CENTRALE MARSEILLE

École Doctorale : Mathématiques et Informatique de Marseille (ED184)

Laboratoire d'Informatique et des Systèmes

THÈSE DE DOCTORAT

pour obtenir le grade de

DOCTEUR de l'ÉCOLE CENTRALE de MARSEILLE

Discipline : Informatique

Multivariate group analyses for functional neuroimaging: conceptual and experimental advances

par

WANG Qi

Directeur de thèse : ARTIÈRES Thierry, TAKERKART Sylvain

Soutenue le 14 Février 2020

devant le jury composé de :

ACHARD Sophie	Directeur de Recherche, Grenoble, LJK	Rapporteur
HABRARD Amaury	Professeur, Université de St Etienne, LHC	Rapporteur
BONASTRE Jean-François	Professeur, Université d'Avignon, LIA	Examinateur
KADRI Hachem	Associate Professeur, Aix-Marseille Université, LIS	Examinateur
ARTIÈRES Thierry	Professeur, Ecole Centrale de Marseille, LIS	Directeur de thèse
TAKERKART Sylvain	Ingénieur de Recherche, CNRS, INT	Co-directeur de thèse

Abstract

In functional neuroimaging experiments, participants perform a set of tasks while their brain activity is recorded, e.g. with electroencephalography (EEG), magnetoencephalography (MEG) or functional magnetic resonance imaging (fMRI). Analysing data from a group of participants, which is often denoted as group-level analysis, aims at identifying traits in the data that relate with the tasks performed by the participant and that are invariant within the population. This allows understanding the functional organization of the brain in healthy subjects and its dysfunctions in pathological populations. While group-level analyses for classical univariate statistical inference schemes, such as the general linear model, have been heavily studied, there are still many open questions for group-level strategies based on multivariate machine learning methods. This thesis therefore focuses on multivariate group-level analysis of functional neuroimaging and brings four contributions.

The first contribution is a comparison of the results provided by two classifier-based multivariate group-level strategies: i) the standard one in which one aggregates the performances of within-subject models in a hierarchical analysis, and ii) the scheme we denote as inter-subject pattern analysis, where a population-level predictive model is directly estimated from data recorded on multiple subjects. An extensive set of experiments are conducted on both a large number of artificial datasets - where we parametrically control the size of the multivariate effect and the amount of inter-individual variability - as well as on two real fMRI datasets. Our results show that the two strategies can provide different results and that inter-subject analysis both offers a greater ability to small multivariate effects and facilitates the interpretation of the obtained results at a comparable computational cost.

We then provide a survey of the methods that have been proposed to improve inter-subject pattern analysis, which is actually a hard task due to the largely heterogeneous vocabulary employed in the literature dedicated to this topic. Our second contribution consists in first introducing an unifying formalization of this framework, that we cast

as a multi-source transductive transfer learning problem, and then in reviewing more than 500 related papers to offer a first comprehensive view of the existing literature where inter-subject pattern analysis was used in task-based functional neuroimaging experiments.

Our third contribution is an experimental study that examines the well-foundedness of our multi-source transductive transfer formalization of inter-subject pattern analysis. With fMRI and MEG data recorded from numerous subjects, we demonstrate that between-subject variability impairs the generalization ability of classical machine learning algorithms and that a standard multi-source transductive learning strategy improves the generalization performances of such algorithms. Based on these promising results we further investigate the use of two more advanced machine learning methods to deal with the multi-source problem.

The fourth contribution of this thesis is a new multivariate group-level analysis method for functional neuroimaging datasets. Our method is based on optimal transport, which leverages the geometrical properties of multivariate brain patterns to overcome inter-individual differences impacting the traditional group-level analyses. We extend the concept of Wasserstein barycenter, which was initially meant to average probability measures, to make it applicable to arbitrary data that do not necessarily fulfill the properties of a true probability measure. For this, we introduce a new algorithm that estimates a barycenter and provide an experimental study on artificial and real functional MRI.

Keywords: Functional neuroimaging, group analysis, multivariate pattern analysis, machine learning

Résumé

Dans les expériences de neuroimagerie fonctionnelle, les participants effectuent un ensemble de tâches pendant que leur activité cérébrale est enregistrée, par exemple en utilisant l'électroencéphalographie (EEG), la magnétoencéphalographie (MEG) ou l'imagerie par résonance magnétique fonctionnelle (fMRI). L'analyse des données d'un groupe de participants, souvent appelée analyse de groupe, vise à identifier des invariants de population qui se rapportent aux tâches accomplies par les participants. Ceci permet de comprendre l'organisation fonctionnelle du cerveau chez les sujets sains et ses dysfonctionnements dans les populations pathologiques. Tandis que les analyses de groupes univariées, basées sur le modèle linéaire généralisé, ont fait l'objet d'études approfondies, de nombreuses questions restent ouvertes pour les analyses de groupe fondées sur des méthodes d'apprentissage machine multivariées.

Cette thèse étudie donc sur les analyses de groupe multivariées pour les expériences de neuroimagerie fonctionnelle. Nous nous focalisons sur un schéma d'analyse de groupe multivarié sous utilisé, que nous désignons "analyse de motifs inter-sujet", qui consiste à entraîner un modèle sur des données d'un ensemble de sujet et à évaluer sa capacité à généraliser sur des données enregistrées dans d'autres sujets. Nous effectuons d'abord une comparaison des résultats fournis par l'analyse de motifs inter-sujet avec ceux obtenus en utilisant la méthode standard. L'analyse inter-sujet offre à la fois une plus grande capacité de détection et facilite l'interprétation des résultats obtenus à un coût de calcul comparable.

Dans ce contexte, notre deuxième contribution introduit une formalisation unifiée de l'analyse de motifs inter-sujet, que nous modélisons comme un problème d'apprentissage par transfert transductif multi-sources. Ensuite, nous produisons une revue de la littérature des méthodes développées pour l'analyse de motifs inter-sujet.

Notre troisième contribution est une série d'études expérimentales qui examine le bienfondé de la formalisation par transfert transductif multi-sources de l'analyse de motifs

inter-sujet.

La quatrième contribution de cette thèse est une nouvelle méthode d'analyse multivariée au niveau du groupe pour les expériences de neuroimagerie fonctionnelle. Notre méthode est basée sur le transport optimal, qui tire parti des propriétés géométriques des cartes d'activité cérébrales pour surmonter les différences inter-individuelles qui ont un impact sur les analyses de groupe traditionnelles.

Mots clés: Neuroimagerie fonctionnelle, analyse de groupe, analyse de modèle multivariée, apprentissage machine

Acknowledgement

Firstly, I would like to express my deepest gratitude to my advisors Thierry ARTIÈRES and Sylvain TAKERKART, who provided generous help and continuous support for my PhD study. Their patience, motivation and knowledge helped me in research and writing this thesis. I want to thank Thierry for his support, he is a very kind professor who provided insightful and practical suggestions for my research and presentation. I am also very grateful to Sylvain, who guided me through conducting scientific research, on preparing presentations and led me to various areas of knowledge. It would have been impossible for me to go so far in this study without his help and encouragement. I also would like to thank the members in my thesis committee and in my PhD defense committee: Pr. Amaury HABRARD , Pr. Liva RALAIVOLA, Dr. Sophie ACHARD, Pr. Jean-François BONASTRE and Pr. Hachem KADRI, for evaluating my work.

My sincere thanks goes to the members of Banco and Qarma, especially to the team leaders Pascal BELIN and Hachem KADRI, it is an honor for me to work with these brilliant and kind researchers in these two teams. Thanks also to the co-authors of work described in the thesis: Thierry Chaminade, Bastien CAGNA and Ievgen REDKO, who helped me solve problems from neuroscience and machine learning.

Furthermore, I would like to thank Alex, Birgit, Clementine, Charly, Manon, Marina, Kepkee and Sophie etc., for all enjoyable time we spent together. I will keep learning French, hope next time when we have chance getting together, I could speak fluent French. Also to my friends who have already finished their PhD or are still working on one, Lu REN, Xiaoxi PAN, Dan FENG, Ziyu GUO, Guochao GAO, Jie ZHANG, Qi WANG, Yuxi ZHAO and Tianqi SONG etc., I want to express my thanks for those happy and unforgettable moments, for those warm company and help.

At the end, I want to express my gratitude to my family, especially to my husband Xiaowei LI. Thanks to him, I had the courage to apply for this PhD program and finish my study, his continuous support and care has greatly motivated me.

Bref Résumé

Neuroimagerie fonctionnelle : une introduction

Lors d'une expérience de neuroimagerie fonctionnelle, les participants effectuent un ensemble de tâches pendant que leur activité cérébrale est enregistrée, par exemple en utilisant l'électroencéphalographie (EEG), la magnétoencéphalographie (MEG) ou l'imagerie par résonance magnétique onctionnelle (IRMf). L'analyse des données d'un groupe de participants, souvent appelée analyse de groupe, vise à identifier des invariants de population qui se rapportent aux tâches accomplies par les participants. Ceci permet de comprendre l'organisation fonctionnelle du cerveau chez les sujets sains et ses dysfonctionnements dans les populations pathologiques.

Afin d'identifier des invariants de population dans le signal IRMf, la méthode conventionnelle d'analyse de groupe est une méthode univariée: un modèle linéaire généralisé (GLM) est appliqué indépendamment à chaque voxel des images du cerveau pour localiser les régions du cerveau qui ont un décours temporel en corrélation avec les tâches.

Bien que les analyses de groupe utilisant cette méthode conventionnelle univariée aient été largement étudiées, il reste de nombreuses questions ouvertes pour les stratégies au niveau du groupe basées sur les méthodes d'apprentissage automatique multivariées. L'analyse de patrons multi-voyels (MVPA en anglais) a gagné en popularité pour l'analyse des données de neuroimagerie. En passant d'une approche univariée à une approche multivariée, la MVPA se concentre sur les informations contenues dans les patrons d'activation distribués dans les voxels plutôt que dans les voxels individuels.

Cette thèse se concentre donc sur l'analyse multivariée au niveau du groupe pour la neuroimagerie fonctionnelle.

L'analyse des patrons inter-sujets : un schéma simple et puissant pour la MVPA au niveau du groupe

Au cours de la dernière décennie, la MVPA est devenue un outil très populaire pour

extraire des connaissances à partir de données de neuroimagerie fonctionnelle. L'analyse des patrons multivariés consiste généralement à décoder l'information contenue dans les patrons fonctionnels à l'aide d'un classifieur qui vise à deviner la nature de la tâche cognitive effectuée par le participant lorsqu'un patron d'activation cérébrale donné a été enregistré.

La performance de décodage est classiquement estimée séparément chez chacun des participants. Au niveau du groupe, ces mesures intra-sujet sont ensuite combinées, -souvent à l'aide d'un test de Student, -t pour fournir une inférence groupe. Malgré plusieurs critiques de cette stratégie hierarchique soulevées dans la littérature , celle-ci reste largement utilisée.

Une autre stratégie travaille directement au niveau du groupe en exploitant les données de tous les individus disponibles dans une seule analyse. Dans ce cas, la performance de décodage est évaluée sur la base des données de nouveaux participants, c'est-à-dire de participants qui n'ont pas fourni de données pour l'entrainement du classifieur, en veillant à ce que la nature des informations soit cohérente pour tous les individus de la population qui a été échantillonnée pour l'expérience. Cette stratégie prend plusieurs dénominations dans la littérature. Nous conservons ci-après la dénomination d'analyse de patrons inter-sujets (ISPA en anglais).

Dans ce chapitre, nous décrivons une comparaison des résultats fournis par ces deux stratégies de décodage au niveau du groupe basé sur la classification avec des ensembles de données d'IRMf artificielles et réelles. Cette étude expérimentale a été soigneusement conçue pour se concentrer exclusivement sur les différences induites par la nature intra-sujet et inter-sujets du décodage, c'est-à-dire en rendant toutes les autres étapes du processus d'analyse strictement identiques. Nous fournissons des résultats à la fois pour deux ensembles de données d'IRMf réelles et pour un grand nombre d'ensembles de données artificielles où les caractéristiques des données sont contrôlées de manière paramétrique.

Les deux stratégies se sont révélées efficaces, mais elles ne fournissent que des résultats partiellement concordants. L'analyse des patrons inter-sujets offre un pouvoir de détection plus élevé pour détecter les effets distribués de faible amplitude et facilite l'interprétation, tandis que les résultats fournis par l'approche hiérarchique nécessitent une investigation plus approfondie pour lever les ambiguïtés potentielles.

L'analyse de patrons inter-sujets (ISPA) pour la neuroimagerie fonctionnelle: une formalisation et une revue de la litérature

Dans la littérature, de nombreuses études utilisent l'ISPA comme outil pour répondre aux questions neuroscientifiques. Cependant, un vocabulaire largement hétérogène est employé dans la littérature consacrée à ce sujet, par exemple le décodage intersujets, l'apprentissage entre participants, la généralisation à de nouveaux individus. En outre, il manque une définition formelle de l'ISPA, ce qui constitue un obstacle à la conception d'algorithmes spécialisés qui pourraient être mis en œuvre dans des méthodes normalisées et proposés à la communauté dans des progiciels conviviaux.

Motivés par l'absence de définition unifiée de l'ISPA et par l'absence de revue de la littérature associée, nous présentons donc dans le présent chapitre une formalisation de l'ISPA et une revue de littérature exhaustive résumant les différentes stratégies méthodologiques proposées pour réaliser l'ISPA. Nous présentons d'abord une formalisation unifiée du cadre ISPA qui formalise l'ISPA en tant que problème d'apprentissage par transfert transductif multi-sources, puis nous détaillons les critères que nous avons appliqués pour rassembler la littérature sur l'ISPA, et enfin nous fournissons une analyse de la littérature existante sur l'ISPA.

Sur le bien-fondé de la formalisation du transfert transductif multi-sources de l'ISPA

Après avoir examiné les études de l'ISPA, nous avons soulevé plusieurs questions. Question 1 : Le cadre transductif multi-sources est-il plus performant que le cadre classique pour la réalisation de l'ISPA ? Question 2 : Après avoir réalisé la construction d'un espace commun et la conception des caractéristiques, y a-t-il encore des changements de distributions entre les sujets ? Question 3 : La mise en commun de plusieurs ensembles de données augmente à la fois la taille et le degré d'hétérogénéité de l'ensemble de formation, comment ces deux facteurs affectent-ils les performances de généralisation du modèle d'apprentissage machine ? Question 4 : La différence entre l'ensemble d'entrainement et l'ensemble de test nuit à la puissance de généralisation des modèles d'apprentissage machine, les méthodes d'apprentissage par transfert peuvent-elles améliorer la performance de généralisation du modèle ?

En ce qui concerne ces questions, nous réalisons quatre études expérimentales. Dans la première étude, nous examinons le cadre transductif multi-source de l'ISPA en appliquant une nouvelle normalisation multi-subjets des données, sujet par sujet. Dans cette étude, l'ISPA est réalisé dans le cadre du cadre transductif multi-sources ainsi que dans le cadre de l'apprentissage machine classique, puis les performances de ces deux cadres sont comparées. La deuxième étude expérimentale vise à examiner le adre de l'apprentissage par transfert pour l'ISPA. Nous comparons les performances d'un classifieur généralisant

à deux ensembles de test, l'un est constitué de nouvelles données provenant de sujets de l'ensemble d'entrainement l'autre est composé de données provenant de nouveaux sujets invisibles. Dans les deux premières études, plusieurs ensembles de données sont agrégés pour former l'ensemble d'entrainement. L'agrégation des ensembles de données augmente la taille de l'ensemble d'entrainement, ce qui est utile pour apprendre un algorithme d'apprentissage automatique, ainsi que le degré d'hétérogénéité, qui augmente la difficulté d'apprentissage de l'algorithme. C'est pourquoi l'étude 3 examine comment ces deux facteurs influencent les performances d'un classifieur. Dans la quatrième étude, nous proposons des méthodes pour exploiter le cadre de transfert transductif multi-sources de l'ISPA, en adoptant deux techniques d'apprentissage machine. La première stratégie est l'alignement subspatial qui est une stratégie d'apprentissage par transfert, l'autre stratégie, la généralisation par empilement, se concentre sur la construction d'un espace de caractéristiques invariantes dans l'ensemble de formation et de test. Ensuite, les performances en généralisation de ces deux stratégies sont comparées.

Analyse multivariée au niveau du groupe utilisant le transport optimal basé sur la distance $\mathbf{L}^p$

Dans les chapitres précédents, nous avons présenté plusieurs méthodes d'analyse de groupe largement utilisées, à savoir l'analyse de groupe univariée avec le GLM, l'analyse de groupe multivariée telle que l'ISPA. Toutefois, pour mettre en œuvre les méthodes d'analyse de groupe conventionnelles avec le GLM sur les ensembles de données de l'IRMf, une opération de lissage spatial est généralement appliquée aux données pour surmonter les différences potentielles qui existent dans les emplacements des foyers d'activation entre les individus. Cela permet d'améliorer le chevauchement spatial entre les sujets. Mais il en résulte une attenuation en amplitude qui nuit à la puissance de détection des méthodes d'analyse de groupe standard basées sur la moyenne euclidienne calculée indépendamment à chaque emplacement du cerveau.

Motivés par ces limitations, nous avons pour objectif de fournir une nouvelle méthode d'analyse multivariée au niveau du groupe pour a neuroimagerie fonctionnelle. Notre méthode est basée sur le transport optimal, qui exploite les propriétés géométriques des modèles cérébraux multivariés pour surmonter les différences interindividuelles qui ont un impact sur les analyses traditionnelles au niveau du groupe. Nous étendons le concept de barycentre de Wasserstein, qui était initialement destiné à faire la moyenne des mesures de probabilité, pour le rendre applicable à des données arbitraires qui ne remplissent pas nécessairement les propriétés d'une véritable mesure de probabilité. Pour cela, nous introduisons un nouvel algorithme qui estime un barycentre et nous four-

nissons une étude expérimentale sur des données IRMf artificielles et réelles. Ensuite, un test statistique non paramétrique est utilisé sur le barycentre pour détecter les régions significativement activées au niveau du groupe.

Table of Contents

Abstract i

Résumé iii

Acknowledgement v

Bref Résumé vii

1 Functional Neuroimaging: a primer 1

1.1 Functional neuroimaging: introduction . . . 1

1.1.1 Electroencephalography and magnetoencephalography . . . 1

1.1.2 Functional magnetic resonance imaging . . . 2

1.2 fMRI data analysis . . . 6

1.3 Univariate techniques . . . 7

1.3.1 The General Linear Model . . . 7

1.3.2 Group analysis with the GLM for fMRI data . . . 10

1.4 Multivariate techniques . . . 11

1.4.1 Multi-Voxel Pattern Analysis for fMRI data . . . 11

1.4.2 Multivariate Group analysis for fMRI . . . 12

1.5 Aim and scope of the thesis . . . 13

2 Inter-subject pattern analysis: a straightforward and powerful scheme for group-level MVPA 15

2.1 Introduction . . . 15

2.2 Methods . . . 17

2.2.1 Group-MVPA (G-MVPA) . . . 17

2.2.2 Inter-Subject Pattern Analysis (ISPA) . . . 17

2.2.3 Artificial data . . . 18

2.2.4 fMRI data . . . 21

2.2.5 fMRI data analysis . . . 22
2.2.6 Classifiers, Statistical inference and performance evaluation . . . 23
2.3 Results . . . 24
2.3.1 Results for artificial datasets . . . 24
2.3.2 Results for fMRI datasets . . . 27
2.4 Discussion . . . 31
2.4.1 G-MVPA and ISPA provide different results . . . 31
2.4.2 ISPA: larger training sets improve detection power . . . 32
2.4.3 ISPA offers straightforward interpretation . . . 33
2.4.4 G-MVPA is more sensitive to idiosyncrasies . . . 33
2.4.5 A computational perspective . . . 34
2.5 Conclusion . . . 35
2.6 Supplementary materials . . . 36
2.6.1 Influence of the number of subjects and number of samples per subject . . . 36
2.6.2 Influence of the within-subject covariance . . . 41
2.6.3 Influence of spatial smoothing on the results obtained on the real fMRI datasets . . . 44
2.6.4 Bias and variance analyses for G-MVPA and ISPA . . . 50
2.6.5 Assessing false positive rates in G-MVPA and ISPA . . . 52

3 Inter-subject pattern analysis for functional neuroimaging: a formalization and a review 55
3.1 Introduction . . . 55
3.2 Formalization of inter-subject pattern analysis . . . 56
3.2.1 Data description . . . 56
3.2.2 Cross-validation, training and test set . . . 57
3.2.3 A multi-source transductive transfer problem . . . 58
3.2.4 Relationships with similar problems . . . 59
3.3 Searching the literature for ISPA papers . . . 60
3.4 Inter-subject pattern analysis: a review . . . 62
3.4.1 Common space construction . . . 65
3.4.2 Feature construction or/and feature selection . . . 68
3.4.3 Predictive models/machine learning algorithms . . . 74
3.4.4 Performance evaluation . . . 76
3.4.5 Statistical assessment . . . 76
3.5 Discussion . . . 78

3.5.1 Handling inter-subject variability in the training set 78
3.5.2 Explaining the transductive nature of ISPA 79
3.5.3 Transferring from the training to test set 79
3.5.4 The gap between methods and applications in ISPA 80
3.5.5 Statistical models for ISPA 80
3.6 Conclusion 80

4 On the well-foundedness of the multi-source transductive transfer formalization of ISPA 83
4.1 Introduction 83
4.2 Experimental setting 85
4.2.1 Experimental data 85
4.2.2 General experimental setting 85
4.3 Study 1: multi-source transductive setting vs. classical machine learning setting 86
4.3.1 Experimental setting 86
4.3.2 Results 88
4.4 Study 2: Generalizing to new data vs. new individuals 89
4.4.1 Experimental setting 89
4.4.2 Results 90
4.5 Study 3: Number of subjects vs. sample size in the training set 92
4.5.1 Experimental setting 92
4.5.2 Results 93
4.6 Study 4: Performing ISPA with multi-source transductive transfer strategies 94
4.6.1 Methods 94
4.6.2 Experiment setting 97
4.6.3 Results 98
4.7 Discussion 101
4.7.1 The multi-source transductive transfer setting is valuable 101
4.7.2 Guidelines and suggestions for ISPA studies. 102
4.7.3 The dimensionality of data affects generalization performance . . . 103
4.7.4 Subspace alignment and stacked generalization provide different results 104
4.8 Conclusion 105

5 Multivariate group-level analysis using L^p distance-based optimal transport 107

5.1 Introduction . . . 107
5.2 Background of optimal transport . . . 109
5.2.1 Wasserstein barycenter of probability measures . . . 109
5.2.2 Barycenter of unnormalized measures . . . 110
5.3 Methods . . . 111
5.3.1 TL^p distance . . . 112
5.3.2 Barycenter with the TL^p distance . . . 112
5.3.3 Artificial fMRI data . . . 113
5.3.4 Real fMRI data . . . 115
5.3.5 Experimental setting . . . 115
5.3.6 Statistical assessment . . . 116
5.4 Results . . . 116
5.4.1 Results on artificial data . . . 122
5.4.2 Results on real fMRI data . . . 125
5.5 Discussion . . . 128
5.5.1 Influence of the noise on the optimal transport methods . . . 128
5.5.2 From artificial to real fMRI data . . . 128
5.5.3 Criterion for choosing parameters . . . 129
5.5.4 KBCM and TL^p-BI . . . 129
5.6 Conclusion . . . 130

6 Conclusion **131**

Chapter 1

Functional Neuroimaging: a primer

Neuroimaging or brain imaging is the use of various techniques to image the structure, function, or pharmacology of the central nervous system. In order to understand the relationship between brain activity in certain brain areas and specific mental functions, functional neuroimaging is used to measure the brain function. In functional neuroimaging experiments, participants perform a set of tasks while their brain activity is recorded, e.g. with electroencephalography (EEG), magnetoencephalography (MEG) or functional magnetic resonance imaging (fMRI), then analysis is performed on functional neuroimaging data to study the correlations between brain activity and tasks.

1.1 Functional neuroimaging: introduction

1.1.1 Electroencephalography and magnetoencephalography

Electroencephalography (EEG) and Magnetoencephalography (MEG) are both functional neuroimaging techniques. 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. With multiple electrodes placed on or over the scalp, EEG and MEG respectively measure electric and magnetic fields produced by synchronized neuronal currents generated from a large amount of active neurons with similar orientations. Although EEG and MEG signals originate from the same neurophysiological processes, magnetic fields are less distorted than electric fields by the skull and scalp, which results in a better spatial resolution of the MEG. Despite limited spatial resolution, EEG continues to be a valuable tool for

research and diagnosis. Furthermore, EEG is one of the few mobile techniques available and offers millisecond-range temporal resolution.

1.1.2 Functional magnetic resonance imaging

Since its development in the early 1990s, functional magnetic resonance imaging (fMRI) has become a popular technique to study human brain function. As a non-invasive imaging technique, performing fMRI scanning does not require to be injected with any tracer or exposed to X-rays, which means a broad range of individuals, including children, could be scanned repeatedly if necessary. Furthermore, fMRI is able to image brain activity with good spatial resolution (1-2 mm) and relatively good temporal resolution (1-2 seconds), which provides the opportunity to study brain function with multiple fine-grained images.

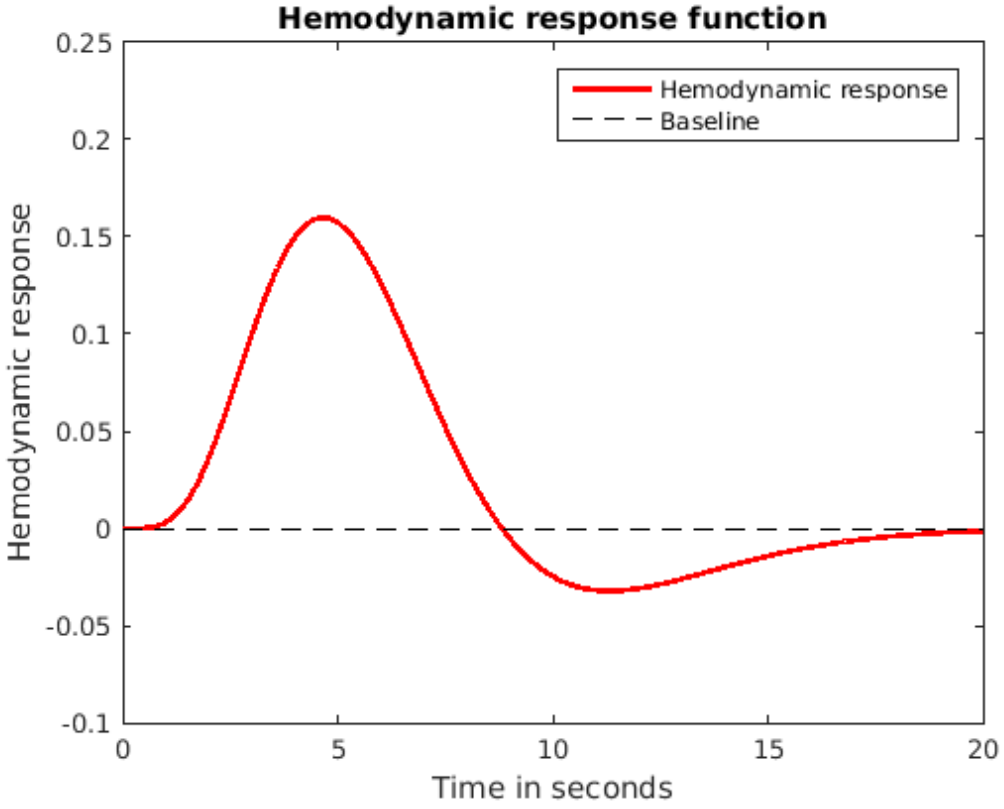


Figure 1.1: Hemodynamic response function

Blood flow and neuronal activity

When participants perform specific tasks or are presented with specific stimuli, neurons involved in processing the stimuli are activated and more oxygen is needed for the activity of neurons. fMRI measures the signal based on the change of blood flow, which is known as blood-oxygen-level-dependent (BOLD) signal. After the onset of a single short stimulus, the neuronal activity from the stimulus has very short duration (milliseconds), but the BOLD response generated from the activity, known as the hemodynamic response function (HRF), is slow. Figure 1.1 shows an illustration of HRF. As illustrated, the hemodynamic response first takes about 5 seconds to reach its peak, which is the result of an inflow of oxygenated blood supplying more oxygen and increasing the local concentration of oxyhemoglobin. The magnetic properties of oxyhemoglobin create local field homogeneities, which results in an increase in $T2^{\star}$-weighted MRI signal. When blood flow returns to normal, the peak of hemodynamic response falls and the decrease in concentration of oxyhemoglobin leads to poststimulus undershoot of fMRI signal. 15-20 seconds after the onset of stimulus, the hemodynamic response returns to baseline.

Experiment designs for fMRI

During an fMRI experiment, a group of participants perform a specific task, e.g. in order to identify voice-sensitive 'temporal voice areas' (TVA) of human auditory cortex which shows particular sensitivity to sounds of voice, subjects passively listen to a series of vocal and non-vocal stimuli while their brain activity is recored [32]. Stimuli belonging to several categories are usually presented to a subject at specific time, fMRI data are recorded simultaneously, i.e. if the repetition time (TR) is two seconds, then every two seconds one fMRI scan is obtained, which is a three-dimensional volume capturing the current brain activity. In order to obtain the brain activity corresponding to a specific stimulus, two major types of experimental designs are studied for fMRI studies: block and event-related designs.

