 1999], which are the extrema of a pyramid constructed from differences of Gaussian bandpass filtered versions of the original image. Once the points of interest have been extracted with their surrounding image descriptor (which are invariant in scale and illumination changes as well as local distortions), they can be used in a bag-of-words framework: one can construct a vocabulary from a training dataset by unsupervised clustering of the SIFT descriptors extracted from the entire set of training images, each cluster then defines a word. The SIFT descriptors extracted from an unseen image are each assigned to one of the words of the vocabulary, and the image is represented by a set of features that are each defined as the frequency of occurrence of each word of the vocabulary in the image. Note that the localization information within the image is totally lost with this method, which has nonetheles proved to be very effective [Lowe 1999]. A large number of variants on the descriptors themselves and on the feature set construction have been proposed based on these ideas [Tuytelaars et al. 2007].

Other image transformations have been developed in order to construct invariant representations. One such transformation is the *scattering transform* [Bruna et al. 2013], which is constructed by using convolutional networks on top of a wavelet transform, thus producing translation-invariant representations. It is to be noted that the first layer provides SIFT-like features, while the following ones provide additional invariant descriptors. When used with a generative PCA model and a SVM classifier, these scattering representations yield state-of-the-art classification results on standard datasets [Bruna et al. 2013].

3.1.2 Structured representations

One of the major limitations of the *feature engineering* approaches described above is that they totally ignore the spatial structure, i.e the relative locations, of the extracted features within the images. A common way to take the spatial organization of images into account goes through the design of objects such as strings, trees or graphs, which all provide *structured representations*. In order to design such objects, one has to define their vertices, their edges, and optionnally attributes that can carry aditional information about the vertices and edges.

When dealing with images, one can define the vertices of a structured representation as two main types of objects extracted in a image: points or regions. We have explored in the previous paragraph the usual ways of extracting points

of interest; when designing graphs from such points of interest, an extra step consists in filtering those points to keep the ones that will be reproducible across examples, as in [Kisku et al. 2007]. The definition of regions inside an image is known as the segmentation problem in image processing; it consists in finding regions which satisfy some criterion. This criterion can be *low-level*, for instance the most basic one being that each region present a homogeneous intensity level. It can also be of higher level, in order to obtain regions which are more meaningful for the task at hand, such as semantic constraint for an object recognition task [Arbeláez et al. 2012]. Then, one need to define the edges. Here again, this can be done through very low-level criteria (such as the spatial adjacency of regions or the proximity of points), and go to higher level information (using semantic, such as part-based models which define the relationships between regions and sub-regions [Felzenszwalb et al. 2010] [B. Yao et al. 2010]). Finally, adding attributes on the edges and/or vertices might be useful to define structured representations that are more representative of the initial images, and might therefore be more informative for the considered problem [Sanromà et al. 2010].

We will design and use such structured representation to handle inter-subject variability in two different problems, as described in Chapter 4 and Chapter 5.

3.1.3 Representation learning

The emerging field of *representation learning* develops *machine learning* methods to produce representations of the input data containing information that will hopefully be more easily usable when attempting to solve a task later on. The goal is therefore identical as in the *feature engineering* and *structured representation design* approaches presented above, i.e to achieve invariant representations with respect to the sources of variability in the input data, but the methodology is different in the sense that we hope that an algorithm can *learn* such representation by itself, instead of using *a priori* knowledge to engineer appropriate representations as in both the other approaches.

Since the breakthrough paper of [Hinton et al. 2006], representation learning methods have allowed to beat state-of-the-art performances in numerous classical learning problems such as natural language processing [Dahl et al. 2012], speech recognition [Mikolov et al. 2011], image classification [Ciresan et al. 2012] and object recognition [Krizhevsky et al. 2012]. A lot of the work has focused on deep architecture of neural networks, *deep* meaning that a large number of layers are included in the network. One can still make use of a priori knowledge by enforcing some properties in the network, such as the expected smoothness of the representation, the sparsity of the information in the hidden layers, the hierarchy of organization of the layers and their respective sizes (for instance by using knowledge on the brain processing system that performs the equivalent task), the type of invariance that is looked after etc.

More recently, working with the assumption that going from one domain to another is a source of variability that could be handled in a *representation learning* framework, these *deep learning* methods have also been used on *transfer learning* problems. They have, yet again, been successfull [Bengio 2012], in particular to multi-source and multi-view learning questions [Ngiam et al. 2011; Zhuang et al. 2014].

It should be noted that the first applications of deep learning to neuroimaging problems have recently been published, with for instance a study dedicated to inter-subject decoding of fMRI data [Koyamada et al. 2015].

3.2 Domain adaptation

In this section, we review the various types of approaches available for the classical *domain adaptation* case where the training examples are all realisations of a single source domain, and the test examples are realisations of another domain. This problem, called *domain adaptation*, is a branch of *transfer learning* that tries to transfer the knowledge available in the source domain to apply it to the target domain, namely by *adapting* the joint probability $P_s(x, y)$ of the source domain to the one of the target domain $P_t(x, y)$. We will not attempt to fully review the *domain adaptation* literature here because it is far too large, but we will summarize the most prominent approaches that are used. The interested reader can turn him/herself to several technical reports that review the literature: [Margolis 2011] for a general review, [Li 2012] for applications in the fields of natural language processing and [Beijbom 2012] for computer vision-related questions; we mention these two fields – natural language processing and computer vision – because they provide questions that have driven a lot of work in *domain adaptation* and led to the design of numerous new methods.

3.2.1 Looking for shared representations

The first type of techniques we describe here aim at finding representations for which the distributions in the source and target domains are similar, or at least more similar than in the native feature space. We start by mentioning the quasi-standard practice that consists in normalizing each feature so that it has zero mean and unit variance on the training set; this heuristic contributes to making the source and target distributions more similar to each other. In practice, it does help a lot, and frustratingly, it is often hard to beat even when comparing its benefits to methods that are far more sophisticated and theoretically grounded. Another class of methods start with the rather intuitive idea that the divergence between the source and target distributions might be caused by the existence of some domain-specific features within the common feature set; a solution for this consists in selecting features that behave the same across domains, i.e that allows

minimizing a divergence statistic between distributions, as implemented using a soft-weighting procedure and conditional random fields in [Satpal et al. 2007] or in the CODA (Co-Training for Domain Adaptation) algorithm [M. Chen et al. 2011]. Finally, other methods look for transformations of the data to achieve the same goal, for instance using mappings to a RKHS that minimize the Maximum Mean Discrepancy (MMD) between distributions [Pan, Tsang, et al. 2011].

3.2.2 Instance weighting

Given the fact that the source and target distributions are different, another intuition leads to reweighting the source distribution, via an *importance weight function* in order to make the reweighted source distribution as close as possible to the target distribution. When placed in the classical *risk minimization* framework, given a loss function ℓ and a vector of parameters θ, this idea becomes:

$$\begin{aligned} R(P_t, \theta, \ell(.,.,\theta)) &= \mathbb{E}_{(x,y)\sim P_t}(\ell(x,y,\theta)) \\ &= \mathbb{E}_{(x,y)\sim P_t}(\frac{P_s(x,y)}{P_s(x,y)}\ell(x,y,\theta)) \\ &= \mathbb{E}_{(x,y)\sim P_s}(P_s(x,y)\frac{P_t(x,y)}{P_s(x,y)}\ell(x,y,\theta)) \\ &= \mathbb{E}_{(x,y)\sim P_s}(P_s(x,y)\beta(x,y)\ell(x,y,\theta)) \\ &= R(P_t, \theta, \beta(.,.)\ell(.,.,\theta)) \end{aligned} \tag{3.1}$$

Empirically, this leads to weighting each instance of the source domain using the following weight:

$$\beta(x,y) = \frac{P_t(x,y)}{P_s(x,y)}.$$

A typical assumption used when trying to solve this problem is the so-called *covariate shift* [Shimodaira 2000], which states that the conditional probabilities are identical in the source and target domains, i.e $P_t(y \mid x) = P_s(y \mid x)$ and that only the marginal distributions differ. This leads to

$$\begin{aligned} \beta(x,y) &= \frac{P_t(x,y)}{P_s(x,y)} \\ &= \frac{P_t(y \mid x)P_t(x)}{P_s(y \mid x)P_s(x)} \\ &= \frac{P(y \mid x)P_t(x)}{P(y \mid x)P_s(x)} \\ &= \frac{P_t(x)}{P_s(x)} \end{aligned} \tag{3.2}$$

This result, which is clearly appealing by its simplicity, nevertheless raises the challenge of estimating these weights from the data since in most cases, the marginal distributions are unknown. A large number of methods have been proposed to solve this problem, as for instance by directly using density estimation techniques ([Sugiyama et al. 2005]) or by solving an annex classification problem to determine whether an instance belongs to the source or the target domain, which provides an estimate of the marginal distributions ([Zadrozny 2004]). In order to avoid having to estimate the marginal distributions, one can also use techniques that allow to directly estimate their ratio, like the Kernel Mean Matching (KMM) [Gretton et al. 2009].

In Chapter 6, we will describe in detail the Kernel Mean Matching approach and propose a KMM extension for the multi-source problem.

3.2.3 Iterative approaches

If some labeled instances from the target domain are available at training time, methods from the *semi-supervised learning* literature ([Chapelle et al. 2006]) can be directly used to perform domain adaptation. When it is not the case, another family of methods consists in guessing the labels of some target samples, and then include them to either train another model or use semi-supervised methods; this process can be iterated to progressively estimate a better classifier. Amongst these *self-labelling* methods, we can for instance mention the work by [Dai et al. 2007] which is based on the Expectation Maximisation (EM) principle, and the DASVM algorithm which trains an SVM classifier at each iteration and selects which instance to keep for the next iteration [Bruzzone et al. 2010].

Another approach is based on a mixture model of the probability distributions, that allows to label the available samples as belonging to the source domain, the target domain or being share between domains. The parameters of the mixture model are estimated iteratively with the Conditional EM algorithm [Daumé III et al. 2006].

3.3 Multi-source-specific methods

When samples from more than one source domain are available at training time, the most straightforward strategy is to pool them all together and to consider the resulting set of data as a regular training dataset, ignoring the differences between training sources. This calls for two remarks. First, this model is clearly sub-optimal and one hopes to be able to use the different sources in a more clever way that can efficiently combine their respective contribution; this is the essence of *multi-source learning*. Secondly, combining the different sources in a simplistic way can lead to worsening the performances of the model, as compared to using only a single source for training (as empirically seen in [Schweikert et al. 2009]

for instance). This phenomenon is referred to as *negative transfer* in the *transfer learning* literature.

Categorizing *multi-source learning* methods is a difficult task because, as defined in 2.1, it is tightly linked with several other machine learning sub-fields that include *domain adaptation*, but also *multi-task*, *multi-view* and *semi-supervised* learning, and *ensemble methods*. Multi-source methods indeed feed themselves from these other domains of research, sometimes borrowing elements from several of them. In what follows, we present a taxonomy of *multi-source* methods in a somehow artificial manner; it is therefore not fully accurate because a given method often use elements from several other classes of methods presented in other paragraphs.

Before describing the array of *multi-source* methods, we first mention the rare theoretical studies available for the *multi-source learning* problem. Assuming they have a distance matrix between the sources, [Crammer et al. 2008] computes a general bound on the expected loss of the multi-source model by using the nearest k sources from the target domain, thus using a source selection scheme. [Mansour et al. 2009] studied the question of combining single-source models and demonstrated that a distribution-weighted combination rule can guarantee a bound on the loss while the standard convex combination rule does not offer such guarantee and therefore can yield negative transfer. Finally, [Ben-David et al. 2010] introduces a distance metric between domains, which makes it possible to compute two learning bounds for the empirical risk minimization in the target domain.

3.3.1 Multi-source domain adaptation

Multi-source learning can be viewed as a generalization of the *domain adaptation* problem: one simply has several, instead of one, source domains that we need to *adapt* to the target domain. A review of *multi-source domain adaptation* method has been recently published in [S. Sun et al. 2015].

Most of the existing methods attempt to either select some sources or weight the contribution of each source domain. This is the case in [Chattopadhyay et al. 2012] which attempts to match the conditional probability distributions across domains through a regularized weighting procedure and define the decision function in the target space accordingly; note that it requires having a small number of labeled samples in the target domain. The method introduced in [Q. Sun et al. 2011] combines the previous one with a standard domain adaptation instance-weighting scheme; this instance weighting is used as a first step to match the marginal distributions from each of the source domain to the target one, and a second step conceptually similar to the method of [Chattopadhyay et al. 2012] attempts to match the conditional probability distributions.

Another way to combine the contributions of the different source domains is to use a two-stage procedure: first, one can use the labeled instances of each of

the source domains to train a single-source model, and then one can combine these models to be adapted to the target domain. The early work of [Fromont et al. 2004] actually follows this principle. The more recent Domain Adaptation Machine described in [Duan et al. 2009] also uses *auxiliary* classifiers (trained on a single source) before a data-dependent regularizer enforces that the target classifier acts similarly as the pre-computed auxiliary classifiers from relevant sources.

The method introduced by [Tan et al. 2014] focuses on the cases where the different sources might also include different *views*. The different views from different sources are iteratively combined by a co-training step, a target instance selection step and a reweighting of the source instances that allows learning a full model.

3.3.2 Boosting-based methods

Several groups have investigated the use of ensemble methods for *multi-source learning*, and in particular boosting. Adaboost is an iterative boosting algorithm that selects a weak classifier at each iteration and adds it to the ensemble of classifiers, while down-weighting the samples that have been correctly classified so that the focus is placed at the next iteration on the samples for which the prediction was not accurate. [Y. Yao et al. 2010] introduces two boosting algorithms that attempt to leverage the knowledge from the multiple available sources at training time. The first one consists in selecting at each iteration the weak classifier from the source that is the closest to the target domain. The second one operates in two stages: in the first one, single-source classifiers are estimated while in the second one, they are used as the weak classifiers in order to boost the target classifier to be estimated. [Huang et al. 2012] define a concept of *view* as a combination of one or several of the training sources; at each iteration, their boosting algorithm selects the best view by computing a distance to the target domain. [Shi et al. 2012] attempts to work in a more general framework, depicted as *heterogeneous learning* in which the feature spaces of each of the sources do not have to be shared, which places this work in the *multi-view* setting; it builds a boosting ensemble by looking to maximize the decision consensus across the instances that are shared across several sources.

3.3.3 Multi-task models

Another way to frame the *multi-source learning* problem is to associate a specific *task* to each source, or combination of sources, and to use a *multi-task* framework considering that these different tasks are somehow related. [Lin et al. 2013] considers the image auto-annotation problem using annotated images from multiple sources as training samples; once again, it builds upon single-source models to construct a multi-task model with inter-sources structural regularization and an

additional set of constraints that the parameters need to verify across sources. In neuroimaging applications, two studies have implemented *multi-task* models to tackle *multi-source* problems. The first one ([Yuan et al. 2012]) studies the group discrimination task described in ?? using multi-modal (i.e multi-view) data; a task is defined for each combination of available modalities; the multi-task model makes it possible to handle observations with missing views, i.e subjects for which not all modalities are available thanks to the design of a shared feature set estimated using a sparse learning technique. The second study ([Marquand et al. 2014]) addresses the inter-subject decoding problem from functional imaging; a task is assigned for each subject available in the training set and a Bayesian multi-task framework is proposed to build a classifier on the test subject.

3.3.4 Other approaches

We can also mention some other approaches that fall in the realm of *multi-source learning*. The work in [Fang et al. 2015] is based yet again on the training of single-source classifiers; the concatenation of the predictions of each of these classifiers form a multi-label vector for the training examples and the proposed method consists in finding a shared subspace of labels that allows defining a classifier in the target domain. [Zhao et al. 2008] introduces a *multi-source* feature selection scheme that examines the covariance structure of the data across sources to define a subset of informative features. Finally [Geras et al. 2013] examines the use of the structure of the sources in a cross-validation procedure, and proposes new variance estimators that have better theoretical ground in the *multi-source* setting and that yield more accurate confidence intervals to perform model selection.

3.4 Other approaches for inter-subject learning

3.4.1 Hyperalignment

Another approach that has been recently introduced in an attempt to improve *inter-subject predictions* in fMRI experiments is the so-called hyperalignment [Haxby, Guntupalli, et al. 2011]. It consists in using a calibration experiment performed by all the subjects in order to compute a transformation of the subject's functional space into a space that is common to all subjects. This common space can be seen as a template, i.e a functionally averaged subject which will then serve as a target. In practice, the authors advocate the use of passive movie viewing in order to provide a maximum amount of information to the algorithm that estimates the transformation to the common space. Once this common space has been built, it can be used to perform *inter-subject predictions* for a second experiment performed by each subject during the same acquisition session.

In other words, the hyperalignment can be seen as a semi-supervised representation learning approach. Indeed, all the subjects perform a calibration experiment for which the values of the output variable y is available to the learning algorithm. And the estimation of the common space is in fact a way to find representations that is invariant across subjects. In the original paper [Haxby, Guntupalli, et al. 2011], the transformation from one subject s_1 to another subject s_2 is estimated by solving the following problem:

$$C(s_1, s_2) = \underset{C \; orthogonal}{\arg\min} \|X^{s_1} C - X^{s_2}\|^2 \tag{3.3}$$

A more general kernelized version of the hyperalignment has been presented in [Lorbert et al. 2012]. Another improvement has been introduced by [Rustamov et al. 2013]: instead of looking for an orthogonal transformation matrix C, which can introduce unrealistic distortions between subjects, the authors proposed to add a regularization term that minimizes the metric distortion across subjects. Their method yields transformations that are more biologically plausible while improving the performances in an inter-subject decoding task.

3.4.2 Spatial regularization

The inter-subject variability can be modeled as a spatial variability across subjects, i.e an inaccurate match of cortical locations across subjects. The spatial smoothing used commonly in group-level univariate analyses is a way to overcome such variations, but such filtering alters the potentially informative content available at fine spatial scales. A more sophisticated way to tackle this problem is to add a spatially informed regularization term to standard classification or regression methods, on top of more standard regularization techniques which are necessary to cope with the high dimensionality of the data. This has been successfully implemented with various approaches and we here briefly describe a few of these studies.

[Michel, Gramfort, Varoquaux, Eger, and Thirion 2011] used a Total Variation regularization term on top of a regression to perform inter-subject predictions with functional MRI experiment. Using Total Variation, which is in fact the ℓ_1 norm of the image gradient, promotes piecewise constant weight maps, and thus well localized brain regions that contributes to the regression. [Grosenick et al. 2013] developed some implementations of the GraphNet – the graph-constrained version of Elastic Net in order to use the spatial structure of the images. This allowed for the combination of a sparsity-inducing regularization and a spatially structuring regularization, which therefore yield interpretable maps in fMRI experiments. Finally, this time working with anatomical MRI data, [Cuingnet et al. 2013] introduced a spatial regularization term into the SVM, using the graph Laplacian, which allowed to obtain state-of-the-art classification performances.

4 Graph-based Support Vector Classification for inter-subject decoding of fMRI data

Contents

Context

In the first contribution of this thesis, we propose a framework that enables to perform *inter-subject predictions* from fMRI data by examining the fine-scale patterns of activation within a localized region of interest of the cortex. The input spaces $\mathcal{X}^s$ of each subject s are not assumed to be identical because of the large variability that exist across subjects at such fine spatial scale. The *multi-source* contribution thus lies in constructing a common space to all subjects by using graphical representations of the input patterns and a graph kernel that implicitely performs the embedding into a reproducing kernel hilbert space. We demonstrate that this framework allows to significantely improve the accuracy obtained in an inter-subject decoding fMRI task aimed at studying the tonotopic organization of the auditory cortex.

Note that this work is the main methodological outcome of the GRABBR (GRAph-Based Brain Reading) project, for which I was the Principal Investigator. This project was funded by the CNRS interdisciplinary program dedicated to computational neuroscience (Neuro-IC) in 2010-2011. The contributors to this project were Daniele Schön (construction of the experimental paradigm and data acquisition), Guillaume Auzias (anatomical data processing), Bertrand Thirion (general methodological advice) and Liva Ralaivola (kernel design and more). The fMRI data was acquired at the *Centre d'IRM fonctionnelle de Marseille*, where Muriel Roth set up a pulse sequence allowing to acquire high resolution fMRI data.

This work was published in the following conference and journal articles:

S. Takerkart, G. Auzias, B. Thirion, D. Schön, et al. "Graph-Based Inter-subject Classification of Local fMRI Patterns". In: *Machine Learning in Medical Imaging*. Ed. by F. Wang et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 184–192

S. Takerkart, G. Auzias, B. Thirion, and L. Ralaivola. "Graph-based inter-subject pattern analysis of fMRI data". In: [2014]

This chapter is therefore mostly composed of the content of the published journal paper, but it includes a few additional figures. We have also added some extra results in the appendices.

4.1 Introduction

Functional Magnetic Resonance Imaging (fMRI) is a modality that has proved extremely useful for understanding brain function as it offers the possibility to map cognitive processes to brain activation patterns. Traditional univariate analysis methods of fMRI data process each voxel separately to perform *forward inference* [Friston 2007], that is, identify those voxels that show an activation profile significantly associated with a given task. With the recently proposed application of multivariate pattern recognition methods to fMRI data, one can also make *reverse inference*, that is, predict a behavioral variable directly from the imaging data, as in the pioneering work described in [Haxby, Gobbini, et al. 2001]. This new approach, often referred to as *multi-voxel pattern analysis* (MVPA), *brain decoding* or *mind reading*, has received an increasing amount of attention over the last few years. The vast majority of papers published so far (see reviews [Norman et al. 2006; Haynes et al. 2006; Mahmoudi et al. 2012]) study the organization of cortical representations, a problem particularly suited for MVPA since such representations arise from neural activity distributed across networks that can cover a large number of fMRI voxels. Another problem that can be addressed through MVPA techniques is to examine the consistency of patterns across tasks by testing whether patterns observed in a given task may arise in different tasks, as in [Meyer et al. 2010; Knops et al. 2009]. Finally, one can also use MVPA to characterize patient groups from fMRI data, in order to identify putative fMRI biomarkers that could be used in diagnosis tools [Coutanche et al. 2011; L. Zhang et al. 2005; Honorio et al. 2012]. All these applications ask for constructing *group-invariant characterizations*. Most studies published until today address this question with a two-level inference, performing MVPA within subject, and testing the consistency of within-subject classification scores across individuals. However, this limits the interpretability of the results because within-subject MVPA often relies on sub-voxel idiosyncratic information [Kamitani et al. 2005]. It is therefore of the highest interest to address this question more directly by performing inter-subject MVPA, i.e. by looking for features that are common across subjects and learning a decision rule on data recorded in a set of subjects to use it on data from different subjects.

Challenge. The potentially large inter-individual variability represents a major challenge to construct group-invariant representations that will allow for successfull inter-subject MVPA. Only few studies have directly addressed this question. Most rely on full brain analysis, using large-scale features that are stable across subjects after spatial normalization [Friston 2007]. While a recent study proposes to use a multi-task framework to handle large scale inter-subject variability [Marquand et al. 2014], all the others focus on the feature construction / selection: several papers use univariate feature selection with different criteria (relative entropy in [Poldrack et al. 2009], most active or discriminative voxels in [Shinkareva, Mason, et al. 2008] and [Cabral et al. 2012]); others

summarize the signal present in a set of regions by their mean, using, e.g., cubic regions [Davatzikos et al. 2005], anatomically defined regions [Mitchell et al. 2004] and [Wang et al. 2004], or functionally defined parcels [Mitchell et al. 2004]; [Mourão-Miranda et al. 2005] uses principal component analysis; finally [Ryali et al. 2010] and [Grosenick et al. 2013] use sparse learning methods that automatically select features. When examining patterns at a finer spatial scale, inter-individual variability is yet larger and performing such inter-subject predictions becomes even more challenging. At such scale, the alignement between cortical folding and the underlying functional organization vary between subjects [Essen et al. 2007; Sabuncu, Singer, et al. 2010], in a way that the potentially poor voxel-to-voxel correspondance provided by spatial normalization procedures limits the generalization power of classifiers that use voxel values as features [Clithero et al. 2011]. To our knowledge, only two studies describe methods specifically designed for inter-subject classification without the need for spatial normalization. The first one [Haxby, Guntupalli, et al. 2011] uses Procrustes transformations to maximally align, in a high-dimensional space, the spatio-temporal patterns recorded during a specific training experiment. The second one [Abdi et al. 2012] is a discriminant analysis that projects the data (through a generalized PCA) onto multiple-subjects factorial maps designed to maximize class separation. Both these techniques do not enforce the preservation of the spatial organization of the input patterns to construct their latent space.

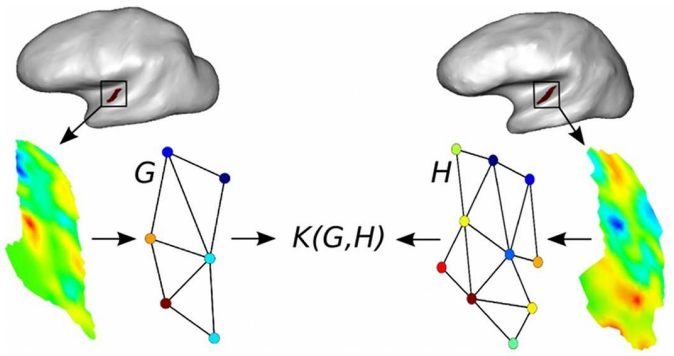

Figure 4.1: Our contributions: i) attributed graphs are learnt in an unsupervised manner to represent local functional patterns observed in unregistered subjects; ii) graphs similarities are evaluated by a custom-designed kernel, allowing to solve various problems (classification, regression, clustering).

Structured learning. In order to tackle the challenge posed by inter-subject variability, structural analysis schemes have proved efficient for forward inference group studies in neuroimaging, as described in [Coulon et al. 2000; Thirion,

Pinel, Tucholka, et al. 2007]. However, no such structural approach has been developed to perform reverse inference. Our goal here is to develop a learning framework that specifically aims at predicting a behavioral variable from imaging data while overcoming inter-subject variability by exploiting the structural properties of the input patterns. Such a framework should address three problems:

- What are the structures of interest? In neuroimaging, two main classes of elementary objects are used in such approaches: points (local maxima of activation [Thirion, Pinel, and Poline 2005]) or regions (clusters of activation [Coulon et al. 2000], parcels [Flandin et al. 2002]). These functional features can be represented into a graph to encode their relationships, as it is now classically done with connectivity-based models of functional or anatomical networks.

- How is the inter-individual variability conveyed? Regardless of the chosen feature type, models of inter-individual functional variability let their location [Thirion, Pinel, Tucholka, et al. 2007] and intensity [Lashkari et al. 2012] vary across subjects.

- What learning method to use? Learning from structured data can be done with a wide variety of methods, among which, neural/deep belief networks [Frasconi et al. 1997], probabilistic/graphical models (such as Markov fields [Coulon et al. 2000], hierarchical Dirichlet processes [Lashkari et al. 2012] or Conditional random fields [Lafferty et al. 2001]), or large margin kernel-based methods with appropriately engineered kernels (see [Mahé, Ralaivola, et al. 2006; Ralaivola, Swamidass, et al. 2005]).

Contributions. In the present paper, we introduce a Graph-based Support Vector Classification (G-SVC) framework that respectively addresses the previous questions by i) using unsupervised learning to construct attributed graphs that represent fMRI patterns of activation; the nodes are patches of activation given by a parcellation algorithm; the graph edges carry the spatial relationships between nodes and their relevant characteristics (location and activation) are encoded as attributes of the graph nodes; ii) assuming that both attributes of the nodes can vary across subjects, i.e. that the inter-individual variability can be characterized along these two dimensions; iii) designing a graph similarity measure (a graph kernel) that is robust to inter-individual variability, and that makes it possible to perform supervised learning directly in graph space, for instance by using support vector classification. These contributions are summarized in Fig. 4.1.

While the use of graphical representations of fMRI data has seen a tremendous boost in the last decade with the fast development of connectivity analyses (see for instance [Sporns et al. 2005; Richiardi et al. 2013]), graph kernels have

only recently been introduced in the neuroimaging field. The few studies that make use of graph kernels to solve neuroimaging learning problems address different sorts of questions (subject classification based on resting-state functional connectivity [Jie et al. 2014] or task-based fMRI [Gkirtzou et al. 2013], characterization of the mental state of the subject from its connectivity [Mokhtari et al. 2013] or activation [Vega-Pons and Avesani 2013; Vega-Pons, Avesani, et al. 2014] patterns), showing their potential versatility.

Our framework falls in the latter category. It is specifically aimed at overcoming the fine scale *functional variability* observed in a given region of interest for a task-based fMRI experiment, which is a key issue in understanding local neural representations [Shinkareva, Malave, et al. 2012; Haxby, Guntupalli, et al. 2011]. In such case where using spatial normalization is the bottleneck, our framework allows to explicitely take into account the different sources of inter-individual variability without requiring perfect cross-subject matching of brain anatomy, hereby alleviating the dependecy of the method to the registration accuracy. Furthermore, it can easily be tuned to address numerous problems provided one may have at hand a meaningful parcellation for the question of interest, as for instance in full brain resting state studies (see a review in [Blumensath et al. 2013]) or diffusion weighted-based segmentation of grey matter regions (as for instance in [Behrens et al. 2003]).

4.2 Materials and methods

4.2.1 Graph-based Support Vector Classification (G-SVC)

The defining task that is usually addressed in inter-subject MVPA might be stated as the learning of a classifying function f able to reliably predict a categorical experimental variable y from fMRI data X recorded in a given set of subjects. In order to gain invariance with respect to the inter-subject variability and be able to generalize to data from new subjects, we use a graphical representation X of the input data (described below). Effective methods have recently emerged to learn from such structured data ([Frasconi et al. 1997; Coulon et al. 2000; Lashkari et al. 2012; Lafferty et al. 2001; Mahé, Ralaivola, et al. 2006; Ralaivola, Swamidass, et al. 2005], and among those, similarity-based learning approaches (nearest-neighbors methods, kernel machines, relevance vector machines, ...) have received much attention. We focus here on *support vector classifiers* [Cortes et al. 1995], or *SVC*, because of their well-foundedness and their effectiveness in various application domains, including neuroimaging. Without entering into too much detail, the most prominent way to perform binary support vector classifi-

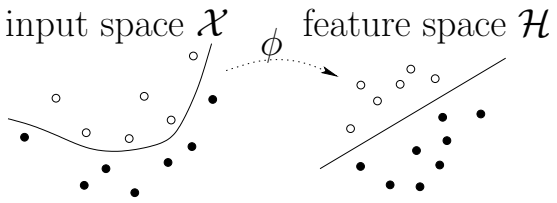

Figure 4.2: The kernel trick: the use of a positive kernel K implicitely maps data from some input space $\mathcal{X}$ into a Hilbert space $\mathcal{H}$ —thanks to the canonical mapping $\phi : X \mapsto \phi(X) = K(X, \cdot)$— where linear separation is possible.

cation works by solving the following quadratic problem:

$$\begin{array}{rl} \min\limits_{\boldsymbol{\alpha}\in\mathbb{R}^n} & \frac{1}{2}\sum\limits_{ij}\alpha_i\alpha_j y_i y_j K(X_i, X_j) - \sum\limits_i \alpha_i \\ \text{s.t.} & \sum\limits_i \alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq \mathcal{C},\ \forall i \end{array} \tag{4.1}$$

The solution $(\boldsymbol{\alpha}^*, b^*)$ of this problem (where b^* is given by the optimality conditions of the problem) defines a classifier f given by:

$$f(X) = \text{sign}\left(\sum_i y_i \alpha_i^* K(X_i, X) + b^*\right), \tag{4.2}$$

where $\{(X_i, y_i)\}_{i=1}^N$ is the *training data* of labeled pairs (X_i, y_i), made of pattern X_i and associated (binary) target y_i, $\mathcal{C} > 0$ is a user-defined (soft-margin) parameter and K is a *positive kernel* function. The kernel function implicitly allows one to map the training patterns X_i's into a relevant Hilbert space (an idea, known as the 'kernel trick', that dates back to [Aizerman et al. 1964]) where large-margin linear classification is possible (see Fig. 4.2). Choosing/designing an appropriate kernel for the data at hand, is therefore the crux of using support vector classification for real-world applications, knowing that dealing with structured inputs merely requires the design of a sound kernel. Note that we limited ourselves in describing the binary case, but well known composition methods such as the *one-vs-all* or *one-vs-one* strategies make it possible to directly build multiclass predictors from the binary method.

In what follows, we describe how we build a graphical representation of functional patterns (section 4.2.2) and a graph kernel (section 4.2.3). With these tools, one can define numerous classifiers to perform inter-subject fMRI prediction (illustrated on Fig. 4.3); here, without loss of generality, we instantiate the support vector classification framework. Therefore, our graph construction scheme and graph kernel fully define our Graph-based Support Vector Classifier (G-SVC) framework.

We assume that for the considered task-based fMRI experiment, we have at our disposal for each subject: i) a pre-defined contiguous *region of interest* (ROI) $\mathcal{R}$, ii) the function ϕ describing the BOLD activation at each experimental trial (each trial providing a different observation), and iii) a coordinate system ω (in practice, we use a 2d cortex-based set of coordinates, which is more meaningful than working in the 3d volume [Van Essen, Drury, et al. 1998]). Furthermore, we assume that the functional organization of $\mathcal{R}$ with respect to our experiment is consistent across subjects.

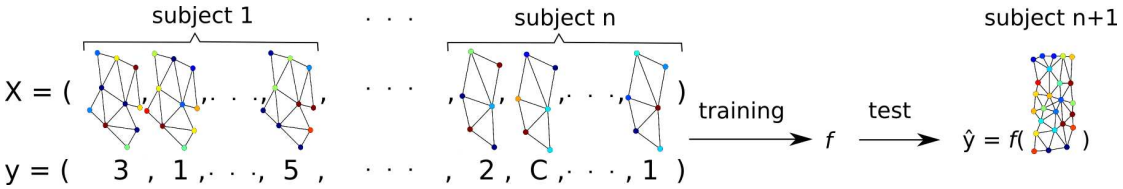

Figure 4.3: Inter-subject graph-based learning. Equipped with a graph construction scheme and a graph similarity function such as those designed in this paper, one can define numerous classifiers. The decision function f is learnt on a training set composed of labeled graphs (X_i, y_i), with $y_i \in \{1 \dots C\}$ from a set of subjects, and can be used on graphs from another subject, potentially with a different number of nodes. We here use support vector machines to demonstrate the soundness of this approach when dealing with inter-subject variability of fMRI activation patterns.

Note that the way we use support vector classification in what follows departs a little bit from what is suggested by the theory that supports the use of SVM. Indeed, the classical framework assumes the training set $\{(X_i, y_i)\}_{i=1}^{N}$ be made of identically and independently distributed random pairs, whereas the pairs we are going to work with may be dependent as different training pairs (X_i, y_i) could relate to the same subject. Carefully characterizing and taking into account these dependencies is an important challenge posed by many inter-subject prediction problem (see [Takerkart and Ralaivola 2014]) that goes beyond the scope of the present paper. Ideas taken from [Janson 2004; Ralaivola, Szafranski, et al. 2010], may lay the theoretical ground to build relevant and original approaches and constitute the main axis of our future researches. In any case, our use of SVM is frequently encountered in the literature, in e.g. information retrieval problems [Cao et al. 2006; Joachims 2002; Liu 2009], where no particular care of such dependencies exist and still very good classification results are achieved.

4.2.2 Graphical representation of fMRI patterns

Here, we detail the unsupervised representation learning scheme that we use to derive graphical representations from fMRI activation patterns.

4.2.2.1 Parcellation of the ROI

Assuming that the ROI $\mathcal{R}$ admits an underlying subdivision into a set of smaller and functionally meaningful sub-regions, the first step to construct our graphical representation consists in estimating a partition of $\mathcal{R}$ into a set of sub-regions or *parcels* [Flandin et al. 2002]. Specifically, a parcellation $\mathcal{P}$ of $\mathcal{R}$, is a set $\mathcal{P} = \{P_i\}_{i=1}^{q}$ of q parcels so that the parcels verify: $\cup_{i=1}^{q} P_i = \mathcal{R}$ and $P_i \cap P_j = \emptyset$ whenever $i \neq j$.

We use Ward's hierarchical clustering algorithm to learn the parcellation in an unsupervised manner. The algorithm starts with one parcel at each point $v \in \mathcal{R}$, and iteratively merges two parcels into one so that the variance across all parcels is minimal, with the added constraint that two parcels can be merged only if they are spatially adjacent [Michel, Gramfort, Varoquaux, Eger, Keribin, et al. 2012]. The input vector that we used is $\{\phi(v), \omega(v)\}_{v \in \mathcal{R}}$; it combines the anatomical information provided by the coordinate system ω, and the full functional information available for a given subject (i.e. the activation maps recorded at each trial for all experimental conditions). Incorporating the anatomical information on top of the functional features acts as a spatial regularization process in the search for functionally meaningful units, which makes the parcellation more robust to the low contrast-to-noise ratio of the functional data usually encountered when using MVPA. The resulting parcels constitute the elementary functional features that are the base elements of our approach.

4.2.2.2 Graph nodes and edges

We use $\mathcal{P}$ as the set of nodes of our graphical representation, i.e. each parcel defines an elementary functional feature of the pattern that is represented by a node of the graph. The set of edges is represented by a binary adjacency matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{q \times q}$, where $a_{ij} = 1$ if parcels P_i and P_j are connected, and $a_{ij} = 0$ otherwise. In this work, since we assume that the topological properties of the patterns are stable across subjects, we use spatial adjacency as the criterion to decide whether two nodes are connected (i.e. $a_{ij} = 1$ if $\exists v_i \in P_i, \exists v_j \in P_j$ so that v_i and v_j are neighbors). This defines a region adjacency graph [Pavlidis 1977] where the structure of the graph encodes the spatial organization of the parcels. Note that our method is also fully valid if one had used other criteria (for instance functional connectivity) to define the edges of the graph.

4.2.2.3 Activation attributes

In a parcel P_i and for a given observation (i.e. experimental trial), the activation values $\{\phi(v)\}_{v \in V_i}$ are summarized by their mean inside the parcel, that we note $\Phi(P_i)$. We note $\mathbf{\Phi}$ the vector $\mathbf{\Phi} = [\Phi(P_1) \cdots \Phi(P_q)]$ of activation attributes. Note that, more generally, we may summarize the activation values measured in P_i

using a feature vector $\Phi(P_i)$; this vector could include the mean, the variance, the skewness, ..., or any other summary statistics.

4.2.2.4 Geometric attributes

The geometric information of parcel P_i is summarized by a feature vector $\Omega(P_i)$, computed from the locations $\{\omega(v)\}_{v \in V_i}$. In this study, we use the coordinates of the center of mass of the ROI, computed within the coordinate system ω. We note $\mathbf{\Omega}$ the matrix of geometric features $\mathbf{\Omega} = [\Omega(P_1) \cdots \Omega(P_q)]$. As before, richer geometric information, accounting for instance for the shape of the parcel, may be considered.

4.2.2.5 Full graphical model

Using these definitions, we have defined an attributed relational graph [Eshera et al. 1986] $G = (\mathcal{P}, \mathbf{A}, \mathbf{\Phi}, \mathbf{\Omega})$ and described how to learn such graphical representations in an unsupervised manner. These graphs fully represent the functional patterns recorded within the ROI $\mathcal{R}$ and carry activation, geometric and structural information, as illustrated in Fig.4.4.

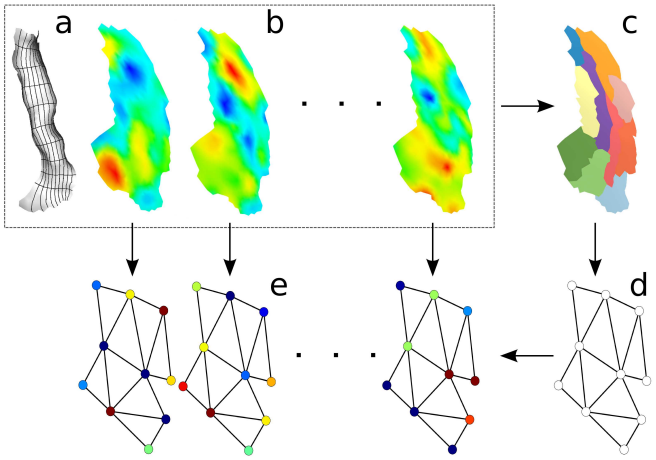

Figure 4.4: Construction of our graphical representation of functional patterns. Starting from a local coordinate system ω (illustrated in $\mathbf{a}$ as a grid on the local cortical mesh) and the functional activation maps ϕ (displayed in $\mathbf{b}$ as overlays on the flattened mesh), we produce a parcellation $\mathcal{P}$ (shown in $\mathbf{c}$) which gives the graph structure $\mathbf{d}$ with the location of the nodes. The activation values of each functional instance are then mapped onto the nodes to produce the attributed relational graphs shown in $\mathbf{e}$, which carry all the necessary information through its structural, geometric and activation features.

4.2.3 Graph similarity

Among the various frameworks that exist to learn from structured data, what makes similarity-based methods popular is that the difficulty of the learning process is transferred to that of defining a similarity function on the space of structured objects at hand. It turns out that a plethora of graph similarity functions exist, defined with respect to the number of edit operations needed to transform one graph to another [Bunke 1997], the number of common subgraphs of certain type (walks [Gärtner et al. 2003], trees [Mahé and Vert 2009], graphlets [Przulj et al. 2004]), or the similarity between vector representations of graphs (see for instance [Riesen et al. 2009]).

Here, we decide to use a positive kernel as a similarity measure between graphs: this makes it possible to envision the use of kernel-based machine learning algorithms (such as the *support vector classifiers* described above), which have proved efficient in this context [Mahmoudi et al. 2012] and offer solid theoretical guarantees. When choosing or designing a graph kernel for a given application, one need to find a good compromise between expressivity (i.e. the ability of the kernel to capture the features of interest in the available graphs) and computational efficiency [Ramon et al. 2003]. The recently developed Weisfeiler-Lehman graph kernel [Shervashidze et al. 2011] offers such properties (which has made it the kernel of choice for several recent neuroimaging applications [Vega-Pons and Avesani 2013; Vega-Pons, Avesani, et al. 2014; Gkirtzou et al. 2013]) but its applicability is limited to labeled graphs. Since the key features of our graphical representations are carried by the *real-valued* attributes of the nodes, we want to avoid having to quantify the values of these attributes into discrete labels, which would imply loosing both some precision and the also the structure provided by real-valued attributes. We therefore decided to construct a dedicated kernel. In order for our kernel to provide a good balance between the two aforementioned criterion, we followed two directions: on the one hand, our design builds upon the work of [Gärtner et al. 2003] which layed the ground for the construction of efficient walk-based kernels computable in polynomial time; on the other hand, the expressivity of our kernel is based on an intuitive design scheme which aims at exploiting each type of graph features available in our representation, and in particular its real-valued node attributes. Below, we describe our design step by step as an instantiation of the generic family of R-convolution kernels [Haussler 1999], which are defined as:

$$K(G, H) = \sum_{g \subseteq G, h \subseteq H} \prod_{t=1}^{\tau} k_t(g, h), \tag{4.3}$$

where G and H are two graphs, $\tau \in \mathbb{N}^*$ is the number of base kernels k_t, which act on subgraphs g and h (for simplicity reasons, we here use walks of length one; note that the definitions below are directly extendable to other types of

subgraphs).

Given the nature of our graphical representation, we define $\tau = 3$ elementary kernels k_s, k_g and k_a, respectively acting on structural, geometric and activation information, and thus covering all characteristics of the graphs.

For two graphs $G = (\mathcal{P}_G, \mathbf{A}_G, \mathbf{\Phi}_G, \mathbf{\Omega}_G)$ and $H = (\mathcal{P}_H, \mathbf{A}_H, \mathbf{\Phi}_H, \mathbf{\Omega}_H)$, we note g_{ij} and h_{kl} two pairs of nodes (i.e. walks of length one) in G and H, respectively; let q_G and q_H be the number of nodes in G and H, respectively — note that q_G and q_H may be different.

4.2.3.1 Structural kernel

Because the structure of our graphical representations encodes the spatial adjacency of the parcels and because we assume that the spatial organization of the functional patterns is consistent across subjects, we include a first base kernel k_s which aims at valuing the structural similarity of G and H. We simply adopt the linear kernel on binary entries a_{ij}^G and a_{kl}^H of the adjacency matrices $\mathbf{A}_G$ and $\mathbf{A}_H$:

$$k_s(g_{ij}, h_{kl}) = a_{ij}^G . a_{kl}^H \tag{4.4}$$

It gives 1 if $a_{ij}^G = 1$ and $a_{kl}^H = 1$, and 0 otherwise, meaning that the other base kernels are only taken into account if g_{ij} and h_{kl} are both actual edges. Our kernel in fact compares each edge of a graph to all edges of the other graph.

4.2.3.2 Geometric kernel

Kernel k_g acts on the geometric attributes, i.e. the location of the graph nodes within the coordinate system ω. The goal of this kernel is to match edges across graphs. To allow for inter-individual differences, we implement a soft matching by using the following product of Gaussian kernels:

$$k_g(g_{ij}, h_{kl}) = e^{-\|\Omega_i^G - \Omega_k^H\|^2 / 2\sigma_g^2} \cdot e^{-\|\Omega_j^G - \Omega_l^H\|^2 / 2\sigma_g^2}, \tag{4.5}$$

where $\sigma_g \in \mathbf{R}_+^*$, and Ω_m^G (resp. Ω_m^H) is the mth column of $\mathbf{\Omega}_G$ (resp. $\mathbf{\Omega}_H$). The contribution of the following base kernel is therefore be weighted by this soft matching term, and quasi-zero if the considered edges are not close to each other.

4.2.3.3 Activation kernel

Finally, base kernel k_a is the heart of the functional pattern comparisons since it handles the functional activation information which carries the discriminative power for our classification task. This kernel measures the similarity of the activation levels recorded in parcels of g_{ij} and h_{kl}. As with k_g, we use a product of

Gaussian kernels:

$$k_a(g_{ij}, h_{kl}) = e^{-\|\Phi_i^G - \Phi_k^H\|^2/2\sigma_a^2} \cdot e^{-\|\Phi_j^G - \Phi_l^H\|^2/2\sigma_a^2}, \tag{4.6}$$

where $\sigma_a \in \mathbf{R}_+^*$ and Φ_m^G (resp. Φ_m^H) is the mth column of $\mathbf{\Phi}_G$ (resp. $\mathbf{\Phi}_H$). Using such kernel allows for variations in the activation attributes across subjects.

4.2.3.4 Resulting kernel.

With the definitions of k_s, k_g and k_a, we may define the resulting kernel:

$$K_{sga}(G, H) = \sum_{i,j=1}^{q_G} \sum_{k,l=1}^{q_H} k_s(g_{ij}, h_{kl}) \cdot k_g(g_{ij}, h_{kl}) \cdot k_a(g_{ij}, h_{kl}), \tag{4.7}$$

This kernel includes two parameters σ_a and σ_g, that are the bandwidths of the activation and geometrical base kernels. In standard learning problems working with vectorial inputs, a classical heuristic to estimate the value of the bandwidth of a Gaussian kernel consists in choosing the median euclidean distance between all observations in the training dataset [Scholkopf et al. 2001]. Here, we use this heuristic by choosing the median euclidean distance between activation and geometric (respectively) attributes of all nodes and all observations (i.e. all graphs) in the training set.

4.2.4 Datasets

4.2.4.1 Artificial data

The generative model described here creates artificial datasets that allows to precisely evaluate G-SVC. Also not designed to simulate patterns with a spatial organization as complex as in real data, the important point here is that these patterns contain variations across subjects with respect to two dimensions, the location of functional features and their activation levels, which makes these artificial datasets realistic for that matter and challenging for inter-subject learning algorithms. By parametrically choosing the amounts of variability along these two dimensions, we are able to study the robustness of G-SVC to such functional variability.

As illustrated on Fig. 4.5, the artificial patterns are created on a rectangular support (20×100 pixels) corresponding to a given ROI $\mathcal{R}$ in the brain. We therefore use the trivial coordinate system $\omega = (\omega_1, \omega_2) \in [1 \cdots 20] \times [1 \cdots 100]$. For a subject s, we directly simulate an observation i of the function ϕ as $\phi_i^s(\omega) = p_i^s(\omega) + n(\omega), \omega \in \mathcal{R}$. The *pixel noise* $n(\omega)$ is added to make the patterns more realistic; it is generated by drawing values from a normal distribution $\mathcal{N}(0, 1)$, which are then smoothed by a 2D Gaussian filter, with FWHM of 2.35 pixels. The

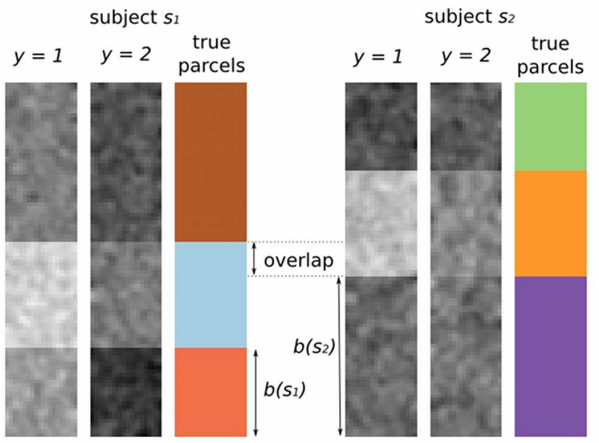

Figure 4.5: Artificial datasets: example patterns, with the true underlying parcellations used to generate the data. The inter-individual variability is controlled by two parameters: i) the locations of the middle parcels, which induces their overlap, here 33% (variability in geometric attributes); and ii) the value of σ_ϵ, which induces the amount of variation in the mean intensities observed in the parcels (variability in activation attributes). Although the parcels are matched across subjects by construction, the colors of the true parcels are chosen different in the two subjects to illustrate that G-SVC does not need an a priori match.

true underlying pattern p_i^s is a piecewise constant function given by

$$p_i^s(\omega) = \begin{cases} a(y_i) + \epsilon(s) & \text{if } b(s) \leq \omega_2 - 30 < b(s) + 30, \\ \epsilon(s) & \text{otherwise,} \end{cases} \tag{4.8}$$

These patterns are therefore composed of three rectangular parcels in a vertical layout and cover the full region. The top and bottom parcels are *not active* (activation level close to zero), while the middle one is *active*; for trial i, its level of activation $a(y_i)$ depends on the experimental condition y_i; two conditions were simulated, with $a(y_i) = 1$ or 2 respectively, producing two classes of patterns. The variability between the two simulated subjects s_1 and s_2 is introduced by i) changing the position of the middle parcel: we used $b(s_1) = 20$ in all cases and $b(s_2) \in \{20, 30, 40, 50\}$, resulting in different amounts of overlap (100%, 67%, 33%, 0%) of this parcel across subjects; and ii) adding a Gaussian random variable ϵ with distribution $\mathcal{N}(0, \sigma_\epsilon)$ to the activation levels of each parcel, using $\sigma_\epsilon \in \{0, 0.25, 0.5, 0.75\}$. This respectively induced variability in the location and intensity of the discriminant functional characteristics of these patterns, i.e. the middle pattern. With four levels for each of these two types of variability, we obtain sixteen quantified cases, hereafter called *variability cases*; for each of these, we generate twenty datasets (each corresponding to a random draw of

the values of ϵ in each parcel and each condition, for each subject) comprising fourty trials (ten trials per condition per subject, each trial being the result of a different realization of the pixel noise $n(\omega)$ on the full pattern).

4.2.4.2 Real data

In order to test our framework on real data, we need an fMRI dataset for which it is known that functional patterns exhibit strong inter-individual differences at fine spatial scale but present a consistent functional organization across subjects. With that goal in mind we here study data recorded in an experiment designed to study the functional organization of the human auditory cortex. The experimental protocol was approved by the local ethics committee of Aix Marseille Université (CCPPRB 01035), and written informed consent was obtained from all participants before the experiment. While the tonotopic property of the auditory cortex should result in a reproducible topographical organization of the activation maps across subjects, it has been shown that these functional maps suffer from a large variability across subjects [Humphries et al. 2010a; Formisano et al. 2003a]. Such dataset therefore represents an adequate challenge for our framework.

Data acquisition took place at the *Centre d'IRM Fonctionnelle de Marseille*. For each of the ten subjects, a T1 image was acquired (1mm isotropic voxels). Each stimulus consisted of a 8s sequence of 60 isochronous tones covering a narrow bandwidth around a central frequency ν. There were five types of sequences (i.e. five conditions $y \in [1, 5]$ corresponding to five classes of patterns), each one centered around a different frequency $\nu \in \{300Hz, 500Hz, 1100Hz, 2200Hz, 4000Hz\}$, with no overlap between the bandwidths covered by any two types of stimuli. Five functional sessions were acquired, each containing six sequences per condition presented in a pseudo-random order. Echo-planar images (EPI) were acquired with slices parallel to the sylvian fissure (repetition time=2.4s, voxel size=2x2x3mm, matrix size 128x128).

The preprocessing of the functional data, carried on in *SPM8* [Friston 2007], consisted of slice timing correction and realignment. Then, a generalized linear model was performed (using *nipy* [Brett et al. 2009]) with one regressor of interest per stimulus. The weights of these regressors (beta coefficients) were used to compute the inputs of the classifiers because they provide a robust estimate of the response size for each stimulus [Mumford et al. 2012]. We then used *freesurfer* [A. Dale 1999] to extract the cortical surface from the T1 image and automatically delineate the primary auditory cortex (Heschl's gyrus) as it is defined in the Destrieux atlas [Destrieux et al. 2010], thus obtaining two cortical ROIs $\mathcal{R}$ for each subject (one in each hemisphere). Note that this definition of $\mathcal{R}$, which implictly uses a spatial normalization, was chosen because it is fully automatic; other strategies working in the subjects' native spaces (manual drawing on the anatomy, functional definition) could also have been used.

In order to compute the graphical representations and apply our G-SVC framework, one need to define the function ϕ and a coordinate system ω. For this, we fully work in the subject's native space, i.e. without having to perform spatial normalization of the data into a common space. The function ϕ is the result of two operations executed in *freesurfer*: first, the beta maps obtained above are projected onto the subject's individual cortical mesh, and second, a slight spatial smoothing is performed along the cortical surface (equivalent FWHM of 3mm). The values of ϕ are then linearly normalized to the [0,1] interval within the ROI $\mathcal{R}$. Several examples (for different observations) of the values of ϕ within the region $\mathcal{R}$ are shown on Fig. 4.4. Furthermore, since Heschl's gyrus has a rectangular-like shape (with another region of the Destrieux atlas on each side), we can define a 2D local coordinates system through a conformal mapping of $\mathcal{R}$ onto a rectangle (with a local version of the method described in [Auzias, Lefevre, et al. 2013]), defining $(\omega_1(v), \omega_2(v),)_{v\in\mathcal{R}} \in [0,1]^2$. It is illustrated as a coordinate grid on Fig. 4.4. Note that forcing $\omega \in [0,1]^2$ separately for each subject allows dealing with the case where the size and shape of $\mathcal{R}$ is different across subjects.

4.2.5 Evaluation framework

In this section, we briefly describe the experiments that we perform, the state of the art methods chosen to benchmark our G-SVC framework, as well as the methodology used to compare the performances of the different algorithms.

4.2.5.1 Experiments

The first set of experiments consists in testing G-SVC and state-of-the-art vector-based methods on the artificial datasets. Since in these datasets, the amount of inter-individual variability is parametrically controlled along two dimensions (the location and activation levels of functional features in the pattern), studying the differential performances of G-SVC and benchmark methods for each *variability case* allows identifying the type(s) and amount(s) of functional variability for which G-SVC offers improved robustness. Then a series of experiments conducted on the real fMRI dataset makes it possible to i) overall compare the performances of G-SVC to those produced by state-of-the-art vector-based methods; ii) evaluate the influence of the number of graph nodes, compare the performances of G-SVC to those produced by standard parcel-based methods and examine the influence of working with individual vs. group parcellations; iii) test the robustness of G-SVC when its inputs are graphs with different numbers of nodes; iv) examine the usefulness of each of the three base kernels; and v) assess the stability in the estimation of the values of the kernel hyper-parameters.

4.2.5.2 State-of-the-art vector- and parcel-based methods

In order to benchmark our G-SVC framework, we compare its performances to state-of-the-art inter-subject classification methods. The standard strategy to solve such inter-subject problem is to obtain a point-to-point mapping across individuals for the spatial domain of interest. This indeed allows flattening a pattern into a feature vector that is used as input for the classification algorithms. In our artificial datasets, the rectangular regions are matched across subjects by construction. In the real experiment where the regions might be slightly different from subject to subject, such feature vector can be constructed using a spatial normalization procedure, i.e. by bringing the data from all subjects into a common standard space. We here used the surface-based normalization process available in *freesurfer*, which provides a vertex-to-vertex correspondance across all subjects in the common *fsaverage* space. The function ϕ that was defined in each individual subject's space (see section 4.2.4) is resampled into this common space, and its values within the ROI (defined as the intersection of all the individual regions projected in the common space) makes up the common feature vector. Several classification algorithms, chosen because they have shown to be efficient for MVPA, are then tested, each time with a large set of values for their respective hyper-parameters: *1)* linear SVC; *2)* nonlinear SVC, with Gaussian (with $\gamma \in \{2^{-n}\}_{n\in[0\cdots25]}$) and polynomial (of order $n \in \{2,3,4\}$) kernels; *3)* k-nearest neigbors (with $k \in \{3,5,7,9,15,20\}$); *4)* logistic regression with l_1 and l_2 regularization (with weight $\lambda \in \{2^n\}_{n\in[-5\cdots10]}$). The parameter sets were empirically selected to ensure capturing the optimal performances of each of these algorithms for all the experiments described above.

Moreover, we also tested parcel-based methods as benchmarks for G-SVC. Once the data is projected in the standard space described above, one can compute a group parcellation for all subjects by using the anatomo-functional parcellation algorithm described previously, but with functional input features coming from all available subjects at training time. The parcels are thus naturally mapped to one another across subjects, and we can use a feature vector composed of the mean activation within each parcel (equivalent to the graph activation attributes $\mathbf{\Phi}$) as input to any of the classification methods described above. Furthermore, using this group-parcellation, one can also construct graphical representations and use our kernel to perform inter-subject predictions; we denote this method as G-SVC$^{\text{g}}$, as opposed to G-SVC when using individual parcellations. Comparing G-SVC with the results obtained with G-SVC$^{\text{g}}$ and the other parcel-based benchmark methods was sued to clarify the role of using graphical representations learnt from individual parcellations and the usefulness of our graph metric itself.

4.2.5.3 Performance evaluation and algorithms comparison

Amongst the wide range of metrics available for measuring the performances of classification methods, we selected the global classification accuracy, i.e. the fraction of correct predictions amongst all attempted predictions (one reflecting a perfect prediction score). Indeed, the design of both the artificial and real datasets used in this study yield balanced classes (identical number of observations in each class and each subject), and it has been shown that the global classification accuracy is perfectly adapted to such case [Ferri et al. 2009]. For G-SVC, we report the global classification accuracy for different values of its hyper-parameters; for the benchmark methods, we report the highest accuracy obtained across all values of their hyper-parameters, thus putting G-SVC in the hardest possible case for performance comparison.

Since the different observations recorded in a given subject can be correlated, it is crucial to use a testing dataset composed of observations from subjects that were not part of the training dataset. We perform a leave-one-subject-out cross-validation and look at the average classification accuracy across folds, which is a natural strategy to measure the performance of an inter-subject classification algorithm. Assessing the significance of such an average classification accuracy and comparing different methods using the same scheme require a carefully elaborated evaluation method, that should for instance avoid employing Student's *t test* [Dietterich 1998]. A solution is to use non-parametric tests. Here, we focus on comparing algorithms, and we apply two permutation tests that allow estimating the distribution of the null hypothesis that the algorithms perform identically. The first test (hereafter called *test1*), described in [Menke et al. 2004], allows to compare the performance scores of two algorithms by generating random sign permutations of the paired performance differences. The second one (hereafter called *test2*), described in [Piater et al. 1998], is a randomized ANOVA that allows to compare curves describing the performance of several algorithms in function of the value of one hyper-parameters shared by the different methods.

4.3 Results

4.3.1 Results on artificial data: G-SVC vs. vector-based methods

In the artificial datasets, the true number of parcels composing the patterns is known by construction (three). Therefore we used the corresponding number of nodes, $q = 3$, in the graph construction phase of our G-SVC framework. We compared the results given by G-SVC with the performances of the different vector-based benchmark methods. For each *variability case*, we used *test1* to

assess whether G-SVC performs at a different level than each of the vector-based benchmark methods. The mean results (across the twenty datasets randomly generated for each *variability case*) are presented in Fig. 4.6.

When no geometrical variability is present (100% overlap of the middle parcels, lower plot on Fig. 4.6), all methods performed similarly. In these cases, the accuracy decreases when the variability in the activation levels increases, for all methods. This can be explained by the fact that when σ_ϵ increases, the discriminability of the patterns decreases; it is even possible that, depending on the sample values drawn for ϵ, the characteristic contrast (the fact that the middle parcel is more activated when $y = 1$ than when $y = 2$) becomes inverted in one subject compared to the other one; in this case, patterns from the two conditions in a given subject are not discriminable from what was learnt in the other subject.

When the level of geometrical variability increases (i.e. when the percentage of overlap of the middle parcels decreases), the accuracy levels of the vector-based method decreases, whereas the performance of G-SVC is not affected. Indeed, statistical differences (*test1*, $p < 0.05$) between G-SVC and all vector based methods are observed for $\sigma_\epsilon \in \{0, 0.25\}$ when the overlap is lower than 100%; for $\sigma_\epsilon = 0.5$, G-SVC also outperforms all methods for a 0% overlap, and some of the vector based methods for intermediate overlap levels (33% and 67%).

Note that we also conducted experiments with $q > 3$ for G-SVC, which in some cases produced slightly higher accuracy levels. But the mean differences with the results obtained with $q = 3$ (black curves of Fig. 4.6) were not significant; the equivalent curves were not distinguishable from the black ones, and are therefore not displayed.

Overall, we can conclude that G-SVC is the only method that deals with geometrical variability in the functional features of the patterns (i.e. in this case, when the middle parcels do not fully overlap), and that it handles variability in the activation levels as well as the vector-based methods.

4.3.2 Results on real data: G-SVC vs. vector-based methods

Since in this case, we do not know the exact number of underlying parcels to be used to model the patterns, we ran G-SVC with a fixed number of nodes q for all subjects, and repeated the analysis for $q \in \{5, 10, 15, 20, 25, 30, 35, 40\}$. We chose the smaller value of q to be five because according to the functional architecture of the primary auditory cortex, the five stimuli used in our experiment should result in at least five different activated regions [Humphries et al. 2010a; Formisano et al. 2003a]. Tab. 4.1 contains the performances of G-SVC vs. the vector-based benchmark methods. For the benchmark methods, the reported score is the highest accuracy obtained across all values of their hyper-parameters; for G-SVC, we report the highest and lowest accuracy levels obtained across all

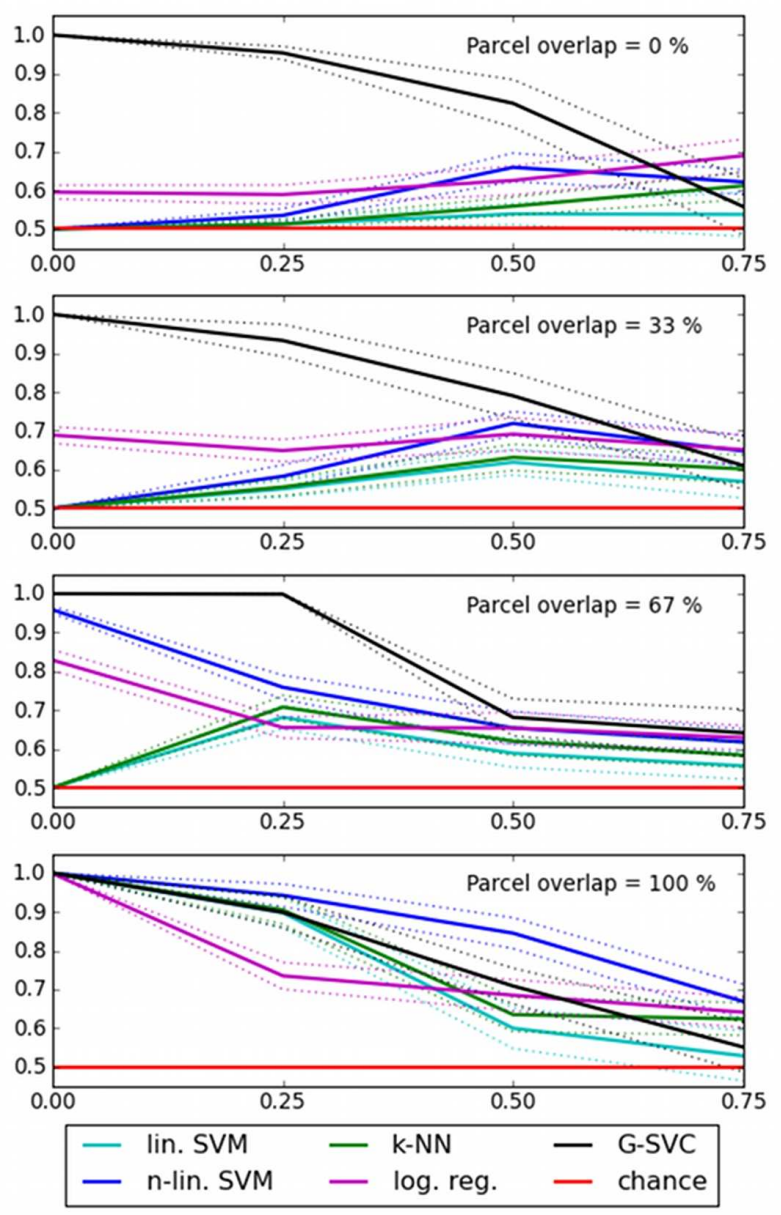

Figure 4.6: Artificial data: influence of the nature and amount of inter-subject variability on the mean accuracy (± standard error). Chance level is 0.5. From top to bottom, the variability in geometrical atributes (parcel location, which controls the parcel overlap) decreases. On the x axis of each subplot, the variability in activation levels increases from left to right. G-SVC is the only method that can handle geometrical variability (the four black curves are almost identical).

values of q. All vector-based methods performed fairly similarly, with accuracy levels between 0.27 and 0.31, i.e. slightly higher than chance level (0.2). G-

SVC yielded higher level of accuracies in all cases, with performances between 0.39 and 0.56, depending on q. In the right hemisphere, the performance differences between both the highest and lowest accuracies obtained with G-SVC and any benchmark methods is statistically significant (*test1*, $p < 0.05$). In the left hemisphere, the highest and lowest accuracies of G-SVC are significantly higher (*test1*, $p < 0.05$) than the best accuracies produced by linear SVM, and the l1- and l2-logistic regressions; the mean differences between the accuracy of G-SVC and the ones of nonlinear SVM and k-nearest neighbors are large (the lowest G-SVC score is 0.39, compared to 0.30 for nonlinear SVM and k-NN), but not significant.

	G-SVC	lin. SVC	n-lin. SVC	k-NN	log. reg.
right HG	**0.56 / 0.44**	0.31	0.30	0.28	0.28
left HG	**0.47 / 0.39**	0.27	0.30	0.30	0.28

Table 4.1: Real data: inter subject mean accuracy of G-SVC (highest / lowest, across q) vs. benchmark vector-based methods (best case). Chance level is 0.2.

4.3.3 Results on real data: G-SVC vs parcel-based methods

In this experiment, we compared the results given by G-SVC, for which an individual parcellation is computed on each subject to estimate our graphical representations, to the ones obtained with methods where the parcellation is identical in all subjects (G-SVC^{g} when using graphical representations computed from the group-parcellation, and other parcel-based benchmark methods). All methods share a common parameter, the number q of parcels. Fig. 4.7 shows the accuracy curves obtained with the different methods in fonction of q (the maximum and minimum values of the black G-SVC curves correspond to the values reported in Tab. 4.1). We used *test2* to assess whether the accuracy curve of G-SVC is different from the ones given by the benchmark parcel-based methods.

As clearly visible in Fig. 4.7, the accuracy curves of both G-SVC and G-SVC^{g} are above the ones given by all other methods; using G-SVC as reference, this difference is significant for all methods in the right hemisphere ($p < 0.05$); in the left hemisphere, the difference with the accuracy of the logistic regression is significant ($p < 0.05$), shows a non-significant trend with linear SVM ($p = 0.07$, but is not significant with the other methods. These results clearly demonstrate the added-value of using our graphical representations associated with our graph kernel to handle inter-subject variability. Furthermore, the accuracy curves of G-SVC are slightly above the ones of G-SVC^{g}. Even though this difference is not statistically significant, this trend might suggest that using individual parcellations

to construct our graphical representations yields more accurate representations of the underlying activation patterns.

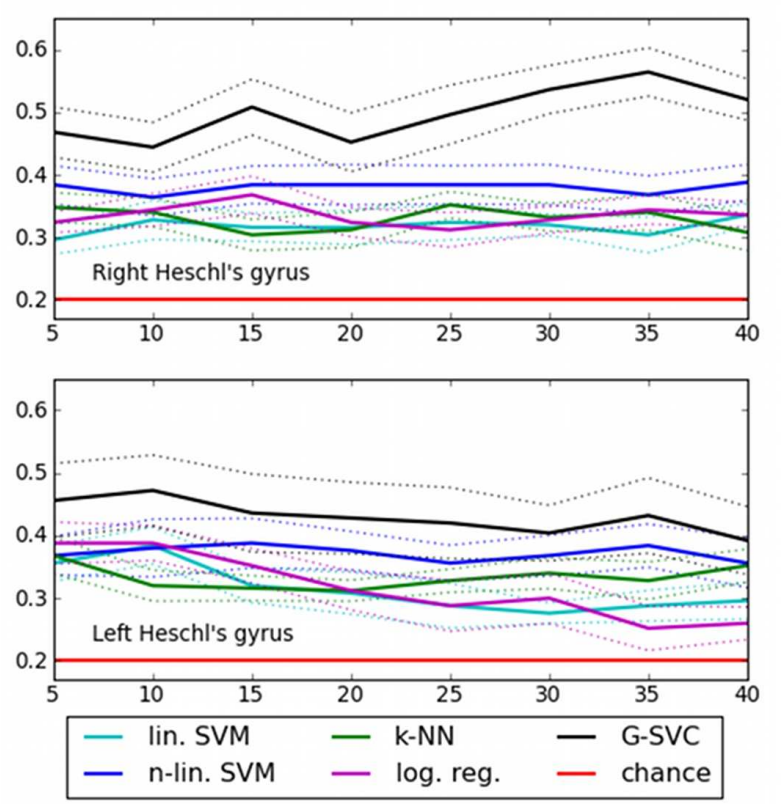

Figure 4.7: Real data: mean accuracy levels ($\pm$ standard error) as a function of the number of parcels q, for G-SVC and other parcel-based approaches. Chance level is 0.2. G-SVC clearly outperforms the other methods, and displays good robustness to the choice of the number of graph nodes.

4.3.4 Results on real data: G-SVC with variable number of nodes

We performed another set of experiments in order to test the ability of our G-SVC framework to learn from a set of graphs having different number of nodes. For each experiment, we randomly draw the number of nodes for each subject, in $\{5, 10, 15, 20, 25, 30\}$. Therefore, in each fold of the cross-validation, the training set contained graphs with different number of nodes, and the generalization power was measured on graphs from the left-out subject, i.e. with yet a different number of nodes. Twenty such experiments were conducted. The average accuracies over these twenty experiments were 0.47 ± 0.02 for the right Heschl's gyrus, and 0.42 ± 0.02 for the left Heschl's gyrus. These numbers are between the

highest and lowest accuracies obtained with fixed q for all subjects (see Tab. 4.1). In only 6 of the 40 cases (20 for each hemispheres) was the accuracy significantely lower than the highest one obtained with fixed q (*test1*, $p < 0.05$). This set of experiment shows that G-SVC can handle graphs with different number of nodes, and therefore is robust to some structural variation between the graphical representations learnt from different observations / subjects.

4.3.5 Results on real data: influence of each graph attribute

The graphical representations $G = (\mathcal{P}, \mathbf{A}, \mathbf{\Phi}, \mathbf{\Omega})$ used in our G-SVC framework comprise information about the activation levels, the location and the spatial structure of the functional features extracted from the patterns, respectively carried by the nodes attributes $\mathbf{\Phi}$ and $\mathbf{\Omega}$, and the adjacency matrix $\mathbf{A}$. Here we want to study whether these three types of features are necessary to achieve accurate classification. Similarly to the definition of our full kernel K_{sga} given in Eq. (5.4), one can define three kernels K_{sg}, K_{sa} and K_{ga} for which one of the three types of graph attributes is ignored by removing the corresponding base kernel from Eq. (5.4). For instance, $K_{sg} = \sum\sum k_s k_g$. Fig 4.8 shows the global classification accurary of G-SVC when using all features (i.e. using kernel K_{sga}) vs when using two out of the three types of features (i.e. using kernel K_{sg}, K_{sa} or K_{ga}). If one does not use the activation attributes (i.e. when using K_{sg}), keeping only the geometric and structural features, the performances are systematically at chance level, showing that, as expected, the activation attributes are necessary to achieve classification. If one does not use geometric or structural information (i.e. by using K_{sa} or K_{ga}), the mean performance curve is lower than when using all three types of information; this difference is statistically significant (*test2*, $p < 0.05$) in the right Heschl's gyrus, but not in the left one.