In block design stimuli are continuously presented for several seconds, followed by a state of rest or baseline for several seconds. Because the BOLD signal generated from a block of stimuli is the summation of several responses, the BOLD signal evoked by the block design has larger amplitude than the signal evoked by a single brief stimulus (Figure 1.2 shows an illustration of a block design). Therefore with block design, subtle differences between different experimental conditions are enlarged and enable to be de-

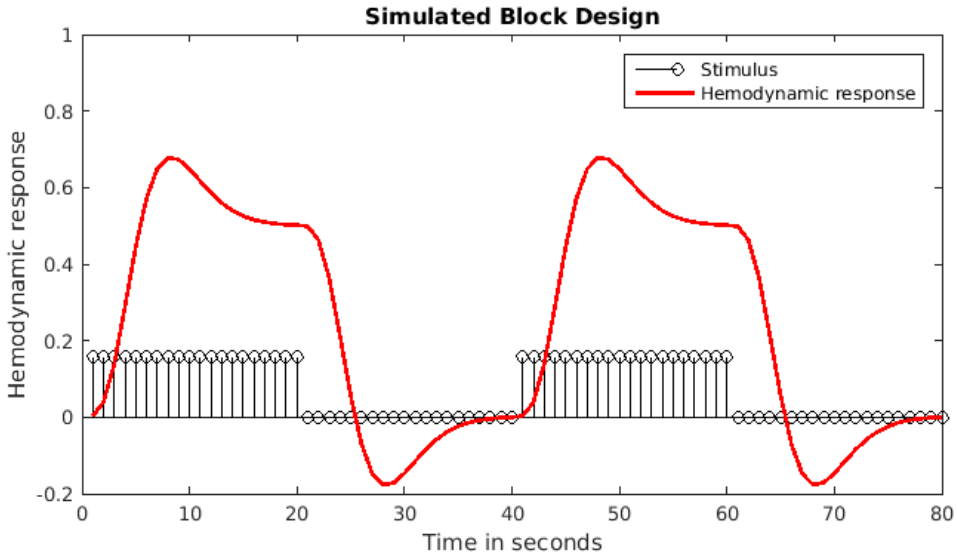


Figure 1.2: BOLD signal generated from block design. A simulated block design in which a 20-second block of stimulus is followed by a 20-second rest state, the amplitude of the BOLD signal is larger than the amplitude of HRF of a single stimulus.

tected. However, the BOLD signal from block design is the combination response of the summation of stimuli, which provides little information in estimating HRF for a single stimulus .

An alternative to block design is event-related design. Instead of presenting several stimuli in a duration of several seconds, event-related design involves presenting discrete and short-duration stimuli with inter-stimulus interval (ISI) (Figure 1.3 illustrates an event-related design). Compared to block design, event-related design generates signals with smaller amplitude. Based on the length of an ISI, event-related design can be categorized into slow or rapid design. A slow event-related design employs an ISI longer than the duration of HRF to avoid the overlapping of individual hemodynamic responses. Taking into account the post-stimulus delay of the BOLD signal, slow event-related design allows the hemodynamic response of a single stimulus to rise and return to baseline completely, therefore it enables estimating individual hemodynamic responses. However, the long ISI is time inefficient thus results in more scanning time. For that reason smaller ISI is employed to include more stimuli in limited time, which is referred to as rapid event-related design. In rapid event-related design, different types of stimuli are placed in fixed or randomized order. In order to estimate less variable response of stimulus, ISI jitters so that time duration between stimuli is not always the same. Because ISI is shorter than the duration of HRF in rapid event-related design, there exists overlapping of individual responses in the BOLD signal.

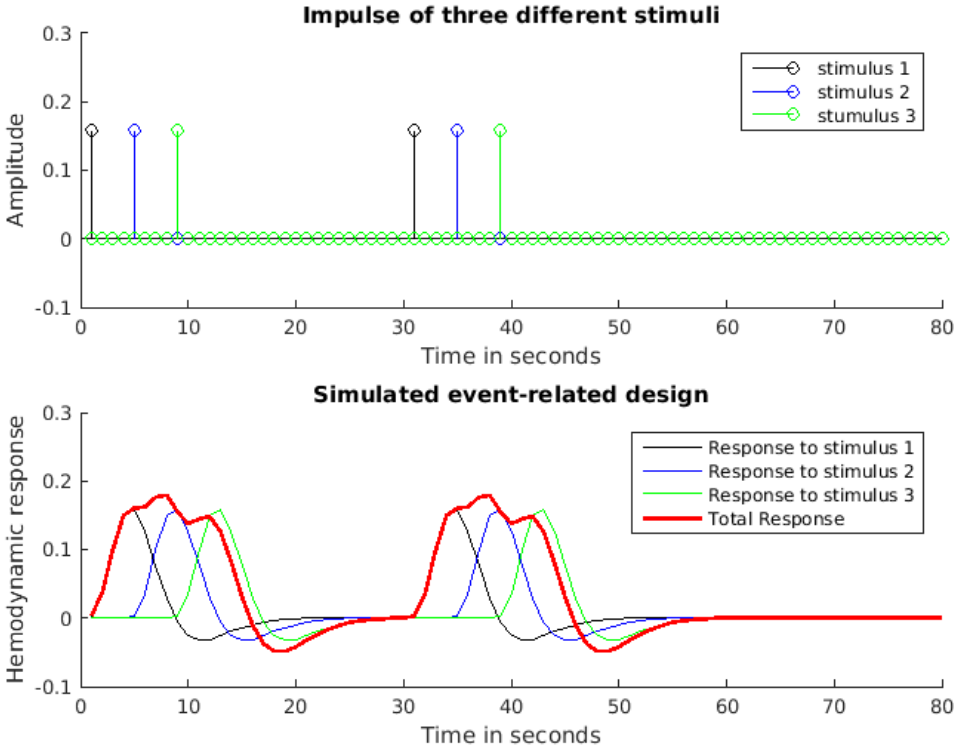


Figure 1.3: BOLD signal generated from event-related design. A simulated event-related design in which three stimuli are presented with ISI of five seconds. The amplitude of the BOLD signal is close to the amplitude of HRF of a single stimulus.

1.2 fMRI data analysis

fMRI dataset obtained from MRI scanner consists of a time-series of three-dimensional volumes each of which contains numerous cubes referred to as voxels (Figure 1.4 shows an example of an fMRI volume). To detect the brain function corresponding to mental processes, e.g. localizing voxels which correlate to tasks, response from each experimental condition requires to be estimated from fMRI data.

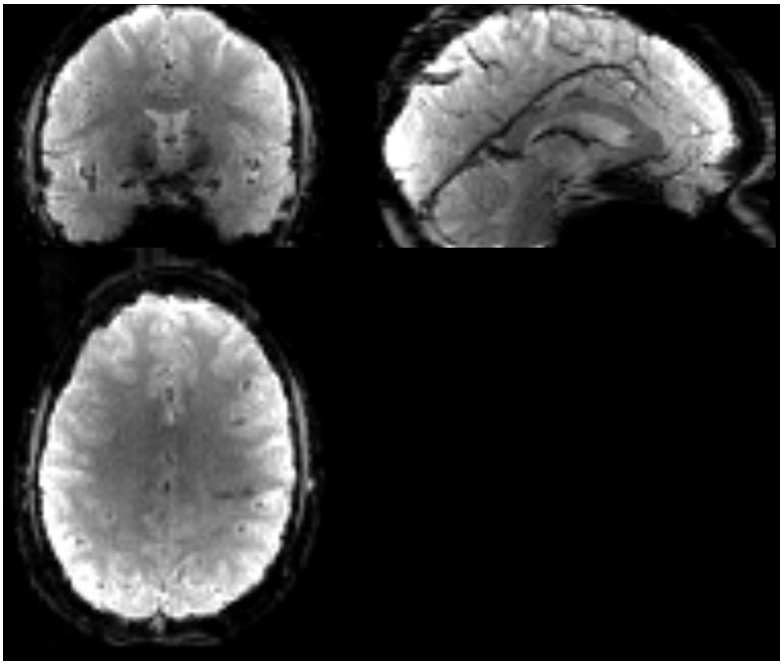


Figure 1.4: An example of 3D fMRI volume

However, fMRI data are noisy and contain a number of artifacts, in order to deal with these problems, a stream of operations, referred to as *preprocessing*, are performed on fMRI data. Preprocessing of fMRI data usually consists of some of the following parts [110]:

- Quality control: MRI scanner may generate several artifacts, e.g. spikes caused by the electrical instability and ghosting due to heartbeat or respiration of the subject performing tasks. Several methods, e.g. principal components analysis (PCA) and independent components analysis (ICA), are employed to remove artifacts, so that data are not corrupted by artifacts.

- Distortion correction: Echo-planar imaging (EPI) is the most common technique used for the acquisition of fMRI data. Magnetic field inhomogeneities in EPI is able to cause spatial distortion which increases variability between subjects and

shifts the location of activation. The correction of distortion, e.g. using magnetic field maps to quantify shifted distance, is capable to reduce the warp of image.

- Motion correction: When research subjects have head motion during the scanning, the location of brain in subsequent images will be mismatched. The misalignment between images is reduced by motion correction, also referred to as *realignment*, in which each image in the fMRI time series is aligned to a common reference scan.
- Slice timing correction: The acquisition of an fMRI volume consists in i) acquisition of several two-dimensional slices, one slice at a time ii) stacking slices to form three-dimensional volume. Therefore different slices are acquired at different time, which causes time mismatch between slices in one volume. In order to adjust time mismatch, the most common approach is choosing a slice as reference and then match all the other slices to the timing of the reference slice.
- Spatial normalization: Difference between individual brains makes it hard to study brain function in the population. In order to detect common traits across subjects, it requires to align data from multiple subjects into a common template, e.g. Montreal Neurological Institute (MNI) template.
- Spatial smoothing: In order to reduce noise in fMRI data and increase the signal-to-noise ratio (SNR), high-frequency information is removed by a filter to smooth small-scale changes in the image. Additionally, spatial smoothing reduces the difference between individuals.

After preprocessing, fMRI data are less noisy and the SNR is improved. However, the magnitude of signal evoked by task is still subtle, e.g. in block design percent signal change (PSC) of task activated voxels has mean values between 0.4% and 1%, while in event-related design PSC was even subtler, only about 0.1% [47]. For this reason, statistical models are exploited to estimate the signal and assess changes between experiment conditions, both univariate and multivariate techniques are used for analyzing data from participants.

1.3 Univariate techniques

1.3.1 The General Linear Model

The conventional statistical method for analyzing fMRI data is a univariate method. The method is performed independently on each voxel with the general linear model

(GLM) [88], and then iteratively on all brain locations to locate which brain regions have a time-course that correlates with tasks.

Estimating the GLM parameters

The GLM is defined as follows:

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

where $Y = [y_1, ..., y_N]^T$ is the BOLD signal time-series within the given voxel, representing the dependent variable which contains N observations recorded at the given location; X is an $N \times m$ design matrix, which contains m regressors each of which represents an explanatory variable; $\beta = [\beta_1, ..., \beta_m]^T$ is a column vector of size m that requires to be estimated, the i-th value weighting the i-th regressor of X; $\epsilon = [\epsilon_1, ..., \epsilon_N]^T$ is an error vector of size N, which contains the value of each observation that is not explained by the weighted sum of explanatory variables. Figure 1.5 illustrates an example of the GLM.

In order to estimate parameter β, the squared differences between Y and the estimate $\hat{Y}$ is minimized, where $\hat{Y} = X\hat{\beta}$. $\hat{\beta}$ is obtained by

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{1.2}$$

Hypothesis testing

After the estimate of β is obtained, hypothesis tests are performed on contrasts. The null hypothesis is defined as $H_0 : c\beta = 0$, where c is a vector or a matrix of constants that represents one or multiple contrasts. For instance, if $\beta = [\beta_1, \beta_2]^T$ and the null hypothesis is defined as $H_0 : \beta_1 = \beta_2$, which is also expressed as $H_0 : \beta_1 - \beta_2 = 0$, the contrast to test whether β_1 is different from β_2 will be $c = [1, -1]$. In order to test whether the null hypothesis is true, a t test is performed to obtain t value and the associated p-value. Besides performing single statistical test using t test, one can also assess multiple contrasts with F-test. For instance, if $\beta = [\beta_1, \beta_2]^T$, to test simultaneously the null hypotheses $H_0 : \beta_1 = \beta_2 = 0$, c would be a matrix:

$$c = \begin{Bmatrix} 1 & 0 \\ 0 & 1 \end{Bmatrix}$$

Then F test is performed with c and the estimate $\hat{\beta}$ to obtain statistic value and p-value.

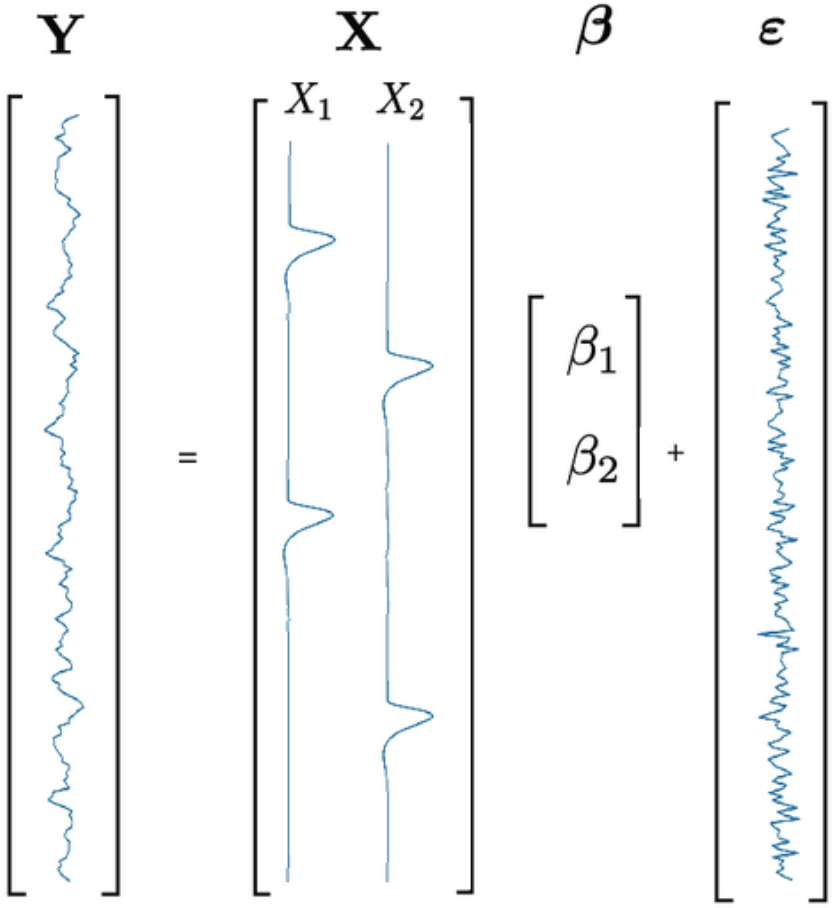


Figure 1.5: Illustration of the GLM. Y is the time series in one voxel, X is the design matrix representing tasks, β is the parameter to be estimated, and ϵ represents values which is not explained by the weighted design matrix.

Statistical inference

The statistical inference on fMRI data aims at detecting which regions in the brain reject the null hypothesis. The inference is performed at each voxel individually by examining whether the statistic at the voxel exceeds a threshold, if it does, the voxel is declared to be significant (the null hypothesis is rejected). However, there exists the situation where the null hypothesis is rejected when the null hypothesis is true, which is referred to as *false positive*. Even the tolerance of false positive is low, e.g. $p < 0.05$, and the null hypothesis is true everywhere, with the number of tests equal to the number of voxels, e.g. 100000 voxels, around 5000 voxels will be false positives. This problem is referred to as multiple comparison problem [94] and the probability of making one or more false discoveries in multiple hypothesis tests is referred to as family-wise error rate (FWE). In order to correct multiple comparison problem, one simple strategy is Bonferroni correction which raises the threshold to control FWE through dividing the p-value by the number of tests. Additionally, because of the fact that signal spreads across multiple voxels, it enables to make statistical inference on clusters of voxels instead of on individual voxels. One method to make cluster statistical inference is gaussian random field theory (RFT) [143]. Under the assumption of the spatial smoothness on the statistical map, RFT assigns each cluster a p value, which reduces the number of statistical tests.

1.3.2 Group analysis with the GLM for fMRI data

In fMRI experiments, usually the same experiment is performed by multiple participants. Analysing data from a group of participants, which is often denoted as group-level analysis, aims at identifying significant effects that are common within the population. However, datasets recorded from several participants each consists of multiple images, furthermore each image contains a large number of voxels. For this reason, a hierarchical two-level GLM is implemented.

The two-level GLM is expressed as follows:

$$\begin{cases} Y^{v,s} = X\beta^{v,s} + \epsilon^{v,s} & (1^{st}level) \\ Y^{v} = X_g\beta_g^{v} + \epsilon_g^{v} & (2^{st}level) \end{cases} \tag{1.3}$$

where $Y^{v,s}$ is the BOLD signal recored from voxel v in subject s, Y^v is the concatenation of first-level outputs across subjects.

The first level is the subject level during which the GLM is implemented independently

for each subject. After the parameter $\hat{\beta}^{v,s}$ is estimated for voxel v of subject s, contrast c is applied to $\hat{\beta}^{v,s}$ to compute $c^{v,s}$–the output of the first level–where $c^{v,s} = c\hat{\beta^{v,s}}$. Then $c^{v,s}$ is carried forward to the second level for group-level analysis.

In the second level, the GLM is performed across participants. For a group of S subjects, Y^v is a column vector consisting of S subject-specific contrasts computed at voxle v, where $Y^v = [c^{v,1}, ..., c^{v,S}]^T$. By introducing a group-level design matrix X_g which contains only one regressor, i.e. $X_g = [1, ..., 1]^T$, the second-level GLM model estimates a mean for the group of subjects, i.e. β_g^v represents the mean contrast across subjects at voxel v. The null hypothesis is defined as $H_0 : \beta_g^v = 0$, which is assessed through a one-sample t-test. After t value and associated p-value are iteratively assigned to each voxel on the brain, the correction for multiple comparison problem is performed to detect regions where the contrast c is significantly non-null across subjects. Besides t test, non-parametric strategies are also used as statistical model. For instance using permutation-based approach [103] to exchange labels of observations in the group level, the null distribution is created for learning the significance level.

1.4 Multivariate techniques

1.4.1 Multi-Voxel Pattern Analysis for fMRI data

Multi-voxel pattern analysis (MVPA) has gained increasing popularity on analyzing neuroimaging data. By switching from a univariate to a multivariate approach, MVPA focuses on information contained in distributed patterns of activation across voxels ([77]; [74]; [57]), rather than in individual voxels.

In general, MVPA uses supervised learning which consists in learning a decision function f that maps an input space $\mathcal{X}$ to a target space $\mathcal{Y}$. We denote a dataset composed of N labeled examples as $\mathcal{D}$, $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{n=N}$, where $x_n \in \mathcal{X}, y_n \in \mathcal{Y}$. x_n is a spatial pattern recording activations across voxels at specific time, e.g. β values estimated from the GLM. y_n is the corresponding label representing the experimental condition or the category of stimulus. The function learned from $\{(x_n, y_n)\}_{n=1}^{n=N}$ is used to predict y for unseen samples. When the value of y is discrete such as $y \in \{1, ..., C\}$, this procedure is known as *classification*. When the value of y is continuous, it is known as *regression*.

After learning a decoder, the performance of the decoder is usually evaluated by a metric which represents the decoder's generalization power to unseen data. Since fMRI data consist of limited observations, a strategy named *cross-validation* [34] is performed to

acquire robust evaluation for a decoder. Performing cross-validation requires to split dataset into subsets and repeatedly learn a decoder from several subsets, which involves steps as follows: i) data from one subset form the *test set* while data from all other subsets form the *training set*; ii) a decoder is learned from the training set; iii) the decoder makes predictions for data in the test set. Then predictions from all test subsets are integrated to measure the performance of the decoder.

In order to assess whether the performance of a classifier is significantly different from the chance level, a non-parametric statistical test [103] is commonly used. This strategy assumes that if there is no difference between experimental conditions, shuffling the labels of examples does not affect the performance of a classifier. Under this assumption labels are shuffled multiple times, each time an accuracy is computed. The accuracies learned from shuffled labels form an empirical null distribution which enables to obtain the p-value of the accuracy learned from true labels. If the p-value is under specific threshold, e.g. $p < 0.05$, it means the accuracy computed from true labels is significantly different from the chance level, which furthermore indicates there exists information in spatial patterns distinguishing experimental conditions.

Unlike the univariate technique which analyzes each voxel independently, it is challenging for MVPA to make inference about which voxels are involved in distinguishing tasks. For this reason a *searchlight* strategy [75] gains popularity in characterizing regional classifier performance. In this strategy a sphere is centered at each voxel, local patterns of voxels within the sphere are used to compute an accuracy which is assigned to the center voxel of the sphere. After sliding the sphere across the full brain, the searchlight ends up with an accuracy map of the full brain. Accuracy significantly above chance level indicates that the local cluster of voxels spatially contains information correlating with the task, whereas accuracy which is not significant implies no such information.

1.4.2 Multivariate Group analysis for fMRI

When provided with a group of subjects that perform the same experiment, multivariate group analysis aims at detecting the consistency of the individual measurements across the population.

Most multivariate group analysis strategies consist in following steps: i) performing cross-validation on single-subject dataset to obtain one measurement per subject; ii) aggregating individual measurements in the second level; iii) detecting the consistency of individual measurements with a statistical test.

Similar to group analysis with the GLM, multivariate group analysis is also capable to generate a statistical map which indicates regions correlating with tasks. The procedure usually involves employing searchlight sphere to produce accuracy maps, the commonly used method consists in: i) sliding searchlight sphere on single subject's dataset to obtain one accuracy map per subject; ii) using a statistical model, e.g. t statistic, to assess the consistency of accuracies across subjects, voxel by voxel; iii) correcting the multiple comparison problem to generate a thresholded statistical map.

1.5 Aim and scope of the thesis

This thesis focuses on group level analysis of functional neuroimaging data that aims at identifying invariant traits in the data within the population. While group-level analyses with classical univariate techniques, such as the GLM, have been massively studied, group-level strategies based on multivariate machine learning methods have a open prospect. Multivariate group analysis is important from both neuroscientific and methodological perspective. It allows understanding the functional organization of the brain in healthy subjects and its dysfunctions in pathological populations. In addition, because multivariate group analysis exploits multiple machine learning techniques, it has important application fields such as human-computer interaction, automated brain disorder diagnosis, etc. Therefore this thesis aims for multivariate group analysis.

The thesis is organized as follows:

Chapter 2 provides a comparison of two multivariate group-level schemes. Because of the difference in the shape and size of brain between subjects, it is challenging to detect common effects across subjects. In the literature, the standard multivariate group analysis strategy is hierarchical, combining the performances of within-subject models in a second-level analysis. The alternative strategy, inter-subject pattern analysis, directly works at the group-level by using, e.g. a leave-one-subject-out cross-validation. It is natural to ask what the difference between effects detected by these two strategies is and why there exists difference. Therefore, we provide a thorough comparison of these two multivariate group-level schemes, using both a large number of artificial datasets as well as two real fMRI datasets.

Compared to the standard hierarchical strategy, inter-subject pattern analysis offers straightforward interpretation of identified information in the population. Additionally, multiple machine learning strategies are exploited for reducing inter-subject variability, which helps inter-subject pattern analysis detect invariant traits across subjects.

Therefore in Chapter 3, we move our focus to inter-subject pattern analysis. Because a formal definition of inter-subject pattern analysis is lacking while a largely heterogeneous vocabulary is employed in the literature dedicated to this topic, we introduce a formalization of inter-subject pattern analysis formalized as a *multi-source transductive transfer* learning problem, using a combination of well defined machine learning concepts. Then we provide a survey reviewing multiple methodological strategies exploited in this topic. Especially, we focus on studies which handle inter-subject variability by solving problems identified in the formalization.

Then in Chapter 4, we aim to examine the well-foundedness of our formalization of inter-subject pattern analysis introduced in Chapter 3. We move to experimental studies with several machine learning strategies which have been reviewed in the survey. fMRI and MEG data recorded from multiple subjects are used in the studies. We demonstrate that the strong between-subject variability hinders obtaining consistent information across subjects. Additionally, the generalization ability of classical machine learning algorithms improves when inter-subject variability is handled.

In most strategies which perform inter-subject pattern analysis, brain patterns are vectorized into high-dimensional vectors, which lose the geometrical properties of brain patterns. Furthermore, in order to overcome inter-individual differences that impact the traditional group-level analyses, in Chapter 5 we propose a new multivariate group-level analysis method for functional neuroimaging datasets based on optimal transport, which leverages the geometrical properties of multivariate brain patterns. The new algorithm extends the concept of Wasserstein barycenter, which was initially meant to average probability measures, to make it applicable to arbitrary data that do not necessarily fulfill the properties of a true probability measure. After the barycenter of multiple brain patterns is estimated, by applying a statistical model on the barycenter, the significant regions in the brain at the group level are detected.

Chapter 2

Inter-subject pattern analysis: a straightforward and powerful scheme for group-level MVPA

Publication associated with this chapter: Wang, Q., Cagna, B., Chaminade, T. and Takerkart, S., 2020. Inter-subject pattern analysis: a straightforward and powerful scheme for group-level MVPA. NeuroImage, 204, p.116205.

2.1 Introduction

Over the past decade, multi-voxel pattern analysis (MVPA, [56]) has become a very popular tool to extract knowledge from functional neuroimaging data. The advent of MVPA has offered new opportunities to examine neural coding at the macroscopic level, by making explicitly usable the information that lies in the differential modulations of brain activation across multiple locations – i.e. multiple sensors for EEG and MEG, or multiple voxels for functional MRI (fMRI). Multivariate pattern analysis commonly consists in *decoding* the multivariate information contained in functional patterns using a classifier that aims to guess the nature of the cognitive task performed by the participant when a given functional pattern was recorded. The decoding performance is consequently used to measure the ability of the classifier to distinguish patterns associated with the different tasks included in the paradigm. It provides an estimate of the *quantity of information* encoded in these patterns, which can then be exploited to localize such informative patterns and/or to gain insights on the underlying cognitive

processes involved in these tasks.

This decoding performance is classically estimated separately in each of the participants. At the group level, these within-subject measurements are then combined – often using a t-test – to provide population-based inference, similarly to what is done in the standard hierarchical approach used in activation studies. Despite several criticisms of this group-level strategy that have been raised in the literature (see herafter for details), this hierarchical strategy remains widely used.

An alternative strategy directly works at the group-level by exploiting data from all available individuals in a single analysis. In this case, the decoding performance is assessed on data from new participants, i.e. participants who did not provide data for the training of the classifier (see e.g. [124, 62, 66, 70, 64, 39]), ensuring that the *nature of the information* is consistent across all individuals of the population that was sampled for the experiment. This strategy takes several denominations in the literature such as across-, between- or inter-subject classification or subject-transfer decoding. We hereafter retain the name inter-subject pattern analysis (ISPA).

In this chapter, we describe a comparison of the results provided by these two classifier-based group-level decoding strategies with both artificial and real fMRI datasets, which, to the best of our knowledge, is the first of its kind. This experimental study was carefully designed to exclusively focus on the differences induced by the within- vs. inter-subject nature of the decoding, i.e. by making all other steps of the analysis workflow strictly identical. We provide results for both two real fMRI datasets and a large number of artificial datasets where the characteristics of the data are parametrically controlled. This allows us to demonstrate that these strategies offer different detection power, with a clear advantage for the inter-subject scheme, but furthermore that they can provide results of different nature, for which we put forward a potential explanation supported by the results of our simulations on artificial data. The chapter is organized as follows. Section 2.2 describes our methodology, including our multivariate analysis pipeline for the two group-level strategies, as well as a description of the real datasets and the generative model of the artificial datasets. Section 2.3 includes the comparison of the results obtained with both strategies on these data, both in a qualitative and quantitative way. Finally, in Section 2.4, we discuss the practical consequences of our results and formulate recommendations for group-level MVPA.

2.2 Methods

2.2.1 Group-MVPA (G-MVPA)

Since the seminal work of [55] that marked the advent of multivariate pattern analysis, most MVPA studies have relied on a within-subject decoding paradigm. For a given subject, the data is split between a training and a test set, a classifier is learnt on the training set and its generalization performance – usually measured as the classification accuracy – is assessed on the test set. If this accuracy turns out to be above chance level, it means that the algorithm has identified a combination of features in the data that distinguishes the functional patterns associated with the different experimental conditions. Said otherwise, this demonstrates that the input patterns contain information about the cognitive processes recruited when this subject performs the different tasks that have been decoded. The decoding accuracy can then be used as an estimate of the *amount* of available information – the higher accuracy, the more distinguishable the patterns, the larger the amount of information.

The group-level extension of this procedure consists in evaluating whether such information is present throughout the population being studied. For this, a second level statistical analysis is conducted, for instance to test whether the average classification accuracy (or any other relevant summary statistic measured at the single-subject level), computed over the group of participants, is significantly above chance level. This can be done using a variety of approaches (see 2.2.6 for references). This hierarchical scheme is the one that is most commonly used in the literature. We denote it as Group-MVPA (G-MVPA) in the rest of the present chapter and illustrate it on Figure 2.1.