The construction of our K_{sga} kernel, described in section 4.2.3, explains the intuition behind the use of each of the three types of information. Here we have demonstrated experimentally that indeed, if any of the three base kernel is removed, the performances decrease. This shows that our full kernel K_{sga} indeed provides the best generalization results, which means that it exploits the information contained in all of the activation, geometric and structural attributes. This result also confirms that our main assumption, namely that the spatial organization of the activation patterns is consistent across subjects, is indeed true, and that our kernel allows exploiting this property efficiently to perform inter-subject predictions.

4.3.6 Kernel parameters

The kernel that we designed includes two hyper-parameters, the bandwidths σ_a and σ_g. We adapted a standard heuristic to estimate their value on the training set, which was used on all experiments. This process implies having different

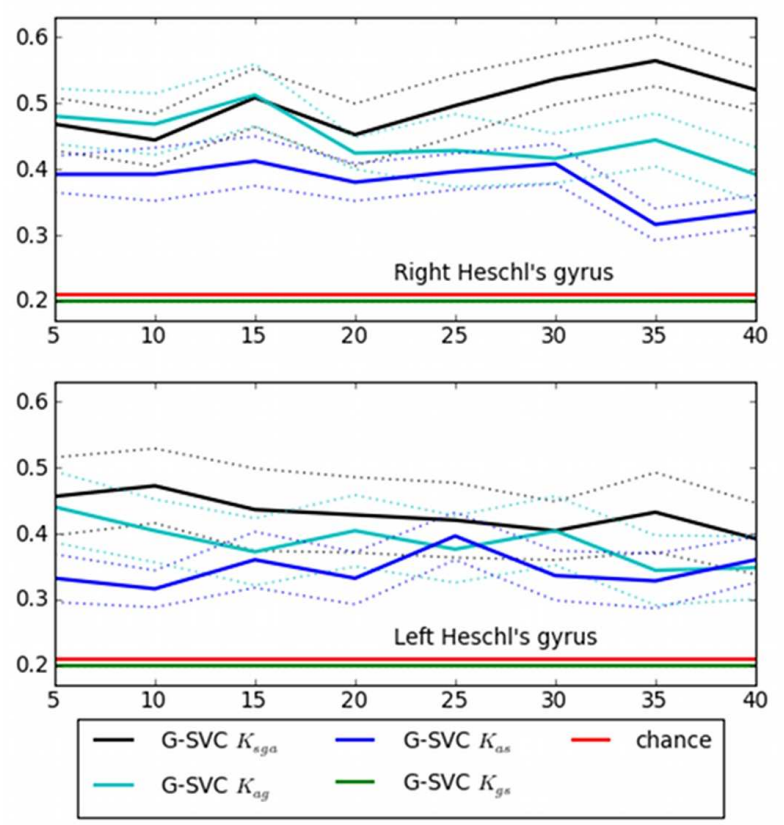

Figure 4.8: Real data: influence of the nature of the exploited graph characteristics. Mean accuracy ($\pm$ standard error) of G-SVC as a function of the number of nodes q. Chance level is 0.2. The best results are obtained when the three types of information are exploited, i.e. with the K_{ags} kernel.

values for σ_a and σ_g in each fold of the cross validation. We therefore computed the mean and standard deviation of the estimated bandwidths across folds of the cross validation, for all experiments performed on the real data with fixed q for all subjects. This yielded $\sigma_a = 0.391 \pm 0.003$ and $\sigma_g = 0.560 \pm 0.01$ for the analyses performed in the right hemisphere, and $\sigma_a = 0.392 \pm 0.002$ and $\sigma_g = 0.556 \pm 0.01$ in the left Heschl's gyrus. We observe that on the one hand, the estimated bandwidth values are very stable across folds, and on the other hand, they are almost identical in the right vs. the left hemisphere. These results combined with the overall positive results given by G-SVC validates the effectiveness of the heuristic that we proposed to choose the values of the kernel parameters.

4.4 Discussion

In this work, we have demonstrated that our G-SVC framework can be used to learn an accurate predictor to perform inter-subject classification of fMRI activation patterns. Experiments conducted on artificial datasets (presented in 4.3.1) showed that G-SVC is the only method that deals with varying locations of the functional features of interest across subjects. Experiments on a real dataset suffering from large inter-individual functional variability showed that G-SVC performed better that all tested vector- and parcel-based methods (respectively in 4.3.2 and 4.3.3); in particular, the latter proved the added-value of a graph-based framework based on individual representations compared to using a common parcellation for all subjects. We also showed that G-SVC is robust to changes in the number of graph nodes q, both if q is identical for all subjects (in 4.3.3) or not (in 4.3.4), the latter demonstrating the robustness of our kernel to some potential structural differences. Furthermore, we showed in 4.3.5 that G-SVC performs best when it uses all the information available in the graph (i.e. the activation, geometric and structural characteristics of the input functional patterns), which demonstrates the soundness of our learning scheme to produce group-invariant graphical representations and the ability of our graph kernel to exploit the consistency of the spatial organization of the activation patterns across subjects when this assumption is verified.

4.4.1 Hyper-parameters estimation

The G-SVC framework comprises three hyper-parameters: the bandwiths σ_a and σ_g of the functional and geometric base kernels used to design the full K_{sga} kernel, and the number of nodes q of the constructed graphs. For the former two, we have proposed a heuristic that estimates the values of σ_a and σ_g on the training dataset. This heuristic selects the median distance between the corresponding attribute values of the observations of the training set. It is simple to implement and the results shown in this paper demonstrate its effectiveness, thus providing an easy way to automatically select the values of these two hyper-parameters.

Regarding the number of graph nodes q, we hereafter describe three potential strategies for choosing q that explore three different directions. First, we have shown that when q is chosen identical for all subjects, G-SVC is robust with regards to the selected value since it significantely outperforms benchmark methods for almost all values of q (see Fig. 4.7). In order to choose the value of q, one can therefore exploit prior knowledge about the functional properties of the studied region, as we did in this study with the architecture of the primary auditory cortex [Humphries et al. 2010a; Formisano et al. 2003a]. Second, in paragraph 4.3.4, we also showed that G-SVC can handle input graphs with different number of nodes q. This suggests that one could attempt to work at the individual level to optimize the graphical representation of such distributed pat-

terns. One simple strategy to do so could be to apply a standard univariate analysis on each subject of the training set, and use the number of significantly activated clusters across all experimental conditions as a lower bound for q. Finally, if one need a fully automatic strategy, the value of q can be chosen in a nested cross validation amongst a given list of values that can be constructed using any of the aforementioned strategies.

4.4.2 Linear vs nonlinear classifiers

Since G-SVC uses a graph kernel to find a nonlinear decision boundary in the original data space, it is in fact a nonlinear SVM classifier. The usefulness of such nonlinear classifiers for neuroimaging applications is the subject of an ongoing debate in the litterature (see the introduction of [Rasmussen et al. 2011] for a summary). The appeal of linear classifiers for fMRI applications is mainly twofold: i) they facilitate the interpretation of the classification results, for instance thanks to the ability to directly visualize weight maps when working with linear SVM [LaConte et al. 2005]; and ii) despite their simplicity, their performances always are amongst the highest [Misaki et al. 2010]. Regarding the first point, [Rasmussen et al. 2011] offers a solution to ease the interpretation of the results given by nonlinear classifiers by visualizing sensitivity maps. As for the second point, our study is a new example where a nonlinear method clearly outperforms linear classifiers.

While it is not the focus of this paper, note that our framework is directly applicable to the equivalent within-subject learning task (see results in Tab. 4.2 in Appendix 4.6). In this easier task, G-SVC produces mean accuracy levels that are not statistically higher than those of vector-based methods.

Therefore G-SVC allows reaching higher accuracy rates than benchmark methods for the inter-subject classification task, but not for the within-subject one. We believe that it is the case because i) we have identified a factor that contributes to the poor performances of standard, and in particular linear, methods (the inter-individual variability) for the selected learning problem (inter-subject classification), ii) we use prior knowledge to model the influence of this factor and exploit it into the design of a nonlinear classifier (here, to construct a graph kernel), and iii) we allow the classifier to work in a fairly low-dimensional space (our graph construction scheme uses a small number of parcels, and therefore acts as a dimensionality reduction process), thus offering a reasonable ratio between the sample size (number of observations available) and the dimensionality of the space in which the classification is performed. These three points could constitute a set of rules that can help identify questions for which it might be worth developing nonlinear classifiers in neuroimaging.

Moreover, this interpretation is consistent with another explanation for the potential usefulness of nonlinear methods, that was for instance described in [Pereira et al. 2009]. Building upon the example of the quadratic kernel, which

is equivalent to adding features equal to the products (i.e. the *interactions*) between the original features, one could understand the added value of nonlinear models when such interactions are related to the experimental variable y to be predicted. In the case of our G-SVC framework, the graphical representation explicitely encodes some these potential interactions by linking spatially adjacent parcels: therefore, instead of using all interactions terms between original features (as with the quadratic kernel), only a subpart of these interactions are considered, those for which there exist an edge between graph nodes. Our G-SVC framework takes advantage of these interactions because the K_{sga} kernel directly uses the information carried by the graph edges.

4.4.3 Examining assumptions and potential applications

The G-SVC framework in its most generic form comprises an unsupervised learning step that constructs attributed graphs built upon a parcellation and a supervised learning step using a carefully designed kernel. It is therefore applicable as soon as it is possible to learn a meaningful parcellation for the problem of interest (classification, regression, clustering etc.). If one wants to extend the framework to other applications than the one described in the present paper, one simply needs to determine what is the information of interest in the parcels (in order to define the attributes of the graph nodes), what criterion to choose to build the graph structure (spatial adjacency, connectivity) and then to define one base kernel for each type of graph attributes.

As implemented in the present paper, our G-SVC framework relies on the main assumption that the spatial organization of the activation patterns is consistent across subjects. Indeed, our model of functional variability lets the intensity levels and locations of functional features vary across subjects, but their relative positions is supposed to be invariant across subjects, which we enforce by looking for region adjacency graphs that have a common structure. The question is then to know under what circumstances this assumption holds true. At the full brain scale, it is well known that the macroscopic organization of the cerebral cortex is reproducible across subjects, as for instance demonstrated by the reproductibility of the respective positions of the Brodmann areas, together with their functional specificity. We therefore believe that our G-SVC framework should be directly applicable to study large scale activation patterns based on full brain individual parcellations such as provided in *freesurfer* [Destrieux et al. 2010].

Studying neural representations at a finer scale is a crucial issue in modern neuroscience [Haxby, Gobbini, et al. 2001; Shinkareva, Malave, et al. 2012; Hanson et al. 2011; Haxby, Guntupalli, et al. 2011]. The topographical organization of primary sensory areas (see for instance [Formisano et al. 2003a; Humphries et al. 2010a] for the auditory cortex) ensures the consistency of the spatial organization of activation patterns across subjects. The successful results provided by our framework show that G-SVC is able to overcome the

large functional variability encountered at such fine spatial scale; furthermore, it constitutes yet another confirmation that fMRI is able to capture the spatial organization of the auditory cortex and it shows that its topography is indeed reproducible across subjects. Furthermore, it is to be noted that G-SVC yields higher classification performances in the right vs. the left Heschl's gyrus. This might reflect a lateralization in the functional specialization of the auditory cortex, such as described in [Tervaniemi et al. 2003], but such a claim would need further investigation.

In general, G-SVC is therefore a tool perfectly suited to study the consistency of representations in all sensory areas, and also the influence of a pathology on the organization of processing in these areas (see an example with macular degeneration in [Baker et al. 2005]). The question whether our main assumption is still valid in higher level cortical areas remains open, and the methods that attempt to deal with functional variability at such fine spatial scale take different routes with respect to this question. The so-called hyper-alignement (hereafter HA, [Haxby, Guntupalli, et al. 2011]), maps the activation patterns of different individuals to a common "high dimensional" space *without* enforcing the preservation of the spatial organization of the input patterns. Even if HA has proved successful in inter-subject decoding tasks, its success does not demonstrate per se the existence of idiosyncrasies, i.e. subject-specific architectures of the activation patterns at the scale offered by fMRI voxels. Indeed, finding common representations across subjects is a learning task that is simply easier to solve when relaxing the constraint on the spatial organization as in HA. Another method, the function-based alignment (hereafter FBA, [Sabuncu, Singer, et al. 2010]) estimates an explicit set of anatomical correspondences between brain voxels of different individuals by matching full-brain functional patterns. The use of a diffeomorphic model to estimate this spatial transformation implicitly requires the spatial organization of activation patterns to be consistent across subjects *over the whole cortex*; the positive results obtained by FBA would tend to validate this assumption, but it remains to be seen whether such method allows to improve prediction accuracy in inter-subject MVPA.

The successful results of the G-SVC framework described in the present paper indicate that it is possible to learn representations that have a common spatial organization across subjects and use them to perform inter-subject MVPA. An interesting feature of our framework is that it bypasses the need of explicit correspondence between brain voxels, which are hard to establish due to the conjunction of shape variability and variable functional organization of anatomical areas. When the delineation of a cortical area is so variable across individuals that it requires the use of functional paradigms known as *localizers* (see examples in [Pinel et al. 2007]), our G-SVC framework will allow studying fine scales representations within these functionally defined regions, thanks to its ability to work with regions that are not matched across subjects. Finally, contrarily to FBA and HA that estimate their respective models on a dedicated experiment

during which the subjects watched a movie, G-SVC does not require such dedicated "calibration" experiment to perform inter-subject predictions. These three methods are therefore somehow complementary, and we believe that comparing their behaviors and results should prove useful to assess the consistency of the spatial organization of brain patterns across individuals, but it is clearly beyond the scope of this paper.

4.4.4 Which graph kernel for fMRI graphs?

In the last two years, the use of graph kernels has emerged as a new tool to handle graphical representations estimated from fMRI data. We here try to summarize the different routes taken in the few published studies and provide directions that could help shape future work. We start by making a clear distinction between the different nature of the input data, namely whether the graphs have been constructed from connectivity (resting state fMRI) or activation (task-based fMRI) studies.

Indeed, connectivity graphs are most often constructed by thresholding a correlation (or other similar criteria) matrix [Richiardi et al. 2013], which makes them inherently unlabeled. In this context, one can expect that it is the topological properties of the graphs that carry most of the predictive power for the learning task at hand. This is the case when trying to characterize populations of patients for which the pathology has affected the connectivity of the brain, which is a major issue in clinical neuroscience. Most existing graph kernels make use of such properties, and when the effect of the pathology is global, one can therefore rely on the vast graph kernels literature that deals with unlabeled graphs. Note that the unifying framework described in [Vishwanathan et al. 2010] has allowed to demonstrate that a large number of previously designed graph kernels are actually instances of the R-convolution kernel family that we instanciated in the present paper. Furthermore, those kernels are often tunable to handle labeled nodes. If one need to add local information to better handle pathologies which result in focal, rather than global, connectivity dysruptions, an easy strategy consists in adding labels on the nodes, as done in [Jie et al. 2014]. In such case, the efficiency of the Weisfeiler Lehman kernel has made it popular in the emerging litterature of graph kernels applications for fMRI data [Jie et al. 2014; Vega-Pons and Avesani 2013; Vega-Pons, Avesani, et al. 2014; Gkirtzou et al. 2013].

In contrast, when working with task-based fMRI data, the predictive power is conveyed by the amplitude differences of the BOLD signal. The most natural way to encode this information in graphical representations is to derive activation features from the BOLD signal (for instance contrast maps estimated with a univariate GLM) and use them as real-valued attributes of the graph nodes. In that case, because the activation differences of interest are often very small, it is crucial to avoid any quantization or discretization of these nodes attributes.

It therefore becomes necessary to use graph kernels that handle real-valued attributes, for which the litterature is somewhat smaller. The most popular kernel in this categorty is the shortest path kernel [Borgwardt et al. 2005], which was successfully used in [Mokhtari et al. 2013]. In the present study, we took another route by constructing a dedicated kernel as yet another instance of the R-convolution kernel family, and demonstrated that such approach can yield accurate inter-subject predictions.

In all cases, fMRI graphs usually have a relatively small number of nodes (at most a few hundreds for graphs generated from full brain parcellations) compared to the more classical applications of graph kernels (world wide web networks, chemo- and bio-informatics) for which graphs often have thousands and sometimes millions of nodes. Even if this allowed us to focus our design scheme on the expressivity of the kernel, rather than its computational efficiency, it remains crucial to use kernels that scale up efficiently to a few hundreds of nodes. The complexity of the classical shortest path kernel is in $O(q^4)$, where q is the number of graph nodes. Our kernel scales up as $O(q^{2n})$, where n is the size of the considered subgraphs, which gives $O(q^4)$ for the implementation given in the present study (with edges as subgraphs, i.e for $n = 2$). Since all kernels scale up linearly with the number of examples, using such kernels in $O(q^4)$ might require several (dozens of) hours, depending on the number of examples. Therefore it remains important to improve the efficiency of the kernel computation. One could envision using recently developed kernels that deal with real-valued attributes and that are significantely more efficient than $O(q^4)$ [Kriege et al. 2012; Feragen et al. 2013]. In the case of our intuitively designed K_{sga} kernel, which compares all of the chosen types of subgraphs to all other subgraphs, one way to gain in efficiency would be to use the location of the nodes to limit the number of comparisons (i.e by comparing a given subgraph only to the ones located close-by). This should also improve its expressivity by avoiding meaningless comparisons of edges that are far away from each other. This strategy could also make it possible to use more complex subgraphs (triplets etc.) while maintaining affordable run times, although it remains to be investigated whether it would improve the prediction accuracy.

4.5 Conclusion

We described a new graph-based structured learning scheme designed to overcome inter-individual variability present in functional MRI data. Our approach constructs attributed graphs to represent distributed functional patterns and performs inter-subject classification to predict an experimental variable from the data. The graph construction scheme that we introduced starts with a parcellation, and then encodes the relevant characteristics of the parcels (their locations, activation levels and spatial organization) into the graph. The classification is

performed with support vector machines using a custom-designed kernel that exploits all the attributes of the graphs. Results on artificial datasets generated to parametrically control the amount of inter-individual variability along two dimensions (the location and intensity of functional features) showed that our G-SVC framework is the only method able to yield satisfactory performances when the locations of functional features vary across subjects. Results on real data showed that G-SVC outperforms both vector- and parcel-based state of the art classification methods, that it is robust to the number of graph nodes in the observations, both if it is chosen constant for all observations in the different subjects, or if it is different across subjects. As implemented in this paper for fMRI data, this framework is a tool of choice to study local neural representations at fine spatial scales in regions that are not well aligned across individuals, which is a crucial problem in modern neuroscience. Moreover, it is easily adaptable to other types of learning problems posed by different neuroimaging modalities.

4.6 Appendix - Within-subject G-SVC decoding results

Even though our G-SVC framework is designed to deal with inter-individual variability, it is directly usable in a within-subject analysis. We here present the results of G-SVC in the within-subject classification task, i.e to predict the class of stimulus that was presented to the subject for a given fMRI pattern. We used the same graph representations as for the inter-subject learning task and repeated the analysis for different number of nodes $q \in \{5, 10, 15, 20, 25, 30, 35, 40\}$. We used a leave-one-session-out cross-validation scheme and report the average global classification accuracy obtained across folds. The results are reported in Tab. 4.2, and compared to results given by the vector-based benchmark methods. The maximum accuracy levels obtained with G-SVC are slightly higher than those given by standard vector-based methods in both hemispheres, but these differences are not statistically significant.

	G-SVC	lin. SVC	n-lin. SVC	k-NN	log. reg.
right HG	0.56 / 0.52	0.51	0.46	0.42	0.50
left HG	0.57 / 0.51	0.48	0.42	0.41	0.51

Table 4.2: Real data: within subject mean accuracy of G-SVC (highest / lowest across q) vs. benchmark vector-based methods (best case). Chance level is 0.2.

4.7 Appendix - Testing pattern symmetry using G-SVC

The auditory cortex has been heavily studied in non human primates using a variety of techniques (see for instance [Morel et al. 1993]). This has produced a model of the spatial organization of the auditory cortex, mostly based on its tonotopic properties, i.e the fact that neigborhing neurons respond preferentially to neigborhing audio frequencies. This forms spatial gradients of low-to-high frequency preference in the recorded neural responses. Using electrophysiology, those gradients have been observed in the different parts of the auditory cortex, both within the central core region – that includes the primary auditory cortex – and its surrounding belt region.

The mapping of this spatial organization onto the human auditory cortex has been a subject of study for many years now. Using neuroimaging, numerous teams have studied the tonotopic organization of the human auditory cortex, as [Formisano et al. 2003b; Humphries et al. 2010b; Talavage 2003] and more. The outcome of these studies usually consists in a description of the observed low-to-high frequency gradients, with their cardinality and their main axis. This has produced results which might appear contradictory, with either one or several gradients, alongside or perpendicularly to Heschl's gyrus – the main anatomical landmark of the human auditory cortex. Although these ambiguities have mostly been raised through a finer description of this spatial organization using high field MRI [Formisano et al. 2003b] and innovative tonotopic mapping paradigms [Moerel et al. 2012], the question of the geometrical properties of the high-to-low gradients can also be examined by studying the pattern symmetry. We here propose a way to quantify the pattern symmetry using our G-SVC framework, by exploiting the fact that G-SVC directly uses the spatial organization of the input pattern.

If a pattern is symmetric along a given axis, then a classifier that would have been trained onto the original patterns of a training set should be able to generalize equally well on the original test patterns or on test patterns transformed by a flip along this axis. In practice, because the actual patterns will never been perfectly symmetric, there will always be a loss in the classification performance when testing on the flipped patterns, and we propose to use this loss as a quantification index of the pattern symmetry, the higher the loss, the less symmetry. By defining a_{orig} and $a_{flipped}$ as the accuracy level obtained by the classifier on the original test patterns and the flipped test patterns, and a_{chance} as the chance level, we define the following symmetry index:

$$I = \frac{a_{flipped} - a_{chance}}{a_{orig} - a_{chance}}$$

. If this index is 1, the patterns are perfectly symmetric along the chosen axis; if

it is 0, the patterns do not present any symmetry along this axis.

In the table below, we present results of inter-subject predictions within Heschl's gyrus when the test patterns have been flipped around four different axes, that we will designate by a symbolic angle:

- the axes that goes alongside the gyrus, which approximately follows the crest of the gyrus (0°);
- the first diagonal (45°)
- the axes perpendicular to the crest (90°)
- the second diagonal (135°)

We provide the classification scores obtained with G-SVC using 15 nodes in the graphs, and those (together with their symmetric index I in parenthesis) obtained when the test patterns are flipped

	true	flip (0°)	flip (45°)	flip (90°)	flip (135°)
right HG	0.464	0.372 (0.65)	0.328 (0.48)	0.324 (0.47)	0.252 (0.20)
left HG	0.424	0.324 (0.55)	0.316 (0.52)	0.324 (0.55)	0.252 (0.23)

Table 4.3: Inter-subject prediction accuracy for G-SVC with 15 parcels, with symmetry index in parenthesis when the test patterns are flipped. Left column: testing on the true patterns. Other columns: testing on flipped patterns. Chance level is 0.2.

These results show that there is a certain degree of symmetry in the frequency response patterns in Heschl's gyrus, alongside several axes. This is consistent with the low-to-high-to-low mirror-symmetric gradient reported in [Formisano et al. 2003b] alongside an axe going closely to parallel of the gyrus, as well as other interpretations given for instance in [Humphries et al. 2010b]. This demonstrates the validity of this approach, without providing a clear cut answer to the original question. A more detailed study would be needed, in particular using different regions of interest or a searchlight approach.

4.8 Appendix - Inter-region decoding using G-SVC

It is also of the highest interest to be able to test whether the functional organization of a region – hereafter ROI1 – is homologous to the one of another region – ROI2. We can address this question by training a classifier on patterns recorded in ROI1 and testing whether the classifier is able to generalize

to patterns recorded in ROI2. Unfortunately, this is difficult to achieve using standard classifiers because the input spaces of the patterns of the two regions are different. It would be possible by matching the two regions and performing a re-sampling of ROI1 to ROI2 (or the opposite), which implicitely requires a dense point-to-point matching of all the points of the two regions.

We here propose a solution that enables such analysis without having to re-sample ROI1 to ROI2 and that uses our G-SVC framework. Because within our framework, the regions are mapped onto a rectangular coordinate system, we only have to map the four corners of ROI1 to the four corners of ROI2, for instance using a priori knowledge on the anatomical organization of the cortex. Once this is done, the graphical representations of patterns of ROI1 and ROI2 do live in the same graph space and it is then totally straigthforward to train a classifier on patterns recorded in ROI1 and test its generalization ability on patterns in ROI2.

We here test this approach by performing inter-hemispheric decoding within the auditory cortex. We define ROI1 and ROI2 as being Heschl's gyri in the left and right hemispheres. We present in Tab 4.4 the results of such task in the inter-subject case, i.e the classifier is trained on ROI1 patterns from a set of subjects before being tested on ROI2 patterns of a different subject. Note that adding the inter-subject aspect on top of the inter-hemispheric one makes this task extremely challenging.

train \ test	right HG	left HG
right HG	0.464	0.292
left HG	0.352	0.424

Table 4.4: Inter-subject mean accuracy of G-SVC (15 parcels), including inter-hemispheric classification performances (off-diagonal terms), which both are above chance level (0.2).

The classification accuracy levels obtained in the inter-region cases (0.352 and 0.292) are both statistically above chance level, which validates G-SVC as being a adequate tool to assess the consistency of the functional organization across several regions. The loss of accuracy compared to the within-region performances (0.464 and 0.424) – which is also significant, can be attributed to the well known lateralization effect which has for consequence that the functional organization of the auditory cortex is indeed overall similar in the left and right hemispheres, but not strictly identical.

5 Mapping cortical shape differences using a searchlight approach based on classification of sulcal pit graphs

Contents

Context

The second contribution of this thesis introduces a new method that enables to identify local cortical shape differences between several populations of subjects. The objects considered are local patterns of sulcal pits – the deepest points of cortical sulci – modeled as graphs in order to deal with the inter-subject variability. A kernel-based mapping to a reprucing kernel hilbert space yields an implicit feature space that is common across subjects, where the classification of subjects within populations can be easily addressed. A searchlight mapping technique – a sliding window strategy that consists in repeating such classification analysis around all cortical locations – followed by a non-parametric statistical inference framework enables to perform the detection, i.e to localize the portions of the cortex that exhibit differences across the considered groups. We show that this framework makes it possible to detect cortical hemispheric asymmetries, as well as gender differences, within a large population of control subjects.

The contributors to this project were Lucile Brun (anatomical data processing), Guillaume Auzias (general methodological advice) as well as Olivier Coulon (general methodological advice). This work was published in the following conference paper:

S. Takerkart, G. Auzias, L. Brun, O. Coulon. "Mapping Cortical Shape Differences Using a Searchlight Approach Based On Classification of Sulcal Pit Graphs". *Proceedings of IEEE ISBI Conference [2015]*

On the day of the defense, a full length paper had been submitted two months earlier to the *Medical Image Analysis* journal with the same authors, under the title *Structural Graph-Based Morphometry (SGBM). A multiscale searchlight framework based on sulcal pits.* At this time when this manuscript is being finalized, this paper is now under revision.

5.1 Introduction

In the past few years, the topography of the cortical surface has raised a lot of interest, in particular to find biomarkers of pathologies [Im, Pienaar, Paldino, et al. 2012; Auzias, Viellard, et al. 2014] or to detect features associated with functional specificities [Z. Y. Sun, Klöppel, Rivière, Perrot, R. S. J. Frackowiak, et al. 2012]. Behind the large apparent variability of cortical folding patterns, a specific attention has been brought to the deepest part of sulci, either to elaborate theoretical models of cortical anatomy and development [Régis et al. 2005], or to automatically extract robust cortical landmarks [Auzias, Brun, et al. 2015]. For the latter, the work of Im has been particularly important in defining the concept of *sulcal pits*. Sulcal pits are defined as local maxima of depth within each cortical fold [Im, Jo, et al. 2010]. They can be extracted, together with their

associated sulcal basins, via a watershed algorithm performed on a sulcal depth map defined on the mesh of the cortical surface [Im, Jo, et al. 2010; Auzias, Brun, et al. 2015]. Sulcal pits have been linked with genetic factors [Im, Pienaar, Lee, et al. 2011] or developmental pathologies [Im, Pienaar, Paldino, et al. 2012; Im, Raschle, et al. 2015]

The challenge in using sulcal pits is to find stable patterns within a population, despite the apparent inter-subject variability. To do so, it is essential to have a good representation of folding patterns. Sulcal pits define sulcal basins that parcellate the surface of the neo-cortex (Fig. 5.1), and the adjacency of neighboring sulcal basins can be used to define sulcal pit graphs. Such representations of local anatomical patterns have been successfully analysed at the brain lobe level using a spectral graph-matching technique [Im, Pienaar, Lee, et al. 2011; Im, Pienaar, Paldino, et al. 2012].

In this work we introduce a multi-scale multivariate technique that enables to precisely localize differences between populations based on local patterns of sulcal pits. It relies on three main contributions: i) the design of graph kernel that allows the group classification to be performed directly in graph space; this kernel has very few parameters that can be efficientely inferred from the data; ii) the definition of a structural searchlight scheme that yields information maps estimated from patterns constructed at different spatial scales and iii) a non parametric multi-scale inference framework that enables the localization of spatially contiguous clusters of patterns that present statistically significant differences between populations. In what follows, we describe in details these three contributions. We then demonstrate the power of our framework on two classical brain mapping problems for which complex patterns of anatomical differences have previously been reported in the literature: the mapping of asymmetries between the left and right brain hemispheres and the detection of cortical shape differences between individuals of different gender.

What's a searchlight?

In this section, we set up a searchlight scheme aimed at mapping the discriminative power of local patterns formed by sulcal pits across populations. The so-called searchlight method, introduced by [Kriegeskorte et al. 2006] and only used so far for fMRI data analysis, consists in using a multivariate statistical model (e.g a classifier) in a sliding window that defines a local neighborhood and moves along the space of interest – the cortex. A summary statistic computed from this model (for instance the accuracy of a classifier) is then assigned to the center of the window, thus yielding a spatial map that allows the localization of the informative patterns.

5.2 Methods

5.2.1 Extracting sulcal pits

In order to localize the sulcal pits, we used a modified version of the procedure initially proposed in [Im, Jo, et al. 2010], designed to yield reproducible sulcal pits in every cortical region and not only in the deepest sulci [Auzias, Brun, et al. 2015], which is of major importance for the presently described method that will examine pits pattern centered around all cortical locations. First, we estimated the sulcal depth map for each subject using the depth potential function, a measure that integrates both curvature and convexity information [Boucher et al. 2009]. Then, we applied a *watershed by flooding* algorithm to extract depth maxima and their corresponding sulcal basins on the mesh (Fig. 5.1). A merging of sulcal basins was performed during the flooding in order to filter spurious extrema caused by noise in the depth map. Merging parameters were optimized as described in [Auzias, Brun, et al. 2015].

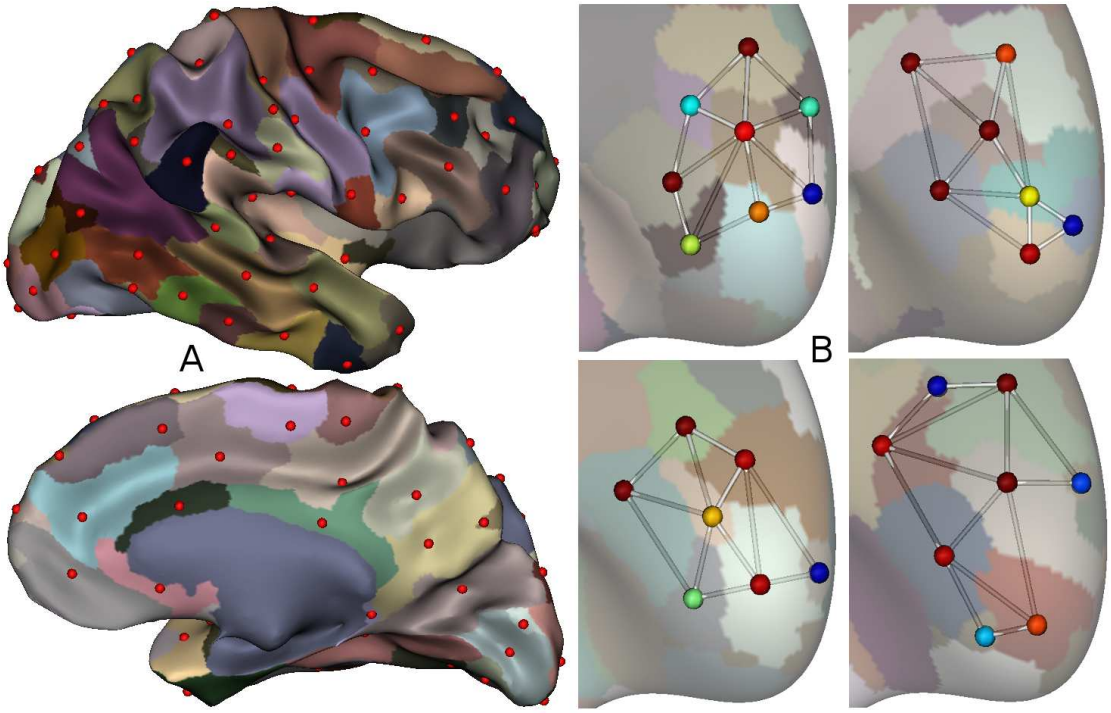

Figure 5.1: A. Illustration of the sulcal pits and their basins on one subject. B. Examples of sulcal pit graphs in the right frontal lobe for four subjects (the color of the nodes encodes the relative depth of the pit, from red (deep) to blue (shallow).

For each subject s, we obtain a set of pits $\Pi^s = \{\pi_i^s\}_{i=1}^{i=N^s}$ and their basins $\{\beta_i^s\}_{i=1}^{i=N^s}$ ($\forall i$, π_i^s is the deepest point of β_i^s). This set of sulcal basins actually forms a complete parcellation of the cortical surface. Also note that the number of pits N^s can vary across subjects.

5.2.2 Representing patterns of sulcal pits as graphs

A natural way to formally represent a pattern formed by a set of pits, including their spatial organization, is to construct a graph. We here use the method proposed in [Im, Pienaar, Lee, et al. 2011], which consists in building a region adjacency graph [Pavlidis 1977] using the set of basins associated with each pit.

For a given subset $\bar{\Pi}^s \subset \Pi^s$ of M pits ($M \leq N^s$), we define a node of the graph for each pit $\pi_i^s \in \bar{\Pi}^s$. The graph edges are then given by the spatial adjacency of their associated basins: this defines a binary adjacency matrix $\mathcal{A} = (a_{ij}) \in \mathbb{R}^{M \times M}$ ($a_{ij} = 1$ if B_i and B_j are adjacent, and 0 otherwise), that encodes the spatial organization of neighboring pits. In order to better characterize the pattern of pits, we add two attributes to each graph node: i) its depth d_i, because it is an intrinsic characteristic of the pit, and ii) the coordinates X_i of the corresponding vertex on the sphere after surface based-alignment to a common template performed in *freesurfer*, so that we can compare the locations of the graph nodes across subjects. Let $\mathcal{D} = \{d_i\} \in \mathbb{R}^M$ be the vector of depth values and $\mathcal{X} = \{X_i\} \in \mathbb{R}^{M \times 3}$ be the matrix of coordinates of all graph nodes.

Any pattern of sulcal pits can therefore be fully represented by an attributed graph defined as $G = (\bar{\mathcal{P}}, \mathcal{A}, \mathcal{D}, \mathcal{X})$. Examples of sulcal pit graphs are shown on Fig. 5.1.