2.2.2 Inter-Subject Pattern Analysis (ISPA)

Besides the hierarchical G-MVPA solution, another classifier-based framework exists to evaluate multivariate effects at the group level. Considering the data from all available individuals, one can train a classifier on data from a set of subjects – the training subjects – and evaluate its generalization capability on data from the others – the test subjects. One can then use a cross-validation scheme that shuffles the subjects between the training and test sets, such as leave-one-subject-out or leave-n-subjects-out. In this setting, obtaining an average classification accuracy – this time across folds of the cross-validation – significantly above chance level means that a multivariate effect has been identified and that it is consistent across individuals. We denote this strategy as

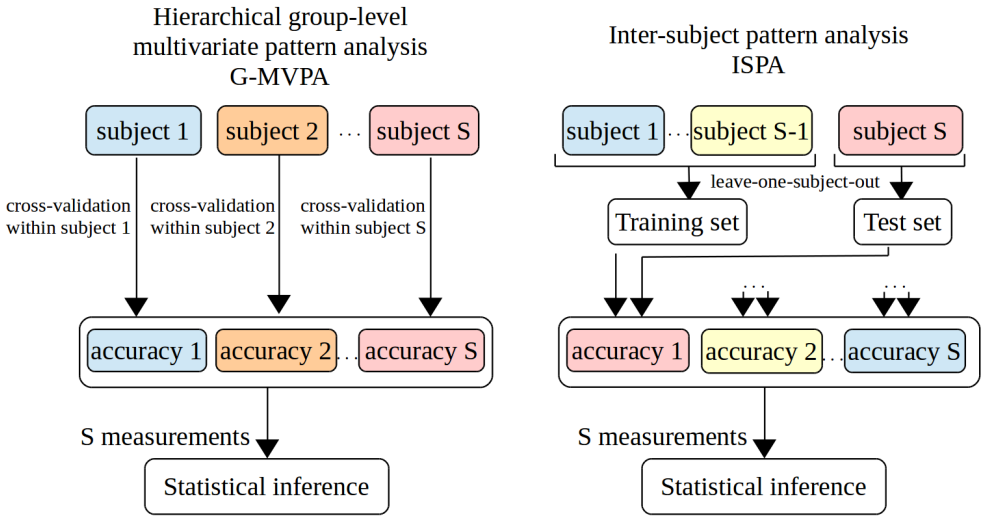


Figure 2.1: Illustration of the two approaches available to perform classifier-based group-level multivariate analysis. Left: hierarchical group-MVPA (G-MVPA). Right: inter-subject pattern analysis (ISPA). Note that if a leave-one-subject-out cross validation is used for ISPA (as illustrated), the two approaches yield the same number of measurements (equal to the number of subjects S), which allows for an unbiased comparison using the same statistical inference method.

Inter-Subject Pattern Analysis (ISPA).

In this study, we use a leave-one-subject-out cross-validation in which the model accuracy is repeatedly computed on the data from the left-out subject. Even if other schemes might be preferable to multiply the number of measurements [134], this choice was made to facilitate the comparison of the results obtained with ISPA and G-MVPA, as illustrated on Figure 2.1.

2.2.3 Artificial data

The first type of data we use to compare G-MVPA and ISPA is created artificially. We generate a large number of datasets in order to conduct numerous experiments and obtain robust results. Each dataset is composed of 21 subjects (for ISPA: 20 for training, 1 for testing), with data points in two classes labeled as $\mathcal{Y} = \{+1, -1\}$, simulating a paradigm with two experimental conditions. For a given dataset, each subject $s \in \{1, 2, ..., 21\}$ provides 200 labeled observations, 100 per class. We denote the i-th observation and corresponding class label (x_i^s, y_i^s), where $x_i^s \in \mathbb{R}^2$ and $y_i^s \in \mathcal{Y}$. The pattern x_i^s is created as

$$x_i^s = \begin{pmatrix} \cos\theta^s & -\sin\theta^s \\ \sin\theta^s & \cos\theta^s \end{pmatrix} \tilde{x}_i^s,$$

where

- $\tilde{x}_i^s$ is randomly drawn from a 2D Gaussian distribution, $\mathcal{N}(C^+, \Sigma)$ and $\mathcal{N}(C^-, \Sigma)$ if $y_i^s = +1$ or $y_i^s = -1$, respectively, which are defined by their centers $C^+ = (+\frac{d}{2}, 0)$ and $C^- = (-\frac{d}{2}, 0)$, where $d \in \mathbb{R}^+$ and their covariance matrix Σ, here fixed to $\begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}$ (see Supplementary Materials for results with other values of Σ);
- θ^s defines a rotation around the origin that is applied to all patterns of subject s; the value of θ^s is randomly drawn from the Gaussian distribution $\mathcal{N}(0, \Theta)$, where Θ defines the within-population variance.

Let $X^s = (x_i^s)_{i=1}^{200}$ and $Y^s = (y_i^s)_{i=1}^{200}$ be the set of patterns and labels for subject s. A full dataset D is defined by

$$D = \bigcup_{s=1}^{s=21} \{X^s, Y^s\}.$$

The characteristics of such a dataset are in fact governed by two parameters:

- d, which defines the distance between the point clouds of each of the two classes, i.e. the multivariate effect size;
- Θ, which controls the amplitude of the rotation that can be applied to the data, separately for each subject: when Θ is small, all the θ^s angles remain small, which means that the data of all subjects are similar; when Θ increases, the differences between subjects become larger; therefore, Θ quantifies the amount of inter-individual variability that exists within the group of 21 subjects for a given dataset.

Figures 2.2a and 2.2b illustrate the influence of each of these two parameters. Figure 2.2c shows different datasets generated with the same values of d and Θ.

In our experiments we used 13 values for d and 11 values for Θ, $d \in \{0.1, 0.12, 0.14, 0.16, 0.18, 0.2, 0.22, 0.24, 0.26, 0.28, 0.3, 0.4, 0.6\}$, $\Theta \in \{0.2\pi, 0.25\pi, 0.3\pi, 0.35\pi, 0.4\pi, 0.45\pi, 0.5\pi, 0.55\pi, 0.6\pi, 0.65\pi, 0.7\pi\}$, which gives 143 points in the two dimensional parameter space spanned by d and Θ. Note that by changing the value of Θ while keeping Σ constant, we control the relative amounts of within- and between-subject variance, which have been shown to be critical in group-level decoding situations [82]. For each pair (d, Θ), we

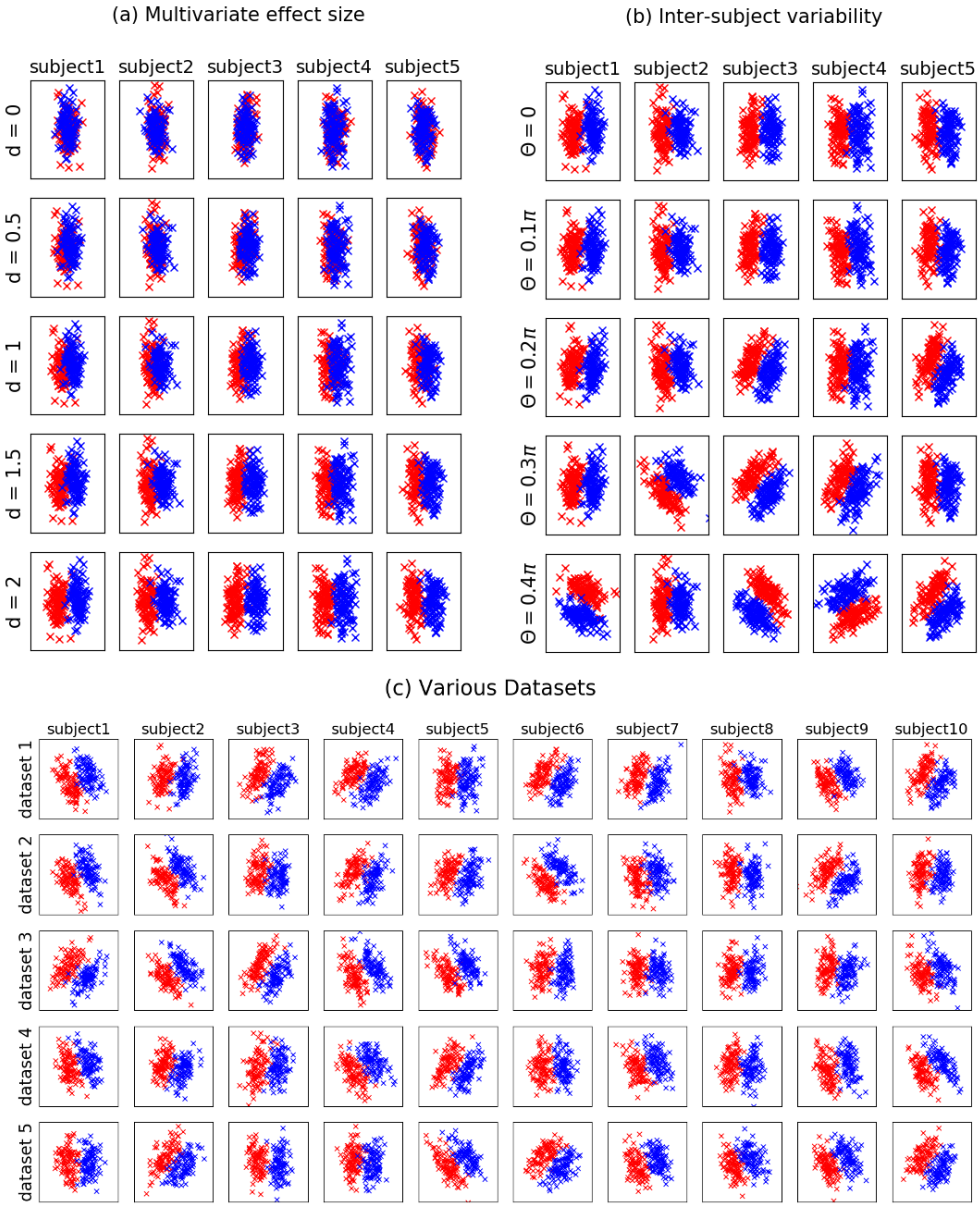


Figure 2.2: Illustration of the artificial datasets generated with the model described in 2.3. Each line is a subpart of a single dataset (5 subjects shown amongst 21 in (a) and (b), 10 subjects shown in (c)). The data points belonging to the class $y = +1$ and $y = -1$ are shown in blue and red, respectively. (a): influence of the d parameter (increasing effect size from top to bottom). (b): influence of the Θ parameter (increasing inter-individual variability from top to bottom). (c) five datasets obtained with the same values of the two parameters ($d = 2$ and $\Theta = 0.2\pi$).

generated 100 datasets. This yields 14300 datasets, each comprising 21 subjects and a total of 4200 data points. The code for generating these datasets (as well as per-

forming the experiments detailed hereafter) is available online at the following URL: http://www.github.com/SylvainTakerkart/inter_subject_pattern_analysis.

2.2.4 fMRI data

We also used two real fMRI datasets that were acquired at the *Centre IRM-INT* in Marseille, France. For both experiments, participants provided written informed consent in agreement with the local guidelines of the South Mediterranean ethics committee.

In the first experiment (hereafter *Dataset1*), fifteen subjects participated in an investigation of the neural basis of cognitive control in the frontal lobe, largely reproducing the experimental procedure described in [71]. Participants lying supine in the MRI scanner were presented with audiovisual stimuli that required a button response, with the right or left thumb. Four inter-stimulus intervals were used equally in a fully randomized order (1.8, 3.5, 5.5, 7.1 seconds), with an average of 4.5 seconds over a session. Data was collected with a 3 Tesla (3T) Bruker Medspec 30/80 Avance scanner running ParaVision 3.0.2. Eight MRI acquisitions were performed. First, a field map using a double echo Flash sequence recorded distortions in the magnetic field. Six sessions with 60 trials each were recorded, each comprising 133 volumes (echo-planar imaging (EPI) sequence, isotropic resolution of $3 \times 3 \times 3$ mm, the echo time (TE) of 30 ms, flip angle of 81.6°, field of view of 192×192 mm, 36 interleaved ascending axial slices acquired within the TR of 2400 ms) encompassing the whole brain parallel to the anterior commissure - posterior commissure (AC-PC) plane. Finally, we acquired a high-resolution T1-weighted anatomical image of each participant (magnetization prepared rapid gradient echo (MPRAGE) sequence, isotropic voxels of $1 \times 1 \times 1$ mm, field of view of $256 \times 256 \times 180$ mm, $TR = 9.4$ ms, $TE = 4.424$ ms).

In the second experiment (*Dataset2*), thirty-nine subjects were scanned using a *voice localizer* paradigm, adapted from the one analyzed in [109]. While in the scanner, the participants were asked to close their eyes while passively listening to a set of 144 audio stimuli, half of them being voice sounds, the other half being non-vocal. Most of the stimuli were taken from a database created for a previous study [18], while the others were extracted from copyright-free online databases. The paradigm was event-related, with inter-stimulus intervals randomly chosen between 4 and 5 seconds. The images were acquired on a 3T Prisma MRI scanner (Siemens, Eerlangen, Germany) with a 64-channels head coil. A pair of phase-reversed spin echo images was first acquired to estimate a map of the magnetic field. Then, a multi-band gradient EPI sequence with a factor of 5 was used to cover the whole brain and cerebellum with 60 slices during the

TR of 955 ms, with an isotropic resolution of $2 \times 2 \times 2$ mm, a TE of 35.2 ms, a flip angle of 56 degrees and a field of view of 200×200 mm for each slice. A total of 792 volumes were acquired in a single run of 12 minutes and 36 seconds. Then, a high resolution 3D T1 image was acquired for each subject (isotropic voxel size 0.8 mm^3, $TR = 2400$ ms, $TE = 2.28$ ms, field of view of $256 \times 256 \times 204.8$ mm). *Dataset2* is part of the InterTVA data set [3], which is fully available online [1].

2.2.5 fMRI data analysis

The two datasets were processed using the same sets of operations. The pre-processing steps were performed in SPM12[2]. They included co-registration of the EPIs with the T1 anatomical image, correction of the image distortions using the field maps, motion correction of the EPIs, construction of a population-specific anatomical template using the DARTEL method, transformation of the DARTEL template into MNI space and warping of the EPIs into this template space. Then, a general linear model was set up with one regressor per trial, as well as other regressors of non interest such as motion parameters, following the least-squares-all approach described in [100]. The estimation of the parameters of this model yielded a set of beta maps that was each associated with a given experimental trial. The beta values contained in these maps allowed constructing the vectors that serve as inputs to the decoding algorithms, that therefore operate on single trials. We obtained 360 and 144 beta maps per subject for *Dataset1* and *Dataset2* respectively. No spatial smoothing was applied on these data for the results presented below (the results obtained with smoothing are provided as Supplementary Materials).

For these real fMRI datasets, we performed a searchlight decoding analysis [75], which allows to map local multivariate effects by sliding a spherical window throughout the whole brain and performing independent decoding analyses within each sphere. For our experiments, we exploited the searchlight implementation available in nilearn[3] to allow obtaining the single-fold accuracy maps necessary to perform inference. For *Dataset1*, the decoding task was to guess whether the participant had used his left vs. right thumb to answer during the trial corresponding to the activation pattern provided to classifier. For G-MVPA, the within-subject cross-validation followed a leave-two-sessions-out scheme. For *Dataset2*, the binary classification task consisted in deciphering whether the sound presented to the participant was vocal or non-vocal. For G-MVPA, because a single session was available, we used an 8-fold cross-validation. Finally, all experiments

[1] https://openneuro.org/datasets/ds001771
[2] https://www.fil.ion.ucl.ac.uk/spm/
[3] http://nilearn.github.io/

were repeated with five different values of the searchlight radius ($r \in \{4$ mm, 6 mm, 8 mm, 10 mm, 12 mm$\}$).

2.2.6 Classifiers, Statistical inference and performance evaluation

In practice, we employed the logistic regression classification algorithm (with l_2 regularization and a regularization weight of $C = 0.1$), as available in the scikit-learn[4] python module, for both artificial and real fMRI data. The logistic regression has been widely used in neuroimaging because it is a linear model that enables neuroscientific interpretation by examining the weights of the model, and because it provides results on par with state of the art methods while offering an appealing computational efficiency [116].

In order to perform statistical inference at the group level, the common practice is to use a t-test on the decoding accuracies. Such a test assesses whether the null hypothesis of a chance-level average accuracy can be rejected, which would reveal the existence of a multivariate difference between conditions at the group level (note that, as detailed in 2.2.1 and 2.2.2, the rejection of this null hypothesis provides different insights on the group-level effect depending on whether G-MVPA or ISPA is used).

However, several criticisms have been raised in the literature against this approach, namely on the nature of the statistical distribution of classification accuracies [106, 5], on the non-directional nature of the identified group-information [49] or on the fact that the results can be biased by confounds [128]. This has led to the development of alternative methods (see e.g. [106, 14, 123, 38, 5]), all dedicated to G-MVPA.

Because our objective is to compare the results given by G-MVPA and ISPA in a fair manner, i.e using the same statistical test, we used a permutation test [103] which allows overcoming some of the aforementioned limitations. This test assesses the significance of the average accuracy at the group-level in a non-parametric manner for all the experiments conducted in the present study, whether conducted with G-MVPA or ISPA. Furthermore, this choice allowed maintaining the computational cost at a reasonable level, a condition that other alternatives, e.g inspired by [123], would not have met (see the Supplementary Materials for a longer discussion on this matter). In practice, for the real fMRI experiments, we used the implementation offered in the SnPM toolbox[5] to analyse the within-subject (for G-MVPA) or the single-fold (for the inter-subject cross-validation of ISPA) accuracy maps, with 1000 permutations and a significance threshold ($p < 0.05$, FWE corrected). For the simulations that used the artificial datasets, we

[4] https://scikit-learn.org/stable/
[5] http://warwick.ac.uk/snpm

used an in-house implementation of the equivalent permutation test (also available in our open source code; see 2.2.3), with 1000 permutations and a threshold at $p < 0.05$. Critically, it should be noted that the same number of samples were available for this statistical procedure when using G-MVPA or ISPA, as shown on Figure 2.1.

In order to compare the group-level decoding results provided by G-MVPA and ISPA, we use the following set of metrics. For the artificial data, we generated 100 datasets at each of the 143 points of the two dimensional parameter space spanned by the two parameters d and Θ. For each of these datasets, we estimate the probability p to reject the null hypothesis of no group-level decoding. We then simply count the number of datasets for which this null hypothesis can be rejected, using the $p < 0.05$ threshold, which we denote N_G and N_I for G-MVPA and ISPA, respectively. For the two real fMRI datasets, we examine the thresholded statistical map obtained for each experiment. We then compare the maps obtained by G-MVPA and ISPA by computing the size and maximum statistic of each cluster, as well as quantitatively assessing their extent and localization by measuring how they overlap.

2.3 Results

In this section, we present the results obtained when comparing G-MVPA and ISPA on both artificial and real fMRI datasets. With the artificial datasets, because we know the ground truth, we can unambiguously quantify the differences between the two strategies. Our focus is therefore on the characterization of the space spanned by the two parameters that control the characteristics of the data to evaluate which of these two strategies provides better detection power. For the real datasets, we do not have access to a ground truth. After having assessed the consistency of the obtained results with previously published work, we therefore focus on describing the differences between the statistical maps produced by G-MVPA and ISPA, examine the influence of the searchlight radius, and try to relate these results to the ones obtained on the artificial datasets.

2.3.1 Results for artificial datasets

The results of the application of G-MVPA and ISPA on the 14300 datasets that were artificially created are summarized in Tables 2.1, 2.2 and 2.3. In order to facilitate grasping the results on this very large number of datasets, we proceed in two steps.

First, we represent the two-dimensional parameter space spanned by d and Θ as a table where the columns and lines represent a given value for these two parameters, respectively. The values in these tables are the number of datasets N_G and N_I (out of the 100 datasets available for each cell) for which a significant group level decoding accuracy ($p < 0.05$, permutation test) is obtained with G-MVPA (Table 2.1) or ISPA (Table 2.2). In order to analyse these two tables, we performed a multiple linear regression where d and Θ are the independent variables used to explain the number of significant results in each cell. In Table 2.1, the effect of d is significant ($F_d^{G-MVPA} = 941.75$, $p_d^{G-MVPA} < 10^{-10}$), but the effect of Θ is not ($F_\Theta^{G-MVPA} = 0.83$, $p_\Theta^{G-MVPA} = 0.36$), while in Table 2.2, both factors have significant effects ($F_d^{ISPA} = 135.20$, $p_d^{ISPA} < 10^{-10}$; $F_\Theta^{ISPA} = 774.27$, $p_\Theta^{ISPA} < 10^{-10}$). This means that, as expected, the effect size d has an effect on the detection power of both G-MVPA and ISPA, i.e the smaller the effect size, the more difficult the detection. But the amount of inter-individual variability, here quantified by Θ, influences the detection capability of ISPA, but not the one of G-MVPA. This produces the rectangle-like area visible in green on Table 2.1 and the triangle-like area visible in red on Table 2.2. When the inter-individual variability is low, ISPA can detect significant effects even with very small effect sizes. When the variability increases, the detection power of ISPA decreases – i.e. for a given effect size, the number of datasets for which ISPA yields a significant result decreases, but the one of G-MVPA remains constant.

Secondly, in order to easily depict the compared behaviors of G-MVPA and ISPA, we *overlapped* the results of the two strategies into Table 2.3. In this third table, the blue cells indicate that $N_G \geqslant 50$ and $N_I \geqslant 50$ (i.e. that both G-MVPA and ISPA produce significant results in more than half of the 100 datasets), while the green and red regions contain cells where it is the case only for G-MVPA or ISPA respectively (i.e. $N_G \geqslant 50$ and $N_I < 50$ in green cells; $N_I \geqslant 50$ and $N_G < 50$ in red cells). Note that for completeness, we also used different values of this arbitrary threshold set at 50, which did not qualitatively change the nature of the results described below (hence these results are not shown). We observe a large blue region in which both strategies provide concordant detections, for the largest values of the effect size d and with a moderately low amount of inter-individual variability. Interestingly, the green and red regions, where one strategy detects a group-level effect while the other does not, also take an important area in the portion of the parameter space that was browsed by our experiments, which means that the two strategies can disagree. G-MVPA can provide a positive detection when the inter-individual variability is very large, while ISPA cannot (green region). But ISPA is the only strategy that offers a positive detection for very

Table 2.1: Number of datasets N_G (out of 100) for which G-MVPA provides significant group decoding (in green: cells where $N_G \geq 50$)

variability \ effect size	0.1	0.12	0.14	0.16	0.18	0.2	0.22	0.24	0.26	0.28	0.3	0.4	0.6
0.7π	7	11	15	22	27	36	40	44	53	59	68	96	100
0.65π	9	14	16	23	29	36	39	45	55	63	70	98	100
0.6π	13	16	17	25	27	35	44	53	58	71	76	95	100
0.55π	10	14	17	21	25	32	37	46	50	61	76	96	100
0.5π	7	8	15	17	22	30	36	46	55	59	64	94	100
0.45π	4	7	13	14	22	32	39	46	51	62	69	96	100
0.4π	8	11	18	22	26	31	40	49	54	63	72	97	100
0.35π	4	7	15	24	34	37	46	51	61	61	68	96	100
0.3π	10	14	24	26	31	40	46	57	58	71	78	95	100
0.25π	9	13	20	25	32	43	45	53	60	62	69	94	100
0.2π	9	12	15	19	28	34	39	50	59	67	75	95	100

Table 2.2: Number of datasets N_I (out of 100) for which ISPA provides significant group decoding (in red: cells where $N_I \geq 50$)

variability \ effect size	0.1	0.12	0.14	0.16	0.18	0.2	0.22	0.24	0.26	0.28	0.3	0.4	0.6
0.7π	8	8	9	10	11	10	10	10	8	10	10	9	13
0.65π	5	8	8	8	9	10	9	10	14	14	14	13	19
0.6π	8	13	12	13	15	15	14	17	16	18	20	20	21
0.55π	13	13	14	15	20	19	23	22	25	26	26	25	27
0.5π	16	15	16	22	24	24	28	27	29	29	30	38	45
0.45π	19	21	22	26	29	32	34	42	47	53	53	59	67
0.4π	22	28	33	34	39	41	42	49	49	50	54	69	85
0.35π	21	27	32	37	44	50	55	61	68	69	74	85	93
0.3π	25	30	39	46	60	67	70	74	82	83	90	98	99
0.25π	44	56	63	69	73	79	84	90	94	96	96	100	100
0.2π	42	55	63	71	80	86	91	93	94	98	98	100	100

Table 2.3: Visual comparison of G-MVPA vs. ISPA (in blue: cells where both $N_G \geq 50$ and $N_I \geq 50$; in green: cells where $N_G \geq 50$ and $N_I < 50$; in red: cells where $N_G < 50$ and $N_I \geq 50$).

variability \ effect size	0.1	0.12	0.14	0.16	0.18	0.2	0.22	0.24	0.26	0.28	0.3	0.4	0.6
0.7π													
0.65π													
0.6π													
0.55π													
0.5π													
0.45π													
0.4π													
0.35π													
0.3π													
0.25π													
0.2π													

small effect sizes, requiring a moderate inter-individual variability (red region).

2.3.2 Results for fMRI datasets

Qualitative observations

For our two real datasets, the literature provides clear expectations about brain areas involved in the tasks performed by the participants and the associated decoding question addressed in our experiments. The active finger movements performed by the participants during the acquisition of *Dataset1* are known to recruit contralateral primary motor and sensory as well as secondary sensory cortices, ipsilateral dorsal cerebellum as well as the medial supplementary motor area [95]. In *Dataset2*, the participants were passively listening to vocal and non-vocal sounds. The contrast, or decoding, of these two types of stimuli is classically used to detect the so-called temporal voice areas, which are located along the superior temporal cortex (see e.g. [109]). We now describe the results obtained with G-MVPA and ISPA with respect to this priori knowledge.

The searchlight decoding analyses performed at the group level were all able to detect clusters of voxels where the decoding performance was significantly above chance level ($p < 0.05$, FWE-corrected using permutation tests) with both G-MVPA and ISPA, for *Dataset1* and *Dataset2* and with all sizes of the spherical searchlight. The detected clusters were overall consistent across values of the searchlight radius, with an increasing size of each cluster when the radius increases. In *Dataset1*, both strategies uncovered two large significant clusters located symmetrically in the left and right motor cortex. Additionally, ISPA was able to detect other significant regions located bilaterally in the dorsal part of the cerebellum and the parietal operculum, as well as a medial cluster in the supplementary motor area (note that some of these smaller clusters also become significant with G-MVPA with the larger searchlight radii). These areas were indeed expected to be involved bilaterally given that button presses were given with both hands in this experiment. In *Dataset2*, both G-MVPA and ISPA yielded two large significant clusters in the temporal lobe in the left and right hemispheres, which include the primary auditory cortex as well as higher level auditory regions located along the superior temporal cortices, matching the known locations of the temporal voice areas. Figure 2.3 provides a representative illustration of these results, for a radius of 6 mm.

Quantitative evaluation

Our quantitative evaluation focuses on the two largest clusters uncovered in each dataset, i.e. the ones in the motor cortex for *Dataset1* and the ones in the temporal lobe for *Dataset2*. We first examine the size of these clusters, separately for each hemisphere

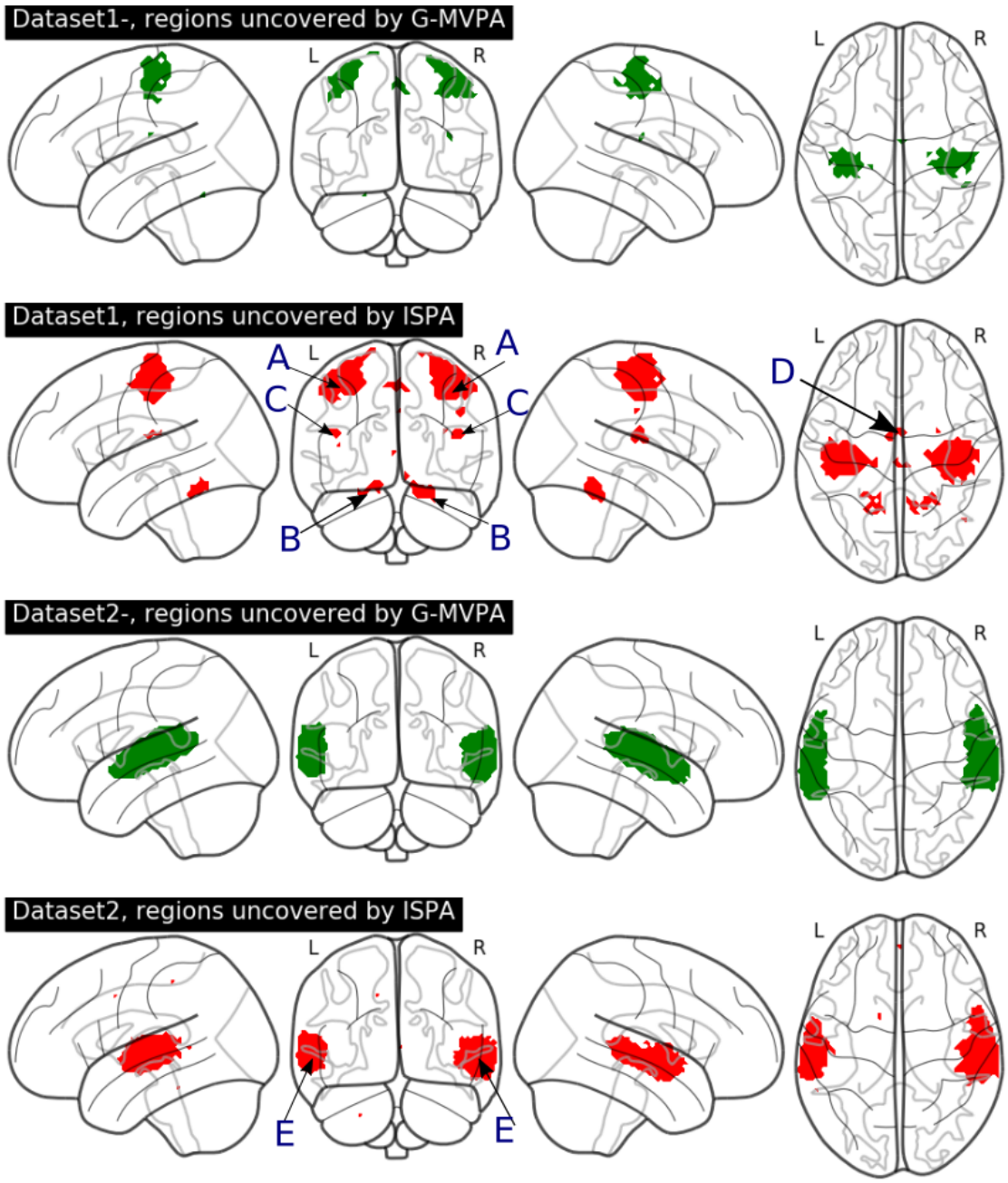


Figure 2.3: Illustration of the results of the group-level searchlight decoding analysis for a 6 mm radius. Top two rows: *Dataset1*; bottom two rows: *Dataset2*. Brain regions found significant using G-MVPA and ISPA are depicted in green and red, respectively. A: primary motor and sensory cortices B:ipsilateral dorsal cerebellum C: medial supplementary motor area D: secondary sensory cortices E: superior temporal cortex

and each of the five values of the searchlight radius. The results are displayed in the left column of Figure 2.4. In almost all cases, the size of the significant clusters increased with the searchlight radius (left column). Moreover, the cluster located in the

right hemisphere is consistently larger than the one on the left. In *Dataset1*, the cluster detected by ISPA is larger than the one detected by G-MVPA, regardless of the hemisphere, while in *Dataset2*, it is G-MVPA that yields larger clusters (except for a 4 mm radius where the sizes are similar). Then, we study the peak value of the t statistic obtained in each cluster (right column of Figure 2.4). In *Dataset1*, the peak t value is higher for ISPA than G-MVPA, for all values of the radius. In *Dataset2*, ISPA yields higher peak t values than G-MVPA for the searchlight radii smaller or equal than 8 mm, and lower peak t values for the larger radius values.