5.2.3 Graph-based support vector classification

Our first contribution is to introduce a new graph kernel that measures the similarity between two graphs $G = (\bar{\mathcal{P}}^G, \mathcal{A}^G, \mathcal{D}^G, \mathcal{X}^G)$ and $H = (\bar{\mathcal{P}}^H, \mathcal{A}^H, \mathcal{D}^H, \mathcal{X}^H)$. Equipped with such a kernel, one can use the *kernel trick* to perform classification directly in graph space using a support vector machine (see for instance [Scholkopf et al. 2001]). Similarly to previous work dedicated to inter-subject multi-voxel pattern analysis of functional MRI [Takerkart, Auzias, Thirion, and Ralaivola 2014], we designed a kernel that exploits all the features of the sulcal pit graphs defined in 5.2.2.

Kernel definition. Our kernel compares all pairs of nodes (i.e potential edges) g_{ij} and h_{kl}, respectively in G and H, without trying to perform a one-to-one pit-matching as in [Im, Pienaar, Lee, et al. 2011]. As such, it belongs to the class of walk-based graph kernels [Gärtner et al. 2003] and uses the most elementary walks, of length one. It combines (see eq. 5.4) the different features of the graphs by using several sub-kernels within the convolution kernel framework [Haussler 1999]. A first sub-kernel aims at ensuring that the comparisons are performed only if g_{ij} and h_{kl} are actual edges. This is done with the linear kernel on the binary entries of the adjacency matrices:

$$k_a(g_{ij}, h_{kl}) = a_{ij}^G . a_{kl}^H, \tag{5.1}$$

which takes the value 1 if $a^G_{ij} = a^H_{kl} = 1$ and 0 otherwise. A second sub-kernel uses a product of gaussian kernels on the coordinates of the nodes of g_{ij} and h_{kl}:

$$k_x(g_{ij}, h_{kl}) = e^{-\|X^G_i - X^H_k\|^2/2\sigma_x^2} \cdot e^{-\|X^G_j - X^H_l\|^2/2\sigma_x^2}. \tag{5.2}$$

In practice, k_x acts as a spatial filter that weights the comparisons of edges with their proximity, thus eliminating the comparisons of edges that are far away from eachother and allowing for inter-subject variability (if edges are close, but not perfectly matched across subjects). Finally, the last sub-kernel compares the depth attributes using the same principle:

$$k_d(g_{ij}, h_{kl}) = e^{-\|d^G_i - d^H_k\|^2/2\sigma_d^2} \cdot e^{-\|d^G_j - d^H_l\|^2/2\sigma_d^2} \tag{5.3}$$

The full kernel is defined as the combination of the three sub-kernels applied on all pairs of nodes of G and H:

$$K(G, H) = \sum_{i,j=1}^{M_G} \sum_{k,l=1}^{M_H} k_a(g_{ij}, h_{kl}) \cdot k_x(g_{ij}, h_{kl}) \cdot k_d(g_{ij}, h_{kl}) \tag{5.4}$$

Note that the number of nodes M_G and M_H in G and H can be different.

We then perform the following normalization procedure:

$$\tilde{K}(G, H) = K(G, H)/\sqrt{K(G, G)K(H, H)}, \tag{5.5}$$

This normalization ensures that for any graph G, $\tilde{K}(G, G) = 1$ (i.e that the diagonal terms of the Gram matrix are equal to one), which will enable an easier interpretation of the results by computing median graphs (see 5.2.6).

Estimating hyper-parameters. Our graph kernel has two hyper-parameters, which are the two bandwidths σ_x and σ_d of the gaussian sub-kernels that respectively act on the coordinates and the depth features. In order to choose the values of these parameters, we use heuristic which is an extension of the standard practice used with vector inputs and that consists in selecting the median euclidean distance between all observations in the training dataset. The extension used for such graphical representations, that we proposed [Takerkart, Auzias, Thirion, and Ralaivola 2014], consists in choosing for σ_x and σ_d the median euclidean distance between the coordinates and depth attributes of the nodes in all graphs available at training time. The values of these parameters are therefore estimated directly and easily from the training dataset, using this heuristic which was proved efficient [Takerkart, Auzias, Thirion, and Ralaivola 2014].

5.2.4 Searchlight mapping

In this section, we set up a searchlight scheme aimed at mapping the discriminative power of local patterns formed by sulcal pits across populations. The so-called searchlight method, introduced by [Kriegeskorte et al. 2006] and only used so far for fMRI data analysis, consists in using a multivariate statistical model (e.g a classifier) in a sliding window that defines a local neighborhood and moves along the space of interest – the cortex. A summary statistic computed from this model (for instance the accuracy of a classifier) is then assigned to the center of the window, thus yielding a spatial map that allows the localization of the informative patterns. More specifically, in order to construct a searchlight scheme for a given task, one needs to define the five following items

- the space of interest,
- the spatial sampling strategy of this space,
- how to define local patterns at each location,
- the statistical model that addresses the task itself,
- the summary statistic to be mapped onto the space of interest.

In what follows, we instantiate step by step each of these five items in order to fully define our pit-based searchlight method. This constitutes the first *structural searchlight* scheme, which allows studying the morphology of the cortex by finding differences in local folding patterns.

5.2.4.1 Searchlight space

Since sulcal pits are defined as local depth extrema on the cortex, it is natural to perform our searchlight strategy along the cortical surface (similarly to what was done in [Y. Chen et al. 2011] when analyzing fMRI patterns). Specifically, we use the interface between the grey matter and the white matter, as identified by *freesurfer*, as our space of interest. In order to obtain a match across subjects, we use *freesurfer*'s cortical registration algorithm, which consists in aligning the curvature information of the cortices of individual subjects onto a template subject (*fsaverage*). This process uses an inflation of the cortex onto a sphere (because the topology of the cortical surface of each hemisphere of the brain is homological to the one of a sphere), projects the curvature of the cortex onto this sphere before deforming the mesh (while keeping its sphericity) so that the curvature information is matched across subjects. For each subject s, the triangulation of the final sphere has been preserved throughout this process (inflation + spherical registration); each vertex v^s of the sphere directly corresponds to the vertex that has the same number on the original cortical mesh for this subject. We can therefore use the unit sphere as a common space for all subjects.

5.2.4.2 Spatial sampling

In the standard searchlight scheme applied to fMRI data [Kriegeskorte et al. 2006], the brain volume is sampled in a dense manner by examining patterns center around ALL brain voxels. In our case, because the size of the sulcal patterns of interest are far larger (by at least two or three orders of magnitude) than the triangles composing the cortical mesh, it is not necessary to sample the cortex using each vertex of the mesh. We therefore need a coarser set of points that are evenly distributed on the cortex. Because we need those points to be matched across subjects, a simple solution is to define them on the unit sphere that serves as the common space. However, there is no perfect solution to define evenly distributed points on the sphere. We therefore resort to a process that yields pseudo-evenly distributed points on the sphere by regularly sampling the golden spiral defined on the sphere [Niederreiter et al. 1994]; this provides a set of Q points $\{(x_q, y_q, z_q)\}_{q=1}^{q=Q}$ which is also called the spherical Fibonacci point set. The choice of the value of Q is detailed later and the resulting sampling is illustrated on Fig. 5.2.

5.2.4.3 Defining local patterns

For a given subject s and one of the searchlight locations q corresponding to the vertex v_q^s, our goal is here to describe the spatial organization of the sulcal pits present in a surrounding neighborhood. Several options are available to define such neighborhood around the vertex v_q^s:

- finding the closest sulcal pit from v_q^s and including its 1-neighbors (i.e the pits whose basin are directly adjacent to the basin of the central pit), its 2-neighbors, ..., its n-neighbors;
- selecting the k nearest pits from v_q^s;
- including the pits that are located within a given distance r from v_q^s.

In all three cases, there is a parameter (n, k, or r) that directly controls the size of the neighborhood, i.e the spatial scale of the pattern. In the two latter options, a spatial distance is required to define this neighborhood. We can use the geodesic distance on the original cortical mesh of the subject, or work directly on the sphere and in this case using the euclidean distance is adequate. Using the geodesic distance on the original cortical mesh is an appealing solution because it takes into account the true geometry of the cortex; but it is fairly complex to compute and it is influenced by the overall brain size, which is not desirable when using it across subjects. The euclidean distance applied in spherical space is more simple to compute, it is directly comparable across subjects, and even if it is influenced by the distortions introduced by the spherical registration, it still allows to respect the true cortical anatomy to a reasonable degree because

the mapping to the sphere and the spherical deformations are estimated while attempting to minimize the metric distortions [Fischl et al. 1999].

For simplicity reasons, and because it is a widely used strategy when defining searchlight mapping frameworks, we will define the neiborhood of vertex v_q^s as all the points located within a radius r of the vertex v_q of the sphere. In particular, we are interested in the set of pits of subject s situated within a radius r of v_q, and we name this set $\mathcal{P}_{q,r}^s$. Given this set of pits, we can directly apply the graph construction scheme described in 5.2.2 to obtain the attributed graph that, for subject s, location q and radius r, we note $G_{q,r}^s$.

5.2.4.4 Statistical model

Equipped with our graph kernel defined in 5.2.3, we can use any kernel method to address a wide range of problems including regression, clustering or classification. In this study, we will address two supervised classification tasks to find gender differences (males vs. females) and asymmetries (right vs. left hemisphere). We will therefore estimate a non linear Support Vector Classifier using our graph kernel and assess its generalization power by measuring the classification accuracy on some data that had not been used to train the classifier. In practice, for a given location q and a given value of r, we have a fully labeled dataset $\{(X_s, y_s)\}_{s=1}^{s=S}$ at our disposal, with $X_s = G_{q,r}^s$. In order to estimate the average accuracy of the model $\bar{a}_q^r$, we resort to a 10-fold cross-validation because it is known to offer a good estimate of the true accuracy [Kohavi 1995].

5.2.4.5 Mapping a point-wise summary statistic

We now need to define an appropriate summary statistic to define an information map over the full cortex using this statistical model. We could directly use the average classification accuracy $\bar{a}_q^r$ – as done in [Y. Chen et al. 2011; Stelzer et al. 2013] with searchlight applications to fMRI data, but it presents two major drawbacks that we would like to circumvent: i) it is drawn from an unknown distribution, which implies that it is difficult to perform direct inference (for instance by thresholding), and ii) its value does not weight the classification scores according to their rareness (i.e having $\bar{a}_q^r = 0.97$ might be five times more rare than $\bar{a}_q^r = 0.96$, but the differences in the values of $\bar{a}_q^r$ is very small). It will become more clear in what follows (5.2.5) why we would like to avoid these two undesirable properties.

We therefore adopt the following strategy to define an adequate point-wise statistic:

- first, we resort to a permutation scheme, as advocated in neuroimaging in general by [Bullmore et al. 1999; Nichols et al. 2002b] and more particularly in statistical analysis of searchlight information maps by [Kriegeskorte

et al. 2006; Stelzer et al. 2013], in order to estimate the cortex-wise distribution of $\bar{a}_q^r$ under the null hypothesis H_0 of no differences between classes;

- then, knowing the p-value p_q^r that corresponds to $\bar{a}_q^r$, we compute the normal statistic that corresponds to the same p-value, which we note z_q^r; this transformation is in fact given by the so-called Inverse Error Function $erf^{-1} : [0,1] \rightarrow \mathbb{R}$ [Wikipedia 2015], i.e we have $z_q^r = erf^{-1}(p_q^r)$.

We therefore know the empirical distribution of z because it is – by construction – the same as the distribution of the accuracy score which was estimated in the first step. And thanks to the non linear transformation performed in the second step, this z statistic favors more strongly the high rare values than the original average classification accuracy; therefore, the z map will show an enhanced contrast compare to the original accuracy map. These two properties make it an adequate statistic to build our information map that is well suited for the spatial inference framework described hereafter.

We describe below (Algorithm 1) the algorithm used to compute our z information map (in which we can temporarilly drop the r index because its value does not change throughout the course of the algorithm).

Algorithm 1: Computing searchlight information maps for a given r

Input : a labeled dataset $\{(X_s, y_s)\}_{s=1}^{s=S}$
Input : a set of T permutations $\{\Gamma_t\}_{t=1}^{t=T}$, Γ_0 being the identity
foreach *searchlight location* q **do**
  foreach *permutation* t **do**
    compute $\bar{a}_t(q)$ from the re-labeled dataset $\{(X_s, y_{\Gamma_t(s)})\}_{s=1}^{s=S}$
The empirical null distribution is $p(u) = \frac{1}{Q \times T}\, card(\{(t,q) \text{ s.t. } u \leq \bar{a}_t(q))\})$
foreach *searchlight location* q **do**
  foreach *permutation* t **do**
    $z_t(q) = erf^{-1}(p(\bar{a}_t(q))$
Output: a set of T z-maps Z_t

This yields the true information map $Z = \{z_0(q)\}_{q \in [1,\cdots,Q]}$ – because the first permutation Γ_0 is the identity, i.e the labels are the true labels – as well as a set of $T-1$ information maps Z_t computed with permuted labels. Note that in order to enable the use of the cluster-based inference techniques described hereafter, it is necessary to maintain the local spatial dependency in the information maps computed with permuted labels. This is ensured by using the same set of permutations for all cortical locations q (see 5.2.5), as done in the algorithm above.

5.2.4.6 Choosing the number Q of searchlight locations

In order to choose the size Q of this point set, we need to fullfill two criterions: on the one hand, Q needs to be as small as possible in order to minimize the computational cost of the entire framework (which increases linearly with Q); on the other hand, Q needs to be large enough to fully capture the spatial processes at hand. In order to assess this second condition, let us consider the following qualitative reasoning. Moving from a location q of the sphere to another close by location q' will result in a discrete change in the dataset: a certain number of subjects will see their surrounding sulcal pattern change; others will not. If the distance between q and q' is small enough, the number of subjects for which there will be a change will be very small compared to the total number of subjects S. This should result in a very small change from $\bar{a}_q$ to $\bar{a}'_q$, which means that $\bar{a}$, and therefore z are continuous functions with respect to q. Consequentely, one should choose Q so that the distance between two neighboring locations is small enough to preserve this intrinsic continuity; this should be the case if this distance is at least an order of magnitude smaller than the mean distance between neighboring sulcal pits. In practice, since the number of pits per hemisphere is in the order of 100, choosing $Q = 2500$ ensures this. This is illustrated on Fig. 5.2.

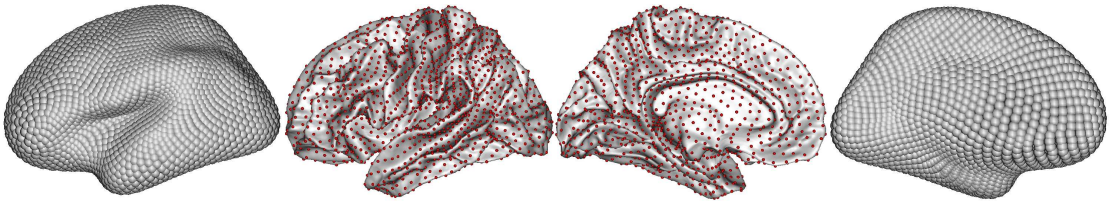

Figure 5.2: Illustrations of the $Q = 2500$ searchlight locations used in our study on the left hemisphere of the *fsaverage* template subject. Each location is represented by a sphere. Internal: small red spheres overlaped on the folded mesh. External: larger spheres localized on the inflated cortex (the underlying mesh of the inflated cortex is not visible).

5.2.5 Multi-scale spatial inference

5.2.5.1 Preliminary

The searchlight framework described above depends on one hyper-parameter: the radius r of the neighborhood in which we define the patterns of sulcal pits. If fixed, this parameter is crucial both in terms of getting enough detection power (if set to a wrong value, we might miss an effect) and also when it comes to the interpretation of the results. However, we have very little $a\ priori$ knowledge on how to choose the value for r: the few published studies examined pits either

over large portions of the cortex (such as brain lobes [Im, Jo, et al. 2010; Im, Pienaar, Paldino, et al. 2012]) or within much smaller regions (single sulcal basins [Auzias, Brun, et al. 2015]). Since both these approaches made it possible to detect significant differences, it seems necessary to study the influence of the radius r, which can be done by embedding our searchlight framework within a multi-scale strategy.

We are therefore faced with two questions:

- at which location(s) of the cortex is there a significant difference between populations?
- at which spatial scale(s) does such effect live?

Now, consider increasing the value of the radius by a small amount from r to $r + \Delta r$ at a given location q. It is possible that for a small number of subjects, a few extra sulcal pits might enter the neighborhood with $r + \Delta r$ on top of those already present with r , but we expect the pit-graphs to stay identical for most subjects. Consequentely, the average classification accuracy $\bar{a}$ and its associated z-score should change only by a small amount. We therefore expect a continuous behavior of $\bar{a}$ with respect to the value of r, which means that z_q^r should be fairly smooth along the scale dimension r. This also means that an observable effect (i.e a location at which the classification score is high) at a given scale should also be observable at closed by scales.

We therefore have to deal with a smooth behavior of z_q^r with respect to both the scale parameter r and the spatial location q. Assessing the statistical significance of the measured statistic over the cortical surface across scales therefore requires an adequate strategy to cope with the multiple comparisons problem that would occur if independently testing at all q and r. An appealing way to deal with the spatial correlation is to resort to cluster-based statistics [Poline et al. 1993]. In that case, it is also possible to correct for the multiple comparisons problem in a non-parametric fashion [Bullmore et al. 1999]. The schematic of such strategy is the following:

- an arbitrary threshold is applied to a statistical map to form contiguous clusters of supra-threshold locations;
- a cluster-wise statistic is defined for each cluster;
- corrected cluster-wise p-values can be obtained by estimating the distribution of the maximum cluster statistics across the full cortex, over a set of maps obtained by permuting the labels of the observations.

In what follows, we introduce two such strategies based on these principles and that deal with the multi-scale nature of the question we have at hand. Both

start from a set of information maps $\{Z^r\}_{r\in\mathfrak{R}}$, where $\mathfrak{R} = \{r_1, \cdots, r_\rho\}$ is a predefined set of discrete scales of size ρ. The first one examines the information map at each individual scale to form single-scale clusters and compute their p-value with a correction for multiple comparisons across space and scales. The second one first construct a multi-scale information map from the full set $\{Z^r\}_{r\in\mathfrak{R}}$, which is then thresholded to obtain multi-scale clusters, for each of which we compute the p-value with a correction for multiple comparisons across space.

Both strategies share the need to define a cluster-wise statistic. The most usual cluster-wise statistic used in the literature is the *cluster size*, which simply consists in counting the number of elements included in each cluster (see [Poline et al. 1993] for the original application in neuroimaging and [Stelzer et al. 2013] for a more recent work using a searchlight framework). Using the cluster size would naturally favor large clusters. Furthermore, it is empirically known that multi-scale methods tend to favor coarser scales [Lindeberg 1993], where spatially larger effects can be expected. In order to counter balance for this, we therefore use the *cluster mass* ([Bullmore et al. 1999; H. Zhang et al. 2009]) – i.e the sum of supra-threshold point-wise statistics, instead of the cluster size, as our cluster-wise statistic. Thanks to the high-contrast information maps built with our z statistic (see 5.2.4), a small cluster with high z-values should be associated with a more significant p-value than if the cluster size had been chosen. Also note that this is a reason why we built the point-wise z statistic; indeed, if we had used the accuracy score instead, this benefitial effect of the cluster mass – as compared to the cluster size – would have been strongly reduced because of the lesser contrast of the accuracy maps.

5.2.5.2 Single-scale clusters with corrections for multiple comparisons across space and scales

The first strategy consists in defining clusters independently at each scale (i.e for each value of r) and perform a correction for multiple comparisons that take into account the repetition of tests both in space and across scales to assess the statistical significance of these clusters. The key here is to estimate the null distribution of the maximum – across clusters of *all* scales – cluster mass though a permutation scheme. It is by taking the *maximum* statistic across clusters estimated at *all* scales for a given permutation of the labels that we are able to correct for the multiple comparisons problem in the multi-scale context [Nichols et al. 2002b]. Here are the details of the algorithm that we propose to compute the corrected – in space and across scales – probabilities that the null hypothesis

of no differences between the populations can be rejected.

Algorithm 2: Computing corrected p-values for single-scale clusters

Input : a set of $T \times R$ spatial maps $\{z_t^r(q)\}$ computed from datasets with permuted labels at all scales
Input : a threshold τ
foreach *permutation* t **do**
  foreach *scale* r **do**
    form cluster sets $C_t^r = \{c_1^{t,r} \dots c_{L^{t,r}}^{t,r}\}$ from thresholded map $z_t^r(q) > \tau$ ($L^{t,r}$ is the number of clusters);
    foreach *cluster* $c_l^{t,r} \in C_t^r$ **do**
      compute cluster mass $m_l^{t,r} = \sum_{q \in c_l^{t,r}} z_t^r(q)$;
    $M^{t,r} = \{m_1^{t,r} \dots m_{L^{t,r}}^{t,r}\}$;
  compute maximum mass$M_t = max_r(max_l(m_1^{t,r} \dots m_{L^{t,r}}^{t,r}))$
foreach *scale* r **do**
  foreach *true cluster* $c_l^{0,r}$ **do**
    compute cluster mass $m_l^r = \sum_{q \in c_l^{0,r}} z_0^t(q)$;
    compute corrected p-value $p_l^r = \frac{1}{T}\, card(\{t \text{ s.t. } M_t \geq m_l\})$
Output: a set of clusters $\{c_l^{0,r}\}$ with their corrected p-values $\{p_l^r\}$

5.2.5.3 Multi-scale clusters with corrections for multiple comparisons across space

Our second strategy consists in agglomerating the results obtained across scales into a single statistic that summarizes the results obtained across all scales. To construct such a statistic, we combine two observations. First, it seems intuitive to look for the maximum statistic across scales at a given searchlight location, which would provide both a statistic value and an estimate of the *best* scale. Secondly, because of the expected smoothness across scales, it seems meaningful to examine the average statistic across a given number of consecutive scales. We therefore propose to use the following statistic:

$$\mathfrak{Z}^R(q) = \max_{r \in \mathfrak{R}} \left(\frac{1}{R} \sum_{r',\, |r-r'| \leq R/2} z^{r'}(q) \right), \tag{5.6}$$

where R is the number of consecutive scales considered.

For a given value of R – index that we temporarily drop in the algorithm below, we describe the details of the method that makes it possible to compute corrected p-values – in space – at the cluster level in the following algorithm. We

note $\mathfrak{Z}_t(q)$ the map obtained with the t-th permutation Γ_t.

Algorithm 3: Computing corrected p-values for multi-scale clusters

Input : a set of T maps $\{\mathfrak{Z}_t(q)\}_{t=1}^{t=T}$ computed from datasets with permuted labels
Input : a threshold τ
foreach *permutation* t **do**
  form clusters $\{c_1^t \dots c_{L^t}^t\}$ from thresholded map $\mathfrak{Z}_t(q) > \tau$ (L^t is the number of clusters);
  foreach *cluster* c_l^t **do**
    compute cluster mass $m_l^t = \sum_{q \in c_l^t} \mathfrak{Z}_t(q)$;
  compute maximum mass $M_t = max(m_1^t \dots m_{L^t}^t)$
foreach *true cluster* c_l^0 **do**
  compute cluster mass $m_l = \sum_{q \in c_l^0} z_q(0)$;
  compute corrected p-value $p_l = \frac{1}{N}\, card(\{t \text{ s.t. } M_t \geq m_l\})$
Output: a set of clusters $\{c_l^0\}$ with their corrected p-values $\{p_l\}$

Furthermore, at any location q within a significant cluster, we can define the *best* scale as the value of r that maximised $\frac{1}{R}\sum_{t=0}^{R-1} z_q^{r+t}$) for the true labels:

$$\bar{r}^R(q) = \arg\max_{r \in \mathfrak{R}} \left(\frac{1}{R} \sum_{r',\, |r-r'| \leq R/2} z^{r'}(q)\right), \tag{5.7}$$

This quantity $\bar{r}$ should be useful for the interpretation of the results as described in the following.

5.2.6 Interpretation-aiding visualization tools

For a given cluster, we are interested in visualizing the features that might have contributed to the effect detected by our spatial inference framework. This is a difficult task because of the complex nature of the objects of interests: indeed, the information that contributed to make this two sets of graphs significantely different might be distributed in a non trivial manner along two dimensions: first spatially (to what region correspond each graph?), and secondly cross-sectionally among the population. In this section, we introduce visualization tools to explore these two dimensions. First, we propose a method to estimate the spatial extent of the region that contributed to this significant difference. Second, we estimate the most representative subject as the geometric median in order to visualize typical graphs associated with each class. The two tools should give us some insights about the differences between classes.

5.2.6.1 Probabilistic density maps of sulcal basins associated with a cluster

In order to provide some interpretation on the nature of the sulcal patterns that contributed to a significant cluster, it is important to know how far these patterns extend over the cortex. The question is not trivial because the pit-graphs were defined in each individual subject's space from the subject's sulcal basins. Therefore, we first use the spherical registration to the template subject to *bring* these individual locations into the common spherical space of the template subject *fsaverage*. Then, for a location q within a significant cluster – for instance the center of mass of the cluster – we count at each cortical vertex v the number of subjects s for which v is part of one of the sulcal basin β_i^s for which the pit π_i^s is closer than the best scale radius ρ. The density map $D(v)$ is then given by

$$D(v) = card(\{\text{subect } s \text{ s.t. } \exists i \text{ s.t. } v \in \beta_i^s \text{ and } \|q - \pi_i^s\|^2 < \rho\}) \tag{5.8}$$

This density map can then be projected on the folded mesh of the template subject *fsaverage* in order to visualize and localize its extent, as shown on the set of figures available in Section 5.3.3.

5.2.6.2 Kernel-based geometric median graph

In order to visualize the graphs that are the most representative of each class, we propose to use the geometric median, which is a generalization of the simple median that is only defined for one-dimensional data. The geometric median provides a typical value (i.e a typical graph in our case) for a distribution sampled by a finite set of points. Given a distance metric d, the general definition of the geometric median x of a set of points $\mathcal{X} = \{x_1, \cdots, x_N\}$ is the following:

$$x = \arg\min_{x \in \mathcal{X}} \sum_{x' \ in \mathcal{X}} d(x, x') \tag{5.9}$$

Given a kernel K that implements some similarity measure, a standard way to define a distance is given in [Phillips et al. 2011]:

$$d_K(x, y) = K(x, x) + K(y, y) - 2K(x, y) \tag{5.10}$$

Assuming that the kernel is normalized, i.e that $\forall x, K(x, x) = 1$, this directly transposes the definition of the geometric median into

$$x = \arg\max_{x \in \mathcal{X}} \sum_{x' \ in \mathcal{X}} K(x, x') \tag{5.11}$$

Following the notations defined in Section 5.2.4, $G_{q,r}^s$ is the pit-graph of subject s at location q for a searchlight radius r. For a given radius r at a location q, we temporarily drop the indices q and r to only keep G^s. For a given class c, let us

define $\mathcal{X}_c$ the set of subjects s of class c. We can then define the median graph G_c^s for the class c using our normalized kernel $\tilde{K}$ (defined in Eqs. 5.4 and 5.5) by looking for the subject s so that

$$s = \arg\max_{s \in \mathcal{X}_c} \sum_{s' \in \mathcal{X}_c} \tilde{K}(G^s, G^{s'}) \tag{5.12}$$

We are then able to compute such representative graph for each class and visualize them to facilitate the interpretation of the results offered by our framework (see in Section 5.3.3 for examples of such median graphs).

5.3 Experiments

5.3.1 Mapping gender and hemispheric differences

The Open Access Series of Imaging Studies[a] (OASIS) cross-sectional database offers a collection of 416 subjects aged from 18 to 96. For each subject, three to four individual T1-weighted MP-RAGE scans were obtained on a 1.5T Vision system (Siemens, Erlangen, Germany) with the following protocol: in-plane resolution = 256x256 (1 mm x 1 mm), slice thickness = 1.25 mm, TR = 9.7 ms, TE = 4 ms, flip angle = 10°, TI = 20 ms, TD = 200 ms. Images were co-registered and averaged to create a single image with a high contrast-to-noise ratio. From this database, we selected two groups of 67 male and 67 female healthy right-handed subjects, aged 18 to 34, matched in age, intra-cortical volume and total cortical surface.

With this large dataset, we studied two different questions by examining local patterns of sulcal pits to characterize cortical morphology. The first one was to map gender differences and the second one to examine cortical asymmetries, i.e differences between the right and left hemispheres[b] (in the males subjects only). At each of the Q locations of the searchlight, a binary classification problem (males vs. females, or right vs. left) was solved using the graph-based classifier defined previously. The average classification accuracy $\bar{a}$ was estimated using a 10-fold cross-validation that ensured class balance in both the training and test datasets for each fold. As with any cluster-based analyses – which represent the vast majority of spatial inference methods used in neuroimaging, the initial cluster-forming thresholding has an influence on the final results. With this in mind, we first repeated our analysis with different threshold values and retained the one that offered clusters which a good compromise between sensitivity and specificity. The results shown hereafter have been obtained with a cluster-forming threshold of $\tau = 3.090$, applied on the z

[a] www.oasis-brains.org

[b] Note that for studying the cortical asymmetries, an extra pre-processing step was carried on using a symmetric cortical template, following [Greve et al. 2013]

information maps, which corresponds to a thresholding at $p < 0.001$ on the uncorrected p-values of our the classification accuracy. Note that the same threshold can be applied at all scales thanks to the properties of z statistic which were induced by construction (see Section 5.2.4). We used the following set of scales in order to study the multi-scale properties of local sulcal patterns: $\mathfrak{R} \doteq \{30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90mm\}$. The values of the smallest and largest scales were chosen to approximately yield patterns of the size of a gyrus and of a brain lobe, respectively.

5.3.2 Results: methodological considerations

All the cortical maps shown in this section take the same format. Each of the $Q = 2500$ searchlight locations used as the center of a neighborhood for the searchlight procedure is represented by a small sphere. Even though these 2500 points are defined on the unit sphere, we display their locations on a partially inflated cortex in order to maintain the possibility to localize them on the cortex, as previously shown on Fig. 5.2. Each sphere receives a color that encodes the information of interest (z-value, corrected cluster p-value, etc.). Note that amongst the full range of scales used to perform spatial inference, we only display the maps for the subset $\{30, 40, 50, 60, 70, 80, 90mm\}$ without any loss of significant results (read below).

For the gender problem, each map is presented under four views, two per hemisphere: the left hemisphere is shown on the left, and the right on the right, with respectively their external and internal faces at the top and the bottom. For the asymmetries, only the left hemisphere is presented, with respectively its external and internal faces at the top and the bottom.

5.3.2.1 Qualitative assessment of single-scale results

In order to assess the results provided by our searchlight framework, one first qualitatively observe the raw searchlight information maps obtained for different values of the radius r. These z-maps are presented on Fig. 5.3 for the gender problem and on Fig. 5.4 for the asymmetry problem. These point-wise statistical maps show contrast, i.e they contain zones of high and low z-values, at all scales and for both problems; furthermore, the regions of high z-values often reach values with $z > 3$ and sometimes reach $z = 5$, which corresponds to very significant uncorrected point-wise p-values ($p < 0.001$ and sometimes $p << 0.001$). This clearly indicates that our framework is able to extract some information about the problems at hand. More specifically, this shows that i) looking at the spatial organization of sulcal pits using a graphical representations is relevant (confirming this result shown by [Im, Pienaar, Lee, et al. 2011]), ii) our graph kernel is an adequate similarity measure to look at such graph (thus offering an alternative to the metric described by [Im, Pienaar, Lee, et al. 2011]), and iii) our z statistic is

well suited to serve as an information measure in the aforementioned searchlight procedure. Furthermore, the points showing high z-values are grouped together, which is consistent with the expected smoothness of the searchlight maps (as explained in Section 5.2.4) and confirms that a value of $Q = 2500$ is large enough to observe such clustered behavior. The spatial smoothness of the maps and the sizes of the groups of searchlight locations with high z-values increase with the value of r, which are also expected behaviors typically observed in searchlight procedures.

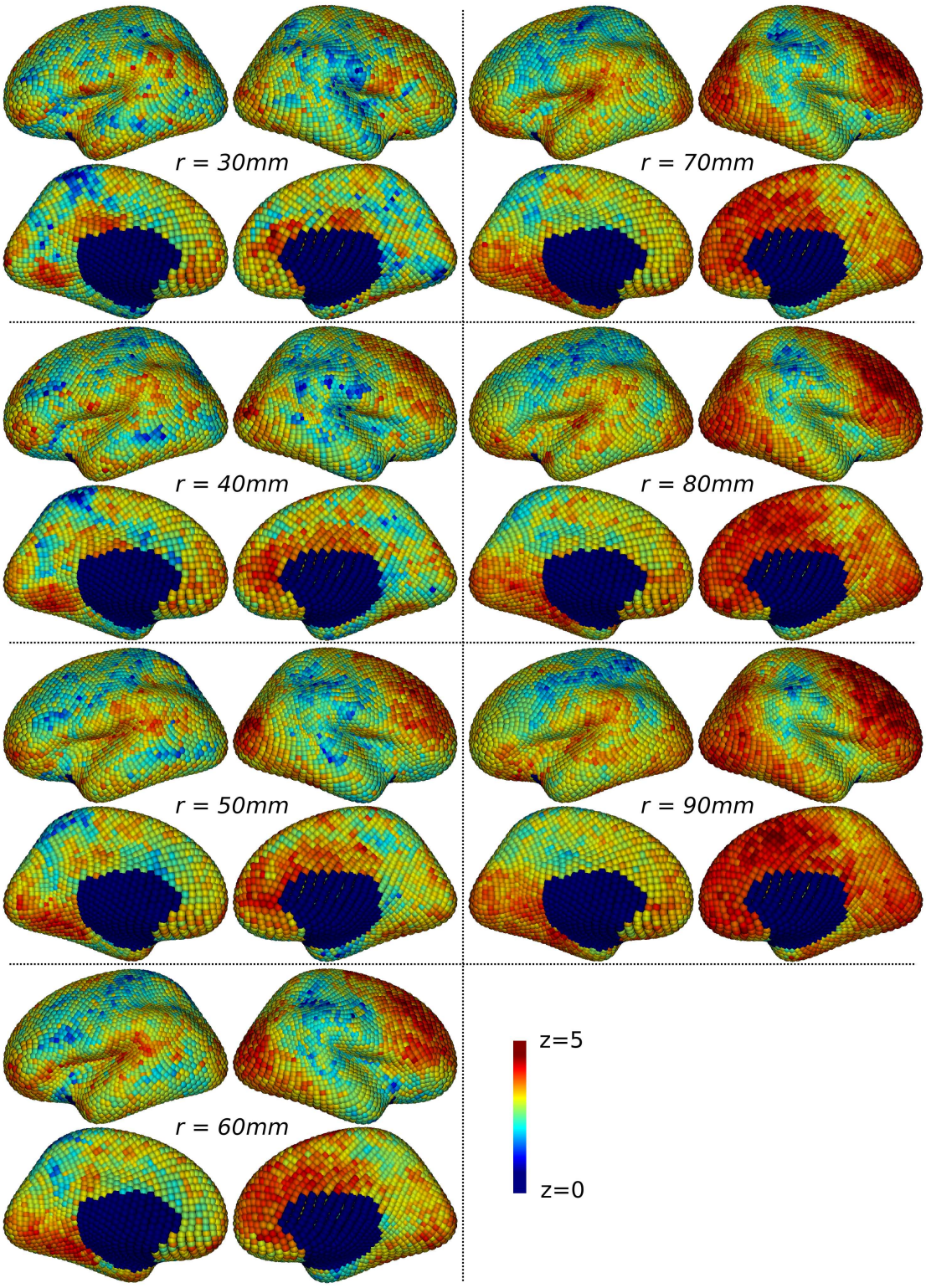

Figure 5.3: Gender differences. Maps of single-scale z-scores $z^r(q)$ for $r \in \{30, 40, 50, 60, 70, 80, 90mm\}$.