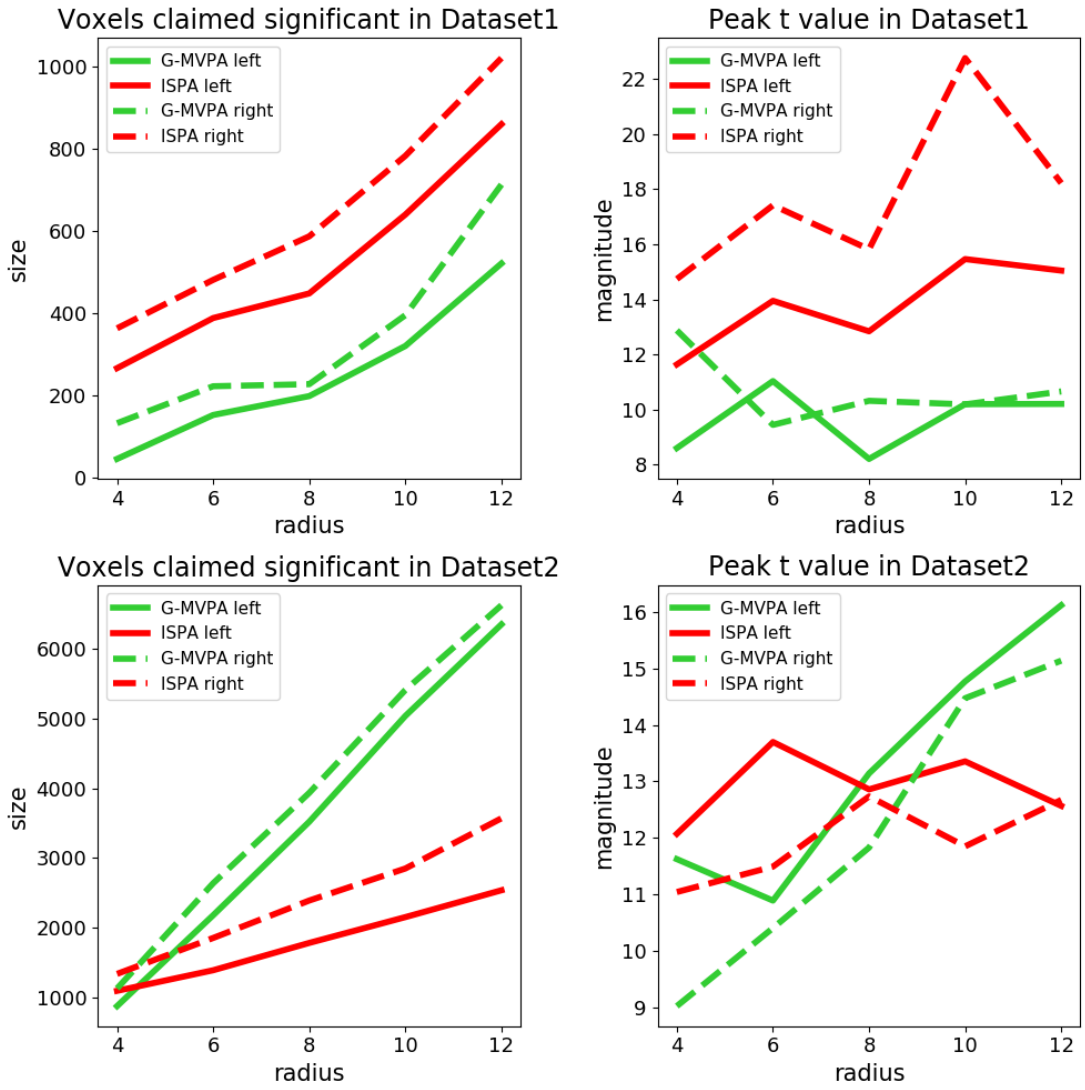


Figure 2.4: Quantitative evaluation of the results obtained on the real fMRI datasets for G-MVPA (green curves) and ISPA (red curves). Solid and dashed lines for the largest cluster in the left and right hemispheres respectively. Left column: size of the significant clusters. Right column: peak t statistic. Top vs. bottom row: results for *Dataset1* and *Dataset2* respectively.

Then, we quantify the amount of overlap between the clusters found by G-MVPA and ISPA, by splitting the voxels into three sub-regions: voxels uncovered only by ISPA, only by G-MVPA or by both strategies (overlap). Figure 2.5 provides an illustration of these sub-regions, which shows that the overlap region (in blue) is located at the core of the detected clusters, while the voxels significant for only one strategy are located in the periphery; these peripheral voxels appear to be mostly detected by ISPA for *Dataset1* (red voxels) and by G-MVPA for *Dataset2* (green voxels). Figure 2.6 shows the voxel counts in each sub-region, which confirms this visual inspection. Overall, the size of the sub-region found by the two strategies increases with the searchlight radius. The ISPA-only sub-region is larger in *Dataset1* than in *Dataset2*, representing between 38% and 83% of all significant voxels. Conversely in *Dataset2*, the G-MVPA-only sub-region is more important – with a percentage of all significant voxels comprised between 18% and 60%.

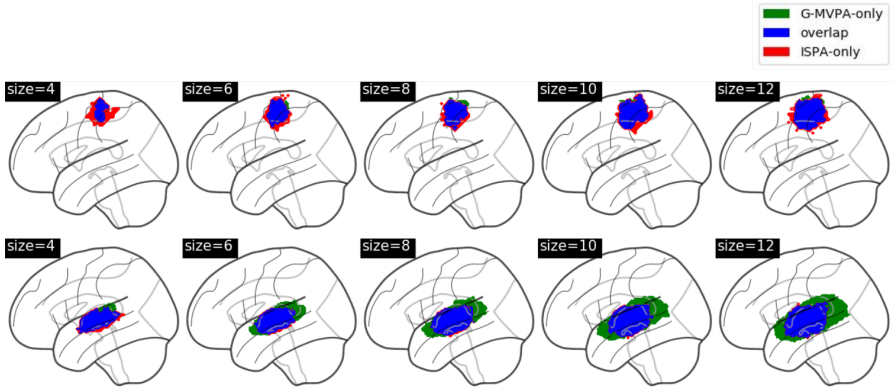


Figure 2.5: Comparison of the clusters detected by G-MVPA and ISPA for the different values of the searchlight radius, in *Dataset1* (top row) and *Dataset2* (bottom row).

We also count the number of voxels in each sub-region for each hemisphere. Figure 2.6 shows that for both datasets, the clusters in the right hemisphere are larger than in the left hemisphere. For both hemispheres, in most cases the number of voxels in each sub-region increases as the searchlight radius increases. However, in *Dataset1*, the number of voxels found only by G-MVPA is much smaller than that of overlap and ISPA-only sub-regions with all five radius values. In contrast, in *Dataset2* the number of voxels in ISPA-only sub-regions decreases for the four smallest values of the searchlight radius, and voxels only uncovered by ISPA are much fewer than those found only by G-MVPA.

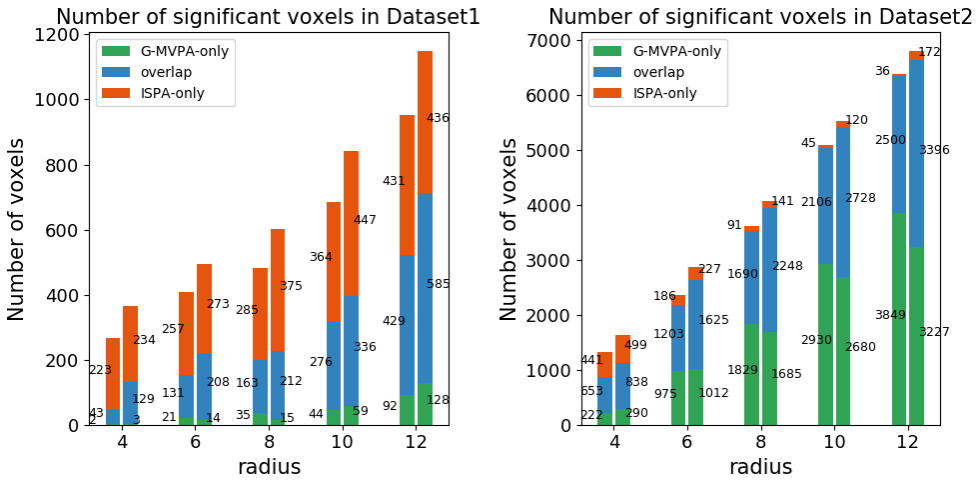


Figure 2.6: Comparison of the voxel counts detected by G-MVPA-only (green), ISPA-only (red) or both (blue) for the different values of the searchlight radius, in *Dataset1* (left) and *Dataset2* (right).

2.4 Discussion

2.4.1 G-MVPA and ISPA provide different results

In this study, we have performed experiments on both real and artificial functional neuroimaging data in order to compare two group-level MVPA schemes that rely on classifier-based decoding analyses: the vastly used G-MVPA, and ISPA. Our results show that both strategies can offer equivalent results in some cases, i.e. that they both detect significant group-level multivariate effects in similar regions of the cortex for our two real fMRI datasets, and in parts of the two-dimensional parameter space browsed using our artificially generated datasets, but that their outcomes can also differ significantly. For instance, in *Dataset1*, ISPA was the only strategy that detected multivariate group-level effects in several regions such as the supplementary motor area, the bilateral parietal operculum and dorsal cerebellum, for most of the searchlight radii that we tested (see an example on Figure 2.3 with a 6 mm radius). Furthermore, when a region is detected by both strategies, it usually differs in its size, extent and/or precise location, resulting in partial overlap; in most cases, the areas of concordance between the two strategies appeared to be centrally located, while the disagreements are located towards the periphery: in some areas, G-MVPA detects a group-level effect while ISPA does not, and inversely in other areas. Note that for *Dataset2*, our results generalize some of the observations reported in [49] with a different, yet comparable, framework of analysis, on a closely related data set.

Surprisingly, the peripheral behaviors were not consistent across the two real fMRI datasets: on *Dataset1*, ISPA only was able to detect effects on the periphery of the core region where both strategies were equally effective, while on *Dataset2*, it was G-MVPA which provided significant results on the periphery. The results of the experiments conducted on the artificial datasets can actually shed some light on these results, thanks to the clear dissociation that was observed in the two-dimensional parameter space browsed to control the properties of the data. Indeed, ISPA is the only strategy that allows detecting smaller multivariate effects when the inter-individual variability remains moderate, which is the case in the largest regions detected in *Dataset1* because they are located in the primary motor cortex, the *primary* nature of this region limiting the amount of inter-subject variability. On the opposite, the peripheral parts of the temporal region detected in *Dataset2* are located anteriorly and posteriorly to the primary auditory cortex, towards higher-level auditory areas where the inter-individual variability is higher, a situation in which G-MVPA revealed more effective in the experiments conducted with our artificial data.

2.4.2 ISPA: larger training sets improve detection power

Our experiments revealed a very important feature offered by the ISPA strategy: its ability to detect smaller multivariate effects. On the one hand, this greater detection power was explicitly demonstrated through the simulations performed on artificial data, where the multivariate effect size was one of the two parameters that governed the generation of the data; we showed that with an equal amount of inter-individual variability, ISPA was able to detect effects as small as half of what can be detected by G-MVPA. Furthermore, on both real fMRI datasets, ISPA was able to detect significant voxels that were not detected using G-MVPA, in a large amount in *Dataset1*, and to a lesser extent in *Dataset2*. This detection power advantage is of great importance, since detecting weak distributed effects was one of the original motivations for the use of MVPA [56].

This greater detection power of ISPA is in fact the result of the larger size of the training set available: indeed, when the number of training examples is small, the performance of a model overall increases with the size of the training set, until an asymptote that is reached with large training sets – as encountered in computer vision problems where millions of images are available from e.g. ImageNet[6]. In the case of functional neuroimaging where an observation usually corresponds to an experimental trial, we usually have a few dozen to a few hundred samples per subject, which clearly belongs to the *small sample*

[6] http://www.image-net.org

size regime, i.e. very far from the asymptote. In this context, ISPA offers the advantage to multiply the number of training samples by a factor equal to the number of subjects in the training set, which is of great value. However, here, the increased sample size comes at the price of a larger heterogeneity in the training set, because of the differences that can exist between data points recorded in different subjects. If these differences are too large, they can represent an obstacle for learning, but if more moderate, the inter-subject variability can reveal beneficial by increasing the diversity of the training set. The fact that we observe a higher detection power with ISPA than with G-MVPA suggests that we are in the latter situation.

2.4.3 ISPA offers straightforward interpretation

When using the ISPA strategy, obtaining a positive result means that the model has learnt an implicit rule from the training data that provides statistically significant generalization power on data from new subjects. Since a cross-validation of the type leave-one-subject-out or leave-n-subjects-out is performed on the available data to quantitatively assess such results, it allows to draw inference on the full population from which the group of participants was drawn, including individuals for which no data was available. As previously pointed out in [73], the interpretation that follows is straightforward: a significant effect detected with ISPA implies that some information has been identified to be consistent throughout the full population. In more details, this means that the modulations of the multivariate patterns according to the experimental conditions that were the object of the decoding analysis are consistent throughout the population, at least at the resolution offered by the modality used for the acquisition. This is the case for all voxels colored blue and red on Figure 2.5.

2.4.4 G-MVPA is more sensitive to idiosyncrasies

A significant result detected by G-MVPA but not ISPA – i.e. the green voxels on Figure 2.5, indicates that there is information at the population level in the input patterns that can discriminate the different experimental conditions, but that the nature of the discriminant information present in the input voxels differs across individuals. In other words, G-MVPA has detected idiosyncratic pattern modulations between conditions, which can be of great neuroscientific interest (see e.g. [21]), that could not have been identified with ISPA. This could be caused by two phenomenons. First, it could mean that the underlying coding strategy is nonetheless invariant across individuals, but that the nature of the data or of the feature space used in this analysis does not allow to

identify it as such, e.g. because of an unperfect inter-subject registration of the functional maps. One would then need to acquire additional data using a different modality ([35]) or to transform the feature space (e.g. using methods such as [58], [124] or [45]) in order to attempt to make this invariance explicit. Secondly, it could also mean that the neural code is simply intrinsically different across subjects, for instance because several strategies had been employed by different individuals to achieve the same task, or because each subject employs its own idiosyncratic neural code. G-MVPA therefore only provides part of the answer, which makes the interpretation much less direct.

Note that beyond searchlight analyses, this potential ambiguity could also occur with decoding performed in pre-defined regions of interest. Although such ROI (region-of-interest) -based decoding are vastly analysed at the group level using G-MVPA, numerous papers interpret the results as if G-MVPA allows identifying population-wise common coding principles, which cannot be claimed with only G-MVPA. These limitations have been pointed out previously in the literature, as in e.g. [128], [5] or [49], and we feel that the community should tackle this question more firmly. This could start by defining what a group-level multivariate analysis should seek – a consistent *amount* or *nature* of the information. Finally, because G-MVPA and ISPA are somehow complementary, one could think about using both types of analysis to better assess neural coding principles.

2.4.5 A computational perspective

Finally, we examine here some practical considerations that are important for the practitioner, by comparing the computational cost of G-MVPA and ISPA, and the availability of ready-to-use software implementations.

To assess the computational complexity of the two approaches, we first compare the number of classifiers that need to be trained for a full group analysis. Using G-MVPA, we need to train K classifiers per subject, where K is the chosen number of within-subject cross-validation folds, so KS classifiers in total (where S is the number of available subjects). For ISPA, the number of cross-validation folds equals to S (for leave-one-subject-out), meaning we need to train a total of S classifiers. The training time of a classifier also depends on the number of training examples: it is linear for classifiers such as logistic regression (when using gradient-based optimizers [6]), and quadratic for e.g. support vector machines [13]. Assuming we have n examples per subject, the number of training examples available for each classifier is $\frac{(K-1)n}{K}$ for G-MVPA and $(S-1)n$ for ISPA. Overall, with linear-time classifiers, the total complexity of a group-level decoding

analysis amounts to $O(nKS)$ for G-MVPA and $O(nS^2)$ for ISPA, which makes them almost equivalent if one assumes that K and S are of the same order of magnitude. With quadratic-time classifiers, the total complexity is $O(n^2KS)$ for G-MVPA and $O(n^2S^3)$ for ISPA, which makes ISPA significantly more costly. We therefore advice to use linear-time classifiers such as logistic regression to perform ISPA analyses, particularly with searchlight decoding where the computational cost is further multiplied by the number of voxels. Furthermore, note that thanks to its hierarchical nature, G-MVPA can be performed in an incremental manner for a low computational cost as participants get scanned: every time data from a new subject becomes available during the acquisition campaign, it suffices to run within-subject decoding for this new subject, which costs $O(nK)$, plus the statistical test. This offers more flexibility for the experimenter than with the inter-subject scheme, for which performing IPSA every time a new subject is scanned amounts to re-doing a full analysis on all the subjects.

In terms of software implementation, because within-subject analyses have been the standard since the advent of MVPA, all software packages provide well documented examples for such analyses which are the base tool for G-MVPA. Even if it is not the case for ISPA, it is easy to obtain an equivalent implementation because to perform inter-subject decoding, one simply need to i) have access to the data from all subjects, and ii) set up a leave-one-subject-out cross-validation, these two operations being available in all software packages. As an example, we provide the code to perform ISPA searchlight decoding from pre-processed data available online, which allows reproducing the results described in the present chapter on *Dataset2*:
http://www.github.com/SylvainTakerkart/inter_subject_pattern_analysis.

2.5 Conclusion

In this chapter, we have compared two strategies that allow performing group-level decoding-based multivariate pattern analysis of task-based functional neuroimaging experiments: the first is the standard method that aggregates within-subject decoding results and a second one that directly seeks to decode neural patterns at the group level in an inter-subject scheme. Both strategies revealed effective but they only provide partially concordant results. Inter-subject pattern analysis offers a higher detection power to detect weak distributed effects and facilitate the interpretation while the results provided by the hierarchical approach necessitate further investigation to raise potential ambiguities. Furthermore, because it allows identifying group-wise invariants from functional neuroimaging patterns, inter-subject pattern analysis is a tool of choice

to identify neuromarkers [46] or brain signatures [73], making it a versatile scheme for population-wise multivariate analyses.

2.6 Supplementary materials

2.6.1 Influence of the number of subjects and number of samples per subject

Aim. Although the two real fMRI datasets used in our experiments have approximately the same total number of observations (5400 observations in *Dataset1*, 5616 observations in *Dataset2*), they differ in the number of subjects that were scanned and the number of trials per subject: *Dataset1* includes 360 trials for each of the 15 subjects, while *Dataset2* offers 144 trials for 39 subjects. We here attempt to investigate whether this could explain some of the differences we observed in the results of G-MVPA and ISPA on these two datasets, using new artificial datasets.

Experiments. We therefore repeat the same set of experiments as in this chapter, using new sets of 14300 datasets generated to maintain the size of the training set constant and modulating the ratio between the number of subjects S and the number of samples per subject N. We used the following parameter values: $(S, N) \in \{(9, 500), (11, 400), (17, 250), (21, 200), (51, 80), (101, 40), (201, 20), (401, 10)\}$, which allows maintaining the size of the group-level dataset approximately constant (and more particularly, the size of the ISPA training set is exactly constant at $(S-1) \times N = 4000$).

Results. We present below the results we obtained, under the same summarized form as in Table 2.3. Tables 2.4 to 2.11 represent the two-dimensional parameter space spanned by d and Θ, for various values of the (S, N) couple. The cells colored in blue are those where G-MVPA and ISPA yielded significant group-level decoding for 50 or more datasets out of the 100, whereas in the green cells it is the case only for G-MVPA and in the red ones it is the case only for ISPA.

The shape of the region where G-MVPA yields 50 or more significant detections, corresponding here to the green and blue cells, remains approximately rectangular for all values of (S, N). But when the number of subject S increases, it is shifted to the right of the table, i.e. G-MVPA is effective only for larger effect sizes. This is a direct consequence of the decreasing value of N, which is critical for within-subject decoding. For ISPA, the shape and location of the colored region (red and blue cells) where it is effective remains fairly constant, showing that ISPA is not or weakly affected by the ratio

N/S when the size of the training set is constant.

We now try to address the question of whether the difference in the N/S ratio between *Dataset1* and *Dataset2* can explain the fact that the most peripheral regions of the main clusters are detected by ISPA for *Dataset1* and G-MVPA for *Dataset2*, in the light of the present results where we vary N/S in artificial datasets. The N/S ratio for *Dataset1* is $360/15 = 24$, while for *Dataset2*, it is $144/39 = 3.7$. We reformulate the previous question as follows: if the N/S ratio of *Dataset1* were smaller (i.e. closer to the one of *Dataset2*), could the peripheral voxels – which are red on Figure 2.5, i.e. are detected by ISPA – *become* green, i.e. be detected only by G-MVPA? For this, consider a red cell in Table 2.6 (for which the N/S ratio is the closest to the one of *Dataset1*): can it become green when the ratio N/S decreases? We clearly see that it is not possible, i.e. that all red cells in Table 2.6 are also red (in almost all cases) in Tables 2.7 to 2.11 where N/S is smaller. We then ask the opposite question: if the N/S ratio of *Dataset2* were larger (i.e. closer to the one of *Dataset1*), could the peripheral voxels – which are green on Figure 2.5, i.e. are detected by G-MVPA – *become* red, i.e. be detected only by ISPA? For this, consider a green cell in Table 2.8 (for which the N/S ratio is the closest to the one of *Dataset2*): can it become red when the ratio N/S increases? We clearly see that it is not possible, i.e. that all green cells in Table 2.8 are also green (in almost all cases) in Tables 2.4 to 2.7 where the N/S ratio is larger. This parallel between the real and the artificial datasets therefore suggests that it is not the difference of N/S ratio between *Dataset1* and *Dataset2* that can explain the different behaviors observed in the periphery of the significant clusters between G-MVPA and ISPA, illustrated on Figure 2.5.

Table 2.4: Visual comparison of G-MVPA vs. ISPA, 9 subjects, 500 data points per subject

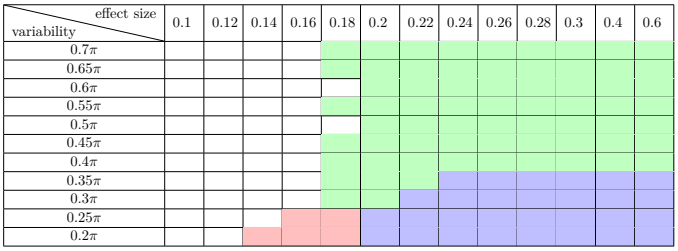


Table 2.5: Visual comparison of G-MVPA vs. ISPA, 11 subjects, 400 data points per subject

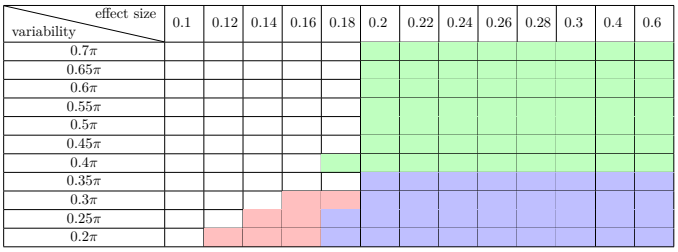


Table 2.6: Visual comparison of G-MVPA vs. ISPA, 17 subjects, 250 data points per subject

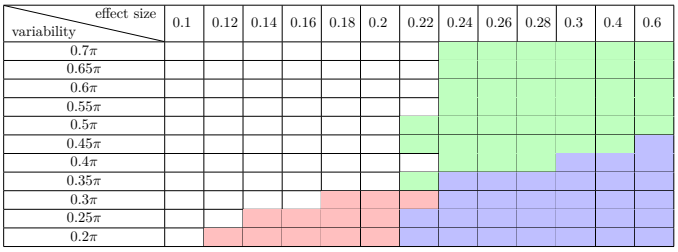


Table 2.7: Visual comparison of G-MVPA vs. ISPA, 21 subjects, 200 data points per subject

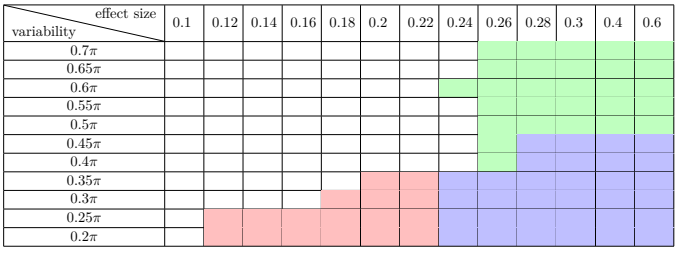


Table 2.8: Visual comparison of G-MVPA vs. ISPA, 51 subjects, 80 data points per subject

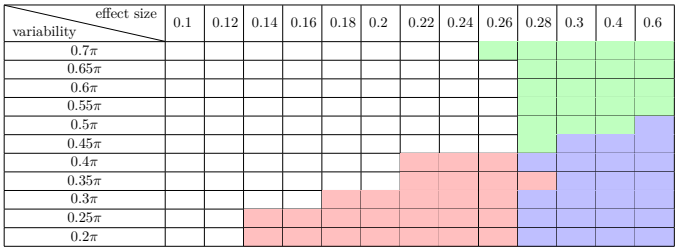


Table 2.9: Visual comparison of G-MVPA vs. ISPA, 101 subjects, 40 data points per subject

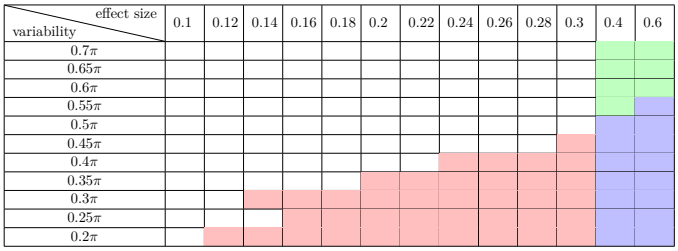


Table 2.10: Visual comparison of G-MVPA vs. ISPA, 201 subjects, 20 data points per subject

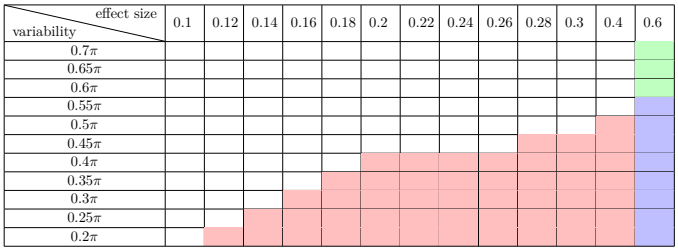


Table 2.11: Visual comparison of G-MVPA vs. ISPA, 401 subjects, 10 data points per subject

variability \ effect size	0.1	0.12	0.14	0.16	0.18	0.2	0.22	0.24	0.26	0.28	0.3	0.4	0.6
0.7π													
0.65π													
0.6π													
0.55π													
0.5π													
0.45π													
0.4π													
0.35π													
0.3π													
0.25π													
0.2π													

2.6.2 Influence of the within-subject covariance

Aim. Group-level decoding is strongly influenced by the ratio of the inter- and within-subject variance. To study the influence of this ratio, we performed experiments in the chapter where the within-subject covariance Σ was fixed and the inter-subject variability was parametrically controlled, using our generative model to create a large number of artificial datasets. Here, we study whether our results hold when we change the within-subject variance.

Experiments. In this subsection we vary both the within- and inter-subject variability. We keep the same range of values for Θ, which controls the amount of inter-subject variability:
$\Theta \in \{0.2\pi, 0.25\pi, 0.3\pi, 0.35\pi, 0.4\pi, 0.45\pi, 0.5\pi, 0.55\pi, 0.6\pi,$
$0.65\pi, 0.7\pi\}$. And we generate five new sets of 14300 datasets using five values for the within-subject covariance matrix:
$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 3 & 0 \\ 0 & 5 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}, \Sigma_4 = \begin{pmatrix} 7 & 0 \\ 0 & 5 \end{pmatrix},$
$\Sigma_5 = \begin{pmatrix} 9 & 0 \\ 0 & 5 \end{pmatrix}$. We fix the number of subjects to 21 and generate 200 data points for each subject. Figure 2.7 illustrates the effect of each of these five covariance matrices on the properties of the generated datasets, for $d = 2$ and $\Theta = 0$. In short, the distinctiveness of the two classes decreases from Σ_1 to Σ_5.

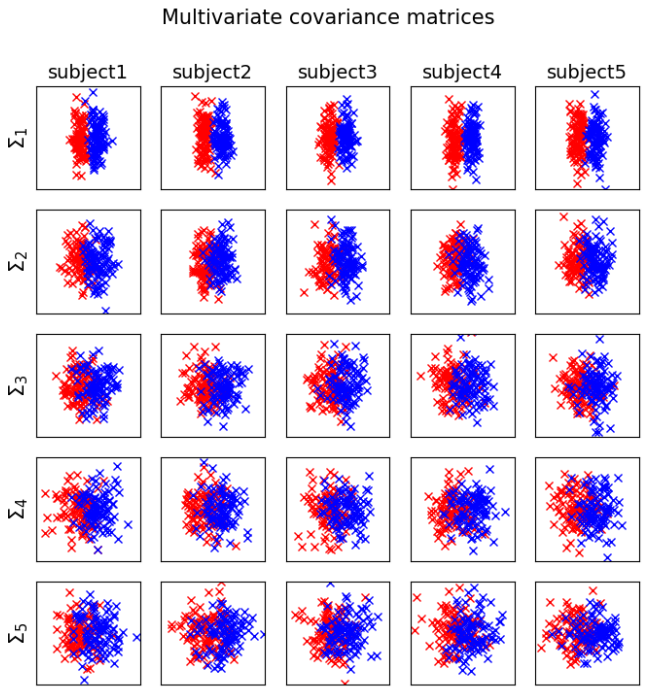


Figure 2.7: Illustration of the artificial datasets generated with five covariance matrices. Each line is a subpart of a single dataset (5 subjects shown amongst 21), $d = 2$ and $\Theta = 0$.

Results. Our results are shown in Tables 2.12 to 2.16. As previously, G-MVPA and ISPA yield more that 50 detections out of the 100 datasets available in each cell in regions that only partially overlap. Both strategies are strongly influenced by the within-subject covariance Σ, i.e. they prove less effective when going from Σ_1 to Σ_5, i.e. when the within-subject distinctiveness of the patterns decreases. However, the shape of the different regions in this parameter space remains constant: G-MVPA is effective in a rectangle area (blue + green cells), showing that it is not affected by the amount of inter-subject variability; and ISPA is effective in a triangle-like area (blue + red cells). This shows that the qualitative nature of our results does not seem to be affected by the value of Σ.

Table 2.12: Visual comparison of G-MVPA vs. ISPA, Σ_1

variability \ effect size	0.1	0.12	0.14	0.16	0.18	0.2	0.22	0.24	0.26	0.28	0.3	0.4	0.6
0.7π													
0.65π													
0.6π													
0.55π													
0.5π													
0.45π													
0.4π													
0.35π													
0.3π													
0.25π													
0.2π													

Table 2.13: Visual comparison of G-MVPA vs. ISPA, Σ_2

variability \ effect size	0.1	0.12	0.14	0.16	0.18	0.2	0.22	0.24	0.26	0.28	0.3	0.4	0.6
0.7π													
0.65π													
0.6π													
0.55π													
0.5π													
0.45π													
0.4π													
0.35π													
0.3π													
0.25π													
0.2π													

Table 2.14: Visual comparison of G-MVPA vs. ISPA, Σ_3

variability \ effect size	0.1	0.12	0.14	0.16	0.18	0.2	0.22	0.24	0.26	0.28	0.3	0.4	0.6
0.7π													
0.65π													
0.6π													
0.55π													
0.5π													
0.45π													
0.4π													
0.35π													
0.3π													
0.25π													
0.2π													

Table 2.15: Visual comparison of G-MVPA vs. ISPA, Σ_4

variability \ effect size	0.1	0.12	0.14	0.16	0.18	0.2	0.22	0.24	0.26	0.28	0.3	0.4	0.6
0.7π													
0.65π													
0.6π													
0.55π													
0.5π													
0.45π													
0.4π													
0.35π													
0.3π													
0.25π													
0.2π													

Table 2.16: Visual comparison of G-MVPA vs. ISPA, Σ_5

variability \ effect size	0.1	0.12	0.14	0.16	0.18	0.2	0.22	0.24	0.26	0.28	0.3	0.4	0.6
0.7π													
0.65π													
0.6π													
0.55π													
0.5π													
0.45π													
0.4π													
0.35π													
0.3π													
0.25π													
0.2π													

2.6.3 Influence of spatial smoothing on the results obtained on the real fMRI datasets

Aim. The preprocessing steps performed in this chapter on the two fMRI datasets did not include any spatial smoothing. Since there is a debate in the literature on the validity and interest of such smoothing when performing MVPA, we here study the influence of the smoothing on our results.

Experiments. We replicate the same experiments as in this chapter by adding some spatial smoothing on the beta maps that are used to construct the inputs of the classifier. We use the Gaussian kernel implemented in SPM to perform the smoothing, with full-width at half-maximum values of 3 mm and 6 mm.

Results. For clarity, we denote as *Dataset1-s3* and *Dataset2-s3* the versions of the two datasets with the 3 mm smoothing, and *Dataset1-s6* and *Dataset2-s6* with the 6 mm smoothing. The results are presented in the same way as in this chapter for the unsmoothed data: Figure 2.8 and 2.9 show the thresholded statistical maps, displaying the significant clusters; Figure 2.10 and 2.11 present the analysis of the size of the main cluster and its peak statistic value, for each dataset, each hemisphere and each size of the searchlight radius; finally, Figure 2.12 and 2.13 describe how the main clusters obtained with G-MVPA and ISPA overlap.

Overall, the results obtained with smoothing are consistent with what we found with the unsmoothed data: clusters were found in the same regions of the brain, their size increased with the value of the searchlight radius and the patterns of overlap between the clusters found by G-MVPA and ISPA were overall similar.

The main effect of smoothing appears to be that the number of significant voxels increases when the size of the smoothing kernel increases, from 0 mm to 3 mm to 6 mm. This can be explained by the fact that a small amount of smoothing helps reduce noise which can slightly improve decoding accuracies on the one hand and also increase the reproducibility across cross-validation folds on the other hand. This therefore facilitates detection as probed by our statistical test on accuracies.

Furthermore, a notable difference observed with 6 mm smoothing is the fact that the size of the activated clusters found by ISPA for *Dataset2-s6* is as large as the ones found by G-MVPA, which is not the case with smaller or no smoothing (see bottom-left graph in Figure 2.11, compared with the one in Figure 2.10 or with Figure 2.4). Congruently, the pattern of overlapping of the clusters found by the two strategies is modified for *Dataset2-s6* when compared with the one found with smaller or no smoothing (see the

right graph of Figure 2.13, compared to the right graph on Figure 2.12 and Figure 2.6 in the chapter): the number of voxels detected only by G-MVPA (green voxels) decreased in an important proportion. This is consistent with the putative explanation that was put forward in the paper, that stated that the large amount of green voxels found in *Dataset2* could be caused by a large inter-individual variability: indeed, one of the effect of spatial smoothing is often to reduce this inter-individual variability. In a complementary manner, this could also mean that the idiosyncrasies that could drive the detection of the green voxels leave at high spatial frequencies. The spatial smoothing, by attenuating these frequencies, would therefore make them vanish as we observe in this experiment.

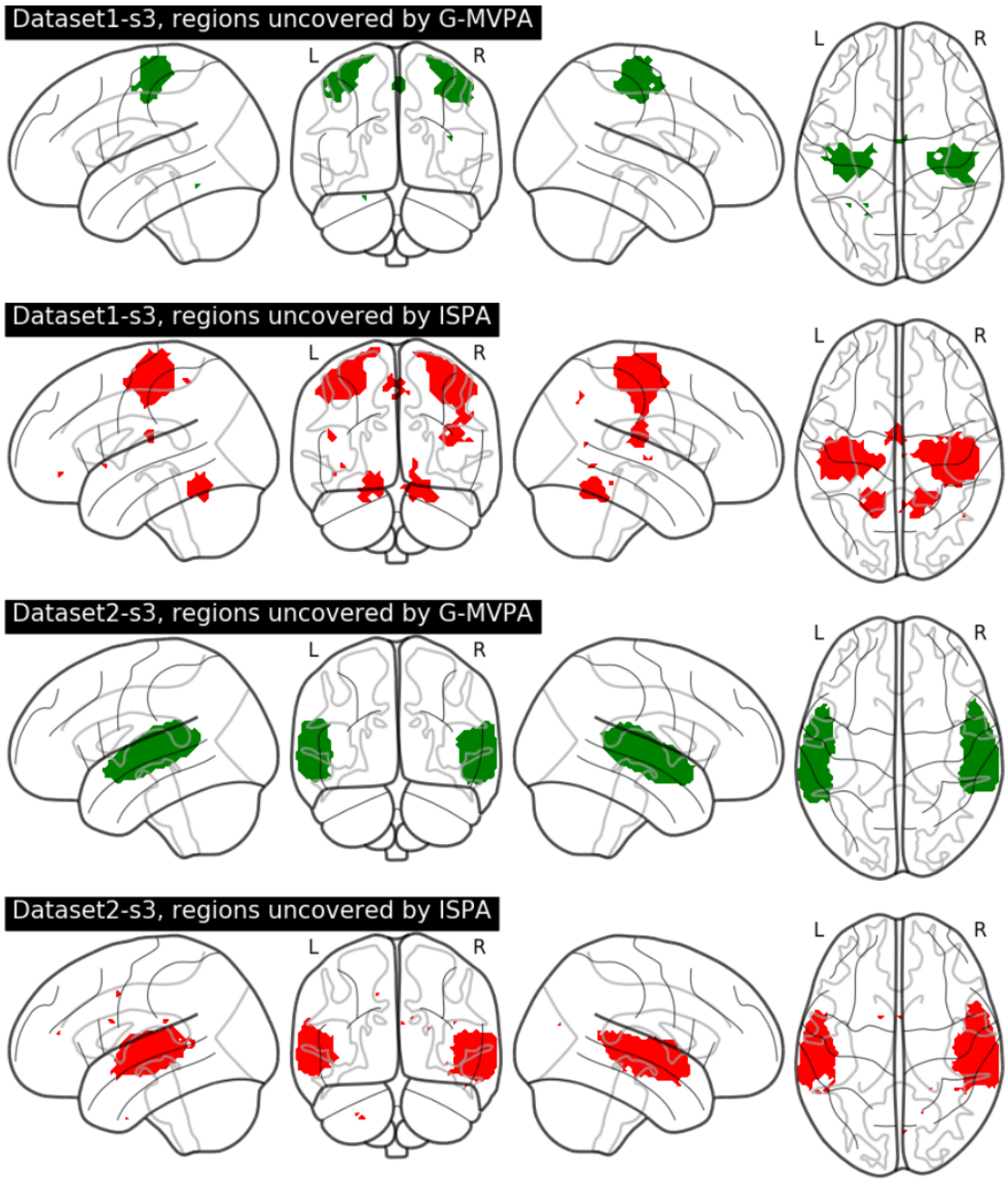


Figure 2.8: Illustration of the results of the group-level searchlight decoding analysis for a 6 mm radius with *Dataset1-s3* and *Dataset2-s3*. Top two rows: *Dataset1-s3*; bottom two rows: *Dataset2-s3*. Brain regions found significant using G-MVPA and ISPA are respectively depicted in green and red.

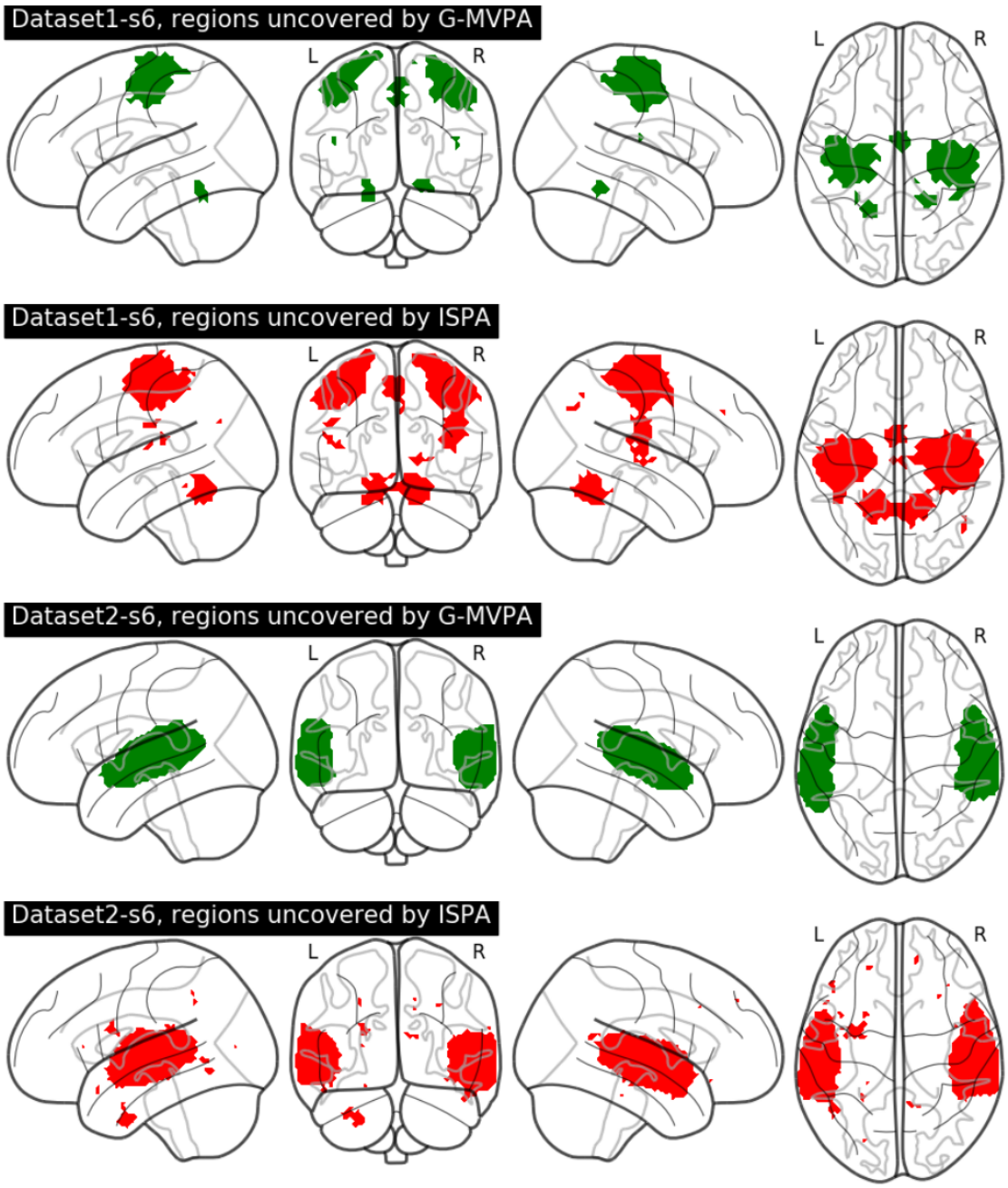


Figure 2.9: Illustration of the results of the group-level searchlight decoding analysis for a 6 mm radius with *Dataset1-s6* and *Dataset2-s6*. Top two rows: *Dataset1-s6*; bottom two rows: *Dataset2-s6*. Brain regions found significant using G-MVPA and ISPA are respectively depicted in green and red.

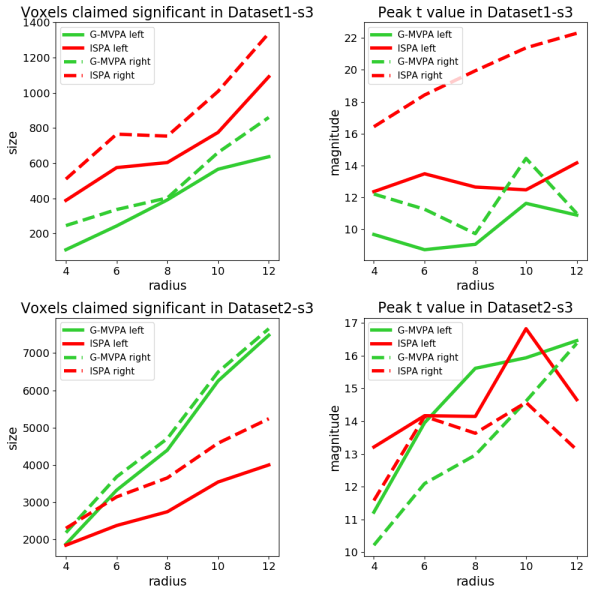


Figure 2.10: Quantitative evaluation of the results obtained on the smoothed fMRI datasets for G-MVPA (green curves) and ISPA (red curves), 3 mm Gaussian kernel. Solid and dashed lines for the largest cluster respectively in the left and right hemispheres. Left column: size of the significant clusters. Right column: peak t statistic. Top vs. bottom row: results for *Dataset1-s3* and *Dataset2-s3* respectively.

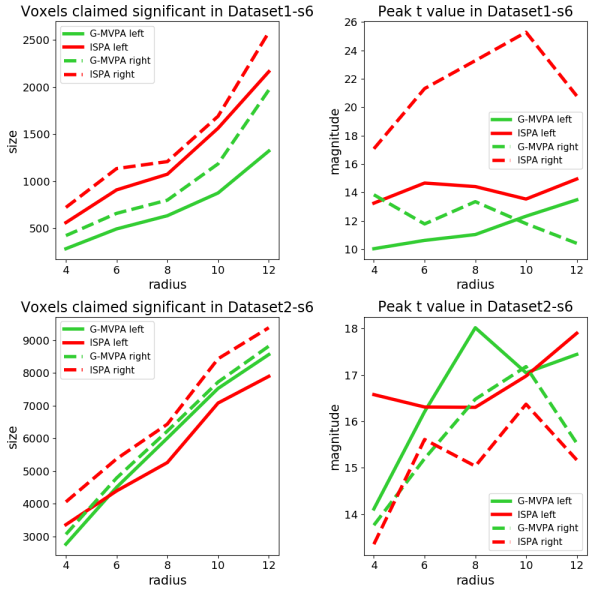


Figure 2.11: Quantitative evaluation of the results obtained on the smoothed fMRI datasets for G-MVPA (green curves) and ISPA (red curves), 6 mm Gaussian kernel. Solid and dashed lines for the largest cluster respectively in the left and right hemispheres. Left column: size of the significant clusters. Right column: peak t statistic. Top vs. bottom row: results for *Dataset1-s6* and *Dataset2-s6* respectively.

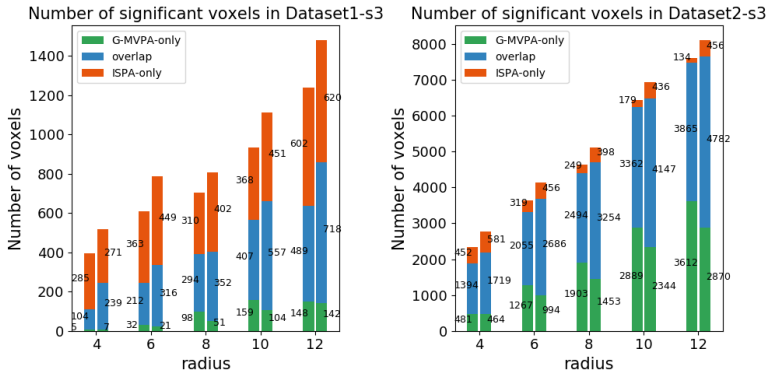


Figure 2.12: Comparison of the voxel counts detected by G-MVPA-only (green), ISPA-only (red) or both (blue) for the different values of the searchlight radius, in *Dataset1-s3* (left) and *Dataset2-s3* (right)

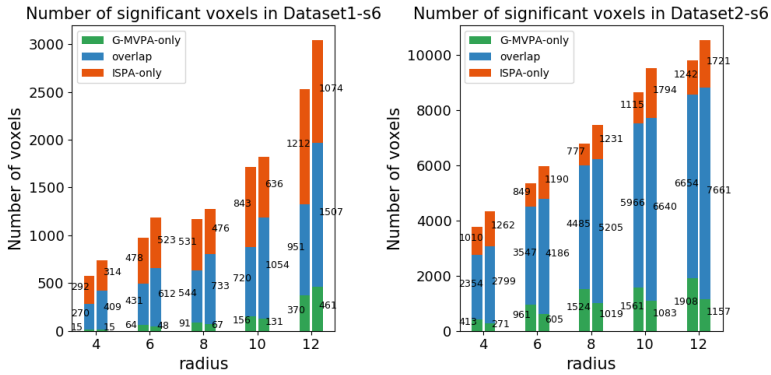


Figure 2.13: Comparison of the voxel counts detected by G-MVPA-only (green), ISPA-only (red) or both (blue) for the different values of the searchlight radius, in *Dataset1-s6* (left) and *Dataset2-s6* (right).

2.6.4 Bias and variance analyses for G-MVPA and ISPA

Aim. In order to further compare G-MVPA and ISPA, we here assess their bias and variance for estimating classifiers. For this, we exploit the generative model used to construct our artificial dataset to define the true classifier, and measure how each strategy deviates from this ground truth.

Experiments. In this subsection we use the same artificial datasets as in Section 2.3 of the paper. Before inducing some random inter-individual variability with a rotation around the origin, the two classes are separated by the the straight line $x = 0$. Because the mean variability (i.e. rotation angle) across the population is equal to zero, the true group-level classifier is defined by $f : x = 0$. Classifiers estimated from the i-th subject in G-MVPA or the i-th cross-validation split in ISPA are denoted as f_G^i and f_I^i, respectively. We use the angle between the estimated and true classifiers to measure the estimation error: the angles between f_G^i and f, f_I^i and f are denoted as θ_G^i and θ_I^i, respectively. We can then compute:

$$bias_G = \frac{1}{S}\sum_{i=1}^{i=S}\theta_G^i$$

$$var_G = \frac{1}{S}\sum_{i=1}^{i=S}(\theta_G^i - bias_G)^2$$

$$bias_I = \frac{1}{S}\sum_{i=1}^{i=S}\theta_I^i$$

$$var_I = \frac{1}{S}\sum_{i=1}^{i=S}(\theta_I^i - bias_I)^2$$

where $bias_G$, var_G and $bias_I$, var_I are the bias and variance of G-MVPA and ISPA respectively. For each point in the parameter space defined by d and Θ, we compute the average bias and variance of the 100 datasets available in each cell of the tables shown in Section 3.1 of the main paper.

Results. The values of the average bias and variance of G-MVPA and ISPA are shown as the intensities of the images presented on Figure 2.14, which represent the parameter space defined by d and Θ, as in Tables 2.1- 2.3 of the main manuscript. Overall, the bias and variance of ISPA appears to be more constant than the ones of G-MVPA, which reach much higher values. ISPA leads to a smaller bias than G-MVPA everywhere, except in a limited part of the parameter space with large effect size and relatively low inter-subject variability, which corresponds to cases where both strategies provide

equivalent detection power. Moreover, the variance of ISPA is smaller than the one of G-MVPA in all cases. In summary, ISPA present smaller bias and variance than G-MVPA in all the challenging part of the parameter space, i.e. where the effect size is small or the variability is large.

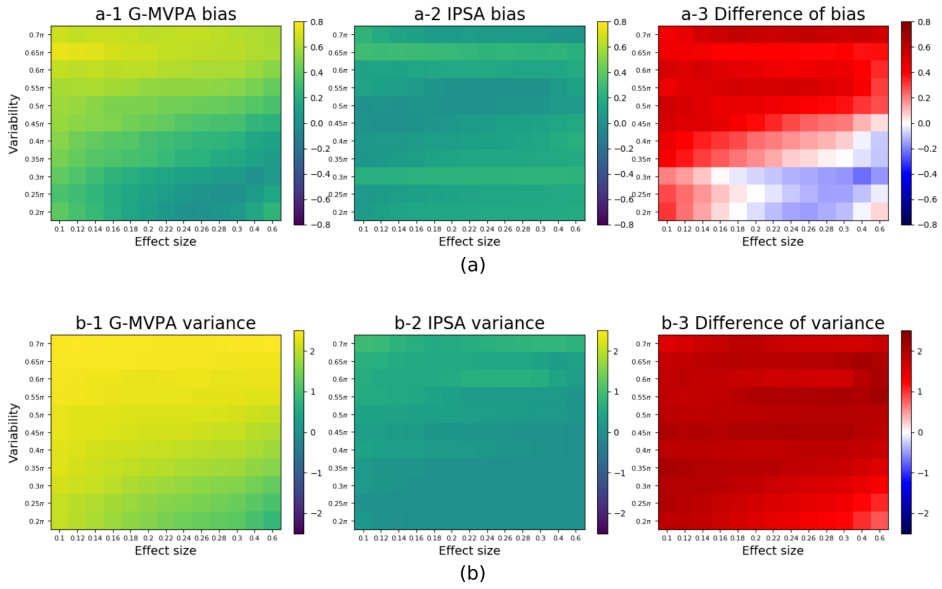


Figure 2.14: Results of the bias and variance analyses. Top, from left to right: bias of G-MVPA, bias of ISPA, difference $bias_G - bias_I$. Bottom, from left to right: variance of G-MVPA, variance of ISPA, difference $var_G - var_I$.

2.6.5 Assessing false positive rates in G-MVPA and ISPA

Aim. In order to complement the sensitivity analyses presented in the main manuscript, we here perform a specificity analysis by measuring the number of false positives produced by G-MVPA and ISPA on artificial data sets where the true effect size is null.

Experiments. We use the same generative model as in the chapter, but we set d to zero, which allows obtaining datasets that contain no effect. Furthermore, we generate datasets with different amounts of inter-individual variability by choosing Θ in $\{0, 0.05\pi, 0.1\pi, 0.15\pi, 0.2\pi, 0.25\pi, 0.3\pi, 0.35\pi, 0.4\pi, 0.45\pi, 0.5\pi, 0.55\pi, 0.6\pi, 0.65\pi, 0.7\pi\}$. For each value of Θ, we generate 1000 independent datasets. The number of datasets for which group-level decoding accuracy is significant is counted using the procedure described in Section 2.6. Because there is no true effect, all these datasets are false positives. With a threshold set at $p < 0.05$ in the permutation test, we should obtain 5% false positives or less, i.e. 50 datasets at most.

Results. The numbers of false detections obtained with G-MVPA and ISPA are shown in Table 2.17. Overall, in all cases, both strategies detect an effect in more than 50 datasets out of 1000 whereas there is none. This indicates an inflated false positive rate (on average 6.4% for G-MVPA and 9.7% for ISPA) when the statistical inference is performed with a permutation test on accuracies.

Discussion. For G-MVPA, as already documented in the literature, this could be caused by the non-symmetric nature of the distribution of accuracies (which violate the exchangeability assumption of the permutation test), or by the lack of use of the uncertainty of single-level measurements (as addressed in (Olivetti et al., 2012)). Several solutions have been proposed, as mentioned in the main manuscript (Olivetti et al., 2012; Brodersen et al., 2013; Stelzer et al., 2013; Etzel, 2015; Allefeld et al., 2016). For ISPA, the same argument about the asymmetry of the distribution holds as a potential cause for this inflated rate of false positives. Furthermore, another potential cause might be the dependence between the accuracies measured on each fold: indeed, even though the test sets are totally independent between the different folds of a leave-one-subject-out cross-validation, the classifiers learnt on each fold are not independent because of the large amount of overlap between the training sets of each fold. To the best of our knowledge, no solution has been proposed in the neuroimaging literature to handle this specific question of statistical inference on inter-subject accuracies. A solution directly applicable could be to use label permutations on the observations (as in Stelzer et al., 2013). We did not use it because of the exponentially high computing time required for this, and also because we believe that these inflated positive rates observed for the two

strategies do not question the nature of our results. Indeed, even if a few percents of the detections are mistagged as positives, i) the shape of Table 2.3 of the main manuscript (that summarizes the comparison of G-MVPA and ISPA on the artificial datasets) will hold, and ii) the qualitative assessment of the statistical maps obtained on real data would barely be affected because it is based on the largest clusters, which are clearly not false positives.

Table 2.17: Number of false positives obtained using the permutation test with G-MVPA and ISPA

variability	G-MVPA	ISPA
0.7π	62	80
0.65π	64	112
0.6π	69	100
0.55π	55	105
0.5π	62	98
0.45π	62	99
0.4π	70	92
0.35π	63	102
0.3π	57	89
0.25π	69	82
0.2π	56	98
0.15π	68	94
0.1π	62	98
0.05π	75	112
0π	71	95

Chapter 3

Inter-subject pattern analysis for functional neuroimaging: a formalization and a review

3.1 Introduction

The vast majority of MVPA studies perform within-subject pattern analysis with multiple classical machine learning techniques. Several MVPA review papers are available (see for instance [104, 108, 86, 56, 59]), however, none of them includes a thorough discussion of the inter-subject case.

At the group level, in Chapter 2 we compared two multivariate group-level analysis– G-MVPA and ISPA– with both artificial and real fMRI data. Results show that compared to G-MVPA, ISPA offers straightforward interpretation that information identified in ISPA is consistent throughout the full population. Furthermore, ISPA directly extends MVPA into inter-subject case by including datasets drawn from multiple subjects. Therefore, in this chapter we focus on ISPA.

In the literature, multiple studies use ISPA as a tool to answer neuroscientific questions, e.g. identifying cognitive states associated with visual perception [120], application to lie detection [33]. However, a largely heterogeneous vocabulary is employed in the literature dedicated to this topic, e.g. inter-subject decoding, between participants learning, generalizing to new individuals. In addition, a formal definition of ISPA is lacking, which is an obstacle to design dedicated algorithms that could be implemented in standardized methods and offered to the community in user-friendly software packages.

Motivated by the lack of unifying definition of ISPA framework and the lack of ISPA review, we therefore provide in the present chapter a formalization of ISPA and a survey summarizing different methodological strategies proposed to perform ISPA. We first introduce a unifying formalization for ISPA framework in section 3.2, then we detail in section 3.3 criteria that we applied to gather the ISPA literature, later in section 3.4 we provide a review of the existing ISPA literature. Finally, in the discussion, we mention several potential avenues to address the methodological challenges raised by ISPA, which have been made explicit thanks to the work described in this chapter.