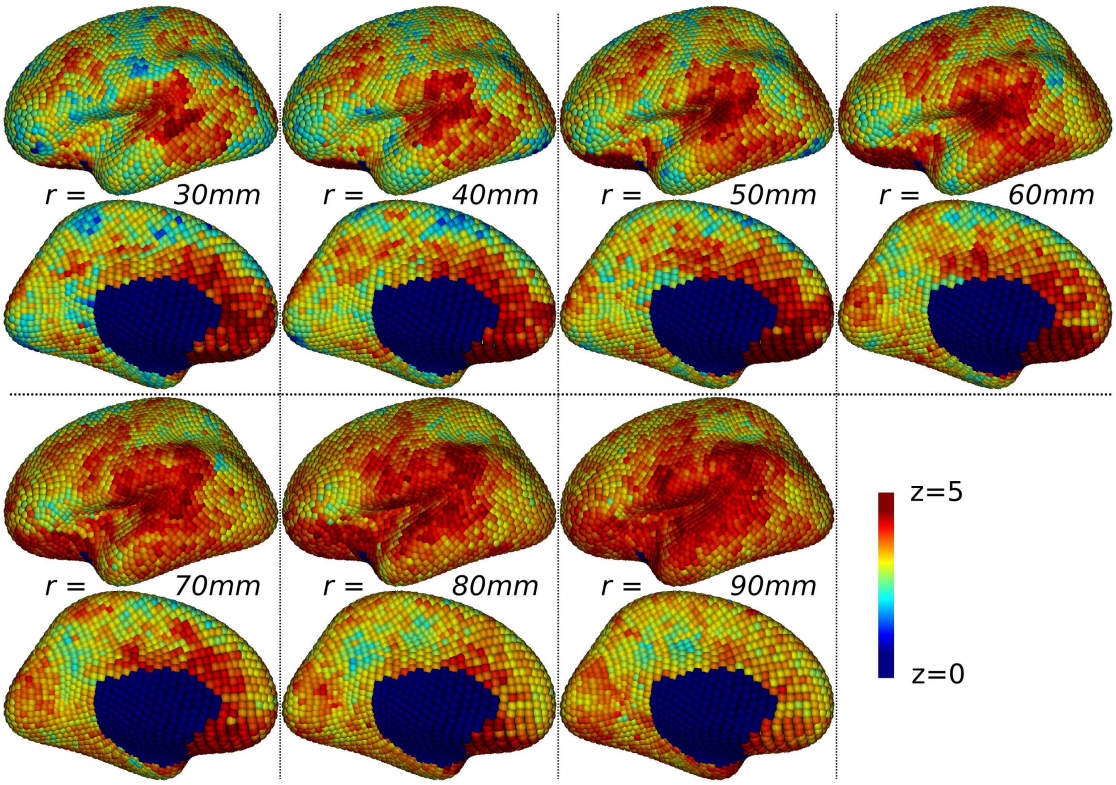

Figure 5.4: Asymmetries. Maps of single-scale z-scores $z^r(q)$ for $r \in \{30, 40, 50, 60, 70, 80, 90mm\}$.

We then examine the single-scale clusters. Figs. 5.5 and 5.6 present the clusters which present a p-value lower than 0.05, after correction for multiple comparisons across space and scale with the method described in Section 5.2.5.2. Each cluster is colored with its corrected p-value, with a color map ranging from yellow to red (the more red the more significant). For both the gender and asymmetry problems, our framework detects significant clusters at all scales with r at least $40mm$, demonstrating its detection power. Almost all clusters are persistent across scales, i.e for a given cluster at scale r, there exist clusters at nearby cortical locations for nearby scales. The clusters detected at finer scales seem overall smaller than the ones at larger scales, but some small clusters do exist at large scales. This demonstrates that our cluster-based spatial inference strategy is effective; in particular, the use of the cluster mass associated with your z statistic makes it possible to detect clusters at fine spatial scale and small clusters at larger scales, which is known to be difficult in such multiscale schemes [Lindeberg 1993].

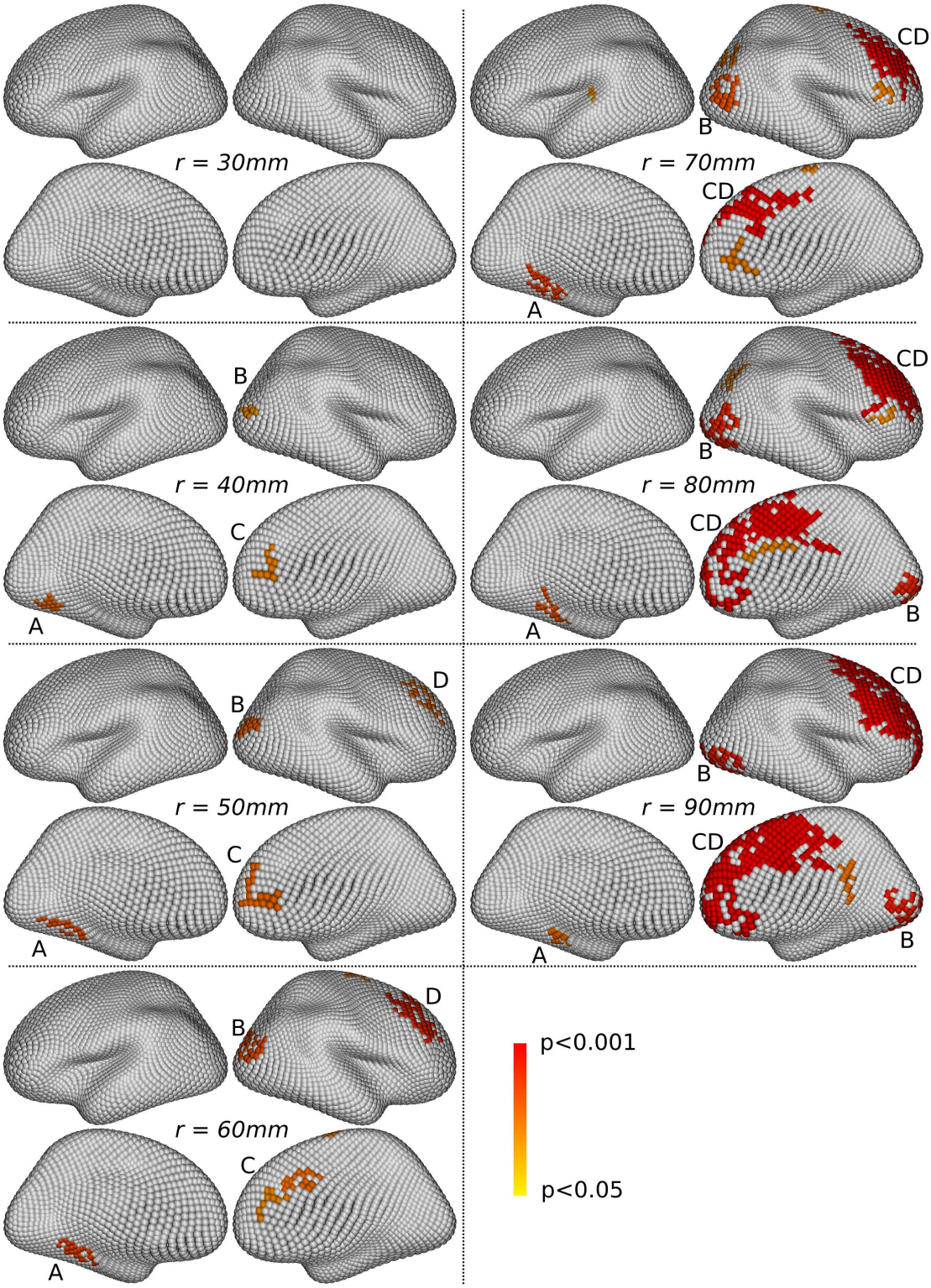

Figure 5.5: Gender. Significant single-scale clusters (corrected $p < 0.05$), with $r \in \{30, 40, 50, 60, 70, 80, 90mm\}$. Four main clusters, tagged A, B, C and D, appear to be persistent across scales, sometimes after having merged.

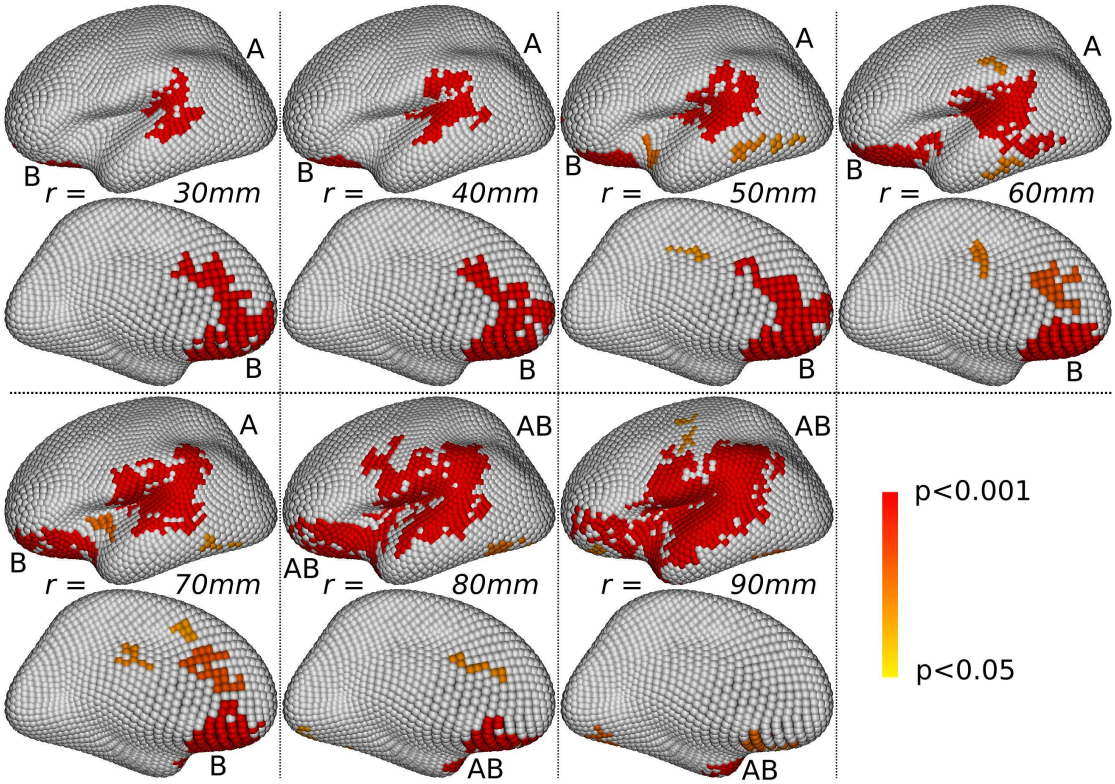

Figure 5.6: Asymmetries. Significant single-scale clusters (corrected $p < 0.05$), with $r \in \{30, 40, 50, 60, 70, 80, 90mm\}$. Two main clusters, tagged A and B, appear to be persistent across scales, sometimes after having merged.

5.3.2.2 Qualitative assessment of multi-scale results

Let us now examine the raw maps of the multi-scale statistic $\mathfrak{Z}^R$ defined in Eq. 5.6. Because this statistic depends on the parameter R, the number of consecutive scales over which the original z^r statistic is averaged, we will first study the results for different values of R. The $\mathfrak{Z}^R$-maps are presented on Figs. 5.7 and 5.8 for $R \in \{1, 3, 5, 7, 9, 11, 13\}$: for $R = 1$, $\mathfrak{Z}^R$ is the max of all z^r across all scales r; for $R = 13$ (13 being the total number of scales used), $\mathfrak{Z}^R$ is the average of z^r across all scales r. Overall, these maps are smoother than the single-scale z^r-maps, which is expected because the construction of the multi-scale statistic $\mathfrak{Z}^R$ consists in a smoothing operation across scales. For the smaller values of R, the contrast of the statistic map is weaker than the one offered by single-scale maps: some very large portions of the cortex show high values of $\mathfrak{Z}^R$, which means that the specificity offered by such low values of R will be limited. This was in fact expected because in that case, $\mathfrak{Z}^R$ carries the effects that exist at any scale, and because we have seen with the single-scale maps that different effects

that are localized across the cortex can exist at different scales. For larger values of R, the maps regain contrast, but the high-$\mathfrak{Z}^R$ values seem to be weaker: this is also expected because in that case, $\mathfrak{Z}^R$ reflects the effect that exist over a large number of consecutive scales, which gets more difficult when R increases.

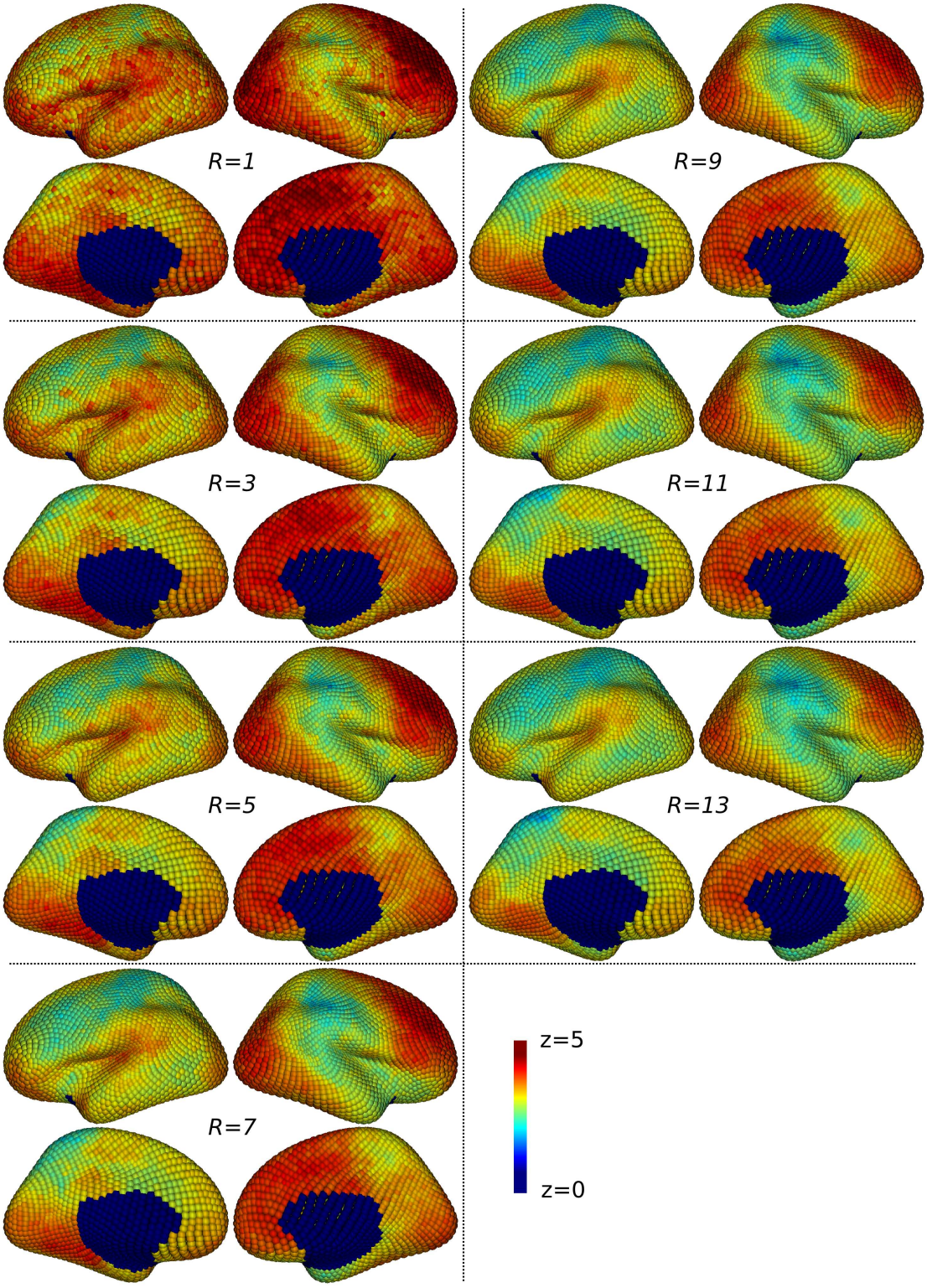

Figure 5.7: Gender problem. Maps of muli-scale z-scores maps for a number of averaged consecutive scales $R \in \{1, 3, 5, 7, 9, 11, 13\}$.

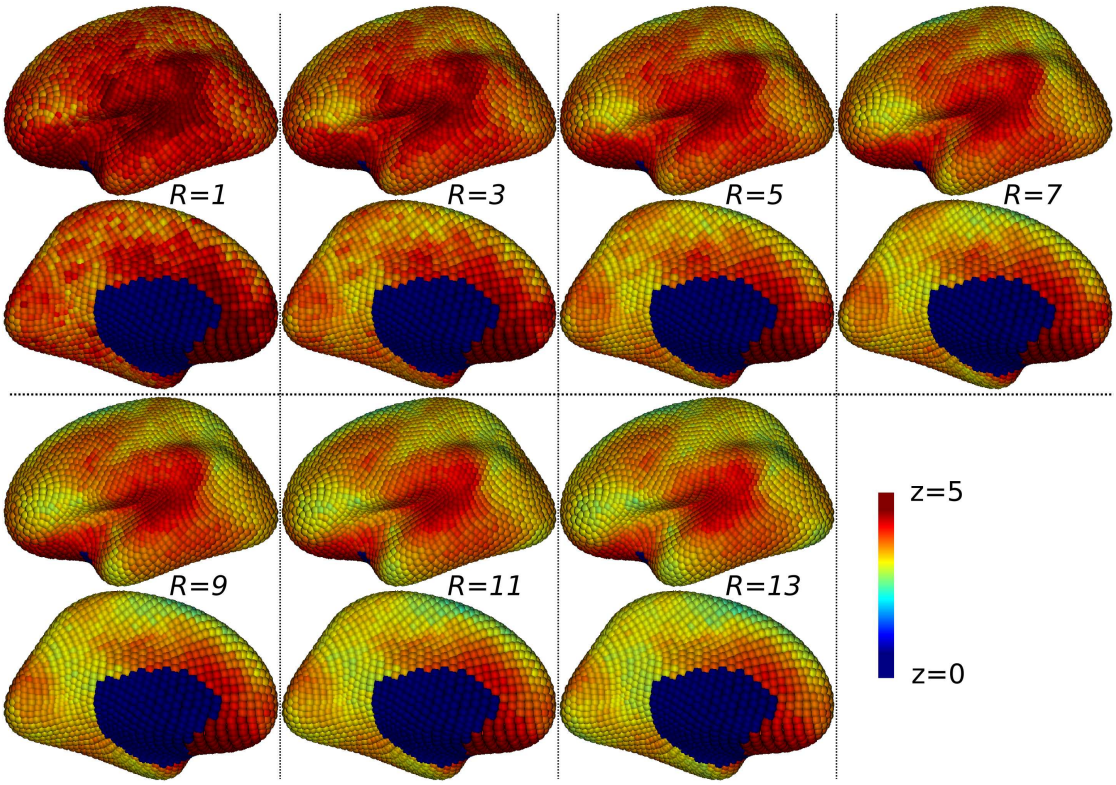

Figure 5.8: Asymmetries. Maps of muli-scale z-scores maps for a number of averaged consecutive scales $R \in \{1, 3, 5, 7, 9, 11, 13\}$.

Figs. 5.9 and 5.10 show the significant clusters (corrected $p < 0.05$) obtained with the method described in 5.2.5.3, for the same values of R. The observations made previously on on the raw statistic map gets confirmed: some overly large clusters are detected for the smaller values of R, while nothing is detected for larger values ($R \geq 11$). This suggests that an intermediary value of R might offer a satisfactory compromise between specificity and sensitivity. This is concordant with the intuition that an effect can live at several consecutive scales (but probably not over a very large number of consecutive scales as sought after with large R values).

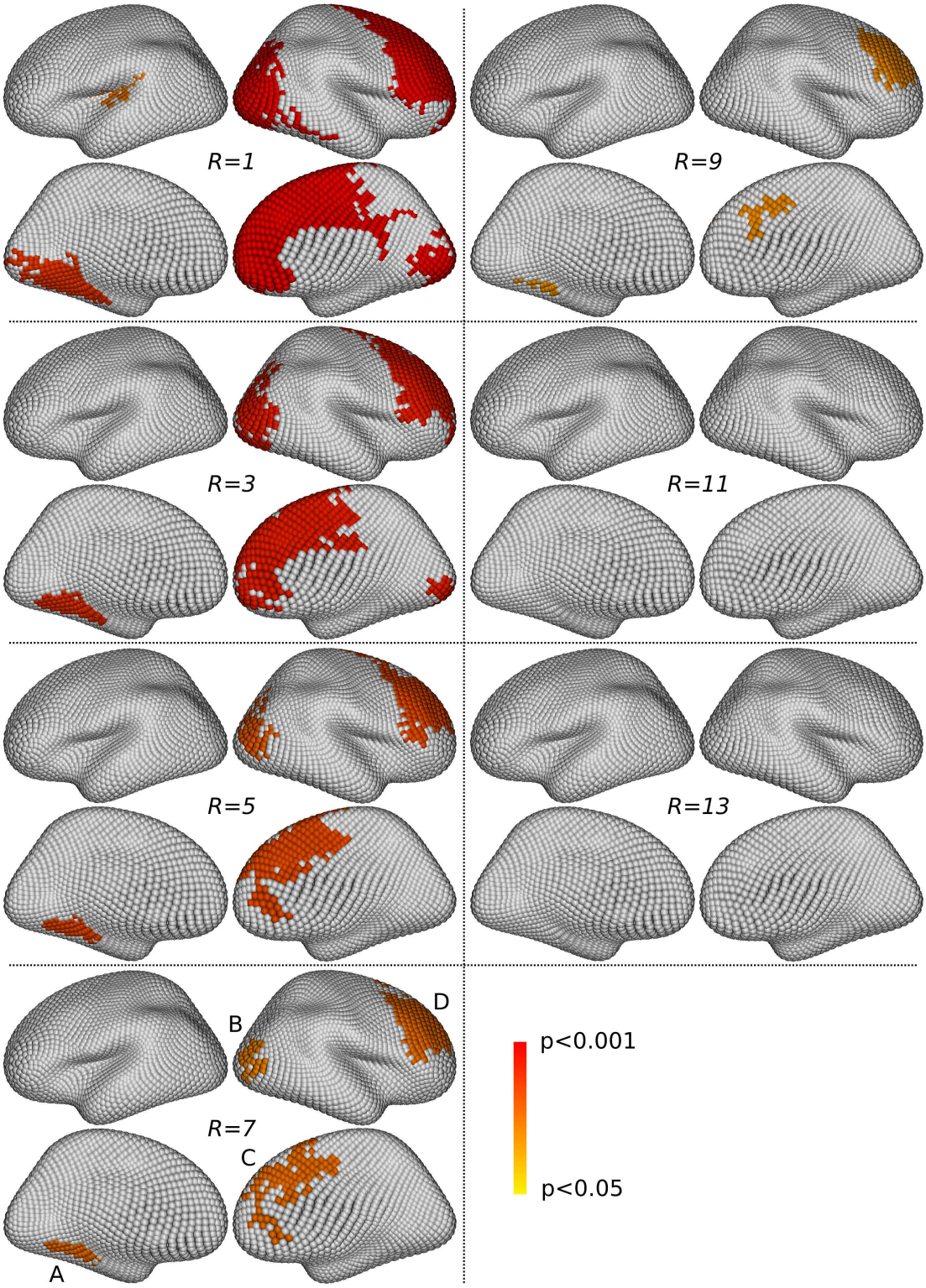

Figure 5.9: Gender. Significant multi-scale clusters (corrected $p < 0.05$), for a number of averaged consecutive scales $R \in \{1, 3, 5, 7, 9, 11, 13\}$. For $R = 7$, the same four clusters as in the single-scale ones are detected. They are tagged A, B, C and D as on Fig. 5.5.

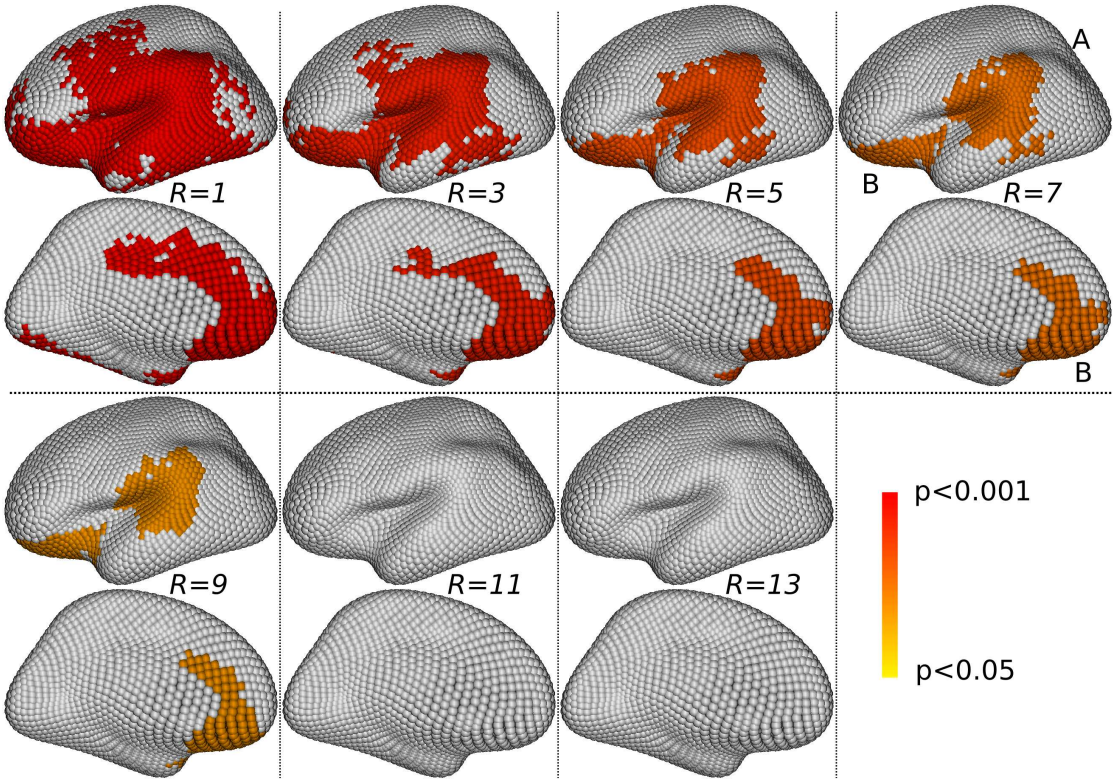

Figure 5.10: Asymmetries. Significant multi-scale clusters (corrected $p < 0.05$), for a number of averaged consecutive scales $R \in \{1, 3, 5, 7, 9, 11, 13\}$. For $R = 7$, the same two clusters as in the single-scale ones are detected. They are tagged A and B, accordingly to Fig. 5.6.

5.3.2.3 Comparing single-scale and multi-scale results

We will now compare the clusters detected by our single-scale and multi-scale approaches. The single-scale analysis offers a description which is more detailed because it is individualized for each scale, while the multi-scale analysis provides a more condensed view on the problem. Our objective is therefore two-fold with this comparison: i) to assess the consistency, or lack thereof, of the results obtained by the two approaches, and ii) to facilitate the selection of the parameter value R of our multi-scale approach based on the single scale results.

For the gender problem, we examine Fig. 5.5 (single-scale clusters) and Fig. 5.9 (multi-scale clusters). The single-scale approach detects clusters in four main locations (tagged A, B, C and D on Fig. 5.5), which are present across a large number of scales. Their precise location and extent vary smoothly across scales. For instance, the cluster at location B slightly moves forward when the scale r increases; some merging phenomena also happen between clusters, as with the clusters at locations C and D, which are separated at scales $r \leq 60mm$ and fused

into one larger cluster at scales $r \geq 70mm$. The multi-scale approach detects clusters in similar locations for various values of the R parameter; however, the results with $R = 7$ are the only ones for which there are four separated clusters for the four locations A, B, C and D.

For the asymmetry problem, we examine Fig. 5.6 (single-scale clusters) and Fig. 5.10 (multi-scale clusters). The single-scale approach detects clusters in two main locations (tagged A and B on Fig. 5.6), which are present across almost the full range of scales from $r = 30mm$ to $r = 90mm$. The clusters present around these locations also fuse together for $r \geq 80mm$. Note that the cluster located on the internal face (location B) is unique for $r \leq 40mm$ but splits into smaller clusters for $r \geq 50mm$ (up to four clusters for $r = 70mm$). The multi-scale approach detects clusters in similar locations for various values of the R parameter; however, only the results obtained with $R = 7$ and $R = 9$ provide two separated clusters for each location A and B.

Overall, the fact that the clusters detected with the single- and multi-scale approaches occupy similar locations clearly demonstrates the consistency of the two approaches, hence the relevance of our multi-scale statistic $\mathfrak{z}^R$. We can also observe that the multi-scale clusters are overall more compact, i.e with less holes, than the single scale clusters; this is clearly a desirable property when it comes to the interpretation of the results and it is a consequence of the smoothing effect of our multi-scale statitic. On another point, as stated previously, the smaller values of R do not provide enough specificity and the larger values lack detection power, and an intermediary value should provide a good compromise between sensitivity and specificity; the results examined here suggest that $R = 7$ is a value that allow finding well separated clusters with the multi-scale approach that fully carry the finer description provided by the single-scale approach. We therefore suggest that the multi-scale results obtained with $R = 7$ is a good choice, providing computational efficiency and a good compromise between sensitivity and specificity.

5.3.3 Results: neuroscience considerations

In this section, we will examine in more details the results obtained by our searchlight framework for the two problems of gender differences and cortical asymmetries. Following the previous observations, we will base our interpretation on the clusters detected by our multi-scale approach with $R = 7$.

Figs. 5.11, 5.12, 5.13, 5.14, 5.15 and 5.16 each describe a significant cluster using the same format, with six panels in each figure. The top left panel shows the searchlight locations (small spheres) included in the cluster, projected on the folded mesh of the *fsaverage* template subject; the slightly larger green sphere is the center of mass of the cluster. The bottow left panel includes the probabilistic density map of basins for the entire population. The top middle and right panels show the local pit graphs for respectively the male and female median

subject for the gender differences problem (Figs. 5.11, 5.12, 5.13 and 5.14) and, respectively, the left and right hemispheres for the cortical asymmetry problem (Figs. 5.15 and 5.16). The bottom middle and right panels show the local basins on the individual cortices of the male and female median subject for the gender differences problem, and the left and right hemispheres for the cortical asymmetry problem.

5.3.3.1 Gender differences

We found four clusters that showed significantly different sulcal patterns between males and female subjects. One is located in the left hemisphere, while the three others are in the right hemisphere.

Cluster A is located in the collateral sulcus of the left hemisphere, as shown on Fig. . It comprises 35 searchlight locations and its corrected p-value is 0.011. Its center of mass is located in the left collateral sulcus. The basins density map goes from the calcarine sulcus in the back, with a superior limit on the parahippocampal gyrus and an inferior limit on the inferior temporal gyrus.

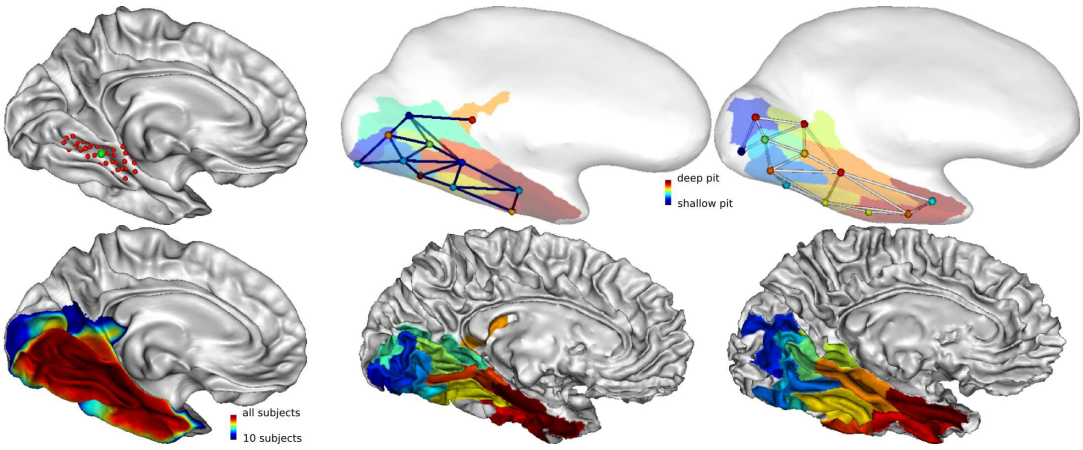

Figure 5.11: Gender differences: left hemisphere, cluster A. Top left: the multi-scale cluster. Bottow left: the probabilistic map of basins. Top middle and right: local pit graph for the male and female median subject. Bottom middle and right: local basins for the male and female median subject.

Cluster B is located in the right hemisphere, centered around the lateral occipital sulcus, as shown on Fig. 5.12. It comprises 29 of the 2500 searchlight locations and its corrected p-value is 0.024. Its associated basins density map covers the lateral occipital lobe.

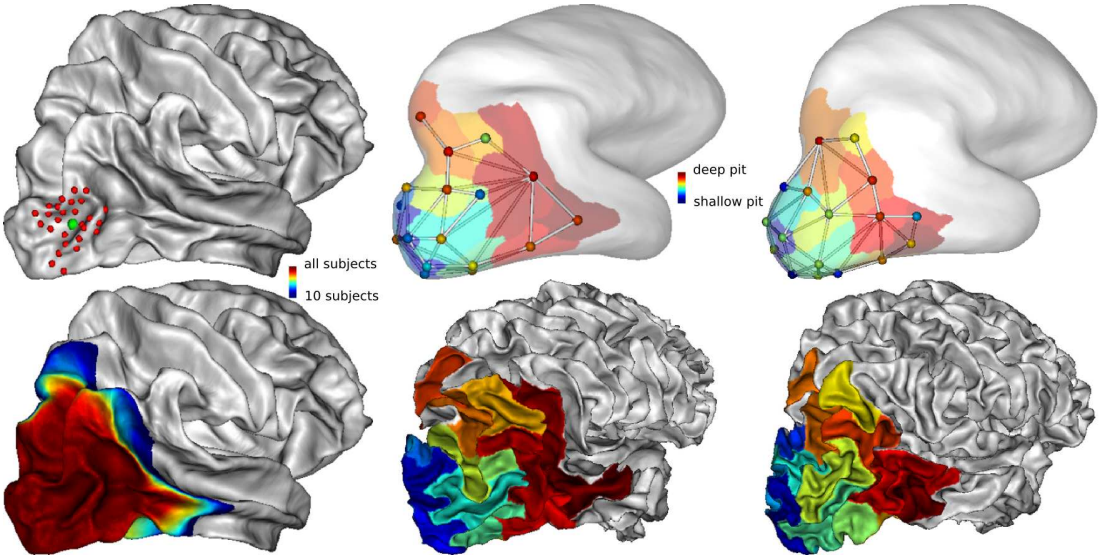

Figure 5.12: Gender differences: right hemisphere, cluster B.

Cluster C, located in the right hemisphere, has its center of mass in the cingulate gyrus, as shown on Fig. 5.13. It is a large cluster with 98 searchlight locations and its corrected p-value is 0.012. Its basins density map covers the median prefrontal lobe, from the marginal ramus in the back, to the gyrus rectus in the front, and from the corpus callosum to the superior frontal paramidline sulcus.

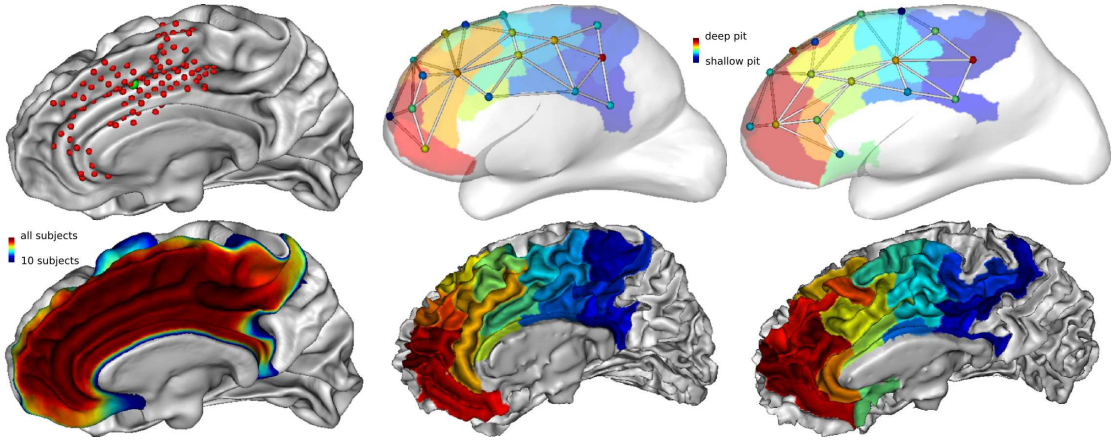

Figure 5.13: Gender differences: right hemisphere, cluster C.

Cluster D is a large cluster centered in the middle frontal gyrus, as shown on Fig. 5.14. This cluster includes 130 searchlight locations and its corrected p-value is 0.012. Its basins density map covers the lateral frontal lobe, from the precentral gyrus in the back to the lateral orbital gyrus.

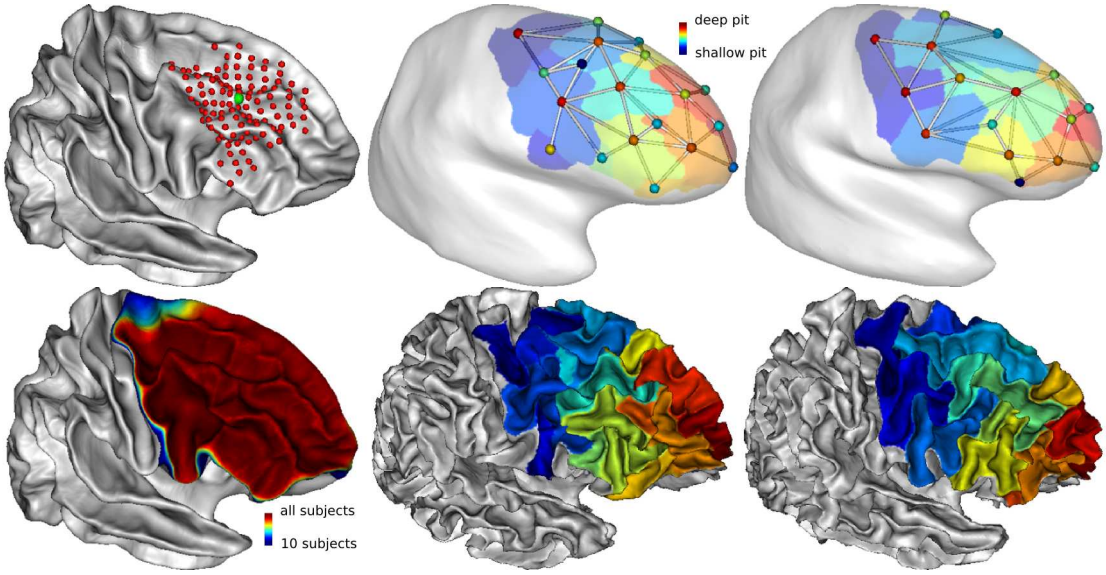

Figure 5.14: Gender differences: right hemisphere, cluster D.

5.3.3.2 Asymmetries

We found two clusters that showed significant differences between the left and right hemispheres in male subjects.

The center of mass of cluster A is clocated in the planum temporale left hemisphere, as shown on Fig. 5.15. It comprises 295 searchlight locations and its corrected p-value is 0.012. The basins density map includes the superior temporal sulcus, the superior temporal gyrus, the full lateral fissure (also known as sylvian fissure) including Heschl's gyrus the planum temporale and the planum polare, and the supra marginal gyrus.

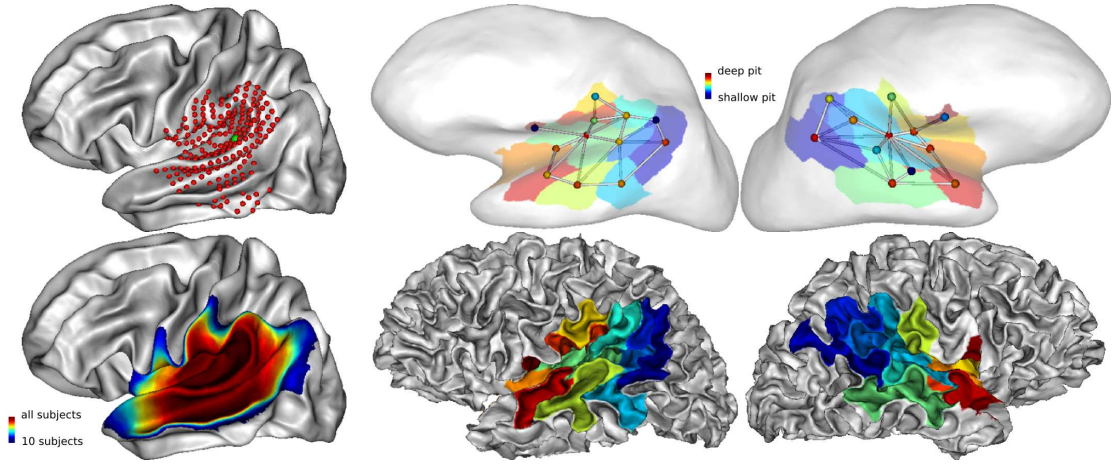

Figure 5.15: Asymmetries: cluster A. Top left: the multi-scale cluster on the symmetrized template. Bottow left: the probabilistic map of basins. Top middle and right: local pit graph for the right and left median hemispheres. Bottom middle and right: local basins for the right and left median hemispheres<.

Cluster B is centered around the frontal pole, as shown on Fig. 5.16. It comprises 205 searchlight locations and its corrected p-value is 0.012. Its basins density maps includes the olfactory sulcus, the lateral orbital gyrus with its anterior, medial and posterior sections, and it extends until the gyrus rectus on the medial side.

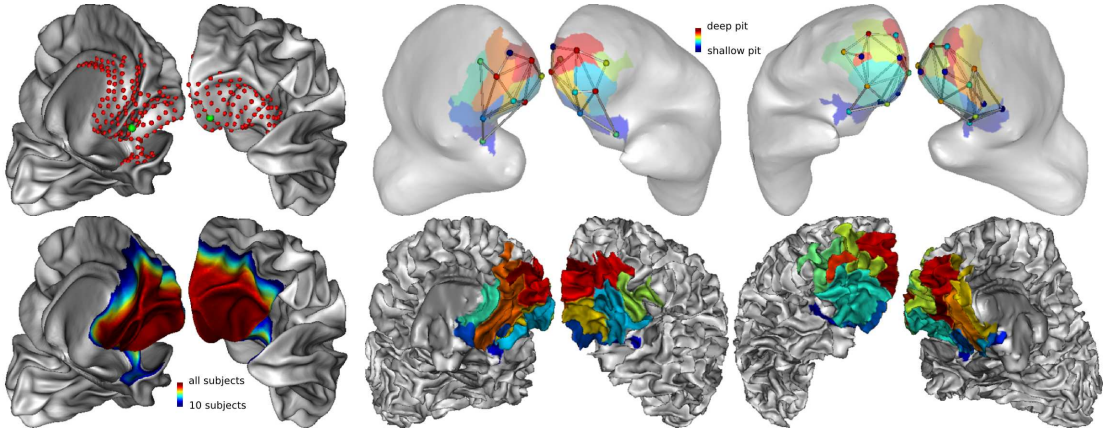

Figure 5.16: Asymmetries: cluster B.

5.4 Discussion

We have introduced a multi-scale searchlight scheme that enables to detect differences in graphs of sulcal pits between populations. This is the first time that a searchlight approach is proposed for studying local patterns of cortical shape. The main strengths of our framework are i) to capitalize on structural pattern recognition tools in order to model and discriminate complex cortical folding patterns, ii) to use non-parametric cluster-based inference in order to achieve high detection power, iii) to embed a multi-scale strategy, thus making it possible to deal with patterns of different sizes, and iv) to overcome the intrinsic limitations of approaches that focus on a pre-defined region thanks to the fine sampling of the cortical surface that implies a strong overlap between patterns.

We start this discussion by examining the neuroscientific relevance of our results before a series of methodological considerations.

5.4.1 Exploring the relevance of our results

Asymmetries are crucial to understand because they are associated with strong functional specialization of the human adult brain, the two most prominent being language lateralization and hand preference. Being able to detect structural markers of these asymmetries is therefore a key point in order to study the processes associated with these function-related lateralization effects, in particular during development and in pathological conditions. Cortical asymmetries have been studied using a large variety of methods, from Voxel-Based Morphometry (VBM, see [Good et al. 2001] for instance) and Surface-Based Morphometry (in [Van Essen, Glasser, et al. 2012]) to sulcal morphometry (in [Duchesnay et al. 2007]) and sulcal pits characterization (in [Im, Jo, et al. 2010; Auzias, Brun, et al. 2015]). By studying the organization of local patterns of sulcal pits, our searchlight framework was able to detect two main asymmetric regions in right-handed males subjects. The first one (cluster A, examined in details in Fig. 5.15) is a large cluster encompassing the posterior part of the sylvian fissure, the supra-marginal gyrus, the superior temporal gyrus and sulcuswas centered in the Planum Temporale. These regions are classical asymmetrical spots which are consistentely reported in the literature ([Van Essen, Glasser, et al. 2012; Duchesnay et al. 2007; Im, Jo, et al. 2010; Auzias, Brun, et al. 2015]). This shows that the multivariate nature of our analysis therefore allows to detect such locally distributed effects among a complex pattern of sulcal basins. Our second cluster (cluster B, Fig. 5.16) is centered around the frontal pole, including part of the lateral prefrontal cortex, until the middle part of the cingulate sulcus. This is consistent with the VBM results of [Good et al. 2001]. This is also consistent with the study of asymmetries in pits frequency obtained in [Auzias, Brun, et al. 2015] (which used the same pit extraction algorithm as ours), but not with [Im, Jo, et al. 2010] which used a different pit extraction algorithm. The large ex-

tent of this cluster might also be the result of the aggregation of several smaller effects, separaterly in the medial and lateral parts of the cortex.

Exploring the cortical gender differences is also important because many neurological diseases are expressed differentely between males and females, with for instance different ages of onset, symptomes and prevalence levels. Being able to identify and localize morphological gender differences is therefore critical to understand their causes and eventually their consequences on neurological diseases. Gender differences have mostly been studied using standard morphometry tools: VBM (with notably a meta-analysis presented in [Ruigrok et al. 2014]) and SBM ([Im, Jo, et al. 2010; Lv et al. 2010]). To our knowledge, the only study that used sulcal morphometry is [Duchesnay et al. 2007]. Our searchlight framework was able to detect four clusters presenting gender differences in their patterns of sulcal pits: one in the left hemisphere and three in the right hemisphere. The left hemisphere cluster (cluster A, Fig. 5.11), centered around the parahyppocampal gyrus, is in a similar location as regions that were significant in [Ruigrok et al. 2014; Im, Jo, et al. 2010; Lv et al. 2010]. Smaller regions located within clusters C and D (right hemisphere, frontal lobe, respectively medial and lateral, Figs. 5.13 and 5.14) were also found significant in these three studies. The cluster centered around the right occipital pole (cluster B, Fig. 5.12) is concordant with very small spots appearing in the surface based studies only ([Im, Jo, et al. 2010; Lv et al. 2010]). The comparison with the results in [Duchesnay et al. 2007] is also interesting because the level of agreement is fairly weak. It means that the shape features which offered the most discriminative power in their paper do not provide the same kind of information as the local pattern encoded in our pit-graphs; even more interesting is the fact that they did use descriptors attempting to capture the local spatial relationships between a sulcus and its neighbors; almost none of these descriptors contributed to their higly discriminative model, probably because these descriptors only encoded very local characteristics of the sulcus relative organization, while our pit-graphs look a the organisation of sulcal patterns at a coarser scales.

Overall, we believe that our graph kernel is able to capture several types of effects, or a combination thereof, that include: i) differences in sulcal depth, distributed across a local folding patternn ii) differences in the local number of pits (as suggested by the concordant results with [Auzias, Brun, et al. 2015], iii) differences in pits localization (as suggested by the concordant results with [Im, Jo, et al. 2010]; this is also consistent with what appears on the median graphs between the two classes: sometimes, an extra pit appears on the outskirt in one class but not in the other, as with the most posterior pit on Figs. 5.13; this pit might actually exist, but it could be localized a bit further than in the first class, thus excluding it from the graph); iv) local organization differences (which might be induced by the two previous types of effects). This is consistent with how this kernel was designed, with its three subkernels, examining respectively the localization of the pits, their depth, and the graph structure. Further posthoc

studies remain needed in order to decipher these effects.

5.4.2 Searchlight statistical analysis

Performing statistical inference on information maps resulting from searchlight-based analysis is a matter that necessitates particular care. Indeed, three major points have to be dealt with: i) the multiple comparison problem induced by the large number of tests performed along the information maps, ii) the spatial correlation of these maps which is naturally produced by the overlapping nature of the sliding window, and ii) the fact that these maps are realizations of unknown distributions. Before the present study, searchlight frameworks have been used to analyse fMRI data. The study that introduced the searchlight concept, [Kriegeskorte et al. 2006] used permutations-based non parametric statistics to deal with the unknown character of the distribution of the statistical score they used – the mahalanobis distance between samples of each class, with a point-wise thresholding and a correction for multiple comparison based on the false discovery rate (FDR). The study in [Y. Chen et al. 2011] was based on a permutation-based framework directly applied on classification scores, with no correction. Finally, [Stelzer et al. 2013] also used a permutation strategy, together with cluster-wise inference to deal with spatial correlations. Here, we adopt a similar strategy, using cluster-based inference via permutation tests that enables to correct for multiple comparisons. Our contribution compared with the previously cited searchlight inference methods is to use the cluster mass as cluster-wise statistic ([Bullmore et al. 1999]), associated with a carefully designed point-wise statistic. This allows small clusters containing locations where the classification accuracy is high to obtain a large cluster-wise statistic value, which should make it possible for such small clusters to be significant.

In practice, our searchligh framework was able to detect significant clusters after correction for multiple comparisons, despite the relatively low classification accuracies achieved (maximum: 0.68, chance level: 0.5). This demonstrates the power of our approach. (Note that, in comparison, the p-values given in Im's papers are uncorrected). Furthermore, some of the clusters were indeed small, which show that the strategy that we employed was efficient in that regard.

5.4.3 On the necessity of the multi-scale approach

Our searchlight framework is the first one that directly embeds a multi-scale approach. In several of the previously published studies that used searchlight frameworks ([Kriegeskorte et al. 2006; Y. Chen et al. 2011]), the scale parameter (most often the radius of the spherical neighborhood defined in the brain volume) was varied but none of these studies actually found a difference in the nature of the extracted significant regions: in short, the larger the sphere, the

larger the detected clusters. This argues for well localized effects where the increased size of the neighborhood acts as a increased smoothing size.

However, these papers worked on fMRI data, and we believe there is a fundamental difference when dealing with cortical morphology: while focal structural effects can be detected – such as changes in grey matter density or cortical thickness that can be revealed with VBM or SBM – it seems more natural to study the shape of the cortex at larger scales, corresponding to the size of cytoarchitectonic regions. Therefore, studying sulcal morphometry and the organization of local sulcal patterns is relevant when looking for biomarkers related to the behavior or to a pathology. Such patterns can have very different sizes, with effects observed in a sulcal basin [Auzias, Viellard, et al. 2014], in a large sulcus composed of a small number of basins [Z. Y. Sun, Klöppel, Rivière, Perrot, R. Frackowiak, et al. 2012] or in a full brain lobe [Im, Jo, et al. 2010]. It is therefore crucial to study the multi-scale nature of these potential differences. In this study, we have demonstrated the detection power of our framework through two multi-scale spatial inference schemes which provide consistent results. Furthermore, our method is able to detect significant effects even at small scales, which is usually difficult in multi-scale inference [Lindeberg 1993]; this has been achieved thanks to i) a point-wise stats that heavily weights the rareness of high classification scores and ii) the use of the cluster mass as our cluster-wise statistic. Our results clearly confirm the importance of the multi-scale approach, with significant clusters which do not only vary in size across scales: some clusters appear at different scales, some others change location across scales, while others change in shape.

Finally, it has been shown that the interpretation of the results offered by searchlight methods can represent a real challenge [Etzel et al. 2013]. Systematically changing the size of the neighborhood in a true multi-scale framework – such as the one we have introduced – is an appealing way to disambiguate the results – as proposed in [Etzel et al. 2013]. Our multi-scale framework therefore represents a good way to avoid erroneous interpretations.

5.4.4 A kernel-based multivariate classification model

The multivariate statistical model that we used to map the information content within our searchlight scheme is based on a support vector machine that attempts to separate two classes in a high-dimensional space defined by the input samples – the pit-graphs – and our graph kernel. The results obtained, which are in line with the existing literature, validate this methodological choice and they notably demonstrate that our graph kernel, which had already proved expressive for dealing with inter-subject variability in fMRI data [Takerkart, Auzias, Thirion, and Ralaivola 2014], is a valid similarity measure for comparing pit-graphs. Furthermore, it provides an alternative to the similarity measure proposed in [Im, Pienaar, Lee, et al. 2011] while being more simple to compute.

While Im's metric includes various attributes associated with the sulcal pits (their depth, the basin area, the node degree in the graph), our kernel only needs the depth of the sulcal pit, which is an intrinsic property of the pits, thought to be related with the folding process that occurs in the early phases of cortical development [Lefevre et al. 2009]. It is to be noted that the choice of the weights that define Im's metric required a lot of dedicated experiments (see [Im, Pienaar, Lee, et al. 2011]) and the result of their follow-up work ([Im, Pienaar, Paldino, et al. 2012; Im, Raschle, et al. 2015]) heavily depends on the values of these seven hyper-parameters. In contrast, our kernel only has two hyper-parameters that we directly estimate from the data, thus avoiding any potential bias linked with the *a priori* selection of such parameters. The only available element of comparison between the two metrics is their ability to detect gender differences. In the work described in [Im, Pienaar, Paldino, et al. 2012], Im et al. reported no gender differences in their control group (first sentence in Results, p.3011), whereas our technique, which similarly attempts to find differences between populations based on graphs of sulcal pits, was able to find fairly large regions that did exhibit differences between males and females subjects. We believe that our method was more sensitive thanks to a combination of the three following reasons: i) our dataset was much larger (134 vs. 26 subjects); ii) our pit detection algorithm is slightly different, producing more pits and including shallower ones; iii) their study focused on predetermined large regions (brain lobes), whereas our framework automatically estimates the localization and the extent of the region, including its size, using a multi-scale searchlight scheme. In theory, it is also possible that our kernel provides a graph similarity measure that might be sensitive to distinct properties of the graphs, but we have no element to concur with this possibility. On the contrary, [Im, Jo, et al. 2010] showed that all elements of the graph (its structure, the pits location and attributes) are necessary to achieve a good performances, and, in another context using fMRI data, we have demonstrated the same about our graph kernel [Takerkart, Auzias, Thirion, and Ralaivola 2014]; we therefore believe that both similarity measures are overall sensitive to the same pattern features. But this remains to be studied empirically.

One of the advantages of using a kernel is that we can benefit from its theoretical properties ([Scholkopf et al. 2001]) that make it usable for a numerous problems such as classification (with support vector machines as in the current paper), regression (with kernel ridge regression), dimensionality reduction (using kernel PCA), ranking or clustering. This means that using two of the key elements of our searchlight framework, the pit-graphs and our graph kernel, it is straightforward to address a wide variety of questions, thus offering a large versatility which is a strong asset for studying such complex objects as cortical folding patterns.

5.5 Conclusion

We have introduced a brain mapping technique dedicated to studying differences in the local organization of cortical anatomy, by the mean of patterns constructed from the deepest part of the cortical sulci – the sulcal pits. Our technique is the first *structural searchlight* framework, and also the first searchlight scheme that embeds a spatial inference framework that uses multi-scale information. It relies on a graph-based support vector classification model, a kernel-based method that has the advantage to be easily generalizable to questions such as regression, ranking or dimensionality reduction. We have demonstrated that our framework provides an interesting detection power on two problems that respectively examined the anatomical gender differences and the cortical asymmetries. This versatile and powerful pit-based searchlight approach should therefore find numerous applications, in particular to study cortical development and clinical populations, two fields where examining the organization of sulcal patterns at different spatial scales is particularly relevant.

6 Multi-source kernel mean matching for inter-subject decoding of MEG data

Contents

Context

The third contribution of this thesis is a multi-source method to perform domain adaptation when several sources of data are available. It builds upon the kernel mean matching procedure, a transductive kernel-based distribution matching

approach that attempts to adapt the distribution of the training dataset to the one of the test dataset by reweighting the training instances. We here introduce an extension of this procedure that is able to deal with the fact that the training dataset is in fact composed of several sources of data. We present preliminary results of our approach on a magneto-encephalography dataset analysed through an inter-subject decoding problem.

This chapter presents the current state of this on-going project, with our first attempt at producing a multi-source extension of the kernel mean matching. This project has mostly involved Liva Ralaivola. The pre-processing of the magneto-encephalography data was performed by Romain Trachel. I would also like to acknowledge fruitful discussions with Hachem Kadri on this topic.

6.1 Introduction

In recent years, the use of machine learning approaches in neuroimaging has gained in popularity. The most prominent applications of machine learning for neuroimaging data analysis are the so-called *multi-voxel pattern analysis* (MVPA), that consists in predicting a behavioral variable from functional MRI data, and the design of *brain computer interfaces* (BCI), often using electro-encephalography (EEG) data. In both cases, the goal is to *decode* the thoughts or the actions of a subject from a recording of its brain activity. The appeal of these multivariate methods relies on their increased sensitivity compared to standard univariate models. However, their generalization power on data recorded in new subjects suffers from the large variability that exists within a population.

We specifically focus on the so-called *inter-subject decoding* problem, which consists in performing predictions on subjects for which only unlabelled data where available at training time. This question is a *multi-source learning* problem where each subject is a source of data. We propose a method that builds upon the *kernel mean matching* procedure ([Gretton et al. 2009]) in order to estimate weights for the training instances of each source, and use these weights with the labels of theses instances in order to learn a relevant classifier. We call this method the Multi-Source Kernel Mean Matching (MSKMM)

The standard decoding paradigm to address the inter-subject prediction task consists in pooling together all samples from the subjects available at training time. A single classifier is estimated in a supervised manner on this dataset to be later tested on data from new subjects. This paradigm, which is used by default in the literature, largely ignores the various sources of inter-subject variability, namely the differences in anatomy and functional organization across subjects, and the fact that all samples are not drawn from the same probability distribution. Therefore there is a crucial need for more elaborate models specifically tuned for the inter-subject prediction task, like those recently proposed in [Barachant et al. 2013; Marquand et al. 2014; Takerkart, Auzias, Thirion, and

Ralaivola 2014] and the new model we introduce in this chapter.

6.2 A reminder on kernel mean matching

In this section, we recall the so-called Kernel Mean Matching procedure procedure(KMM, [Gretton et al. 2009]) which makes it possible to account for the differences in distributions between a single source domain (that we will designate by $\mathcal{D}^s$) and a target domain $\mathcal{D}^t$ under the assumption of covariate shift. We first provide a reminder about the instance weighting scheme for domain adaptation, before describing in detail the KMM and how to solve it in practice

6.2.1 Instance weighting for domain adaptation

When learning on samples taken from one domain (called the source domain s), obtaining a classifier that can generalize effectively on samples from another domain (the target domain t) is a difficult problem which may be tackled with techniques of *domain adaptation*. The goal of *domain adaptation* is to find a transformation of the source domain so that the transformed distribution $P'_s(x, y)$ is closer to the target distribution $P_t(x, y)$ closer than the original source distribution $P_s(x, y)$. This can be for instance performed by finding *importance weights* for the samples from the source domain, giving a higher weights to samples that are closer to the target domain. Within the classical *risk minimization* framework, this idea becomes:

$$\begin{aligned} R(P_t, \theta, \ell(.,.,\theta)) &= \mathbb{E}_{(x,y)\sim P_t}(\ell(x, y, \theta)) \\ &= \mathbb{E}_{(x,y)\sim P_t}(\frac{P_s(x, y)}{P_s(x, y)}\ell(x, y, \theta)) \\ &= \mathbb{E}_{(x,y)\sim P_s}(P_s(x, y)\frac{P_t(x, y)}{P_s(x, y)}\ell(x, y, \theta)) \\ &= \mathbb{E}_{(x,y)\sim P_s}(P_s(x, y)\beta(x, y)\ell(x, y, \theta)) \\ &= R(P_t, \theta, \beta(.,.)\ell(.,.,\theta)) \end{aligned} \tag{6.1}$$

The importance weights of the instances of the source domain are therefore

$$\beta(x, y) = \frac{P_t(x, y)}{P_s(x, y)}.$$

Under the assumption of *covariate shift* [Shimodaira 2000], which supposes that the conditional probability distributions are identical in the source and target domains, i.e $P_t(y \mid x) = P_s(y \mid x)$, the expression of the weights simplifies as

follows:

$$\begin{aligned}\beta(x,y) &= \frac{P_t(x,y)}{P_s(x,y)} \\ &= \frac{P_t(y \mid x)P_t(x)}{P_s(y \mid x)P_s(x)} \\ &= \frac{P(y \mid x)P_t(x)}{P(y \mid x)P_s(x)} \\ &= \frac{P_t(x)}{P_s(x)} \end{aligned} \tag{6.2}$$

The weights are therefore simply the ratio of the marginal distributions of the target and the source domains. However, it is still challenging to estimate these weights properly. A variety of methods have been proposed to do through the estimatation of the marginal distributions, for instance with density estimation techniques [Sugiyama et al. 2005] or by using the results of another learning problem in which we attempt to classify the instances as belonging to the source or the target domain [Zadrozny 2004]. The Kernel Mean Matching procedure offers another method to estimate these weights, which has the advantage not to require the estimation of the marginal distributions.

6.2.2 Kernel Mean Matching

In order to tackle the problem of dissimilar distributions, we may try to "align" the marginal distribution P_s of the training domain $\mathcal{D}^s$ with the distribution P_t of the target domain $\mathcal{D}^t$. To do so, we may use the idea of Hilbert space embedding of distributions proposed by [Smola et al. 2007]. Given a positive kernel k and its associated feature map Φ, a distribution $\mathcal{L}$ acting on the same space $\mathcal{X}$ may be mapped as $\mu_{\mathcal{L}}$ to the Reproducing Kernel Hilbert Space associated with k through the following embedding:

$$\mu_{\mathcal{L}} \doteq \mathbb{E}_{x\sim\mathcal{L}}[\Phi(x)]. \tag{6.3}$$

Given a set of samples $x_1, \ldots, x_N$ drawn independent and identically distributed according to $\mathcal{L}$, the empirical embedding $\hat{\mu}_{\mathcal{L}}$ of $\mathcal{L}$ defined by

$$\hat{\mu}_{\mathcal{L}} \doteq \frac{1}{N}\sum_{n=1}^{N}\Phi(x_n) \tag{6.4}$$

is an estimate of $\mu_{\mathcal{L}}$.

Taking advantage of this idea, we can try to estimate the β weights so that the weighted source distribution is as close as possible to the target distribution by minimizing the objective function $\|\beta\,\mu_s - \mu_t\|^2$. In practice, when working with a finite number of samples N^s and N^t, respectively from the source and the target domains, we define $\bar{\beta}$ the vector with N^s elements – its n-th element being β_n.

This gives rise to the following problem:

$$\min_{\bar{\beta}} \left\| \frac{1}{N^s} \sum_{n=1}^{N^s} \beta_n \Phi(x_n^s) - \frac{1}{N^t} \sum_{n=1}^{N^t} \Phi(x_n^t) \right\|^2 \tag{6.5}$$

subject to

$$\begin{cases} \forall n \in \{1, \cdots, N^s\}, 0 \leq \beta_n \leq B & \text{(6.6a)} \\ |\frac{1}{N^s} \sum_{n=1}^{N^s} \beta_n - 1| \leq \epsilon, & \text{(6.6b)} \end{cases}$$

where B and ϵ are two constants.

The two constraints in Eq. (6.6) enables to find adequate solutions for the β weights: the first constraint, given by Eq. (6.6a), bounds the values of the weights in order to limit the discrepancy between the source and the target distributions, while limiting the potential influence of single samples. The second constraint, given by Eq. (6.6b), ensures that the re-weighted source distribution stays close to an actual probability distribution.

The objective function ψ_{KMM} develops as:

$$\begin{aligned} \psi_{KMM}(\bar{\beta}) &= \left\| \frac{1}{N^s} \sum_{n=1}^{N^s} \beta_n \Phi(x_n^s) - \frac{1}{N^t} \sum_{n=1}^{N^t} \Phi(x_n^t) \right\|^2 \\ &= \frac{1}{N^{s2}} \sum_{i=1}^{N^s} \sum_{j=1}^{N^s} \beta_i \beta_j k(x_i^s, x_j^s) - \frac{2}{N^s N^t} \sum_{i=1}^{N^s} \beta_i \sum_{j=1}^{N^t} k(x_i^s, x_j^t) + const \end{aligned} \tag{6.7}$$

Together with the constraints in Eq. (6.6), this forms a quadratic program that can be solved with standard optimization librairies. The implementation details are presented in Appendix 6.6.

6.2.3 A transductive domain adaptation classifier

In order to use an instance weighting scheme within a learning algorithm, several options are available, as nicely summarized in [Gretton et al. 2009]. We will here resort to the weighted version of the Support Vector Machine in order to maintain some homogeneity with the other chapters of this thesis, which all make use of SVM. The principles of weighted SVM (hereafter WSVM) consists in using the importance weights of the training instances to locally modulate the amount of regularization, i.e to scale the C parameter used in the optimization problem

that needs to be solved. In practice, the β_n weights are introduced as follows:

$$\min_w \frac{1}{2} \|w\|^2 + C \sum_{n=1}^{N^s} \beta_n \ell(w, x_n, y_n) \tag{6.8}$$

The general description of how to solve such problem is given in [Tsochantaridis et al. 2005] and an example of applications of WSVM can be found in [X. Yang et al. 2007], with the goal of down-weighting the outliers in the training dataset.

The end result of the combination of using KMM to estimate importance weights and WSVM to learn a classifier can be seen as a transductive learning algorithm. This method finds a weighting scheme to modify the distribution of the source domain so that it matches more closely the distribution of the target domain before using the importance weights to perform predictions on the instances of the target domain.

6.3 Multi-source kernel mean matching

In this section, we introduce Multi-Source Kernel Mean Matching (MSKMM), an extension of the KMM that is aimed at dealing with the case where the training data is drawn from several distinct distributions, i.e when we have at our disposal several sources of training data. First, we recall the multi-source setting which was defined for the inter-subject decoding task in Chapter 2. Then we present our MSKMM framework in details, before taking a closer look at some particular cases which makes MSKMM equivalent to other models for certain values of its main parameter.

6.3.1 Multi-source setting

We here define the multi-source setting that describes our *inter-subject* decoding problem. First, we denote $\mathcal{S} = \{1, \cdots, S\}$ the set of subjects, where S is the number of available subjects.

Then, for each subject $s \in \mathcal{S}$, we note $\mathcal{D}^s = \{(x_n^s, y_n^s)\}_{n=1}^{n=N^s}$ the labeled training data for subject s, where x_n^s lives in a feature space $\mathcal{X}^s$, the target variables y_n^s are scalar values of $\mathbb{R}$ and N^s designates the number of observations available for this subject. The full training set, of total size $N = \sum_{s=1}^{s=S} N^s$, is then defined as

$$D \doteq \cup_{s=1}^{S} D^s.$$

In addition to the training data, there exist a dataset $\mathcal{D}^t$ of the same nature for a test subject t not in $\mathcal{S}$

$$\mathcal{D}^t = \{(x_n^t, y_n^t)\}_{n=1}^{n=N^t},$$

except the labels $\{y_n^t\}_{n=1}^{n=N^t}$ are not observed. For simplicity reasons, we assume

that the feature spaces of all subjects are identical, i.e that $\mathcal{X}^1 \doteq \cdots \doteq \mathcal{X}^S \doteq \mathcal{X}^t \doteq \mathcal{X}$. Furthermore, we now restrict ourselves to a binary classification problem, i.e the output space is $\{-1, +1\}$

Within this setting, the *inter-subject decoding* task is a *multi-source learning* problem for which each training subject provides a *source* of data. Our goal is to be able to perform realiable predictions on data from a *source* not available at training time, i.e data recorded on the test subject.

6.3.2 Multi-source kernel mean matching

In this multi-source context, we have at our disposal S sources from each of which a set of samples have been drawn from a different subject. The subjects therefore provides some structure on the training dataset. We now aim at exploiting this structure within the KMM framework, i.e in order to estimate importance weights for each of the training instances.

We propose to to do so by adding an extra constraint in the optimization problem that allows estimating the weights. A natural constraint to add in this case is to ask for all the weights acting on the samples from a given source to be close to each other. In other words, we can attempt to find a mean weight for all the samples of a given source, and to bound the deviations of the individual sample weights from this mean. Let us first define $\bar{\beta}$ the vector of $N = \sum_{s=1}^{s=S} N^s$ elements, composed of S blocks, with the s-th block containing the weights of subject s, i.e $\beta_1^s, \cdots, \beta_{N_s}^s$. This then defines the following optimization problem:

$$\min_{\bar{\beta}} \left\| \sum_{s=1}^{S} \sum_{n=1}^{N^s} \beta_n^s \Phi(x_n^s) - \hat{\mu}^t \right\|^2 \tag{6.9}$$

subject to

$$\begin{cases} \forall (s,n), \ \left| \beta_n^s - \dfrac{1}{N^s} \displaystyle\sum_{n'=1}^{N^s} \beta_{n'}^s \right| \leq \eta & \text{(6.10a)} \\ \displaystyle\sum_{s=1}^{S} \frac{1}{N^s} \sum_{n=1}^{N^s} \beta_n^s = 1 & \text{(6.10b)} \\ \forall (s,n), \ \beta_n^s \geq 0. & \text{(6.10c)} \end{cases}$$

The first constraint (Eq. (6.10a)) bounds the difference between the individual weights and the mean weight across samples, for each individual subject, using a positive constant η. The second constraint (Eq. (6.10b)) ensures that the weighted training samples actually follow a probability distribution, similarly to the constraint used in KMM with the case of a single source of training data, given in Eq. (6.6b). The last constraint (Eq. (6.10c)) simply ensures that all weights are positive.

We now examine how to solve the MSKMM in practice. The objective function to be minimize develops as:

$$\begin{aligned}\psi_{MSKMM}(\bar{\beta}) &= \left\| \sum_{s=1}^{S} \frac{1}{N^s} \sum_{n=1}^{N^s} \beta_n^s \Phi(x_n^s) - \frac{1}{N_t} \sum_{n=1}^{N_t} \Phi(x_n^t) \right\|^2 \\ &= \sum_{s_1=1}^{S} \sum_{s_2=1}^{S} \frac{1}{N^{s_1} N^{s_2}} \sum_{i=1}^{N^{s_1}} \sum_{j=1}^{N^{s_2}} \beta_i^{s_1} \beta_j^{s_2} k(x_i^{s_1}, x_j^{s_2}) \\ &\quad - 2 \sum_{s=1}^{S} \frac{1}{N^s N^t} \sum_{i=1}^{N^s} \beta_i^s \sum_{j=1}^{N^t} k(x_i^s, x_j^t) + cst \end{aligned} \tag{6.11}$$

Together with the constraints of Eq. (6.10), this forms a quadratic optimization program. The implementation details used for solving this optimization problem are given in Appendix 6.7.

6.3.3 Limiting cases of the model

The only parameter of the MSKMM optimization problem is the η constant, which determines how different the weights of the individual training samples can be from the mean weight of all samples of a given subject. In other words, the parameter η controls the amount of structure that is imposed on the weights. We here examine the two extreme cases: $\eta = 0$ and $\eta \to \infty$.

When $\eta \to \infty$, Eq. (6.10a) does not enforce any constraint to take into account the multi-source structure of the training dataset $\mathcal{D}$, as if we did not dispose of the knowledge that the different samples were recorded from different subjects. The model is then almost equivalent to the standard Kernel Mean Matching for which we have a single source of training data. We call it MSKMM$_\infty$.

When $\eta = 0$, Eq. (6.10a) in fact imposes that all the weights given to the samples of a given subject are equal. Let us define β^s the value of all the weights for subject s. Our optimization problem is then reduced to

$$\begin{aligned}\min_{\beta^s} \left\| \sum_{s=1}^{S} \frac{\beta^s}{N^s} \sum_{n=1}^{N^s} \Phi(x_n^s) - \frac{1}{N_t} \sum_{n=1}^{N_t} \Phi(x_n^t) \right\|^2 = \\ \min_{\beta^s} \left\| \sum_{s=1}^{S} \beta^s \hat{\mu}_s - \hat{\mu}_t \right\|^2 \end{aligned} \tag{6.12}$$

subject to

$$\begin{cases} \sum_{s=1}^{S} \beta^s = 1 & (6.13) \\ \forall s, \beta^s \geq 0. & (6.14) \end{cases}$$

This in fact means that we model the distribution of the target subject as a mixture of the distributions of the source subjects, through their embedding μ

into the RKHS defined by the kernel k. We call this model MSKMM$_0$.

In that sense, MSKMM$_0$ is similar to the method introduced in [Chattopadhyay et al. 2012], which also estimates one weight per source of data. The main difference is that their weighting scheme is driven by working on the conditional probability distributions, which they have access to by using some labeled samples available in the target domain – such labeled samples being absent in our setting. Our full model MSKMM can therefore be seen as close to [Q. Sun et al. 2011], which uses a two stages weighting scheme, with a first stage that yields one weight per source and a second stage which adds a single-instance importance weight.