3.2 Formalization of inter-subject pattern analysis

In this section, we first introduce a precise formalization of inter-subject pattern analysis. It relies on well-defined machine learning concepts, namely cross-validation, transfer learning and transduction, that we encapsulate in a multi-source learning setting. We detail how each of these concepts contributes to the definition of ISPA as a solution for group analysis of multivariate pattern-like effects. We also emphasize the similarities and differences with two closely related topics: brain-computer interfaces and computer-aided diagnosis systems.

3.2.1 Data description

The goal of ISPA is to identify patterns that are invariant at the group level for a given experiment. For this, we assume that we dispose of a fixed amount of data that has been recorded on a set of participants $\mathcal{S} \doteq \{1, \ldots, S\}$, who performed one or several tasks as in any typical functional neuroimaging experiments. For subject s, let $X^s = \{x_n^s\}_{n \in \{1, \cdots, N^s\}}$ be the data set consisting of N^s observations with corresponding labels denoted as $Y^s = \{y_n^s\}_{n \in \{1, \cdots, N^s\}}$, where x_n^s is the n-th observation with corresponding label y_n^s. Because the participants perform numerous repetitions of the task, we have $N^s \gg 1$. These observations, which will be used to estimate our model, can be constructed from the raw imaging data, or – as in most cases – from the outcome of several data pre-processing operations [40]. Each of these samples $x_n^s \in \mathcal{X}^s$ is labeled with a value $y_n^s \in \mathcal{Y}$, which is associated with the task performed by the subject while x_n^s was recorded. Since the brain of each individual is unique, the input spaces $\mathcal{X}^s$ are different across subjects, but the output space $\mathcal{Y}$ is common for all individuals because they have been scanned using the same experimental paradigm. The y variable can be categorical, as when it describes a finite set of categories from which the stimuli are

chosen (e.g. images of faces and houses), thus defining a classification problem, or continuous, as with behavioral measures such as the reaction time, leading to a regression problem. The full data set D can then be defined as the set of all labeled examples available for the participants in $\mathcal{S}$, i.e.

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

A probabilistic setting that may be associated with this data consists in stating that for a given subject s, the pairs (x_n^s, y_n^s) are realisations of a probability distribution $\mathcal{P}^s$. It is crucial to understand that this distribution can differ across individuals. Indeed any experimenter that has examined neuroimaging data at the single subject level is well aware that the noise level in the data often varies across subjects, and that the characteristics of the signal itself, e.g. the amplitude and location of informative features, can also be different amongst participants.

In order to formalize ISPA, we will first define a cross-validation scheme that allows addressing our objective of studying population-wise invariants, together with the training and test set. We then set up ISPA itself and position it within the machine learning realm using previously defined concepts.

3.2.2 Cross-validation, training and test set

Drawing inference on brain function at the population level requires estimating the generalization capability of a model – for instance a decoder – on data from new subjects. For this, and given the fact that $\mathcal{D}$ is composed of data recorded on a fixed number of subjects, the solution consists in using a cross-validation scheme where some subjects $\mathcal{S}_{test} \subset \mathcal{S}$ are set apart to serve as the test data, while the others – $\mathcal{S}_{train}$ – provide the data on which the model is trained. Therefore, the only cross-validation that fits the objective of ISPA is the leave-P-subjects-out strategy.

In order to simplify the notations, we will now assume that only one subject t is included in $\mathcal{S}_{test}$ – i.e we choose $P = 1$. Therefore, for a partition of the subjects into two disjoint subsets $\mathcal{S} = \mathcal{S}_{train} \cup \mathcal{S}_{test}$, the training and test sets are respectively defined as

$$\mathcal{D}_{train} \doteq \cup_{s \in \mathcal{S}_{train}} \{(x_n^s, y_n^s)\}_{n \in \{1, \cdots, N^s\}}$$

and

$$\mathcal{D}_{test} \doteq \{(x_n^t, y_n^t)\}_{n \in \{1, \cdots, N^t\}}.$$

Classically, the labels of the test examples – which are known to the experimenter – are not made available to the learning algorithm at training time. They are only used to estimate the quality of the prediction of the models in the test phase.

3.2.3 A multi-source transductive transfer problem

Given the training set $\mathcal{D}_{train}$, we would like to be able to perform predictions on the test dataset, i.e to compute targets $\{\hat{y}_n^t\}_{n\in\{1,\cdots,N^t\}}$ associated with the data from $\mathcal{D}_{test}$. Obtaining predicted targets that are close to the real ones $\{y_n^t\}_{n\in\{1,\cdots,N^t\}}$ would provide evidence that the modulations of the patterns with respect to the values of y are consistent across the training and the test subjects, which directly answers the objective of a multivariate group analysis.

Given this definition, we can exploit several existing machine learning settings in order to frame ISPA:

1. first, the nature of the training set, which gathers data from several individuals drawn from their own distributions, make it a *multi-source learning* problem, as defined in [27];

2. secondly, because the test set includes data from new individuals, thus drawn from distribution(s) not represented in the training set, ISPA offers a *transfer* question [107];

3. lastly, because inference on brain function will be drawn from the prediction of *all* labels of the test subject(s), ISPA can naturally be considered in a *transductive* context [48].

Consequently, ISPA can be formulated as a *multi-source transductive transfer* learning problem. Note that in practice, these three machine learning concepts are barely used independently. Indeed most real life problems include two or more of these elements, such as with transductive transfer or multi-source domain adaptation, which has led to strong conceptual overlap between the methods that attempt to solve them. Effectively combining these three concepts is an important current challenge that is met with all applications in life sciences where numerous data samples are recorded in several individuals (ISPA being one instance of such problem), but also in other fields such as natural langage processing (see e.g [50, 63]).

3.2.4 Relationships with similar problems

Link with Computer-Aided Diagnosis (CAD) systems. CAD systems also aim at performing predictions for individuals that were not available at training time. In this setting, a data points is given by an individual: typically, a physician would like to predict whether a new patient is healthy or sick using for instance a single T1 image (i.e $N^s = 1$). In contrast, in ISPA, we have $N^s >> 1$ and a prediction has to be performed for each of the numerous data points available for a new individual.

Link with Brain-Computer Interface (BCI). In a BCI problem, one need to obtain a model that is as reliable as possible in order for the device to be effective. While most BCI-dedicated studies focus on within-individual analyses, there is a clear added value to being able to fuse the data from multiple subjects in order to obtain a system that would be effective on new individuals. However, contrarily to ISPA for which the amount of data is fixed by the time the analysis is performed, a BCI is – by design – aimed at being used on *any* data that will be acquired in the future. Therefore, while ISPA only requires to predict the labels of the available test data $\{x_n^t\}_{n\in\{1,\cdots,N^t\}}$ – making it a *transductive* problem, training a BCI system requires to learn from $\mathcal{D}_{train}$ a predictor $f : \mathcal{X} \to \mathcal{Y}$ with risk $R(f) \doteq \mathbb{E}_{(x,y)\sim\mathcal{P}^t}[f(x) \neq y]$ as small as possible, which is a more difficult *inductive* learning question. Consequently, inductive methods developed for multi-subjects BCI should be directly useful for ISPA (see examples in section 3.4), and eventually improved using transduction.

Link with multi-view and multi-task learning. In multi-view learning, which is also called multi-modal learning, each observation is composed of several *views* of the same sample. For instance, in object recognition, an object can be photographed from several angles, and in a CAD problem, a patient can be scanned with different medical imaging devices (MRI, PET etc.); in both cases, each view provides complementary information which should help solving the learning task. Therefore, multi-view learning is characterized by the nature of the input space, which is a cartesian product of several spaces where each view lives. In multi-task learning, the goal is to solve several learning tasks simultaneously. Assuming these tasks are somehow related, multi-task learning exploits these relationships in order to better solve them jointly than what they would be independently. Thus, multi-task learning is defined by the nature of the output space $\mathcal{Y}$, which is a multi-dimensional vector space where each of the tasks is encoded in a given dimension. Since in ISPA all subjects are observed through a single imaging modality (i.e the input space is not a cartesian product of several spaces) and perform the same set of tasks (i.e the output space is mono-dimensional), it does not fit directly in the

definitions of multi-view and multi-task learning. However, under certain hypotheses, it is possible to reframe ISPA into these settings (see examples in section 3.4).

The different meanings of multi-source. To conclude this section, we would like to bring the attention of the reader to the fact that the *multi-source* term appears in the literature used with very different meanings, sometimes designating heterogeneous data, sometimes being used as a synonymous of multi-view or multi-modal data etc. We insist on the fact that we here use the clearly defined multi-source learning setting proposed in [27], which unambigusously designates the supervised learning problem where a single model has to be learnt from training data drawn from several probability distributions but associated with a common output space $\mathcal{Y}$.

3.3 Searching the literature for ISPA papers

Our goal is to identify all the papers performing ISPA in the literature. However, because there is no agreement on how to designate ISPA in the literature, studies use heterogeneous vocabulary for describing ISPA. We here describe an exhaustive set of keywords that should allow identifying all ISPA related studies. First ISPA is a supervised learning approach that performs "classification" or "regression" (K1) on functional neuroimaing data, i.e. "fMRI", "EEG" or "MEG" (K2). Then ISPA implements "cross validation" (K3), especially "leave-P-subjects-out" (K4) for estimating the performance of classification or regression models. Furthermore, ISPA performs "across subjects" decoding (K5). Since there are heterogeneous vocabularies for describing "leave-P-subjects-out" and "across-subjects", we identified all synonymous of "leave-P-subjects-out"and "across-subjects" in the literature, which are concatenated with 'OR' operator in K4 and K5, respectively. The full list of keywords that we used to identify all ISPA papers is shown in Figure 3.1a. Finally, we applied a search query consisting of all five criteria K1 AND K2 AND K3 AND K4 AND K5, so that we could identify papers which contain all five keywords.

We searched for papers in Google Scholar with a software named *Publish or Perish*[1]. We searched for papers dated between January 2000 and April 2019. Since the limitation for the number of characters of input field in Google scholar is 150 characters, we could not put all five keywords in one query. Therefore we used two search queries, the first one is K1 AND K2 AND K3 AND K4 and the second one is K1 AND K2 AND K3 AND K5. Additionally, Publish or Perish has a limitation that each query obtains at

[1] https://harzing.com/resources/publish-or-perish

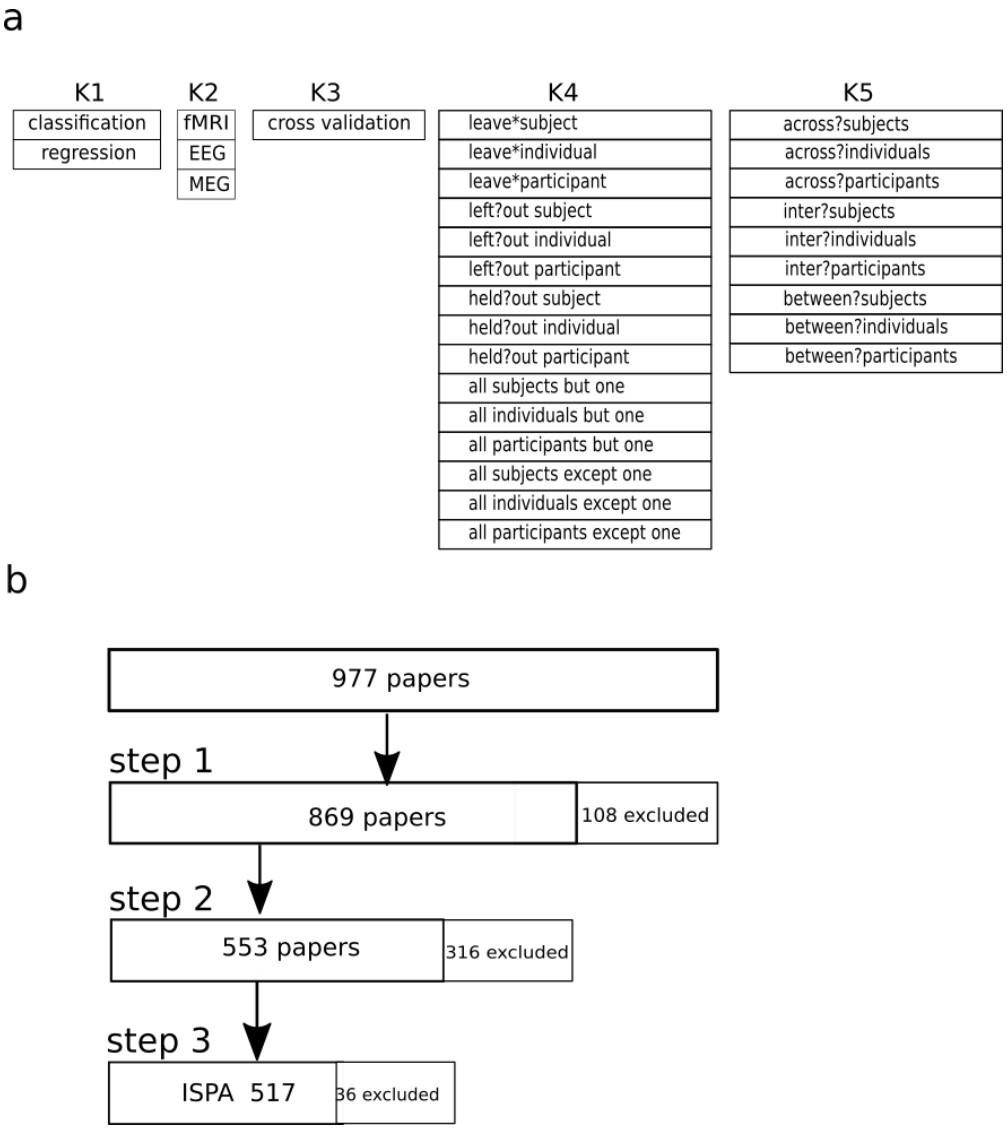


Figure 3.1: Searching criteria for ISPA papers. a: The search query consists of five criteria K1 AND K2 AND K3 AND K4 AND K5, the operator between two items in one criterion is 'OR' searching for content containing either item, e.g. K1= "classification" OR "regression", "*" and "?" are wildcard characters replacing one or more characters. b: We use three steps for filtering identified records. The first step excludes articles by paper type, the second step excludes articles by reading abstracts and the third by thoroughly reading articles.

most 1000 results, we searched for papers with these two queries separately in each year. The search ended up with 1573 papers for the first query and 13798 papers for the other. In order to obtain papers with K1 AND K2 AND K3 AND K4 AND K5, we made an intersection of results obtained by the two queries, which ended up with 977 results. Because there are several topics related to ISPA, these 977 results are not all ISPA studies, therefore a filtering process is needed.

We conducted three steps to identify ISPA studies. The first step was based on article type: book chapters, thesis, reviews and journal articles that had not been peer-reviewed were excluded, which left us 869 records. In the second step, abstracts from all the 869 articles were screened. 80 records dedicated to BCI and 151 records involving CAD were excluded for the reason we have given in section 3.2. Then 85 records which contain one brain pattern per subject were also excluded, because they are in fact methodologically equivalent to a CAD problem. In the third step, 553 papers were thoroughly read, and articles using data from all subjects but one for performing univariate analysis [141] were excluded, thus leading to a total 517 ISPA articles (see illustration in Figure 3.1b).

3.4 Inter-subject pattern analysis: a review

Performing ISPA includes several steps which are also involved in within-subject pattern analysis. Therefore we first describe steps in within-subject pattern analysis and then outline steps in ISPA.

In within-subject pattern analysis, individual dataset acquired from MRI, EEG or MEG device is first carried forward to functional neuroimaging analysis to generate input dataset for follow-up machine learning stages. This analysis usually includes data pre-processing steps, e.g. fMRI data preprocessing described in Chapter 1, and extracts patterns from the outcome of data pre-processing operations, such as beta values estimated from the GLM. Then patterns are split into the training and test set which are carried forward to machine learning stages. Figure 3.2 illustrates the flowchart of within-subject pattern analysis which includes four machine learning steps [104]. The first step is *feature construction or/and selection* which is performed on the training set and later implemented on the test set. In the second step training data are fed into a machine learning algorithm, i.e. classification or regression algorithm, for learning the mapping between input feature space and output space, which is the learned *predictive model*. The third step, *performance evaluation*, aims to evaluate how well the predictive model generalizes to unseen data in the test set with some classical metric, e.g. accuracy.

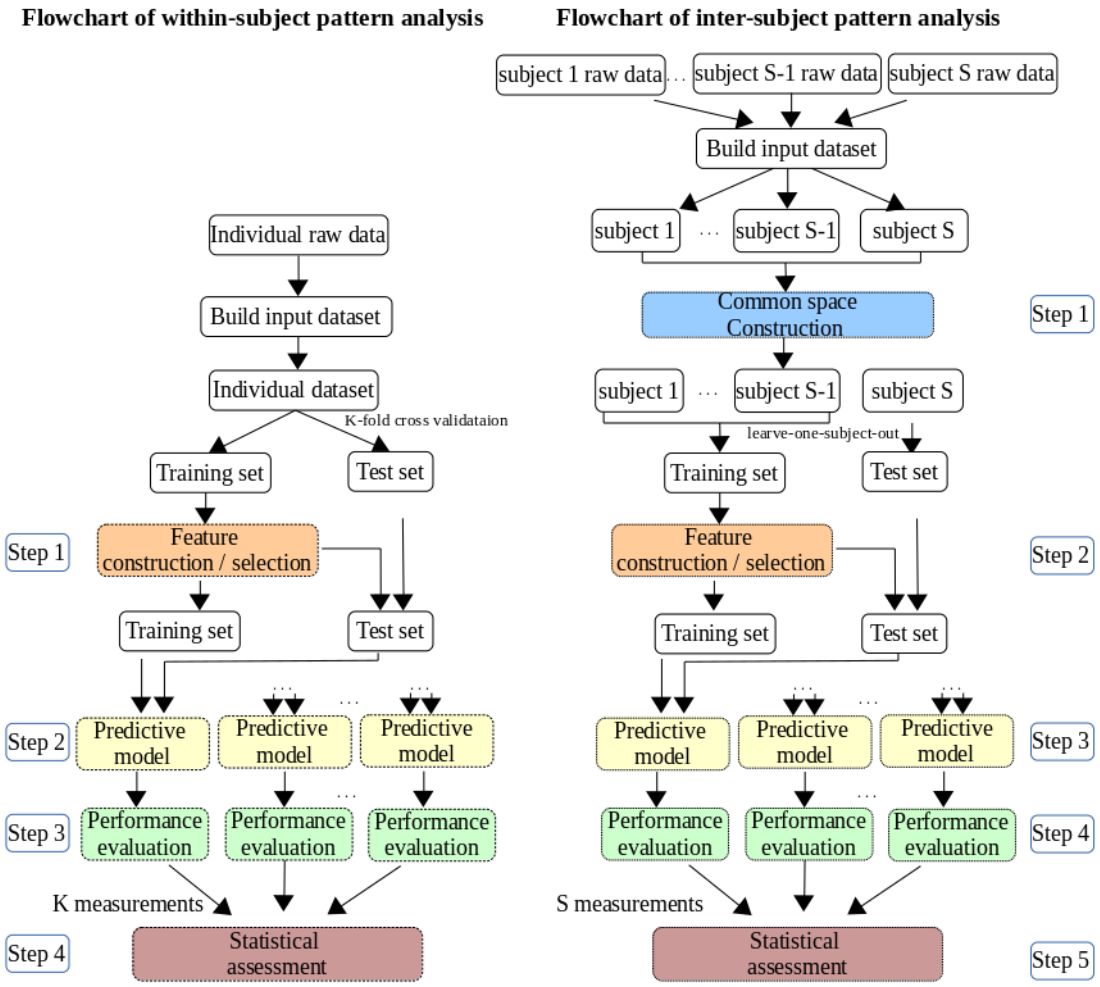


Figure 3.2: The flowchart of within-subject pattern analysis and ISPA. Left: Flowchart of within subject pattern analysis. Right: Flowchart of inter-subject patten analysis. ISPA has similar framework as within-subject pattern analysis, the difference is that a common space is constructed for performing ISPA, which results in five steps for ISPA: common space construction, feature construction/selection, learning predictive model, performance evaluation and statistical assessment.

The fourth step is statistical assessment which exploits statistical models, e.g. t test or permutation test, to assess the significance of model performance.

All four steps in within-subject pattern analysis are also necessary in ISPA (see illustration in Figure 3.2). The main difference between ISPA and within-subject pattern analysis is that ISPA has one more step prior to these four steps. Because ISPA involves datasets drawn from multiple subjects, performing ISPA requires to construct a common space which overcomes the challenge that brains of the different subjects (i.e. the original input spaces) are not identical. Common space construction is usually achieved via data preprocessing, e.g. spatial normalization with Talairach atlas or Montreal Neurological Institute(MNI) templates, or via advanced methods performed on patterns which are drawn from output of preprocessing operations. After the construction of a common space, ISPA studies exploit multiple machine learning strategies for the other four steps, e.g. classical machine learning methods which are also adopted by within-subject pattern analysis, or advanced methods specific for inter-subject decoding.

Overall, with respect to these five steps in ISPA, the most commonly used strategy consists in 1. construction of a common space with spatial normalization 2. classical feature construction / selection on the aggregation of multiple datasets–known as cross-subjects data *pooling*, which ignores the inter-subject variability between subjects and thus assumes data in the pooled training set are drawn from the same distribution 3. learning a predictive model 4. performance evaluation and 5. statistical assessment with classical methods. In fact, most studies (439 papers) out of the 517 identified ISPA results adopt the strategy mentioned above, among which 350 studies exploit spatial normalization to construct a common space without follow-up feature designing strategies, while others (89 papers) perform classical feature designing following common space construction. Furthermore, for these 89 papers which adopt classical feature designing methods, we will summarize their methods in section 3.4.2.

Despite its massive usage in the literature, there are several limitations with the commonly used strategy. For instance, spatial normalization has shortcomings which are well identified in [126], multiple datasets pooling ignores the difference between subjects. Among 517 identified studies, there are 78 ISPA studies proposing advanced methods to handle these limitations, which we will review below. This section is divided into five parts, where each part focuses on one step of the framework of ISPA.

Table 3.1: Commonly used or advanced approaches proposed for common space construction

References	Data type	Calibration dataset	Strategy
	fMRI	Talairach atlas	spatial normalization
	fMRI	MNI template	spatial normalization
Fuchigami et al.(2018)[45]	fMRI	DTI	inter-subject registration
Haxby et al.(2011)[58]	fMRI	fMRI from movie watching	HA
Xu et al.(2012)[144]	fMRI	fMRI from movie watching	regularized HA
Lorbert et al.(2012)[83]	fMRI	fMRI from movie watching	kernelized HA
Chen et al.(2014)[22]	fMRI	fMRI from movie watching	HA with SVD
Guntupalli et al.(2016)[54]	fMRI	fMRI from movie watching	HA with searchlight
Chen et al.(2015)[24]	fMRI	fMRI from movie watching	SRM
Chen et al.(2016)[23]	fMRI	fMRI from movie watching	SRM with searchlight
Zhang et al.(2016)[150]	fMRI	fMRI from movie watching	ICA-based SRM with searchlight
Turek et al.(2017)[131]	fMRI	fMRI from movie watching	semi-supervised SRM
Rustamov et al.(2016)[115]	fMRI	fMRI from movie watching	anatomical constraints in HA
Yousefnezhad et al.(2016) [147]	fMRI	fMRI from movie watching	HA with labeled data
Yousefnezhad et al.(2017)[148]	fMRI	functional template	deep learning-based HA
Xu et al.(2018)[145]	fMRI	fMRI from movie watching	HA with ICA and SGA
Turek et al.(2018)[130]	fMRI	fMRI from movie watching	SRM-based model
Yamada et al.(2015)[146]	fMRI	one fMRI dataset	neural code converter
Morioka et al.(2015)[97]	EEG	resting-state data	common directory

3.4.1 Common space construction

In ISPA the objective is to find regularities that hold across subjects at the population level. This requires data to be integrated across individuals. However, the size and shape of brain vary across subjects, learning a predictive model directly from the aggregation of multiple datasets is infeasible. Therefore numerous studies build a common space to align multiple datasets (see illustration in Figure 3.3). The transformation to this common space is learned in the training set and then applied to the test set, therefore the construction of a common space also offers a solution to *transfer* knowledge from the training to the test subjects.

The construction of a common space usually exploits two independent sets of data. The first experiment, here we call *calibration* experiment, is used to align the brain functional responses of multiple subjects by learning the mapping between individual and common spaces. The other one, designated as the *main* experiment – is dedicated to study the brain function of interest to the researcher, for which he/she aims at studying the activation patterns at the group level using ISPA.

In general, there are two types of calibration datasets, anatomical and functional, which usually correspond to anatomical and functional alignment, respectively.

Anatomical alignment uses anatomical image as calibration dataset, with which datasets of multiple subjects are transformed into a common space based on anatomical features. Other than spatial normalization which takes T1-weighted structural MRI images as

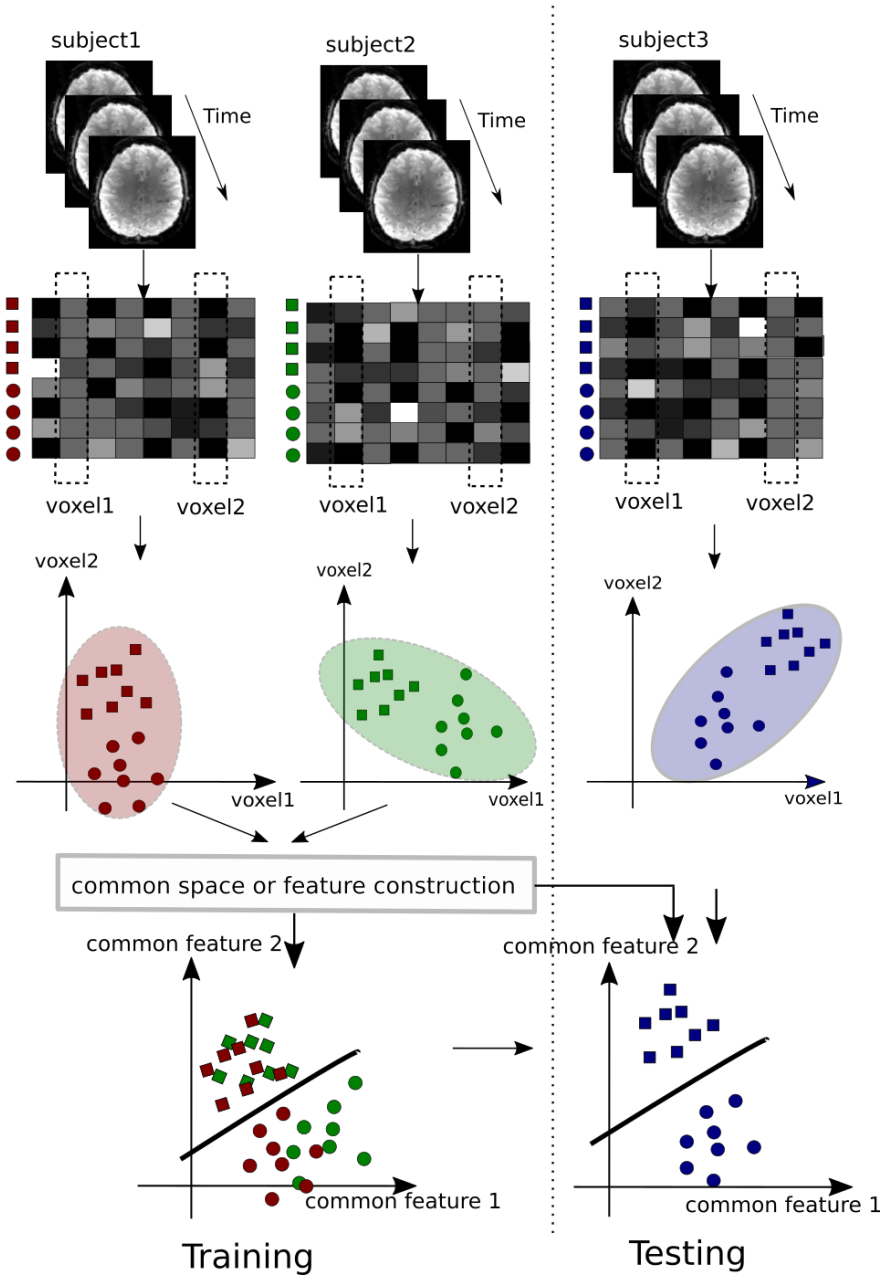


Figure 3.3: Reducing inter-subject variability by common space and feature construction. Individual functional dataset acquired by MRI, EEG or MEG devices is first represented as a set of vectors, one vector per trial with the corresponding label. Squares and circles represent two experimental conditions. We use two features per trial for illustrating in two-dimensional space. Common space construction and/or feature construction are usually performed in the training set then later implemented on test set, which shows the transferring from the training to the test set. A predictive model is learned from the transformed training set later generalizes to the transformed test set.

calibration dataset, Fuchigami et al. proposed Diffusion tensor imaging (DTI) based registration which used DTI instead of T1-weighted image to transform functional maps of each individual to a common anatomical space. Compard to inter-subject registration with T1-weighted templates which captures the structure of grey matter, DTI-based registration captures the structure of white matter and provided comparable inter- and within-subject decoding results [45].

Anatomical alignment is commonly used in the literature, but several studies proposed to align datasets with functional alignment, which precisely aligns the functional images across subjects. In the following, we examine ISPA studies which have introduced specific methods to perform functional alignment.

One popular method for functional alignment is Hyperalignment (HA) [58]. fMRI scans from various time points contain both temporal and spatial variability. In order to align multiple datasets, in [58] and follow-up studies, subjects in both training and test set are presented with identical, time synchronized stimuli, during which the temporal variability across subjects is removed. The alignment protocol usually consists of two steps: i) calibration experiment, which can consist in passively watching a movie, allows to learn the mapping parameters between common space and individual feature spaces with the functional calibration dataset, ii) main experiment, which usually include movie segment matching or image classification, is dedicated to study the brain function of interest at the group level with a new set of datasets aligned by mapping parameters learned from calibration experiment.