6.4 Simulations

6.4.1 Dataset and pre-processing

We perform experiments on the data recorded for a magneto-encephalography (MEG) experiment. The data was made available to the community through the *DecMec2014* competition hosted by *Kaggle*[a]. The full dataset was composed of recordings from 23 subjects. Each sample corresponds to an experimental trial during which the subject was looking either at faces ($y = -1$) or at scrambled images ($y = -1$). The labels were available for 16 subjects, which composed the training dataset of the competition; predictions had to be submitted for the samples of the 7 other subjects for which the labels were hidden.

Each sample is composed of time-series lasting 1.5 seconds for each of the 306 MEG sensors covering the head of the subject. The data is sampled at 250Hz, and had been high-pass filtered at 1Hz to remove the slow instrumental drifts. The spatial layout of the MEG sensors are also provided; three sensors, two orthogonal gradiometers and one magnetometer, were positionned at each of the 102 locations that covers fully the scalp. The magnetometers measure the radial component of the magnetic field, while the gradiometers measure its tangential derivative.

Some extra pre-processing steps were conducted using *mne*[b]. First, outlier trials were removed from each subject's dataset if their mean deviated too much from the mean of all trials. Second, we kept only the signals from the gradiometers and temporally down-sampled the time-series by a factor 16, in order to reduce the dimensionality of the dataset. Finally, the resulting time-series from all 102 locations were concatenated to form a single feature vector that was normalized so that its mean was zero and its standard deviation was one. This resulted in around 215 samples for each of the 15 subjects.

[a] https://www.kaggle.com/c/decoding-the-human-brain
[b] http://www.martinos.org/mne/stable/mne-python.html

6.4.2 Experiments

We restricted ourselves to 15 subjects for which the labels are available. Using a leave-one-subject-out cross-validation, we evaluated the generalization power of different algorithms for the inter-subject prediction task: given a MEG trial from a new subject – i.e the left-out subject in the cross validation, the task of the algorithm was to predict the class of the stimulus that had been presented to the subject, i.e a face or a scrambled image.

We can then compare the performances of our transductive classification algorithm, composed of the weight estimation step using MSKMM and the prediction step using WSVM, to other algorithms. In order to specifically assess the added value of the importance weighting schem, we chose to use the same prediction method applied without weighting as benchmark i.e the regular SVM where the training data from the different subjects is pooled together without any knowledge about the subject structure. For both the weight estimation and the classification steps, we used a Gaussian kernel. In order to gain some insight on the effectiveness of our weighting approach, we used oracles that selected the optimal values for the hyper-parameters of each method among a large array of values. For both classification methods, there were two parameters to be chosen: the SVM regularization constant C and the bandwidth of the Gaussian kernel used for classification γ_c; the C parameter was chosen in $\{10^{-n}\}_{n\in[-5\cdots5]}$ and γ_c in $\{2^{-n}\}_{n\in[-15\cdots5]}$. For our approach, there were two additional hyper-parameters that came with the estimation of the weights with MSKMM, the η parameter and the bandwidth γ_w of the Gaussian kernel; the η parameter was chosen in $[0., 0.01, 0.05, 0.1, 0.2, 0.5, 1., 2., 5., 10.]$ and γ_w in $\{2^{-n}\}_{n\in[-15\cdots5]}$. Each time, we report the maximum performance over the different values taken by these parameters.

6.4.3 Results

6.4.3.1 Examining the weights

On Fig. 6.1, we present a typical example of the weight values attributed to all the samples of the training set. Because we use a leave-one-subject out scheme to evaluate the different algorithms, the training set is composed of samples from 14 subjects. The samples are listed on the x-axis; they are grouped by subjects in order to make the subject structure on the training dataset clearly visible. If $\eta = 0$, i.e for MSKMM$_0$ (Fig. 6.1A), the weights of all samples of a given subject take a constant value, as imposed by the constraint 6.10a when $\eta = 0$; because the samples are grouped by subject on the x-axis, we therefore observe that the weights follow a piecewise constant function. When η takes an intermediairy value (Fig. 6.1B), the individual weights can take values that recede from the mean value for each subject; this maintains some structure on the weights imposed by the different sources, while providing more flexibility than when $\eta = 0$.

Finally, when η is large, i.e for MSKMM$_\infty$ (Fig. 6.1C), there is no such multi-source constraint anymore; we experimentally verified that MSKMM$_\infty$ yields the same weights than the standard KMM – where all training samples are provided as a single source of data to KMM. In that case, as shown by the inset, the weights are sparse, i.e a large number take the zero value, which is expected from weights estimated using KMM [Gretton et al. 2009].

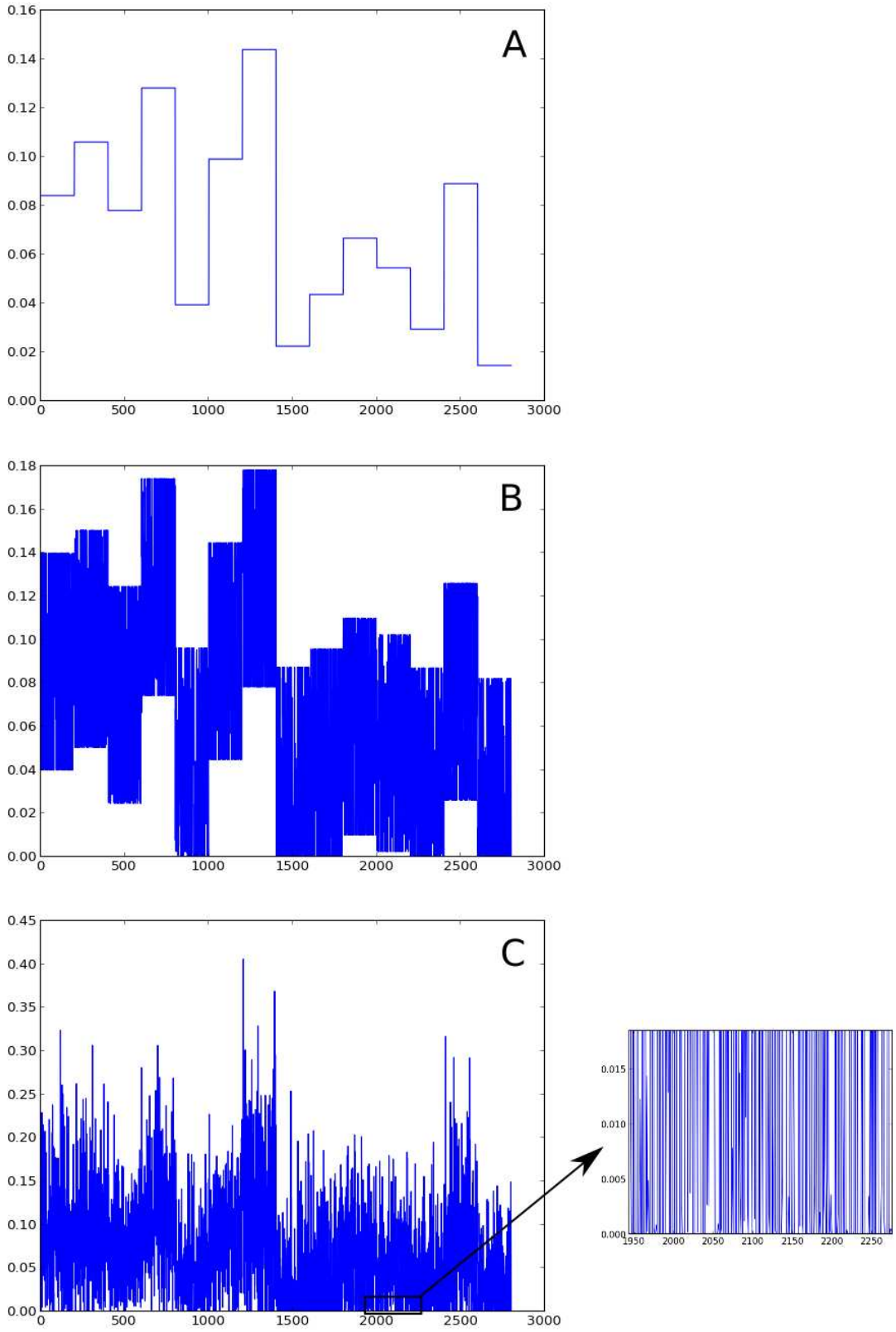

Figure 6.1: Illustration of the weights estimated with MSKMM. For MSKMM_0 (A), the weights are equal for all instances of a subject; for MSKMM_∞ (C), there is no subject structure imposed on the weights, which tend to be sparse; using an intermediary value of η (B) imposes some structure on the weights while offering some flexibility.

Overall, our MSKMM model behaves as expected, showing the validity of our procedure. When using an intermediairy value for η (in practice, between 0.01 and 0.2), the solution is adequate for capturing the multi-source nature of the training dataset while maintaining some flexibility on the importance weights of single instances.

6.4.3.2 Classification performances

We here detail the classification performances, measured by the average accuracy across the folds of the leave-one-subject-out cross-validation. As stated previously, we compare the performances of an oracle that uses the optimal values of the hyper-parameters, in order to gage the maximal performance attainable by each method. The results are presented in Tab. 6.1

SVM	MSKMM + WSVM	MSKMM_∞ + WSVM	MSKMM_0 + WSVM
0.624	0.629	0.629	0.629

Table 6.1: Mean accuracies for standard SVM (without multi-source adaptation) versus weighted SVM using MSKMM (chance level = 0.5)

Overall, we see that there is a small gain in performance between using the simple SVM without any adaptation strategy, and using MSKMM + WSVM. However, this gain is clearly not statistically significant. When it comes to choosing the value of the η parameter used to estimate the weights in MSKMM, we see that this parameter has no influence on the classification performances. This is fairly surprising because we saw previously (see Fig. 6.1) that the values of the individual weights are strongly influenced by the parameter η.

6.5 Discussion and conclusion

We have introduced a multi-source extension of the Kernel Mean Matching procedure. This MSKMM framework attempts to adapt the distribution of the training data to the target domain by reweighting the training samples according to their level of matching with the target distribution. We have described the model and its practical translation into a quadratic optimization problem, with the details that makes it possible to solve it to estimate importance weights associated with each instance of the training dataset. The main hyper-parameter of the model, η, controls the amount of structure that is added by the multi-source nature of the data available for training, on top of the standard KMM procedure. We have shown that when $\eta \to \infty$, our MSKMM is equivalent to the standard KMM where all the training instances are supposed to be drawn from a unique

distribution. When $\eta = 0$, our framework in facts models the target distribution as a mixture of the different source distributions.

We have performed simulations on a real world dataset from a Magneto-Encephalography (MEG) experiment. Each subject that participated in this experiment is an independent source of data. The task, the so-called *inter-subject decoding* problem, consisted in guessing the type of stimulus – amongst two categories – that was presented to a subject whose data was not used during training, from his brain activation pattern recorded using MEG. The initial results of our MSKMM framework on this MEG dataset are clearly disappointing. No significant gain was observed from simply using a Support Vector Classifier that used no information about the training subjects. In what follows, we discuss the potential reasons for this and propose some directions from future work.

The first reason that might explain the non-improvement over the standard SVC results might lie in the assumption that was implicitely made which implies that the original feature spaces of all sources are identical. This translates into the fact that the brain activity measured by MEG in all subjects share most of its specificity with regard to the task, i.e that the features that discriminate a *face* trial from a *scrambled* trial are the same for all subjects. This ignores the inter-subject variability, which – for MEG data – might result in changes in spatial and temporal features, i.e the topographies and the time-series of MEG recordings. We had implicitely assumed that examining whole brain maps – instead of localized patterns as in Chapters 4 and 5, and using a temporal down-sampling might reduce this variability, but our results suggest that this might not be sufficient. Therefore, an extra step of using a representation more robust to such variability might be necessary before attempting to match the distributions with MSKMM. A good example is given by the winner of the DecMeg competition [Barachant et al. 2013]

A second possibility lies in the method itself. The instance weighting principle has already proved effective in the past, but there is room for improvement both in the way to estimage the weights and in the method that uses the weights to perform the predictions. A first avenue might consist in a refinement of MSKMM so that it attempts to match the conditional probability distributions across sources instead of matching the marginals, as done in [Chattopadhyay et al. 2012]. This calls for obtaining labels for instances of the target domain, which can be done in two ways: i) acquiring truely labeled instances, for instance through a calibration experiment, which is a usual practice in BCI experiments and is implicetely done in fMRI experiments when using hyperalignment [Lorbert et al. 2012]; ii) estimating labels of instances of the target domain, for example by using the prediction of a classifier. A second avenue might attempt to improve the use of the weights, for example by using a truely transductive method in order to perform predictions on data of the target domain; this could be done with the work described in [Joachims 2003].

In all cases, the choice of the hyper-parameters remains an important problem

with such methods, which, in view of our results – we have not addressed in this work. In simplistic simulations (not shown), we have encountered large difficulties to reproduce the results presented in the original KMM article [Gretton et al. 2009], largely because the results are heavily influenced by the values of the hyper-parameters. For instance, the kernel used for the estimation of the weights and for the prediction do not have to be identical, nor the values of their hyper-parameters, and it is challenging to select them with a simple cross-validation [Gretton et al. 2009]. The difficulty of obtaining real improvements on practical problems with KMM is also reflected by the very small amount of studies that used it since its introduction in 2007. We therefore believe that a large amount of work remains to make KMM-like methods usable in practice.

6.6 Appendix - Solving the KMM optimization problem using cvxopt

In order to solve the optimization problem defined for KMM, we used the *qp* solver of the *cvxopt* Python module. The user's guide of *cvxopt* provides the following definitions of a quadratic program:

$$\begin{aligned} \text{minimise}(\underline{x}) \quad & \tfrac{1}{2}\underline{x}^T \, \underline{P} \, \underline{x} + \underline{q}^T \, \underline{x} \\ \text{subject to} \quad & \underline{G} \, \underline{x} \preceq \underline{h} \\ & \underline{A} \, \underline{x} = \underline{b} \end{aligned}$$

Note that all the variable names are underlined to differentiate them from the rest of the variable names used throughout the manuscript while keeping the definitions of the *cvxopt* user's guide.

The unknown variables to be estimated are placed in the $\underline{x}$ vector, and we therefore define $\underline{x} = \bar{\beta}$. $\underline{P}$ is a $N^s \times N^s$ matrix which is simply the Gram matrix computed on the elements of the source domain, i.e $\underline{P}_{ij} = k(x_i^s, x_j^s)$. The vector $\underline{q}$ is composed of N^s elements equal to $\underline{q}_i = \frac{N^s}{N^t} \sum_{j=1}^{N^t} k(x_i^s, x_j^t)$.

The two constraints in Eq. (6.6) develop into:

$$(6.6a) \Leftrightarrow \begin{cases} \forall n, & -\beta_n \leq 0 \\ \forall n, & \beta_n \leq B \end{cases} \Leftrightarrow \begin{cases} Id^{N^s} \; \bar{\beta} \preceq \begin{pmatrix} B \\ \vdots \\ B \end{pmatrix} \\ -Id^{N^s} \; \bar{\beta} \preceq \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \end{cases}$$

and

$$(6.6b) \Leftrightarrow \begin{cases} \frac{1}{N^s}\sum_{n=1}^{N^s}\beta_n - 1 \leq \epsilon \\ 1 - \frac{1}{N^s}\sum_{n=1}^{N^s}\beta_n \leq \epsilon \end{cases} \Leftrightarrow \begin{cases} \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \bar{\beta} \leq N^s \times (\epsilon + 1) \\ \begin{pmatrix} -1 & \cdots & -1 \end{pmatrix} \bar{\beta} \leq N^s \times (\epsilon - 1) \end{cases}$$

This leads to having (6.6a) & (6.6b) $\Leftrightarrow \underline{G}\ \underline{x} \preceq \underline{h}$ by defining the matrix $\underline{G}$ and the vector $\underline{h}$ as follows:

$$\underline{G} = \begin{pmatrix} Id(N^s) \\ -Id(N^s) \\ \begin{matrix} 1 & \cdots & 1 \\ -1 & \cdots & -1 \end{matrix} \end{pmatrix} \quad \text{and} \quad \underline{h} = \begin{pmatrix} \begin{matrix} B \\ \vdots \\ \vdots \\ B \end{matrix} \\ \begin{matrix} 0 \\ \vdots \\ \vdots \\ 0 \end{matrix} \\ \begin{matrix} N^s \times (\epsilon + 1) \\ N^s \times (\epsilon - 1) \end{matrix} \end{pmatrix}$$

Finally, $\underline{A}$ and $\underline{b}$ are unused.

6.7 Appendix - Solving the MSKMM optimization problem using cvxopt

We will now detail the different components of this program using the *cvxopt* definitions given in Appendix 1.

- $\underline{x}$ is the vector of size N containing the weights to be estimated, and it is equal to $\bar{\beta}$: it is composed of S blocks, and its s-th block contains the weights of subjects s, i.e $\beta_1^s, \cdots, \beta_{N_s}^s$;
- $\underline{P}$ is a matrix of total size $N \times N$, structured in $S \times S$ blocks; the block (s_1, s_2) is composed of the kernel values between samples of subjects s_1 and s_2, weighted by the inverse product of the sample sizes for these two subjects: on line i and column j, this block takes the following value:

$$\frac{1}{N^{s_1} \times N^{s_2}} k(x_i^{s_1}, x_j^{s_2}) \tag{6.15}$$

- $\underline{q}$ is a vector with N elements organised in S blocks, one for each subject of the training set; the i-th element of the s-th block corresponding to subject

s is given by

$$-\frac{1}{N^s \times N^t} \sum_{j=1}^{j=N^t} k(x_i^s, x_j^t) \tag{6.16}$$

Let us now develop the different constraints. First, we have

$$(6.10c) \Leftrightarrow \forall (s,n), -\beta_n \leq 0 \Leftrightarrow \underline{G_1}\ \underline{x} \preceq \underline{h_1},$$

where $\underline{h_1}$ is the constant zero vector of size N, $\underline{G_1} = -Id^{N^S}$ and, as a reminder, $\underline{x} = \bar{\beta}$. Then we have

$$(6.10a) \Leftrightarrow \begin{cases} \forall (s,n),\ \beta_n^s - \frac{1}{N^s} \sum_{n'=1}^{N^s} \beta_{n'}^s \leq \eta \\ \forall (s,n),\ \frac{1}{N^s} \sum_{n'=1}^{N^s} \beta_{n'}^s - \beta_n^s \leq \eta \end{cases} \Leftrightarrow \begin{cases} \underline{G_2}\ \underline{x} \preceq \underline{h_2} \\ \underline{G_3}\ \underline{x} \preceq \underline{h_3} \end{cases}$$

where $\underline{h_1}$ and $\underline{h_2}$ are vectors of size N with constant value η, and $\underline{G_2} = -\underline{G_3}$ is a matrix of size $N \times N$ given by:

$$\underline{G_2} = Id^N - \begin{pmatrix} \boxed{\begin{matrix} \frac{1}{N^1} & \cdots & \frac{1}{N^1} \\ \vdots & \ddots & \vdots \\ \frac{1}{N^1} & \cdots & \frac{1}{N^1} \end{matrix}} & & 0 \\ & \ddots & \\ & & \ddots \\ & & & \ddots \\ 0 & & & \boxed{\begin{matrix} \frac{1}{N^S} & \cdots & \frac{1}{N^S} \\ \vdots & \ddots & \vdots \\ \frac{1}{N^S} & \cdots & \frac{1}{N^S} \end{matrix}} \end{pmatrix}$$

With these definitions of $\underline{G_1}$, $\underline{G_2}$, $\underline{G_3}$, $\underline{h_1}$, $\underline{h_2}$ and $\underline{h_3}$, we get

$$(6.10a)\ \&\ (6.10c) \Leftrightarrow \underline{G}\ \underline{x} \preceq \underline{h}$$

by defining the block matrix $\underline{G}$ and the block vector $\underline{h}$ as follows:

$$\underline{G} = \begin{pmatrix} \underline{G}_1 \\ \underline{G}_2 \\ \underline{G}_3 \end{pmatrix} \quad \text{and} \quad \underline{h} = \begin{pmatrix} \underline{h}_1 \\ \underline{h}_2 \\ \underline{h}_3 \end{pmatrix}$$

Finally, we examine the last constraint, given by Eq. (??):

$$(6.10b) \Leftrightarrow \sum_{s=1}^{S} \frac{1}{N^s} \sum_{n=1}^{N^s} \beta_n^s = 1 \Leftrightarrow \underline{A}\, \underline{x} = \underline{b},$$

where $\underline{b}$ is the scalar value 1, and $\underline{A}$ is the following vector that contains S constant blocks:

$$\underline{A} = \begin{pmatrix} \frac{1}{N^1} \\ \vdots \\ \frac{1}{N^1} \\ \vdots \\ \vdots \\ \vdots \\ \frac{1}{N^S} \\ \vdots \\ \frac{1}{N^S} \end{pmatrix}$$

We have now defined the full configuration that enables us to solve the MSKMM optimization problem in *cvxopt*.

7 Conclusion

In this thesis, we have introduced a unifying perspective on neuroimaging data analysis when data from multiple subjects are available – which is the most common case (see Chapter 1). We simply stated that each subject is a *source* of data and advocated the use of the *multi-source learning* setting. The term *multi-source* has been used for a long time to characterize data analysis problems, sometimes in different contexts with different meanings. Therefore, we started this thesis in Chapter 2 by clearly recalling the *multi-source* setting defined in machine learning. In the rest of this Chapter, we argued that the *multi-source* framework is the most natural one for dealing with multi-subject datasets such as those commonly available in neuroscience. We therefore precisely instanciated this setting to the *inter-subject* learning questions offered in neuroimaging. Inter-subject learning is the generic problem that we face when attempting to perform predictions on data from a subject that was not available at training time.

A multi-source problem should be tackled from two different angles. First, if the input spaces are different across sources, a common space should be found. Secondly, because the data from the different sources come from different subjects, their probability distribution are instrinsically different and should somehow be matched. We have presented different approaches that address these different problems. First, addressing the *inter-subject decoding* problem for fMRI data in Chapter 4, we have introduced a graphical representation and a graph kernel that implicitly bring local functional patterns recorded in different subjects into a common graph-space, which allows to succesfully overcome the fine scale inter-subject variability. Secondly, we studied how to detect group differences in local cortical shape in Chapter 5. Using graphical representations of patterns of sulcal pits, we introduced a graph kernel to project such patterns into a graph space common to all subjects. We then presented a searchlight scheme and a spatial inference method in order to exploit the properties of this common space, which proved effective to detect cortical asymmetries and sex differences. Finally, in Chapter 6, we introduced a multi-source extension of the kernel mean matching procedure in order to build a multi-source domain adaptation method. However, when applied to an inter-subject decoding task on a MEG dataset, this approach did not outperform standard classification methods.

In view of these results, it seems critical to address the problem of the representation of the neuroimaging data by asking in which case it is valid to assume

that the feature spaces of different subjects are the same. From a very broad perspective, all these problems share a common object of study, the brain, and a common single modality of observation (either anatomical MRI, functional MRI or MEG). Therefore, it seems to make sense to assume that the original feature spaces are identical for all subjects and to attempt to use multi-source domain adaptation techniques. However, the lack of success of this approach (used in Chapter 6) might lead us to accept that the inter-individual differences in the anatomical and functional organization of the brain are too strong for this assumption to hold – but further work is clearly necessary to confirm this claim. In any case, it is clearly relevant to seek representations that are common across subjects, as demonstrated by the success of the methods introduced in Chapters 4 and 5. Provided enough data points are available, the use of multi-source domain adaptation techniques could well be advocated as a second stage of analysis, working in such common space.

Finally, we would like to emphasize yet again that the inter-subject learning problems brought by neuroimaging datasets are natural applications of the multi-source framework defined in machine learning. Actors of the two scientific communities should naturally benefit from eachother through stronger interactions and cross-fertilization. On the one hand, neuroscientists who attempt to grasp the commonalities across subjects and the differences between groups of subjects should gain a finer understanding from using advanced techniques of multi-source learning. On the other hand, computer scientists working on multi-source learning should find an opportunity to validate and improve their methods in the challenging datasets provided by neuroscientists.

Bibliography

[1] H. Abdi, L. J. Williams, A. C. Connolly, M. I. Gobbini, J. P. Dunlop, and J. s. V. Haxby. “Multiple Subject Barycentric Discriminant Analysis (MUSUBADA): How to Assign Scans to Categories without Using Spatial Normalization”. In: *Computational and Mathematical Methods in Medicine* 2012 (2012), pp. 1–15 (cit. on p. 44).

[2] M. Aizerman, E. Braverman, and L. Rozonoer. “Theoretical foundations of the potential function method in pattern recognition learning.” In: *Automation and Remote Control* 25 (1964), pp. 821–837 (cit. on p. 47).

[3] P. Arbeláez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik. “Semantic segmentation using regions and parts”. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3378–3385 (cit. on p. 33).

[4] J. Ashburner. “Computational anatomy with the SPM software”. In: *Magnetic Resonance Imaging* 27.8 (Mar. 2009), pp. 1163–1174 (cit. on p. 19).

[5] J. Ashburner. “SPM: A history”. In: *NeuroImage* 62.2 (Aug. 2012), pp. 791–800 (cit. on p. 16).

[6] G. Auzias, C. Breuil, S. Takerkart, and C. Deruelle. “Detectability of brain structure abnormalities related to autism through MRI-derived measures from multiple scanners”. In: *Proceedings of BHI*. IEEE, June 2014, pp. 314–317 (cit. on p. 23).

[7] G. Auzias, L. Brun, C. Deruelle, and O. Coulon. “Deep sulcal landmarks: Algorithmic and conceptual improvements in the definition and extraction of sulcal pits”. In: *NeuroImage* 111 (May 2015), pp. 12–25 (cit. on pp. 76–78, 86, 108, 109).

[8] G. Auzias, J. Lefevre, A. Le Troter, C. Fischer, M. Perrot, J. Regis, and O. Coulon. “Model-Driven Harmonic Parameterization of the Cortical Surface: HIP-HOP”. In: *Medical Imaging, IEEE Transactions on* 32.5 (2013), pp. 873–887 (cit. on p. 56).

[9] G. Auzias, M. Viellard, S. Takerkart, N. Villeneuve, F. Poinso, D. Da Fonséca, N. Girard, and C. Deruelle. “Atypical sulcal anatomy in young children with autism spectrum disorder”. In: *NeuroImage: Clinical* 4 (2014), pp. 593–603 (cit. on pp. 23, 76, 111).

[10] C. I. Baker, E. Peli, N. Knouf, and N. G. Kanwisher. “Reorganization of Visual Processing in Macular Degeneration”. In: *The Journal of Neuroscience* 25.3 (2005), pp. 614–618 (cit. on p. 68).

[11] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten. “Classification of covariance matrices using a Riemannian-based kernel for BCI applications”. In: *Neurocomputing* 112 (July 2013), pp. 172–178 (cit. on pp. 115, 127).

[12] T. E. J. Behrens, H. Johansen-Berg, M. W. Woolrich, S. M. Smith, C. A. M. Wheeler-Kingshott, P. A. Boulby, G. J. Barker, E. L. Sillery, K. Sheehan, O. Ciccarelli, A. J. Thompson, J. M. Brady, and P. M. Matthews. “Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging”. In: *Nat Neurosci* 6.7 (2003), pp. 750–757 (cit. on p. 46).

[13] O. Beijbom. “Domain adaptations for computer vision applications”. In: *arXiv preprint arXiv:1211.4860* (2012) (cit. on p. 34).

[14] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. “A theory of learning from different domains”. In: *Machine Learning* 79.1-2 (May 2010), pp. 151–175 (cit. on p. 37).

[15] Y. Bengio. “Deep learning of representations for unsupervised and transfer learning”. In: *Unsupervised and Transfer Learning Challenges in Machine Learning, Volume 7* (2012), p. 19 (cit. on p. 34).

[16] M. Bilello, M. Arkuszewski, P. Nucifora, I. Nasrallah, E. R. Melhem, L. Cirillo, and J. Krejza. “Multiple sclerosis: identification of temporal changes in brain lesions with computer-assisted detection software”. In: *The neuroradiology journal* 26.2 (2013), pp. 143–150 (cit. on p. 23).

[17] T. Blumensath, S. Jbabdi, M. F. Glasser, D. C. V. Essen, K. Ugurbil, T. E. J. Behrens, and S. M. Smith. “Spatially constrained hierarchical parcellation of the brain with resting-state fMRI”. In: *NeuroImage* 76 (2013), pp. 313–324 (cit. on p. 46).

[18] K. M. Borgwardt and H.-P. Kriegel. “Shortest-Path Kernels on Graphs”. In: *2013 IEEE 13th International Conference on Data Mining* (2005), pp. 74–81 (cit. on p. 70).

[19] M. Boucher, S. Whitesides, and A. Evans. “Depth potential function for folding pattern representation, registration and analysis.” In: *Medical image analysis* 13.2 (Apr. 2009), pp. 203–14. pmid: 18996043 (cit. on p. 78).

[20] M. Brett, J. Taylor, C. Burns, K. J. Millman, F. Perez, A. Roche, B. Thirion, and M. J. D’Esposito. “NIPY: an open library and development framework for FMRI data analysis”. In: *NeuroImage* 47 (2009), S196 (cit. on p. 55).

[21] J. Bruna and S. Mallat. “Invariant Scattering Convolution Networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (Aug. 2013), pp. 1872–1886 (cit. on p. 32).

[22] L. Bruzzone and M. Marconcini. “Domain Adaptation Problems: A DASVM Classification Technique and a Circular Validation Strategy.” In: *IEEE Trans. Pattern Anal. Mach. Intell.* 32.5 (Apr. 20, 2010), pp. 770–787 (cit. on p. 36).

[23] E. T. Bullmore, J. Suckling, S. Overmeyer, S. Rabe-Hesketh, E. Taylor, and M. J. Brammer. “Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain”. In: *Medical Imaging, IEEE Transactions on* 18.1 (1999), pp. 32–42 (cit. on pp. 17, 83, 86, 87, 110).

[24] H. Bunke. “On a relation between graph edit distance and maximum common subgraph”. In: *Pattern Recognition Letters* 18.9 (1997), pp. 689–694 (cit. on p. 51).

[25] C. Cabral, M. Silveira, and P. Figueiredo. “Decoding visual brain states from fMRI using an ensemble of classifiers”. In: *Pattern Recognition* 45.6 (2012), pp. 2064–2074 (cit. on p. 43).

[26] Y. Cao, J. Xu, T. Y. Liu, H. Li, Y. Huang, and H. W. Hon. “Adapting ranking svm to document retrieval”. In: *SIGIR06: Proceedings of the 29th Annual International Conference on Research and Development in Information Retrieval*. 2006, pp. 186–193 (cit. on p. 48).

[27] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. Adaptive computation and machine learning. Cambridge, MA, USA: MIT Press, Sept. 2006. 508 pages (cit. on p. 36).

[28] R. Chattopadhyay, Q. Sun, W. Fan, I. Davidson, S. Panchanathan, and J. Ye. “Multisource domain adaptation and its application to early detection of fatigue”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6.4 (2012), p. 18 (cit. on pp. 37, 122, 127).

[29] M. Chen, K. Q. Weinberger, and J. Blitzer. “Co-Training for Domain Adaptation”. In: *Advances in Neural Information Processing Systems 24*. Ed. by J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger. Curran Associates, Inc., 2011, pp. 2456–2464 (cit. on p. 35).

[30] Y. Chen, P. Namburi, L. T. Elliott, J. Heinzle, C. S. Soon, M. W. Chee, and J.-D. Haynes. “Cortical surface-based searchlight decoding”. In: *NeuroImage* 56.2 (May 2011), pp. 582–592 (cit. on pp. 81, 83, 110).

[31] D. Ciresan, U. Meier, and J. Schmidhuber. “Multi-column deep neural networks for image classification”. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3642–3649 (cit. on pp. 31, 33).

[32] J. A. Clithero, D. V. Smith, R. M. Carter, and S. A. Huettel. “Within- and cross-participant classifiers reveal different neural coding of information”. In: *NeuroImage* 56.2 (2011), pp. 699–708 (cit. on p. 44).

[33] C. Cortes and V. Vapnik. “Support Vector Networks”. In: *Machine Learning* 20 (1995), pp. 1–25 (cit. on p. 46).

[34] O. Coulon, J.-F. Mangin, J.-B. Poline, M. Zilbovicius, D. Roumenov, Y. Samson, V. Frouin, and I. Bloch. “Structural Group Analysis of Functional Activation Maps”. In: *NeuroImage* 11.6 (2000), pp. 767–782 (cit. on pp. 44–46).

[35] M. N. Coutanche, S. L. Thompson-Schill, and R. T. Schultz. “Multi-voxel pattern analysis of fMRI data predicts clinical symptom severity”. In: *NeuroImage* 57.1 (2011), pp. 113–123 (cit. on p. 43).

[36] K. Crammer, M. Kearns, and J. Wortman. “Learning from multiple sources”. In: *The Journal of Machine Learning Research* 9 (2008), pp. 1757–1774 (cit. on p. 37).

[37] R. Cuingnet, J. A. Glaunes, M. Chupin, H. Benali, and O. Colliot. “Spatial and Anatomical Regularization of SVM: A General Framework for Neuroimaging Data”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.3 (2013), pp. 682–696 (cit. on p. 40).

[38] G. Dahl, D. Yu, L. Deng, and A. Acero. “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition”. In: *Audio, Speech, and Language Processing, IEEE Transactions on* 20.1 (Jan. 2012), pp. 30–42 (cit. on p. 33).

[39] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu. “Transferring naive bayes classifiers for text classification”. In: *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*. Vol. 22. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007, p. 540 (cit. on p. 36).

[40] A. Dale. “Cortical Surface-Based Analysis I. Segmentation and Surface Reconstruction”. In: *NeuroImage* 9.2 (1999), pp. 179–194 (cit. on p. 55).

[41] H. Daumé III and D. Marcu. “Domain Adaptation for Statistical Classifiers”. In: *Journal of Artificial Intelligence Research (JAIR)* 26 (2006), pp. 101–126 (cit. on p. 36).

[42] C. Davatzikos, K. Ruparel, Y. Fan, D. G. Shen, M. Acharyya, J. W. Loughead, R. C. Gur, and D. D. Langleben. “Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection”. In: *NeuroImage* 28.3 (2005), pp. 663–668 (cit. on p. 44).

[43] C. Destrieux, B. Fischl, A. Dale, and E. Halgren. “Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature”. In: *NeuroImage* 53.1 (2010), pp. 1–15 (cit. on pp. 55, 67).

[44] T. G. Dietterich. “Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms”. In: *Neural Computation* 10 (1998), pp. 1895–1923 (cit. on p. 58).

[45] L. Duan, I. W. Tsang, D. Xu, and T.-S. Chua. “Domain adaptation from multiple sources via auxiliary classifiers”. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 289–296 (cit. on p. 38).

[46] E. Duchesnay, A. Cachia, A. Roche, D. Rivière, Y. Cointepas, D. Papadopoulos-Orfanos, M. Zilbovicius, J.-L. Martinot, J. Régis, and J.-F. Mangin. “Classification Based on Cortical Folding Patterns”. In: *IEEE Trans. Med. Imaging* 26.4 (2007), pp. 553–565 (cit. on pp. 108, 109).

[47] M. Eshera and K.-s. Fu. “An Image Understanding System Using Attributed Symbolic Representation and Inexact Graph-Matching”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* PAMI-8.5 (1986), pp. 604–618 (cit. on p. 50).

[48] D. C. V. Essen and D. L. Dierker. “Surface-Based and Probabilistic Atlases of Primate Cerebral Cortex”. In: *Neuron* 56.2 (2007), pp. 209–225 (cit. on p. 44).

[49] J. A. Etzel, J. M. Zacks, and T. S. Braver. “Searchlight analysis: Promise, pitfalls, and potential”. In: *NeuroImage* 78 (Sept. 2013), pp. 261–269 (cit. on p. 111).

[50] M. Fang, Y. Guo, X. Zhang, and X. Li. “Multi-source transfer learning based on label shared subspace”. In: *Pattern Recognition Letters* 51 (Jan. 2015), pp. 101–106 (cit. on p. 39).

[51] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. “Object detection with discriminatively trained part-based models”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32.9 (2010), pp. 1627–1645 (cit. on p. 33).

[52] A. Feragen, N. Kasenburg, J. Petersen, M. de Bruijne, and K. Borgwardt. “Scalable kernels for graphs with continuous attributes”. In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Curran Associates, Inc., 2013, pp. 216–224 (cit. on p. 70).

[53] C. Ferri, J. Hernández-Orallo, and R. Modroiu. “An experimental comparison of performance measures for classification”. In: *Pattern Recognition Letters* 30.1 (2009), pp. 27–38 (cit. on p. 58).

[54] B. Fischl, M. I. Sereno, R. B. Tootell, A. M. Dale, et al. “High-resolution intersubject averaging and a coordinate system for the cortical surface”. In: *Human brain mapping* 8.4 (1999), pp. 272–284 (cit. on p. 83).

[55] G. Flandin, F. Kherif, X. Pennec, G. Malandain, N. Ayache, and J.-B. Poline. “Improved detection sensitivity of functional MRI data using a brain parcellation technique”. In: *Proc. 5th MICCAI*. LNCS 2488 (Part I). INRIA Projet Epidaure, Sophia Antipolis/SHFJ-CEA, Orsay, 2002, pp. 467–474 (cit. on pp. 45, 49).

[56] E. Formisano, D.-S. Kim, F. Di Salle, P.-F. van de Moortele, K. Ugurbil, and R. Goebel. “Mirror-Symmetric Tonotopic Maps in Human Primary Auditory Cortex”. In: *Neuron* 40.4 (13, 2003), pp. 859–869 (cit. on pp. 55, 59, 65, 67).

[57] E. Formisano, D.-S. Kim, F. Di Salle, P.-F. van de Moortele, K. Ugurbil, and R. Goebel. “Mirror-symmetric tonotopic maps in human primary auditory cortex”. In: *Neuron* 40.4 (2003), pp. 859–869 (cit. on pp. 72, 73).

[58] P. Frasconi, M. Gori, and A. Sperduti. “On the efficient classification of data structures by neural networks”. In: *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI 97)*. 1997 (cit. on pp. 45, 46).

[59] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson. “The clinical use of structural MRI in Alzheimer disease”. In: *Nature Reviews Neurology* 6.2 (Feb. 2010), pp. 67–77 (cit. on p. 23).

[60] K. J. Friston. *Statistical parametric mapping: the analysis of funtional brain images*. 1st ed. Amsterdam ; Boston: Elsevier/Academic Press, 2007 (cit. on pp. 43, 55).

[61] É. Fromont, M.-O. Cordier, and R. Quiniou. “Learning from multi-source data”. In: *Knowledge Discovery in Databases: PKDD 2004*. Springer, 2004, pp. 503–505 (cit. on p. 38).

[62] T. Gärtner, P. A. Flach, and S. Wrobel. “On Graph Kernels: Hardness Results and Efficient Alternatives”. In: *Proc. of the 16th Conf. on Computational Learning Theory*. 2003 (cit. on pp. 51, 79).

[63] K. Geras and C. Sutton. “Multiple-source cross-validation”. In: *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. 2013, pp. 1292–1300 (cit. on p. 39).

[64] K. Gkirtzou, J. Honorio, D. Samaras, R. Goldstein, and M. B. Blaschko. “fMRI Analysis with Sparse Weisfeiler-Lehman Graph Statistics”. In: *Machine Learning in Medical Imaging, LNCS vol. 8184*. 2013, pp. 90–97 (cit. on pp. 46, 51, 69).

[65] L. Goldfarb, D. Gay, O. Golubitsky, and D. Korkin. *What is a Structural Representation? (Second Version)*. Faculty of Computer Science, University of New Brunswick, 2004 (cit. on p. 31).

[66] C. D. Good, I. Johnsrude, J. Ashburner, R. N. Henson, K. J. Friston, and R. S. Frackowiak. “Cerebral Asymmetry and the Effects of Sex and Handedness on Brain Structure: A Voxel-Based Morphometric Analysis of 465 Normal Adult Human Brains”. In: *NeuroImage* 14.3 (Sept. 2001), pp. 685–700 (cit. on p. 108).

[67] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. “Covariate shift by kernel mean matching”. In: *Dataset shift in machine learning* 3.4 (2009), p. 5 (cit. on pp. 36, 115, 116, 118, 124, 128).

[68] D. N. Greve, L. Van der Haegen, Q. Cai, S. Stufflebeam, M. R. Sabuncu, B. Fischl, and M. Brysbaert. “A Surface-based Analysis of Language Lateralization and Cortical Asymmetry”. In: *Journal of Cognitive Neuroscience* 25.9 (Sept. 2013), pp. 1477–1492 (cit. on p. 91).

[69] L. Grosenick, B. Klingenberg, K. Katovich, B. Knutson, and J. E. Taylor. “Interpretable whole-brain prediction analysis with GraphNet”. In: *NeuroImage* 72 (2013), pp. 304–321 (cit. on pp. 40, 44).

[70] S. J. Hanson and A. Schmidt. “High-resolution imaging of the fusiform face area (FFA) using multivariate non-linear classifiers shows diagnosticity for non-face categories”. In: *NeuroImage* 54.2 (2011), pp. 1715–1734 (cit. on p. 67).

[71] D. Haussler. *Convolution kernels on discrete structures*. UCSC-CRL-99-10. UC Santa Cruz, 1999 (cit. on pp. 51, 79).

[72] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini. “Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex”. In: *Science* 293.5539 (2001), pp. 2425–2430 (cit. on pp. 43, 67).

[73] J. V. Haxby, J. S. Guntupalli, A. C. Connolly, Y. O. Halchenko, B. R. Conroy, M. I. Gobbini, M. Hanke, and P. J. Ramadge. “A Common, High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex”. In: *Neuron* 72.2 (2011), pp. 404–416 (cit. on pp. 39, 40, 44, 46, 67, 68).

[74] J.-D. Haynes and G. Rees. “Decoding mental states from brain activity in humans”. In: *Nature Reviews Neuroscience* 7.7 (2006), pp. 523–534 (cit. on p. 43).

[75] G. E. Hinton and R. R. Salakhutdinov. “Reducing the Dimensionality of Data with Neural Networks”. In: *Science* 313.5786 (2006), pp. 504–507 (cit. on p. 33).

[76] J. Honorio, D. Tomasi, R. Z. Goldstein, H.-C. Leung, and D. Samaras. “Can a Single Brain Region Predict a Disorder?” In: *Medical Imaging, IEEE Transactions on* 31.11 (2012), pp. 2062–2072 (cit. on p. 43).

[77] P. Huang, G. Wang, and S. Qin. “Boosting for transfer learning from multiple data sources”. In: *Pattern Recognition Letters* 33.5 (Apr. 2012), pp. 568–579 (cit. on p. 38).

[78] C. Humphries, E. Liebenthal, and J. R. Binder. “Tonotopic organization of human auditory cortex”. In: *NeuroImage* 50.3 (2010), pp. 1202–1211 (cit. on pp. 55, 59, 65, 67).

[79] C. Humphries, E. Liebenthal, and J. R. Binder. “Tonotopic organization of human auditory cortex”. In: *NeuroImage* 50.3 (Apr. 2010), pp. 1202–1211 (cit. on pp. 72, 73).

[80] K. Im, H. J. Jo, J.-F. Mangin, A. C. Evans, S. I. Kim, and J.-M. Lee. “Spatial distribution of deep sulcal landmarks and hemispherical asymmetry on the cortical surface.” In: *Cerebral cortex* 20.3 (Mar. 2010), pp. 602–11. PMID: 19561060 (cit. on pp. 76–78, 86, 108, 109, 111, 112).

[81] K. Im, R. Pienaar, J.-M. Lee, J.-K. Seong, Y. Y. Choi, K. H. Lee, and P. E. Grant. “Quantitative comparison and analysis of sulcal patterns using sulcal graph matching: a twin study.” In: *NeuroImage* 57.3 (Aug. 2011), pp. 1077–86. pmid: 21596139 (cit. on pp. 77, 79, 92, 111, 112).

[82] K. Im, R. Pienaar, M. J. Paldino, N. Gaab, A. M. Galaburda, and P. E. Grant. “Quantification and Discrimination of Abnormal Sulcal Patterns in Polymicrogyria.” In: *Cerebral cortex* 23.12 (Sept. 2012), pp. 3007–15. pmid: 22989584 (cit. on pp. 76, 77, 86, 112).

[83] K. Im, N. M. Raschle, S. A. Smith, P. Ellen Grant, and N. Gaab. “Atypical Sulcal Pattern in Children with Developmental Dyslexia and At-Risk Kindergarteners”. In: *Cerebral Cortex* (Jan. 9, 2015) (cit. on pp. 77, 112).

[84] S. Janson. “Large deviations for sums of partly dependent random variables”. In: *Random Struct. Algorithms* 24.3 (2004), pp. 234–248 (cit. on p. 48).

[85] B. Jie, D. Zhang, C.-Y. Wee, and D. Shen. “Topological graph kernel on multiple thresholded functional connectivity networks for mild cognitive impairment classification.” In: *Human brain mapping* 35.7 (July 2014), pp. 2876–97. pmid: 24038749 (cit. on pp. 46, 69).

[86] T. Joachims. “Optimizing search engines using clickthrough data”. In: *KDD02: Proceedings of the 8th international conference on Knowledge discovery and data mining*. 2002, pp. 133–142 (cit. on p. 48).

[87] T. Joachims. “Transductive Learning via Spectral Graph Partitioning”. In: *International Conference on Machine Learning (ICML)*. 2003, pp. 290–297 (cit. on p. 127).

[88] Y. Kamitani and F. Tong. “Decoding the visual and subjective contents of the human brain”. In: *Nat Neurosci* 8.5 (May 2005), pp. 679–685 (cit. on p. 43).

[89] D. R. Kisku, A. Rattani, E. Grosso, and M. Tistarelli. “Face identification by SIFT-based complete graph topology”. In: *Automatic Identification Advanced Technologies, 2007 IEEE Workshop on*. IEEE, 2007, pp. 63–68 (cit. on p. 33).

[90] A. Knops, B. Thirion, E. M. Hubbard, V. Michel, and S. Dehaene. “Recruitment of an Area Involved in Eye Movements During Mental Arithmetic”. In: *Science* 324.5934 (2009), pp. 1583–1585 (cit. on p. 43).

[91] R. Kohavi. “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection”. In: Morgan Kaufmann, 1995, pp. 1137–1143 (cit. on p. 83).

[92] S. Koyamada, Y. Shikauchi, K. Nakae, M. Koyama, and S. Ishii. “Deep learning of fMRI big data: a novel approach to subject-transfer decoding”. In: *arXiv preprint arXiv:1502.00093* (2015) (cit. on p. 34).

[93] N. Kriege and P. Mutzel. “Subgraph Matching Kernels for Attributed Graphs”. In: *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. Ed. by J. Langford and J. Pineau. ICML ’12. New York, NY, USA: Omnipress, July 2012, pp. 1015–1022 (cit. on p. 70).

[94] N. Kriegeskorte, R. Goebel, and P. Bandettini. “Information-based functional brain mapping”. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.10 (2006), pp. 3863–3868 (cit. on pp. 22, 77, 81–83, 110).

[95] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*. 2012, pp. 1097–1105 (cit. on p. 33).

[96] S. LaConte, S. Strother, V. Cherkassky, J. Anderson, and X. Hu. “Support vector machines for temporal classification of block design fMRI data”. In: *NeuroImage* 26.2 (2005), pp. 317–329 (cit. on p. 66).

[97] J. Lafferty, A. McCallum, and F. Pereira. “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”. In: *Proc. 18th Int. Conf. on Machine Learning (ICML 2001)*. 2001, pp. 282–289 (cit. on pp. 45, 46).

[98] D. Lashkari, R. Sridharan, E. Vul, P.-J. Hsieh, N. Kanwisher, and P. Golland. “Search for patterns of functional specificity in the brain: A nonparametric hierarchical Bayesian model for group fMRI data”. In: *NeuroImage* 59.2 (2012), pp. 1348–1368 (cit. on pp. 45, 46).

[99] J. Lefevre, F. Leroy, S. Khan, J. Dubois, P. S. Huppi, S. Baillet, and J.-F. Mangin. “Identification of Growth Seeds in the Neonate Brain through Surfacic Helmholtz Decomposition”. In: *Information Processing in Medical Imaging, 21st International Conference, IPMI 2009, Williamsburg, VA, USA, July 5-10, 2009. Proceedings*. 2009, pp. 252–263 (cit. on p. 112).

[100] Q. Li. “Literature Survey: Domain Adaptation Algorithms for Natural Language Processing”. In: (2012) (cit. on p. 34).

[101] Z. Lin, G. Ding, and M. Hu. “Multi-source image auto-annotation.” In: *ICIP*. 2013, pp. 2567–2571 (cit. on p. 38).

[102] T. Lindeberg. “Effective scale: a natural unit for measuring scale-space lifetime”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 15.10 (Oct. 1993), pp. 1068–1074 (cit. on pp. 87, 95, 111).

[103] T. Y. Liu. “Learning to rank for information retrieval”. In: *Foundations and Trends in Information Retrieval* 3 (2009), pp. 225–331 (cit. on p. 48).

[104] A. Lorbert and P. J. Ramadge. “Kernel Hyperalignment.” In: *NIPS*. 2012, pp. 1799–1807 (cit. on pp. 40, 127).

[105] D. G. Lowe. “Object recognition from local scale-invariant features”. In: *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. Vol. 2. Ieee, 1999, pp. 1150–1157 (cit. on p. 32).

[106] B. Lv, J. Li, H. He, M. Li, M. Zhao, L. Ai, F. Yan, J. Xian, and Z. Wang. “Gender consistency and difference in healthy adults revealed by cortical thickness”. In: *NeuroImage* 53.2 (Nov. 1, 2010), pp. 373–382 (cit. on p. 109).

[107] J. Maclaren, Z. Han, S. B. Vos, N. Fischbein, and R. Bammer. “Reliability of brain volume measurements: A test-retest dataset”. In: *Scientific Data* 1 (Oct. 14, 2014), p. 140037 (cit. on p. 29).

[108] P. Mahé, L. Ralaivola, V. Stoven, and J.-P. Vert. “The Pharmacophore Kernel for Virtual Screening with Support Vector Machines”. In: *J. Chem. Inf. Model.* 46.5 (2006), pp. 2003–2014 (cit. on pp. 45, 46).

[109] P. Mahé and J. P. Vert. “Graph kernels based on tree patterns for molecules”. In: *Machine Learning* 75.1 (2009), pp. 3–35 (cit. on p. 51).

[110] A. Mahmoudi, S. Takerkart, F. Regragui, D. Boussaoud, and A. Brovelli. “Multivoxel Pattern Analysis for fMRI Data: A Review”. In: *Computational and Mathematical Methods in Medicine* 2012 (2012), pp. 1–14 (cit. on pp. 43, 51).

[111] Y. Mansour, M. Mohri, and A. Rostamizadeh. “Domain adaptation with multiple sources”. In: *Advances in neural information processing systems*. 2009, pp. 1041–1048 (cit. on p. 37).

[112] A. Margolis. “A literature review of domain adaptation with unlabeled data”. In: *Rapport Technique, University of Washington* (2011), p. 35 (cit. on p. 34).

[113] A. F. Marquand, M. Brammer, S. C. R. Williams, and O. M. Doyle. “Bayesian multi-task learning for decoding multi-subject neuroimaging data”. In: *NeuroImage* 92 (2014), pp. 298–311 (cit. on pp. 39, 43, 115).

[114] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY, USA: Henry Holt and Co., Inc., 1982 (cit. on p. 31).

[115] J. Menke and T. R. Martinez. "Using permutations instead of student's t distribution for p-values in paired-difference algorithm comparisons". In: *Proc. IEEE Joint Conference on Neural Networks*. 2004 (cit. on p. 58).

[116] K. Meyer, J. T. Kaplan, R. Essex, C. Webber, H. Damasio, and A. Damasio. "Predicting visual stimuli on the basis of activity in auditory cortices". In: *Nature Neuroscience* 13.6 (2010), pp. 667–668 (cit. on p. 43).

[117] V. Michel, A. Gramfort, G. Varoquaux, E. Eger, C. Keribin, and B. Thirion. "A supervised clustering approach for fMRI-based inference of brain states". In: *Pattern Recognition* 45.6 (2012), pp. 2041–2049 (cit. on p. 49).

[118] V. Michel, A. Gramfort, G. Varoquaux, E. Eger, and B. Thirion. "Total Variation Regularization for fMRI-Based Prediction of Behavior". In: *IEEE Transactions on Medical Imaging* 30.7 (2011), pp. 1328–1340 (cit. on p. 40).

[119] T. Mikolov, A. Deoras, S. Kombrink, L. Burget, and J. Černocký. "Empirical Evaluation and Combination of Advanced Language Modeling Techniques". In: *Proceedings of Interspeech 2011*. Vol. 2011. = {Tomáš Mikolov and Anoop Deoras and Stefan Kombrink and Lukáš Burget and Jan Černocký}. Florence, IT: International Speech Communication Association, 2011, pp. 605–608 (cit. on p. 33).

[120] M. Misaki, Y. Kim, P. A. Bandettini, and N. Kriegeskorte. "Comparison of multivariate classifiers and response normalizations for pattern-information fMRI". In: *NeuroImage* 53.1 (2010), pp. 103–118 (cit. on p. 66).

[121] T. Mitchell, R. Hutchinson, R. Niculescu, F. Pereir a, X. Wang, M. Just, and S. Newman. "Learning to Decode Cognitive States from Brain Images". In: *Machine Learning* 57.1-2 (2004), pp. 145–175 (cit. on p. 44).

[122] M. Moerel, F. De Martino, and E. Formisano. "Processing of Natural Sounds in Human Auditory Cortex: Tonotopy, Spectral Tuning, and Relation to Voice Sensitivity". In: *Journal of Neuroscience* 32.41 (Oct. 10, 2012), pp. 14205–14216 (cit. on p. 72).

[123] F. Mokhtari and G.-A. Hossein-Zadeh. "Decoding brain states using backward edge elimination and graph kernels in fMRI connectivity networks". In: *Journal of Neuroscience Methods* 212.2 (2013), pp. 259–268 (cit. on pp. 46, 70).

[124] A. Morel, P. E. Garraghty, and J. H. Kaas. "Tonotopic organization, architectonic fields, and connections of auditory cortex in macaque monkeys". In: *Journal of Comparative Neurology* 335.3 (1993), pp. 437–459 (cit. on p. 72).

[125] J. Mourão-Miranda, A. L. W. Bokde, C. Born, H. Hampel, and M. Stetter. "Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data". In: *NeuroImage* 28.4 (2005), pp. 980–995 (cit. on p. 44).

[126] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett. "Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI)". In: *Alzheimer's & Dementia* 1.1 (2005), pp. 55–66 (cit. on p. 23).

[127] J. A. Mumford, B. O. Turner, F. G. Ashby, and R. A. Poldrack. "Deconvolving {BOLD} activation in event-related designs for multivoxel pattern classification analyses". In: *NeuroImage* 59.3 (2012), pp. 2636–2643 (cit. on p. 55).

[128] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. "Multimodal Deep Learning". In: *ICML*. 2011 (cit. on p. 34).

[129] T. E. Nichols and A. P. Holmes. "Nonparametric permutation tests for functional neuroimaging: a primer with examples". In: *Human brain mapping* 15.1 (2002), pp. 1–25 (cit. on p. 17).

[130] T. E. Nichols and A. P. Holmes. "Nonparametric permutation tests for functional neuroimaging: a primer with examples". In: *Human brain mapping* 15.1 (2002), pp. 1–25 (cit. on pp. 21, 83, 87).

[131] H. Niederreiter and S. I. H. "Integration of nonperiodic functions of two variables by Fibonacci lattice rules". In: *Journal of Computational and Applied Mathematics* 51.1 (1994), pp. 57–70 (cit. on p. 82).

[132] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby. "Beyond mind-reading: multi-voxel pattern analysis of fMRI data". In: *Trends in cognitive sciences* 10.9 (2006), pp. 424–430 (cit. on p. 43).

[133] S. J. Pan, I. Tsang, J. Kwok, and Q. Yang. "Domain Adaptation via Transfer Component Analysis". In: *Neural Networks, IEEE Transactions on* 22.2 (Feb. 2011), pp. 199–210 (cit. on p. 35).

[134] S. J. Pan and Q. Yang. "A Survey on Transfer Learning". In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (Oct. 2010), pp. 1345–1359 (cit. on p. 30).

[135] T. Pavlidis. *Structural pattern recognition*. Springer-Verlag, 1977. xii, 302 p. : (cit. on pp. 49, 79).

[136] F. Pereira, T. Mitchell, and M. Botvinick. "Machine learning classifiers and fMRI: A tutorial overview". In: *NeuroImage* 45.1, Supplement 1 (2009), S199–S209 (cit. on p. 66).

[137] J. M. Phillips and S. Venkatasubramanian. “A gentle introduction to the kernel distance”. In: *arXiv preprint arXiv:1103.1625* (2011) (cit. on p. 90).

[138] J. Piater, P. Cohen, X. Zhang, and M. Atighetchi. “A Randomized ANOVA Procedure For Comparing Performance Curves”. In: *Proc. Fifteenth International Conference on Machine Learning*. 1998, pp. 430–438 (cit. on p. 58).

[139] P. Pinel, B. Thirion, S. Meriaux, A. Jobert, J. Serres, D. Le Bihan, J.-B. Poline, and S. Dehaene. “Fast reproducible identification and large-scale databasing of individual functional cognitive networks”. In: *BMC Neuroscience* 8.1 (2007), p. 91 (cit. on p. 68).

[140] R. A. Poldrack, Y. O. Halchenko, and S. J. Hanson. “Decoding the Large-Scale Structure of Brain Function by Classifying Mental States Across Individuals”. In: *Psychological Science* 20.11 (2009), pp. 1364–1372 (cit. on p. 43).

[141] J.-B. Poline and B. M. Mazoyer. “Analysis of Individual Positron Emission Tomography Activation Maps by Detection of High Signal-to-Noise-Ratio Pixel Clusters”. In: *J Cereb Blood Flow Metab* 13.3 (May 1993), pp. 425–437 (cit. on pp. 86, 87).

[142] S. Polyn, G. Detre, S. Takerkart, V. Natu, M. Benharrosh, B. Singer, J. Cohen, J. Haxby, and K. Norman. “A Matlab-based toolbox to facilitate multi-voxel pattern classification of fMRI data”. In: *Proceedings of HBM*. HBM, June 2005 (cit. on p. 21).

[143] N. Przulj, D. G. Corneil, and I. Jurisica. “Modeling interactome: scale-free or geometric?” In: *Bioinformatics* 20.18 (2004), pp. 3508–3515 (cit. on p. 51).

[144] L. Ralaivola, S. J. Swamidass, H. Saigo, and P. Baldi. “Graph Kernels for Chemical Informatics”. In: *Neural Networks, special issue on Neural Networks and Kernel Methods for Structured Domain* 18.8 (2005), pp. 1093–1110 (cit. on pp. 45, 46).

[145] L. Ralaivola, M. Szafranski, and G. Stempfel. “Chromatic PAC-Bayes Bounds for Non-IID Data: Applications to Ranking and Stationary beta-Mixing Processes”. In: *Journal of Machine Learning Research* 11 (2010), pp. 1927–1956 (cit. on p. 48).

[146] J. Ramon and T. Gärtner. “Expressivity versus efficiency of graph kernels”. In: *Proceedings of the First International Workshop on Mining Graphs, Trees and Sequences*. 2003, pp. 65–74 (cit. on p. 51).

[147] P. M. Rasmussen, K. H. Madsen, T. E. Lund, and L. K. Hansen. “Visualization of nonlinear kernel models in neuroimaging by sensitivity maps”. In: *NeuroImage* 55.3 (2011), pp. 1120–1131 (cit. on p. 66).

[148] J. Régis, J.-F. Mangin, T. Ochiai, V. Frouin, D. Rivière, A. Cachia, M. Tamura, and Y. Samson. “Sulcal root generic model: a hypothesis to overcome the variability of the human cortex folding patterns.” In: *Neurologia medico-chirurgica* 45.1 (Jan. 2005), pp. 1–17. pmid: 15699615 (cit. on p. 76).

[149] J. Richiardi, S. Achard, H. Bunke, and D. Van De Ville. “Machine Learning With Brain Graphs: predictive modeling approaches for functional imaging in systems neuroscience”. In: {*IEEE*} *Signal Processing Magazine* (2013), pp. 58–70 (cit. on pp. 45, 69).

[150] K. Riesen and H. Bunke. “Graph Classification Based On Vector Space Embedding”. In: *International Journal of Pattern Recognition and Artificial Intelligence* 23.06 (2009), pp. 1053–1081 (cit. on p. 51).

[151] A. N. Ruigrok, G. Salimi-Khorshidi, M.-C. Lai, S. Baron-Cohen, M. V. Lombardo, R. J. Tait, and J. Suckling. “A meta-analysis of sex differences in human brain structure”. In: *Neuroscience & Biobehavioral Reviews* 39 (Feb. 2014), pp. 34–50 (cit. on p. 109).

[152] R. M. Rustamov and L. Guibas. “Hyperalignment of Multi-Subject fMRI Data by Synchronized Projections”. In: *Proceedings of the 3rd NIPS Workshop on Machine Learning and Interpretation in Neuroimaging (MLINI)*. 2013 (cit. on p. 40).

[153] S. Ryali, K. Supekar, D. A. Abrams, and V. M. on. “Sparse logistic regression for whole-brain classification of fMRI data”. In: *NeuroImage* 51.2 (2010), pp. 752–764 (cit. on p. 44).

[154] M. R. Sabuncu, E. Konukoglu, and for the Alzheimer’s Disease Neuroimaging Initiative. “Clinical Prediction from Structural Brain MRI Scans: A Large-Scale Empirical Study”. In: *Neuroinformatics* 13.1 (Jan. 2015), pp. 31–46 (cit. on p. 23).

[155] M. R. Sabuncu, B. D. Singer, B. Conroy, R. E. Bryan, P. J. Ramadge, and J. V. Haxby. “Function-based Intersubject Alignment of Human Cortical Anatomy”. In: *Cerebral Cortex* 20.1 (2010), pp. 130–140 (cit. on pp. 44, 68).

[156] G. Sanromà, R. Alquézar, and F. Serratosa. “Attributed graph matching for image-features association using sift descriptors”. In: *Structural, Syntactic, and Statistical Pattern Recognition*. Springer, 2010, pp. 254–263 (cit. on p. 33).

[157] S. Satpal and S. Sarawagi. “Domain adaptation of conditional probability models via feature subsetting”. In: *Knowledge Discovery in Databases: PKDD 2007*. Springer, 2007, pp. 224–235 (cit. on p. 35).

[158] B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001 (cit. on pp. 53, 79, 112).

[159] J. M. Schott, N. C. Fox, C. Frost, R. I. Scahill, J. C. Janssen, D. Chan, R. Jenkins, and M. N. Rossor. "Assessing the onset of structural change in familial Alzheimer's disease". In: *Annals of Neurology* 53.2 (Feb. 2003), pp. 181–188 (cit. on p. 19).

[160] G. Schweikert, G. Rätsch, C. Widmer, and B. Schölkopf. "An empirical analysis of domain adaptation algorithms for genomic sequence analysis". In: *Advances in Neural Information Processing Systems*. 2009, pp. 1433–1440 (cit. on p. 36).

[161] N. Shervashidze, P. Schweitzer, E. J. V. Leeuwen, K. Mehlhorn, and K. M. Borgwardt. "Weisfeiler-Lehman Graph Kernels". In: *Journal of Machine Learning Research* 12 (2011), pp. 2539–2561 (cit. on p. 51).

[162] X. Shi, J.-F. Paiement, D. Grangier, and S. Y. Philip. "Learning from Heterogeneous Sources via Gradient Boosting Consensus." In: *SDM*. SIAM, 2012, pp. 224–235 (cit. on p. 38).

[163] H. Shimodaira. "Improving predictive inference under covariate shift by weighting the log-likelihood function". In: *Journal of Statistical Planning and Inference* 90.2 (2000), pp. 227–244 (cit. on pp. 35, 116).

[164] S. V. Shinkareva, V. L. Malave, M. A. Just, and T. M. Mitchell. "Exploring commonalities across participants in the neural representation of objects". In: *Human Brain Mapping* 33.6 (2012), pp. 1375–1383 (cit. on pp. 46, 67).

[165] S. V. Shinkareva, R. 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". In: *PLoS ONE* 3.1 (2008), e1394 (cit. on p. 43).

[166] A. Smola, A. Gretton, L. Song, and B. Schölkopf. "A Hilbert space embedding for distributions". In: *Algorithmic Learning Theory*. Springer, 2007, pp. 13–31 (cit. on p. 117).

[167] O. Sporns, G. Tononi, and R. Kötter. "The Human Connectome: A Structural Description of the Human Brain". In: *PLoS Comput Biol* 1.4 (2005), e42 (cit. on p. 45).

[168] J. Stelzer, Y. Chen, and R. Turner. "Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control". In: *NeuroImage* 65 (Jan. 2013), pp. 69–82 (cit. on pp. 83, 84, 87, 110).

[169] S. Sugiyama and K.-R. Müller. “Input-Dependent Estimation of Generalization Error under Covariate Shift”. In: *Statistics and Decisions* 23.4 (2005), pp. 249–279 (cit. on pp. 36, 117).

[170] Q. Sun, R. Chattopadhyay, S. Panchanathan, and J. Ye. “A two-stage weighting framework for multi-source domain adaptation”. In: *Advances in neural information processing systems*. 2011, pp. 505–513 (cit. on pp. 37, 122).

[171] S. Sun. “A survey of multi-view machine learning”. In: *Neural Computing and Applications* 23.7-8 (2013), pp. 2031–2038 (cit. on p. 25).

[172] S. Sun, H. Shi, and Y. Wu. “A survey of multi-source domain adaptation”. In: *Information Fusion* 24 (July 2015), pp. 84–92 (cit. on p. 37).

[173] Z. Y. Sun, S. Klöppel, D. Rivière, M. Perrot, R. S. J. Frackowiak, H. Siebner, and J.-F. Mangin. “The effect of handedness on the shape of the central sulcus.” In: *NeuroImage* 60.1 (Mar. 2012), pp. 332–9. PMID: 22227053 (cit. on p. 76).

[174] Z. Y. Sun, S. Klöppel, D. Rivière, M. Perrot, R. Frackowiak, H. Siebner, and J.-F. Mangin. “The effect of handedness on the shape of the central sulcus”. In: *NeuroImage* 60.1 (Mar. 2012), pp. 332–339 (cit. on p. 111).

[175] S. Takerkart and L. Ralaivola. “Multiple Subject Learning for Inter-Subject Prediction”. In: *Pattern Recognition in Neuroimaging (PRNI), 2014 International Workshop on*. June 2014, pp. 9–12 (cit. on p. 48).

[176] S. Takerkart, G. Auzias, L. Brun, and O. Coulon. “Mapping Cortical Shape Differences Using a Searchlight Approach Based On Classification of Sulcal Pit Graphs”. In: *Proceedings of IEEE ISBI Conference* (2015).

[177] S. Takerkart, G. Auzias, B. Thirion, and L. Ralaivola. “Graph-based inter-subject pattern analysis of fMRI data”. In: (2014) (cit. on pp. 42, 79, 80, 111, 112, 115).

[178] S. Takerkart, G. Auzias, B. Thirion, D. Schön, and L. Ralaivola. “Graph-Based Inter-subject Classification of Local fMRI Patterns”. In: *Machine Learning in Medical Imaging*. Ed. by F. Wang, D. Shen, and P. Yan. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 184–192 (cit. on p. 42).

[179] T. M. Talavage. “Tonotopic Organization in Human Auditory Cortex Revealed by Progressions of Frequency Sensitivity”. In: *Journal of Neurophysiology* 91.3 (Oct. 29, 2003), pp. 1282–1296 (cit. on p. 72).

[180] B. Tan, E. Zhong, E. W. Xiang, and Q. Yang. “Multi-Transfer: Transfer Learning with Multiple Views and Multiple Sources”. In: (Aug. 2014) (cit. on p. 38).

[181] M. Tervaniemi and K. Hugdahl. “Lateralization of auditory-cortex functions”. In: *Brain Research Reviews* 43.3 (2003), pp. 231–246 (cit. on p. 68).

[182] B. Thirion, P. Pinel, A. Tucholka, A. Roche, P. Ciuciu, J.-F. Mangin, and J.-B. Poline. “Structural Analysis of fMRI Data Revisited: Improving the Sensitivity and Reliability of fMRI Group Studies”. In: *Medical Imaging, IEEE Transactions on* 26.9 (2007), pp. 1256–1269 (cit. on pp. 44, 45).

[183] B. Thirion, P. Pinel, and J.-B. Poline. “Finding landmarks in the functional brain: detection and use for group characterization”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005*. Springer, 2005, pp. 476–483 (cit. on p. 45).

[184] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. “Large margin methods for structured and interdependent output variables”. In: *Journal of Machine Learning Research*. 2005, pp. 1453–1484 (cit. on p. 119).

[185] T. Tuytelaars and K. Mikolajczyk. “Local Invariant Feature Detectors: A Survey”. In: *Foundations and Trends® in Computer Graphics and Vision* 3.3 (2007), pp. 177–280 (cit. on p. 32).

[186] D. C. Van Essen, M. F. Glasser, D. L. Dierker, J. Harwell, and T. Coalson. “Parcellations and Hemispheric Asymmetries of Human Cerebral Cortex Analyzed on Surface-Based Atlases”. In: *Cerebral Cortex* 22.10 (Oct. 1, 2012), pp. 2241–2262 (cit. on p. 108).

[187] D. C. Van Essen, H. A. Drury, S. Joshi, and M. I. Miller. “Functional and structural mapping of human cerebral cortex: Solutions are in the surfaces”. In: *Proceedings of the National Academy of Sciences* 95.3 (1998), pp. 788–795 (cit. on p. 48).

[188] S. Vega-Pons and P. Avesani. “Brain Decoding via Graph Kernels”. In: *Pattern Recognition in Neuroimaging (PRNI), 2013 International Workshop on*. June 2013, pp. 136–139 (cit. on pp. 46, 51, 69).

[189] S. Vega-Pons, P. Avesani, M. Andric, and U. Hasson. “Classification of inter-subject fMRI data based on graph kernels”. In: *Pattern Recognition in Neuroimaging (PRNI), 2014 International Workshop on*. June 2014, pp. 5–8 (cit. on pp. 46, 51, 69).

[190] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt. “Graph kernels”. In: *The Journal of Machine Learning Research* 11 (2010), pp. 1201–1242 (cit. on p. 69).

[191] X. Wang, R. Hutchinson, and T. M. Mitchell. “Training fMRI Classifiers to Discriminate Cognitive States across Multiple Subjects”. In: *Proc. Sixteenth NIPS Conference*. Cambridge, MA: MIT Press, 2004 (cit. on p. 44).

[192] Wikipedia. “Error function”. In: *https://en.wikipedia.org/wiki/Error_function* (2015) (cit. on p. 84).

[193] K. J. Worsley, A. C. Evans, S. Marrett, and P. Neelin. “A Three-Dimensional Statistical Analysis for CBF Activation Studies in Human Brain”. In: *J Cereb Blood Flow Metab* 12.6 (1992), pp. 900–918 (cit. on p. 17).

[194] X. Yang, Q. Song, and Y. Wang. “A Weighted Support Vector Machine for Data Classification”. In: *International Journal of Pattern Recognition and Artificial Intelligence* 21.05 (2007), pp. 961–976 (cit. on p. 119).

[195] B. Yao and L. Fei-Fei. “Grouplet: A structured image representation for recognizing human and object interactions”. In: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 9–16 (cit. on p. 33).

[196] Y. Yao and G. Doretto. “Boosting for transfer learning with multiple sources”. In: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1855–1862 (cit. on p. 38).

[197] L. Yuan, Y. Wang, P. M. Thompson, V. A. Narayan, and J. Ye. “Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data”. In: *NeuroImage* 61.3 (July 2012), pp. 622–632 (cit. on p. 39).

[198] B. Zadrozny. “Learning and evaluating classifiers under sample selection bias”. In: *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 114 (cit. on pp. 36, 117).

[199] H. Zhang, T. Nichols, and T. Johnson. “Cluster mass inference via random field theory”. In: *NeuroImage* 44.1 (Jan. 1, 2009), pp. 51–61 (cit. on p. 87).

[200] L. Zhang and D. Samaras. “Machine learning for clinical diagnosis from functional magnetic resonance imaging”. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. 2005, pp. 1211–1217 (cit. on p. 43).

[201] Z. Zhao and H. Liu. “Multi-Source Feature Selection via Geometry-Dependent Covariance Analysis.” In: *FSDM*. 2008, pp. 36–47 (cit. on p. 39).

[202] F. Zhuang, X. Cheng, S. Pan, W. Yu, Q. He, and Z. Shi. “Transfer Learning with Multiple Sources via Consensus Regularized Autoencoders”. In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by T. Calders, F. Esposito, E. Hüllermeier, and R. Meo. Vol. 8726. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2014, pp. 417–431 (cit. on p. 34).