Haxby et al. first proposed HA to align multiple datasets using the orthogonal Procrustean transformation [58]. The authors demonstrate that projecting the data from all subjects into this common space allows obtaining significantly higher ISPA generalization performance in contrast with anatomical alignments, confirming the relevance of this functional alignment procedure. Several follow-up studies by the same research group have pursued the development of HA by studying its relationship with Canonical Correlation Analysis and proposing a regularized version of HA [144], or by introducing a kernelized extension which allows grasping non-linear effects [83]. Compared to HA using PCA to reduce the dimensionality of data, Chen et al. employs a joint Singular Value Decomposition (SVD) to perform dimensionality reduction [22]. [54] applied HA together with a searchlight approach, which yielded a cortex-wise function-based common representational space capturing fine-scale topographies. In addition, Chen et al. proposed Shared Response Model (SRM), a derivation of HA, to estimate a response model that is shared across individuals and directly embeds dimensionality reduction,

which allows improving generalization performance and extracting information [24]. In their follow-up studies, Chen et al. proposed a multi-layer convolutional autoencoder (CAE) which is a nonlinear functional alignment model based on SRM and uses searchlight algorithm to preserve spatial locality [23]. Similarly, Zhang et al. proposed to identify shared information in searchlight sphere with ICA-based SRM [150]. Compared to studies above which only employ unlabeled data to learn shared response, [131] introduced a semi-supervised SRM with data and their corresponding labels. A few other groups have also contributed to extensions of the HA. [115] added anatomical constraints in the estimation of the functional alignment. [147] raised the constraint that all participants must have been scanned using a strictly identical calibration paradigm. In their follow-up studies, they presented a deep-learning-based functional alignment procedure thus accommodating for non-linear mappings [148], and proposed to use ICA and Stochastic Gradient Ascent (SGA) to accelerate the computation for large amounts of samples and voxels in HA [145]. Turek et al. introduced a SRM-based model which preserves a component of sparse individual activity not accounted for by the shared representation [130]. In conclusion, all the works above tried to find a better solution for HA.

Similar to HA performed with time synchronized stimuli, the method in [146] consists in choosing one individual as template to define a common functional space: their *neural code converter* is a set of multiple linear regressions where each voxel of the target subject is modeled as a linear combination of several voxels of source subject. After neural code converter was learned from calibration experiment, it's applied in the main experiment to transform source subject's activity patterns into target subject's feature spaces.

Other than task-related functional data, other types of data are also exploited to align multiple datasets, such as resting-state data which is used to learn a common dictionary shared by multiple subjects [97].

3.4.2 Feature construction or/and feature selection

In the literature, numerous methods of feature construction and selection are implemented to improve the performance of machine learning model in ISPA. Additionally, functional neuroimaging datasets usually contain only hundreds of observations, while the number of features in each observation may reach thousands which is much more than the number of observations. Therefore feature construction or/and selection offers to reduce the number of features, decrease the computational complexity and overfitting.

Table 3.2: Approaches for feature construction/selection. Each group corresponds to one category in feature construction/selection

references	**input space**	**feature selection / construction**
Gilron et al.(2017)[49]	fMRI	searchlight
Mourao-miranda et al.(2005)[99]	fMRI	PCA
Mourao et al.(2006)[98]	fMRI	SVD
Duff et al.(2012)[36]	fMRI	RFE
Shinkareva et al.(2011)[119]	fMRI	regression weight map
Reichert et al.(2014)[112]	fMRI	combined SVM weight maps
Michel et al.(2011)[92]	fMRI	bayesian regression
Ryali et al.(2010)[117]	fMRI	l_1, l_2 regularization
Michel et al. (2011)[93]	fMRI	total variation regularization
Grosenick et al.(2013)[52]	fMRI	GraphNet regularization
Koyamada et al.(2015)[72]	fMRI	ROI averaging, DNN
Bashivan et al.(2015)[8]	EEG	recurrent-convolutional neural network
Lu et al.(2016)[84]	EEG	restricted Boltzmann machine
Jang et al.(2017)[65]	fMRI	3D DNN with pretrained deep belief network
Fahimi et al.(2019)[41]	EEG	CNN
Chang et al.(2011) [20]	fMRI	regression model mapping word semantic features to brain activities
Wang et al.(2017)[138]	fMRI	regression model mapping sentence semantic features to brain activities
Van et al.(2018)[133]	fMRI	regression model mapping word2vec to brain activities
Li et al.(2018) [81]	EEG	multi-subject standardization
Mitchell et al.(2004)[96]	fMRI	most active voxels, average in manually-defined ROI
Davatzikos et al.(2005)[33]	fMRI	average in cubic regions
Singh et al.(2007)[121]	fMRI	most active voxels
Shinkareva et al.(2008)[120]	fMRI	most active, most stable voxels
Just et al.(2010)[67]	fMRI	most active, most stable voxels, two-level factor analysis
Chang et al.(2011)[20]	fMRI	most stable voxels
Cabral et al.(2012)[17]	fMRI	most active, discriminative voxels
Michel et al.(2012)[91]	fMRI	supervised parcellation and average
Accamma et al.(2015)[2]	fMRI	voxels coordinate, manifold learning
Barachant et al.(2013)[7]	EEG	alignment in Riemannian manifold
Fatima et al.(2017)[43]	EEG	alignment in Riemannian manifold
He et al.(2018)[60]	EEG	alignment in Euclidean space
Mason et al.(2016)[90]	fMRI	two-level factor analysis
Damarla et al.(2016)[30]	fMRI	common most stable voxels across subjects
Markides et al.(2012)[87]	fMRI	group-level ICA masks
Eisenbarth et al.(2016)[37]	fMRI	weighted feature maps
Van gerven et al.(2009) [132]	EEG	l_1/l_p regularization
Kia et al.(2017)[68]	MEG	$l_{2,1}$ regularization
Takerkart et al.(2014)[124]	fMRI	parcel-based graphs
Vega-pons et al.(2014)[136]	fMRI	parcel-based graphs
Raizada et al.(2012)[111]	fMRI	similarity-based features
Zhang et al.(2015)[151]	EEG	transfer component analysis
Zheng et al(2016) [152]	EEG	transfer component analysis and kernel PCA
Tong et al.(2018)[129]	EEG	domain classifier in DNN
Li et al.(2018)[80]	EEG	deep adaptation network
Li et al.(2018)[79]	EEG	adversarial domain adaptation networks
Luo et al.(2018)[85]	EEG	WGAN
Chai et al.(2018)[19]	EEG	subspace alignment

After the construction of a common space, multiple studies adopt feature designing strategies on the aggregation of multiple datasets, which ignores the difference between subjects and implicitly assumes that common space construction transforms multiple datasets into one same distribution. Some studies explore methods of feature construction and selection, with which inter-subject invariance is able to be achieved. Therefore in this section we attempt to classify these studies into a few categories.

Methods performed on pooled training data

In this category, studies adopt feature designing strategies on the training set which consists of data pooled from multiple subjects. Strategies first learn parameters or select features from the training set then apply parameters or select same features on the test set.

- **Classical feature designing methods**

  In this part, we summarize the feature design methods from the 89 papers that adopted classical schemes for this processing step. These methods are exploited in both within-subject pattern analysis and ISPA, which ignore the multi-subject nature of the training data. Since multiple papers use the same strategies, we only provide a few representative examples.. One popular feature designing method is PCA that is used to both construct features and perform dimensionality reduction. For instance, PCA is allowed to be applied either directly on the BOLD signal [99] or on the probability distribution of fMRI intensity calculated within each ROI [42]. Similarly, SVD could also be used to construct features in ISPA [98]. Instead of constructing new features from the original ones, a few studies opted to frame a searchlight strategy for fMRI data. *Local* patterns – defined in the spherical region where the searchlight beam focuses – can be obtained to make inference in the single-subject or inter-subject setting [49, 25]. Recursive feature elimination (RFE) is another commonly used method for feature selection. For instance, RFE is used for selecting features for a multi-class classification model in [36].

- **Feature selection based on weights of linear models**

  One group of studies adopt linear models on the aggregation of multiple datasets. Since linear regression or classification models have weights the same number as the amount of features, they are used to select features for whole brain data. In [98] support vector machine (SVM) is exploited on the training set to perform feature selection. Similarly, [119] uses a logistic regression to select features that

have the highest weights.

Additionally, in order to obtain consistent features across different subjects which improve the interpretability of the decoding model, some studies add regularization to models. [92] proposed a bayesian regression model for feature selection. [117] exploited l_1 and l_2 norm regularization, the feature selection process was implicitly induced by the learning algorithm through the regularization terms that enforced sparsity on the spatial weights. [52] added a GraphNet regularization which is the graph-constrained l_1 and l_2 norm. [93] adopted total variation as regularization to select informative voxels that are sparse and spatially grouped into connected clusters.

- **Learning invariant representations** Some studies focus on learning invariant representations across subjects. A commonly used method is using neural network. Since deep neural network (DNN) obtains increasing attention in recent years, several methods hails from the recent successes of deep learning for numerous learning tasks. The intuition is here that the representations offered by intermediary layers of such neural networks could offer a higher level of reproducibility across subjects, such as four-layer DNN learned from average of values in predefined ROIs (regions of interest) [72], 3D DNN initialized with a pretrained deep belief network on whole brain data [65], restricted Boltzmann machine proposed for motor imagery classification [84], recurrent-convolutional neural network used for mental workload classification [8], and convolutional neural network (CNN) applied on raw EEG data which reduced information loss caused by feature extraction [41]. Furthermore, in some studies, extra semantic features are required for learning invariant representations. In [20] and [138], each word or sentence is represented by semantic features, then a regression model learns the mapping between semantic features and brain patterns recorded from multiple subjects. After the training phase is finished, the regression model generates one shared brain pattern per stimulus, which is used to examine the category of test data. Similarly, in [133] each word stimuli is first represented by word2vec, then word2vecs are fed into a ridge regression model that learns the mapping between semantic features and HA aligned brain patterns.

Methods exploiting the multi-subject nature of the data

Instead of performing feature designing strategies on the aggregation of multiple datasets, a number of studies take into account the multi-subject nature of the training set and

perform feature designing subject by subject.

One simple strategy for feature construction is multi-subject feature rescaling, e.g. feature standardization which rescales features to have zero-mean and unite variance. Multi-subject feature standardization learns rescaling parameters (mean and variance) subject by subject in both training and test set, e.g. multi-subject feature standardization for EEG datasets [81] and for fMRI datasets [149]. In order to better understand the effects of this feature construction strategy which is often used but rarely explicitly mentioned, an in-depth study of the effect of this operation will be presented in Chapter 4.

Several feature selection and construction methods have been successfully used on fMRI data. Some employ univariate selection on voxel features, using criteria such as, most active or discriminative voxels in the predefined ROIs [120, 17, 121, 67, 96], most stable voxels in [67, 120, 20]. Others summarize the signal present in a set of regions by their mean, using, e.g. cubic regions [33], anatomically defined regions [140, 96], functionally defined parcels [91], or regions selected by univariate analysis [61].

Some studies use features from a manifold. [2] takes an original route by first selecting a fixed number of voxels within a given ROI, then encoding the spatial organization of these voxels using a manifold learning approach applied on the set of 3D coordinates of these voxels, which provides a low dimensional feature space. Interestingly, the decoding is directly performed on these features, which do not contain any information on the amplitude of activation. Additionally, one popular method to construct invariant feature space for multi-channel MEG and EEG trials is using Riemannian space. MEG and EEG trials include both spatial and temporal information which is discriminative between experiment conditions, a commonly used strategy is directly encoding the spatio-temporal characteristics of the signals in covariance matrices, which lie on Riemannian manifold, to obtain population-wise invariants [7], [43]. However, aligning covariance matrices of multiple datasets on Riemannian space has high computational cost, therefore He et al. proposed to directly align each subject's EEG trials with individual average trial, which improved alignment speed [60].

Several studies use two-level strategies which first perform feature designing on individual subject's data, then select common features across subjects, such as two-level factor analysis [67, 90], selection of common stable voxels across individuals [30], group-level independent component masks created from individual independent component masks [87] and group-level feature map computed from weighted combination of individual feature maps [37], [112].

Additionally, in order to obtain group-wise interpretable features which also preserve idiosyncratic properties within each individual, group-wise regularization is added to linear model, e.g. [132] proposed to use l_1 /l_p regularization term, and [68] adopted a $l_{2,1}$ regularization.

Methods that implicitly use transduction

We have framed ISPA as a transductive learning question, therefore researchers have access to unlabeled test data during the training phase. Several studies exploit this transductive nature of ISPA to construct features for both training and test set. First, because of test data available beforehand, above-mentioned multi-subject feature rescaling is able to rescale features using standardization subject by subject for both training and test set before the training phase begins. This method exploits the multi-source transductive nature of ISPA, which will be further studied in Chapter 4.

Then, instead of constructing activation-based features, the method introduced by [111] encodes the data in a similarity space, each example being compared to the prototypes of each category to be decoded. This bears a strong resemblance with the concepts behind Representation Similarity Analysis [76], but exploits similarity-based coding for the purpose of obtaining inter-subject invariance. Because the predictions are performed jointly on all examples of the test subject, it fully exploits the transductive nature of ISPA.

A complementary set of methods aims at exploiting the spatial structure of the information, because structural representations are known to improve the robustness of the information encoding in presence of large variability. [124] and [136] follow similar ideas, by constructing graphical representations from a parcellation of the data. They then use a kernel-based learning method to build a model that is able to produce accurate predictions on data from new subjects. Note that i) they do not require having performed spatial normalization of the data before hand, and ii) the feature space that is common to all subjects is here implicitly defined as the Reproducing Kernel Hilbert Space (RKHS) spanned by the kernel. Since the parcellation and graphical representations are learned from all observations of each subject including the test subject, it also exploits the transductive nature of ISPA.

Methods that include a transfer component

Given the fact that the test set is composed of subjects that are not included in the training set, it is challenging for model learned from the training set to generalize to the test set. In order to reduce the difference between two sets, it requires to transfer knowledge from the training set to the test set.

Several approaches proposed for transfer learning, particularly for domain adaptation, are implemented for performing ISPA. The major computational problem is how to reduce distribution shifts between source and target domain, i.e. training and test set. One commonly used strategy aims to minimize some metrics between the training and test set, for which data from both training and test set are exploited during the training phase. [151] proposed transfer component analysis which learns a transformation matrix to transform the empirical kernel mapping to a low-dimensional latent space where the source and target domain are similar. [152] minimized maximum mean discrepancy of source and target domain to implement personalized EEG-based emotion models. In their follow-up studies they performed sleep quality estimation by maximizing the loss of a domain classifier which is trained along a deep neural network [129]. For emotion recognition they applied deep adaptation network by minimizing multi-kernel maximum mean discrepancies [80]. Furthermore, for vigilance estimation they adopted two adversarial domain adaptation networks and compared these two methods with several traditional domain adaptation methods [79]. In addition, they proposed a domain adaptation framework with Wasserstein generative adversarial network (WGAN) [85]. Instead of considering the training set as one domain, [19] proposed multi-source subspace alignment which transforms each source subject's dataset into target subject's subspace. In this study, first d eigenvectors of the dataset are extracted to compute the transformation matrix, then source subjects whose transformed subspaces are the closest to the target's subspace are chosen to learn the predictive model.

3.4.3 Predictive models/machine learning algorithms

After the construction of a common space and/or feature designing, most ISPA studies pool multiple datasets together and feed data into a classification or regression model.

Instead of pooling multiple datasets and training one predictive model from the training set, some studies learn several models from the training set and make the final decision from ensemble of models. [69] proposed to build an ensemble of classifiers in order to design a multi-class decoder, where each sub-classifier focuses on an optimally se-

Table 3.3: Approaches proposed for learning predictive models

references	**data type**	**model**
Kim et al. (2017) [69]	fMRI	ensemble of SVM classifiers and voting
Fuchigami et al. (2018) [45]	fMRI	ensemble of SVM classifiers and voting
Li et al. (2018)[81]	EEG	ensemble of linear regression models
Zheng et al.(2016)[152]	EEG	regression model generating parameters of classifier
Fatima el al.(2017)[43]	MEG, EEG	two-layer classification model
Olivetti et al. (2014)[105]	MEG	staked generalization
Marquand et al.(2014)[89]	fMRI	coupling between subjects, multi-task learning
Van gerven et al.(2009)[132]	EEG	l_1/l_p regularization, multi-task learning
Kia et al.(2017)[68]	MEG	$l_{2,1}$ regularization, multi-task learning

lected set of local features, the final decision is classically taken using voting. Similarly, in [45] SVM classifiers were trained from individual subject's dataset in the training set, then the category of each test data was predicted based on a majority of votes. [81] also learned one classifier per subject from the training set, but the prediction of each test data was made by weighted predictions of all classifiers. [152] introduced a regression model to learn the mapping between individual datasets and parameters of subject-specific SVM classifiers. The regression model was later used to generate parameters for target classifier. Furthermore, hierarchical classification model is used to make predictions for test data. For instance, two-layer hierarchical classification model is constructed in [43], the probability estimates of several non-linear SVM classifiers are fed into a random forest classifier which makes the final prediction.

Additionally, very few studies take into account the structure of data and directly exploit the multi-source nature of the training data in the learning phase itself. The first one [105] has been proposed for MEG data. It puts together a subject-to-subject transfer scheme that uses a classical domain adaptation scheme based on instance-weighting, with an ensemble-based stacking method to combine the information provided by the different training subjects. The other ones recast the multi-source ISPA problem as a multi-task question where each subject defines a task. In [89], the coupling between subjects is induced through a Gaussian process prior and the learning is achieved in a bayesian setting, while [68] uses a $\ell_{2,1}$ regularization that allows focusing on sparse features that are commonly informative for all tasks / subjects. Similarly, the l_1/l_p group-wise regularization scheme proposed in [132] can also be viewed as a multi-task implementation.

3.4.4 Performance evaluation

After a model is learned from the training set, in order to examine how well the model generalizes to unseen data, the performance of the model is evaluated on the test set.

For 517 identified ISPA papers, there are no specific evaluation metric for ISPA, all of them use same metrics as the ones used in within-subject pattern analysis. Therefore in this section we quickly summarize those metrics exploited in ISPA studies, for each of which we provide a few examples.

The most common metric for evaluating a classification model is an accuracy metric or an error rate, which is the ratio of the number of predictions correctly or wrongly classified to the total of test data, e.g. the error rate is measured to evaluate the performance of SVM in [99]. In binary classification, sensitivity and specificity (recall) allow to further quantify the proportion of predictions correctly identified in one specific class, which are also used in [99]. Similarly, as the ratio of predictions correctly classified, precision and recall are used to compute f1 value, which allows to evaluate the performances of models in ISPA [26], especially when the number of observations in each class is uneven. For imbalanced classes, there exist other metrics which are exploited in ISPA studies, such as Cohen's kappa adopted in [78]. Instead of evaluating model performance with a ratio, receiver operating characteristic (ROC) curve is a graphical alternative, and area under the ROC curve (AUC) provides a visual representation of the relative trade-offs between the benefits and costs of a classification model, which is used in [113]. Metrics above are commonly exploited for binary classification. Furthermore, for multi-class classification, rank accuracy, which is a list of potential classes rank-ordered from most to least likely, is adopted for performance evaluation in [96], [120], [67], [30].

For a regression model, in general, its performance is usually evaluated by a metric which exploits the difference between true values and predictions, such as Root Mean Square Error [81], R-square (R^2) [92] (also known as ratio of explained variance [93], [91]) and Correlation coefficient [137]. However, Correlation coefficient is suggested to be avoided using as it is susceptible to bias [102].

3.4.5 Statistical assessment

The null hypothesis in ISPA is that there is no consistent information that links brain patterns and stimuli in the population level. Therefore under this null hypothesis the classification accuracy across all test subjects should be at the chance level. Significance of the observed accuracy is assessed by computing the probability of observing this result

under the null hypothesis. If the probability, known as p-value, is below some specific threshold, e.g. p-value < 0.05, the null hypothesis is rejected. The rejection means that ISPA obtains an accuracy significantly different from chance level, which further indicates there exists a link between brain patterns and stimuli in the population level.

However, we did not find statistical models which are specifically designed for assessment of ISPA results. All ISPA studies we have identified exploit same statistical models as the ones adopted in within-subject pattern analysis. Therefore we summarize these statistical models below and list some ISPA studies which use these models.

One popular statistical model is binomial distribution, which gives the probability of k correct predictions out of N test data by modeling each test data as an independent trial [119], [101]. Another widely used model is one sample t test, which exploits accuracies computed from all test sets to assess the null hypothesis that average classification accuracy of all test subjects is at the chance level [1]. An alternative is non-parametric permutation test, which assumes there is no class information in the data so that the labels can be permuted multiple times to obtain the null empirical distribution [96], [90], [139]. Because binomial distribution and t test have several limitations, e.g. t-test is designed for data which are normally distributed, permutation test is recommended despite of its high computational cost.

Even though there are no specific statistical models designed for ISPA, in the literature we found several studies which proposed models for better assessing significance of G-MVPA results. Compared to t test which does not take into account the size of individual dataset, [106] introduced a bayesian hypothesis testing based on a Beta-Binomial model, which computes bayes factor with the size of test set and uses the value of bayes factor to accept or reject the null hypothesis. In order to overcome the limitations in binomial distribution and t test, Brodersen et al. proposed a Beta-binomial model and a normal-binomial model with underlying Markov chain Monte Carlo (MCMC) sampling algorithms for inferring on performance measures of imbalanced dataset [16]. In their follow-up work, they proposed a variational bayes algorithm to reduce the computational complexity of MCMC, which works for both balanced and imbalanced datasets [15]. Additionally, [122] proposed a permutation test scheme, which first permutes labels to compute multiple permuted accuracies in within-subject decoding, then forms the null empirical distribution of average accuracy over all subjects by bootstrapping and averaging individual permuted accuracies.

3.5 Discussion

In this chapter, we have presented a formalization that defines inter-subject pattern analysis (ISPA) as a tool suited for group-level multivariate analysis. The multi-source transductive transfer setting that we have introduced for ISPA allowed identifying several sub-problems that each has been tackled in previous work. Problem 1: how to deal with the multi-subject nature of the data in ISPA; Problem 2 : as a transductive learning question, how to exploit the availability of unlabeled test data during the training phase ; Problem 3 : how to transfer information from the training set to new individuals in the test set. Below, we raise questions on those sub-problems and provide some guidelines which could drive future ISPA research from a methodological perspective.

3.5.1 Handling inter-subject variability in the training set

In the first step of ISPA, the construction of a common space reduces the disparity between individual input spaces with anatomy-based or function-based transformations. In the second step, feature construction /selection aims to learn subject-invariant feature space. Most studies handle inter-subject variability through these two steps, then the follow-up schemes used to assess the generalization power of the models in the common representational space are usually very basic (data pooling followed by a simple classifier). This implicitly assumes that the construction of a common space and feature designing transform all datasets in the training set into one same distribution. However, this assumption has not been tested, therefore in Chapter 4 we perform extra experiments to test whether the inter-subject variability is removed by these two steps. Compared to strategies aiming to reduce distribution shifts in the training set, in the third step of ISPA, few studies directly learn machine learning models from multiple distributions, with the notable exceptions of the multi-task recasting of the multi-source problem [132], [89], [68].

Strategies in these three steps handle inter-subject difference from different perspectives. However, in the literature, these strategies have never been integrated in one method. We believe that it could be valuable to merge advanced methods proposed in these three steps in one study. First, one could exploit the most advanced methods developed for common space construction. Since the between- subject differences are reduced, handling inter-subject variability through feature designing and learning directly from multiple distributions would face an easier challenge and should hopefully reveal more effective. We will perform further study in Chapter 4.

3.5.2 Explaining the transductive nature of ISPA

Having access to unlabeled test data during the training phase allows ISPA to exploit the information contained in test set. Test set available beforehand implies that i) inductive methods which exploit only the training set could be adopted for performing ISPA, e.g. methods proposed for BCI ii) methods which exploit information from both training and test set are also applicable. Inductive methods demand to learn a classifier from the labeled training data, for which training data have to be transformed into forms acceptable by the classifier, e.g. high-dimensional vectors. While in ISPA, all data in the test set are available before the training starts, it allows to directly generate values for test data without training a classifier, which means data are able to be transformed into any forms as long as they facilitate making predictions for test data, e.g. in [111] individual dataset is represented as a similarity matrix, one per subject. Additionally, subject-specific parameters can be learned subject by subject for both training and test set, which enables to construct subject-invariant feature patterns by removing subject-specific characteristics, e.g. multi-subject feature rescaling which will be further discussed in Chapter 4, or by preserving idiosyncratic structural properties at the same time [124]. We believe that in the future there will be other forms of feature patterns favoring the transductive nature of ISPA.

3.5.3 Transferring from the training to test set

Because of the multi-subject nature of the data in ISPA, there exists difference between training and test set, which hinders a machine learning model from generalizing to unseen subjects in the test set. Transferring between the training and test set in ISPA is usually achieved by common space construction or transfer learning strategies. In several cases a common space is built with training datasets and then applied to test dataset. Similarly, in transfer learning methods, especially in domain adaptation methods, the training set which consists of multiple datasets is usually considered as one domain, which ignores the difference between training subjects. Therefore, we expect domain adaptation methods which take each subject as one domain and perform transfer learning across multiple domains.

In addition, transfer learning in ISPA is usually performed after the construction of a common space. However, given the fact that the original input space, i.e. the brains of the different subjects are not identical, we wish to directly transfer knowledge across subjects in their original spaces.

3.5.4 The gap between methods and applications in ISPA

Multiple methods are exploited in the literature for performing ISPA, which we have already reviewed in this chapter. However, among 517 ISPA papers we have identified, most of them use classical machine learning methods. Even though multiple methods have been proposed by computer scientists or applied mathematicians to address the limitations of classical methods, only very few of them are used by neuroscientists. For instance, hyperalignment is a highly cited functional alignment method aimed at reducing inter-subject variability, however, it has barely been used in neuroscience papers. Therefore more efforts are required to make these advanced methods accepted and used by the community, e.g. through user-friendly software packages.

3.5.5 Statistical models for ISPA

ISPA papers identified in the literature exploit same statistical models as the ones adopted in within-subject pattern analysis. Even though those models are widely applied in the literature, they have limitations that hinder ISPA from obtaining accurate significance. For instance, exploiting binomial distribution takes assumption that all observations are independent, and t test should be applied on normal distributed data. However, in ISPA the training sets of two cross validation splits have overlapping data, therefore ISPA generates results neither following normal distribution nor independent. Permutation test offers a solution for assessing significance of ISPA results, which is the only solution to control the false positive around 5% in experiments in Chapter 2, but every time labels are shuffled comes to re-doing a full analysis on all the subjects, it takes high computational cost to obtain empirical null distribution. Therefore, we wish to assess the significance of ISPA results with strategies that take into account the nature of ISPA results and the computational cost.

3.6 Conclusion

In this chapter, we have introduced a complete formalization of Inter-Subject Pattern Analysis (ISPA) as a multi-source transductive transfer learning problem and proposed ISPA as a tool of choice for multivariate group analysis. This setting offers a unifying perspective on the existing methods available to perform ISPA. We believe that this formalization should facilitate shaping new methodological developments and we hope that this chapter will draw further attention on the possibilities offered by ISPA for

studying neural representation using functional neuroimaging techniques.

Chapter 4

On the well-foundedness of the multi-source transductive transfer formalization of ISPA

Publication associated with this chapter: Q. Wang, T. Artières, and S. Takerkart, 'Inter-subject pattern analysis for task-based functional neuroimaging data experiments. A unifying formalization', Computer Methods and Programs in Biomedicine, submitted

4.1 Introduction

In Chapter 3 we defined a multi-source transductive transfer formalization for ISPA. Multiple ISPA studies were reviewed in Chapter 3, some take into account the multi-subject nature of the data and exploit strategies to handle inter-subject variability, while others do not. After reviewing ISPA studies, we raised several questions corresponding to three problems listed in section 3.5. Question 1: Is multi-source transductive setting better than classical setting for performing ISPA? Question 2: After performing common space construction and feature designing, is there still distribution shifts between subjects? Question 3: Pooling multiple datasets together increases both the size and the amount of heterogeneity of the training set, how do these two factors affect the generalization performance of machine learning model? Question 4: The difference between training and test set impairs generalization power of machine learning models, may transfer learning methods improve model's generalization performance?

With respect to these questions, we perform four experimental studies.

In the first study we examine the multi-source transductive setting of ISPA by applying multi-subject feature rescaling subject by subject. The reason why we select multi-subject feature rescaling is that this strategy is easy to be implemented but allows to construct feature space from both the multi-source and transductive points of view. In this study ISPA is performed under the multi-source transductive setting as well as under the classical machine learning setting, then performances from these two settings are compared.

The second experimental study aims to examine the transfer setting of ISPA. Numerous ISPA studies perform common space construction or/and feature designing for handling inter-subject variability, then multiple datasets are pooled to train a machine learning model, which ignores between-subject differences and implicitly assumes all datasets have been transformed into the same distribution. In order to examine whether there still exists difference between subjects after performing common space construction and multi-subject feature designing, we compare performances of a classifier generalizing to two test sets, one consists of unseen data drawn from subjects in the training set, the other is composed of data from unseen subjects.

In the first two studies, multiple datasets are aggregated to form the training set. The aggregation of datasets increases the size of the training set, which is beneficial to learn a machine learning algorithm, as well as the amount of heterogeneity, which increases the difficulty in learning the algorithm. Therefore study 3 examines how these two factors impact the performances of a classifier.

In the fourth study, we propose methods to exploit the multi-source transductive transfer setting of ISPA, by adopting two machine learning techniques. The first strategy is subspace alignment which is a transfer learning strategy, the other strategy, stacked generalization, focuses on building an invariant feature space across the training and test set. Both strategies allow to reduce the differences between training and test set. They are applied to data rescaled by multi-subject feature rescaling, then their generalization performances are compared.

4.2 Experimental setting

4.2.1 Experimental data

We use functional neuroimaging data from two previously published studies. The first one includes 100 participants from the fMRI data used in [32] [1] to identify voice sensitive areas in the temporal cortex, using a *voice localizer* paradigm [10]. The participants passively listened to 40 blocks of auditory stimuli (20 vocal blocks and 20 non-vocal blocks). The responses to each of these blocks were estimated using a GLM in SPM12 (beta maps), and projected to the cortical surface using tools from freesurfer. An anatomical region of interest (ROI) was selected to englobe the auditory cortex as well as voice sensitive regions, separately in both hemispheres, using the Desikan parcellation provided in freesurfer. The task was to decode voice vs. non-voice blocks from the brain responses recorded in these regions. This yielded two sets of data, hereafter denominated "fMRI_1" (for the left hemisphere) and "fMRI_2" (for the right hemisphere). The second study is the event-related MEG experiment from [31] [2], where 16 participants viewed one picture per trial, either of a face or of a scrambled face. On average, 580 trials were available per subject. In order to construct our input feature vector, we first selected the $500ms$ window after stimulus onset and we temporally downsampled the data by a factor eight. The task of the decoder was to guess whether the subject was viewing a real face or a scrambled one from the MEG data. This provided us with a third data set, hereafter simply denominated "MEG" data. Overall, these data sets present complementary characteristics – their nature (fMRI vs. MEG), the studied domain (within ROIs vs. on the full brain), the nature of the paradigm (block vs. event-related), the number of subjects available (100 vs. 16), the number of samples available for each subject (40 vs. around 580) – which should help us draw conclusions that are widely applicable.

4.2.2 General experimental setting

All experimental studies consist in four steps of ISPA, i.e. common space construction, feature construction, learning a machine learning model and performance evaluation. For fMRI data, a common space is constructed by surface-based anatomical alignment using freesurfer. Because MEG data were acquired with the same number of sensors across subjects, we assume that the measurements provided by these MEG sensors correspond

[1] data available at http://https//openfmri.org/dataset/ds000158
[2] data available at https://www.kaggle.com/c/decoding-the-human-brain

to each other across subjects, therefore no common space construction is performed for MEG data. In all studies, we use cross-validation schemes where several subjects are left out for the test set in each data split – the exact scheme being detailed hereafter in each case. For feature designing, we adopt feature rescaling strategies to construct new representations for both training and test data, then representations in the training set are fed into a classifier. We choose to use two families of classifiers, logistic regression and linear support vector machine (SVM), because they are largely employed in neuroimaging and are known to perform well in such context. Model selection is performed on the training set, i.e the values of the hyper-parameters of the models are chosen through an inner cross-validation limited to the training data. For both types of models, SVM and logistic regression, the regularization weight C is chosen within *{10, 1, 0.1, 0.01, 0.001, 0.0001, 0.00001}*. For logistic regression, we also select the type of regularization (l_1 or l_2). The generalization performance of the classifier is assessed with the average classification accuracy across all splits, which offers a robust and unbiased estimate because the number of splits offered through such leave-P-subjects-out schemes is very large [135].

4.3 Study 1: multi-source transductive setting vs. classical machine learning setting

4.3.1 Experimental setting

In order to evaluate the relevance of the multi-source transductive setting for ISPA, we here benchmark it versus the classical machine learning setting by quantifying the effect of feature rescaling.

Feature standardization, a feature rescaling strategy, is a data transformation technique which consists in removing the mean across examples of each feature and scaling it to unit variance. Standardization is commonly used in machine learning before feeding the data to the learning algorithm. In a classical machine learning setting, the parameters of the standardization – the means and variances of each feature – are computed on the training data then applied both on the training and test set. The multi-source transductive setting makes it possible to modify this process in two ways. First, knowing the multi-source structure of the training set, it is valid to standardize the training data separately for each subject, i.e source by source, adaptively to the distribution of the data of each subject. Secondly, the transductive nature of the problem allows performing this

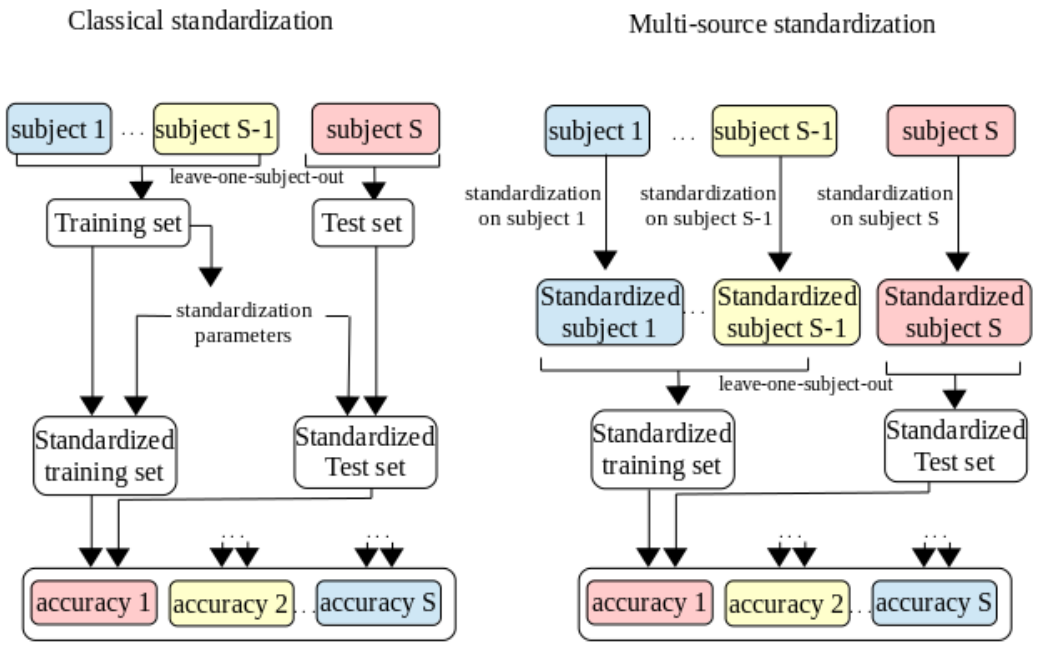


Figure 4.1: Illustration of the two standardization strategies. Left: classical standardization, standardization parameters are learned from the training set and applied both on the training and test set. Right: multi-source standardization, standardization parameters are learned subject by subject from individuals in the training and test set.

standardization separately on the samples from each test subject. We denote as multi-source standardization the procedure that uses these two possibilities, and we here aim at comparing its influence on the generalization performances of the model with the one of classical standardization. The flowchart of these two standardization strategies are illustrated in Figure 4.1.

For the sake of exhaustivity, we will also include the case where no standardization is used. With the fMRI data, we randomly defined 50 splits of the 100 subjects, each time including ten subjects for the test set and the other 90 for training. For the MEG data, we used a leave-two-subjects-out cross-validation scheme, which, for the 16 available subjects, yielded 120 train-test splits, which we all included in our analyses.

4.3.2 Results

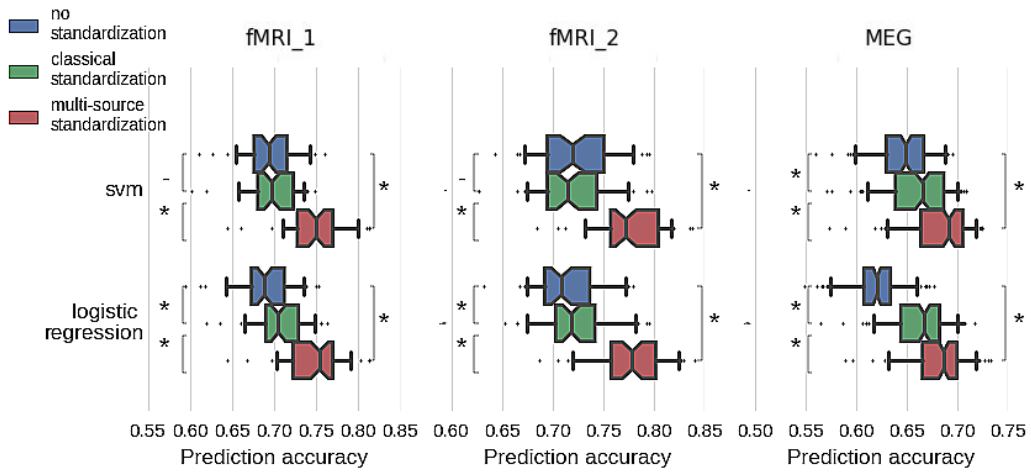


Figure 4.2: Results of Study 1: ISPA vs. classical machine learning. The multi-source feature standardization (red) enabled by our ISPA formalization yields models with higher generalization performances than with classical standardization (green) or no standardization (blue) in all cases (for the three data sets and the two classifier families). * denotes a highly significant difference ($p < 10^{-4}$), while - designates comparison for which no significant difference was found ($p > 0.05$).

The results of this first experiment are summarized on Figure 4.2. Unlike statistical methods we have reviewed in section 3.4.5, of which the null hypothesis is that the classification accuracies across all cross validation splits is at the chance level, the statistical assessment performed in this section assesses whether there is difference between generalization performances obtained from different standardization strategies. We used paired t-tests to compare the mean generalization performances over all splits between different strategies. First, when comparing classical vs. no standardization,

we show that the classical standardization allows improving model accuracy in several cases ($t > 4.31, p < 10^{-4}$ for all cases with the logistic regression, and for the MEG data with SVM), but not all (fMRI_1 and fMRI_2 with SVM, $t = 1.82, p = 0.07$ and $t = 1.29, p = 0.20$ respectively). This might point to some of the limitations of the classical machine learning setting for our ISPA problem. Then, in all cases (fMRI_1, fMRI_2 and MEG, both for SVM and logistic regression), the model accuracy is significantly higher with multi-source standardization than with the classical one ($t > 11.28$ in all cases, i.e $p < 10^{-16}$).

While in a classical machine learning setting, feature standardization insures that the objective function to be minimized during learning is not dominated by features that would have a large mean or variance, multi-source standardization has an additional interest for ISPA: because it is performed separately on each subject, it attenuates the distribution shifts that exist between subjects. This yield both a more homogeneous training set – easing the learning task itself, and a test set that is closer to the training data – thus facilitating the generalization of the model, as evidenced in this experiment. Such multi-source standardization (or other types of feature normalization performed on a subject-by-subject basis) has often been used in the past (see e.g [96]; [121]; [25]; [67]; [17], we also believe it is being used in other studies without being explicitly reported) without having been methodologically motivated. In fact, *multi-source transductive transfer* offers the first explicit justification for this common operation, and this first study demonstrates its importance in practice.

4.4 Study 2: Generalizing to new data vs. new individuals

4.4.1 Experimental setting

As seen in Study 1, after performing feature standardization multiple datasets in the training set are pooled together, which ignores the distribution shifts that might exist between subjects and assumes that all examples are drawn from the same distribution. We here want to quantitatively assess whether there still exist distribution shifts across subjects, after the attenuation offered by multi-source feature standardization. For this, our reasoning is the following: if the distribution shifts were negligible after performing multi-source standardization, the data from all subjects could be modeled as being drawn from a single distribution, and a decoder would perform identically on any new data, whether from new subjects or from training subjects.

We therefore probe this directly by performing experiments where a single model is evaluated on two sets of new data, taken either in new subjects or from independent samples of the training subjects, after multi-source standardization. For this, we first extract two disjoint subsets of subjects $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively of sizes S^1 and S^2, from the full set of available subjects $\mathcal{S}$. For each subject s in $\mathcal{S}_1$, we split the data into $\mathcal{T}_s^{train}$ and $\mathcal{T}_s^{test}$ with a ratio r of the data of this subject going into $\mathcal{T}_s^{test}$. The data set $\bigcup_{s\in\mathcal{S}_1} \mathcal{T}_s^{train}$ defines the training set $\mathcal{D}_{train}$ for this experiment and we will use $\bigcup_{s\in\mathcal{S}_1} \mathcal{T}_s^{test}$ as a first test set $\mathcal{D}_{test}^1$, which thus contains samples not included in the training set, but recorded in training subjects. All the samples of the subjects in $\mathcal{S}_2$ constitute a second test set $\mathcal{D}_{test}^2$. We can then achieve our goal by fitting a model on the training set and comparing its generalization performances on the two different test sets. In order to quantify a potential difference, we repeat this operation with numerous splits of the data. We also study the effect of the size of the training set by changing the number of subjects in $\mathcal{S}_1$.

In practice, because the number of subjects available in our fMRI and MEG data is very different, we used two distinct implementations of this paradigm. For the fMRI data, because the data set includes a large number of subjects (100) we first define $\mathcal{S}_2$ by choosing $S^2 = 50$ subjects. We then define $\mathcal{S}_1$ within the subjects not in $\mathcal{S}_2$, by choosing 30 random sets of S^1 subjects, which each provide an estimate of the model performance. We repeat this with several sizes for $\mathcal{S}_1$, with $S^1 \in \{10, 20, 30, 40\}$. The MEG data includes 16 subjects, which strongly restricts both the number of subjects that can be included in the training set and the combinatorial possibilities. We choose to work with a set of $S^2 = 2$ subjects in $\mathcal{S}_2$, which we will repeatedly draw randomly within the 16 subjects. Then, for each of these draws, we define $\mathcal{S}_1$ within the leftover subjects, with sizes $S^1 \in \{5, 7, 9, 11, 13\}$. For each value of S^1, we therefore obtain 50 measurements of the model accuracy. For classifier, because logistic regression produces similar performances as SVM but with less computational cost, in Study 2-4 only logistic regression is implemented.

4.4.2 Results

The results are shown on Figure 4.3. We used an analysis of variance (ANOVA) with two factors – the nature of the test set ($\mathcal{D}_{test}^1$ or $\mathcal{D}_{test}^2$) and the number of subjects in the training set – in order to study these results. The main effect of the number of subjects was found to be significant in all three cases ($F > 20.96, p < 10^{-11}$). This shows that the predictive power of the model improves when data from more subjects are added

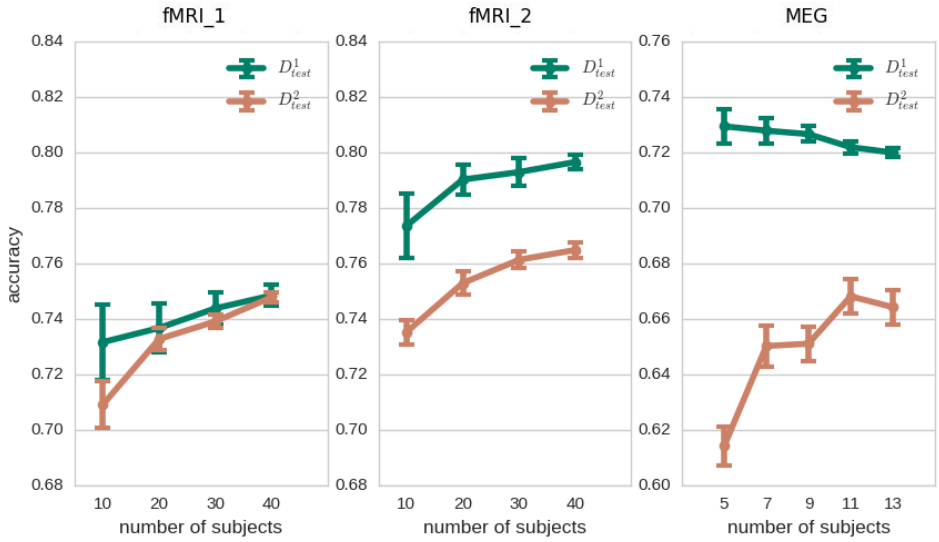


Figure 4.3: Results of Study 2: Evaluating generalization performances to new data and new individuals. The accuracy of a given model is compared on two test sets composed of new data from the training subjects ($\mathcal{D}^1_{test}$, green curves) or data from new subjects ($\mathcal{D}^2_{test}$, orange curves). Two effects are clearly visible: 1. increasing the sample size by adding subjects allows improving the generalization to new subjects' data (orange curves); 2. it is more difficult to generalize to data from new subjects. The latter demonstrates that we cannot ignore the distribution shifts between subjects, i.e. that strategies reducing difference between the training and test set are required.

to the training set. Most importantly, the main effect of the nature of the test set was also significant in all cases ($F = 9.62, p < 10^{-2}$ for fMRI_1, $F = 295.02, p < 10^{-42}$ for fMRI_2 and $F = 1823.35, p < 10^{-166}$ for MEG). This demonstrates that it is indeed more difficult for the model to generalize to data from new subjects than to new data from the training subjects, i.e that the distribution shifts between subjects, especially between training and test set cannot be ignored, even after multi-source standardization. This clearly indicates that pooling multiple datasets in the training set which ignores differences between subjects has limitations. It motivates further methodological work to overcome the limitations of the pooling strategy and improve the transfer to new subjects, which corresponds to the transfer setting of our formalization for ISPA.

Beyond these very significant main effects, other more complex ones appear, as e.g the decrease of the accuracy on $\mathcal{D}^1_{test}$ for the MEG data (green curve). These can be explained by the respective levels of heterogeneity of the data, both within the training set and between the training and test set.

4.5 Study 3: Number of subjects vs. sample size in the training set

4.5.1 Experimental setting

In machine learning, it is well known that the predictive power of a model for a given task increases with the sample size available for training, until reaching a plateau. Neuroscientists using functional neuroimaging are in the small sample size regime – i.e often working with only a few hundreds of samples – meaning that this plateau is not reached (see the orange curves of Study 2 on Figure 4.3). Adding more subjects in the training set has several effects. First, it enhances the sample size, which should help approaching this plateau. Secondly, because of inter-individual differences, it also increases the heterogeneity of the training set, which could have positive effects (exploiting the benefits of added *diversity*) or negative ones (if the *variability* is overwhelming, thus making learning more difficult). We here attempt to decipher these effects by constructing ISPA experiments where the number of subjects and of samples of the training set are controlled separately.

For this, we kept the general principle of the paradigm defined in Study 2 to define the training set $\mathcal{D}_{train}$ as $\bigcup_{s \in \mathcal{S}_1} \mathcal{T}_s^{train}$. But we slightly modified it to obtain training sets with a constant size N when the number of subjects S^1 changed, which simply

implied changing the ratio r to maintain $N = S^1 \times r$ constant. We repeated this with several values of N. In all cases, we used the test set $\mathcal{D}^2_{test}$ defined identically to what it was in Study 2 to evaluate the models on data from unseen subjects. In practice, the choices available for N and S^1 were limited, either by the number of samples available in each subject and/or by the number of subjects. For the fMRI data, we used $N \in \{360, 720, 1080\}$ with a corresponding minimum number of subjects of $10, 20, 30, 40$, and a maximum number of subjects of 45 or 48 depending on the case. For the MEG data, we used $N \in \{1300, 2600, 3900\}$, with a number of subjects varying between 5 and 13 for the first two values, and 9 and 13 for the last one.

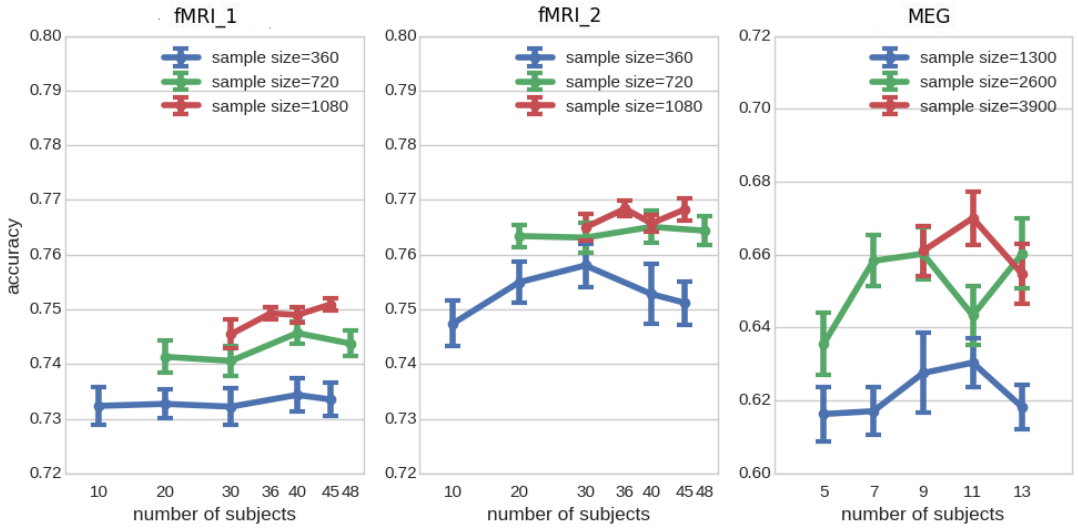


Figure 4.4: Results of Study 3: Number of subjects vs. sample size in the training set. Decoding accuracy shown is obtained when machine learning model generalizes to new subjects. Each colored line corresponds to a given size of the training set, along which the number of subjects and the number of samples per subject vary. The influence of the number of subjects (horizontal axis) depends on the case. On the contrary, for all cases, the generalization power of the model significantly increases with the total sample size.

4.5.2 Results

The results, presented on Figure 4.4, were analyzed using a two-way ANOVA for which the two factors are the number of subjects and the total size of the training set. The main effect of the size of the training set was highly significant in all cases ($F = 113.02, p < 10^{-19}$ in fMRI_1; $F = 51, p < 10^{-19}$ in fMRI_2; $F = 95.41, p < 10^{-19}$ for the MEG data). This demonstrates the expected effect that for a given number of subjects, increasing the size of the training set allows improving the generalization power of the

model. Furthermore, the main effect of the number of subjects was also significant for all three cases ($F = 2.71, p < 0.05$ in fMRI_1; $F = 2.59, p < 0.05$ in fMRI_2; $F = 5.15, p < 10^{-3}$ for MEG), but the influence of the number of subjects appears to be complex, and at least not monotonous. For instance, the blue curves for the fMRI_2 and MEG data sets show a first phase of performance increase until reaching a peak – which could be the beneficial effect of the added diversity – and then a decrease when adding more subjects – which might be caused by the overwhelming variability in the training data.

4.6 Study 4: Performing ISPA with multi-source transductive transfer strategies

In Study 1 and 2, we have demonstrated the well-foundedness of our formalization of ISPA. In this study we aim to perform ISPA with two machine learning methods that exploit this formalization.

4.6.1 Methods

In the first three experimental studies we have shown how multi-source transductive setting influences the generalization ability and the deficiency of the lack of a transfer strategy. How to handle inter-individual differences so that learning algorithms generalize well to new individuals is a challenge. Furthermore, in our experiments each dataset consists of high-dimensional data, therefore the machine learning methods that we choose for this study perform dimensionality reduction and fulfill the transfer setting of ISPA at the same time. We select two machine learning strategies, one is subspace alignment which is a transfer learning strategy [44], the other is stacked generalization that is a hierarchical feature construction strategy [142]. Both of them create new representations for data in the training and test set which are considered to have less distribution shifts than original ones.

Subspace alignment

Subspace alignment identifies a transferable subspace which is a lower-dimensional space than the original one. It first implements PCA to extract the first d eigenvectors from both training and test sets, then the source subspace is aligned to the target subspace

using a transformation matrix that is learned from these eigenvectors. After the transformation, data in the training and test sets are replaced by d-dimensional representations, then machine learning algorithm is learned and evaluated in this d-dimensional feature space. In our experiment, subspace alignment is implemented on data that have already been processed by multi-source standardization. In this case, we exploit the multi-source nature of the data and the transductive nature of ISPA by performing multi-source standardization, and address the transfer question by using subspace alignment. Therefore our multi-source transductive transfer setting is fully exploited.

Stacked generalization

Stacked generalization is an ensemble method which combines multiple low-level models to improve the predictive power of a high-level model [142]. Typically, the outputs of the low-level models form a new feature space which is common across the high-level training and test data. Stacked generalization consists in two steps:

- Level-0 step learns low-level models from original data to create new representations for both training and test data. In level-0 step, the original data, denoted as level-0 data, are split into the training and test set, multiple level-0 models are learned from the training set and make predictions for both the training and test data. Outputs from multiple level-0 model are stacked to form a new representation for each sample in the training and test set, those representations are denoted as level-1 data.
- Level-1 step a learns high-level model from these new representations. Level-1 model is learned from level-1 training set and then generalizes to level-1 test set.

In our experiment, datasets are first processed by multi-source standardization, which exploits the multi-source nature of the data. Then stacked generalization is applied on standardized data by training level-0 classifiers from several subjects and making predictions for data in both training and test sets. Inter-subject variability in level-1 dataset is reduced not only in the training set but also between the training and test sets, which addresses the transductive transfer question in ISPA. Therefore, the multi-source transductive transfer setting is exploited. Furthermore, the dimensionality of level-1 data is decided by the number of level-0 classifiers, compared to the high-dimensional level-0 data, level-1 data is in a lower dimensional space.

In the literature, in order to ensure diversity of the training set Olivetti et al. proposed to create each level-0 model on the data of one subject in the training set [105],

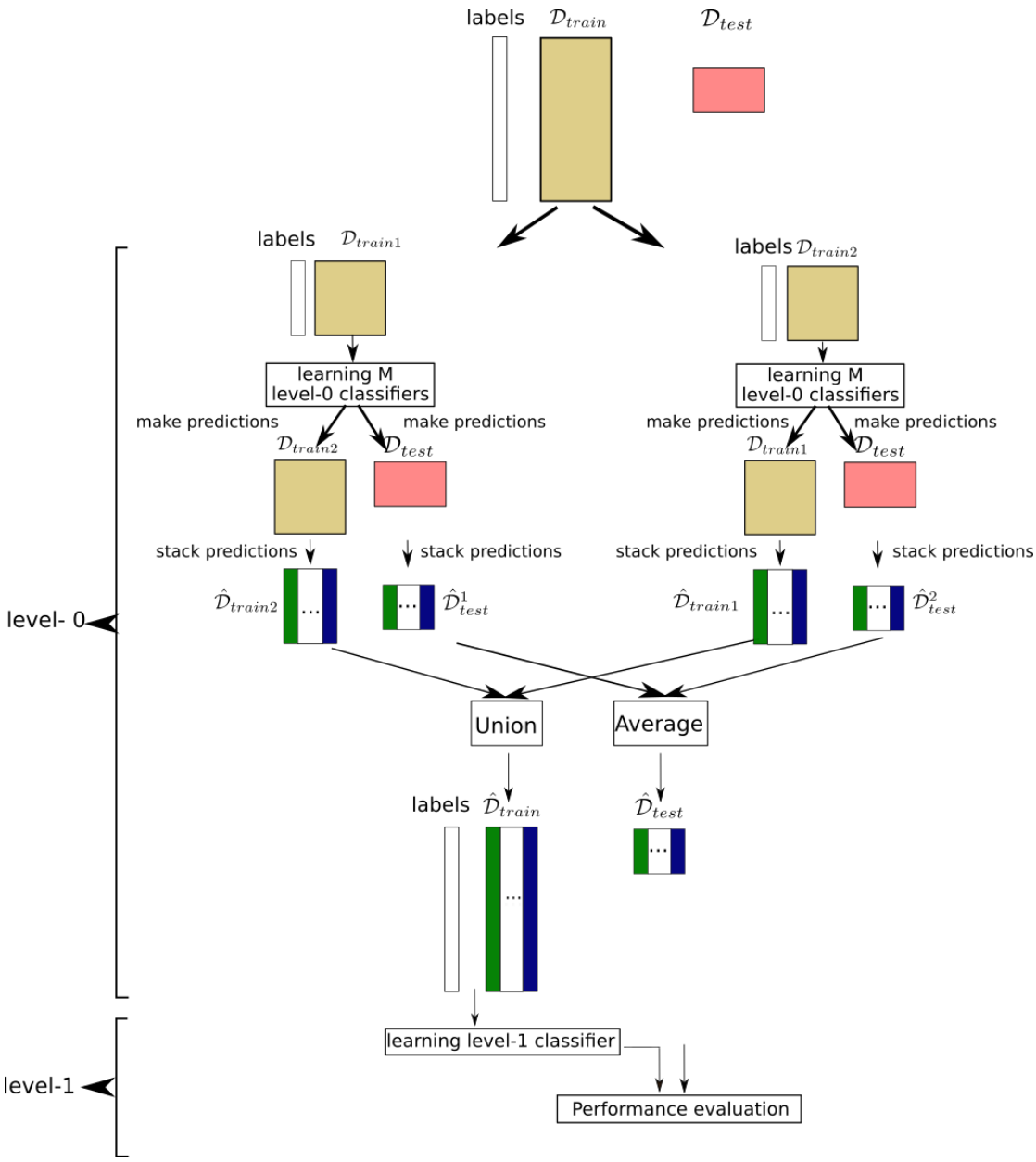


Figure 4.5: Illustration of one stacked generalization framework. Two steps are included in stacked generalization. In level-0 step, the training set (yellow boxes) $\mathcal{D}_{train}$ is split into two subsets $\mathcal{D}_{train1}$ and $\mathcal{D}_{train2}$, the test set (red boxes) is $\mathcal{D}_{test}$. A set of M level-0 classifiers are learned from one subset and used to make predictions for data in another subset and in test set. Predictions generated from level-0 classifiers are stacked to form level-1 representations (green boxes are predictions generated from the first level-0 classifier, and blue boxes are predictions generated from the M-th level-0 classifier). In level-1 step, level-1 classifier is learned from level-1 data in the training set and generalizes to level-1 data in the test set.

which results in only dozens of features in the level-1 representation. In our experiment, we want to probe how the amount of diversity in the training set and the number of features in the level-1 data influence the final generalization ability of machine learning model. Therefore, we use data drawn from different number of subjects to training level-0 models, and level-1 data are constructed by outputs of different number of level-0 models.

We split $\mathcal{S}$ available subjects into two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ of sizes S^1 and S^2, respectively. Data recorded from subjects in $\mathcal{S}_1$ constitute level-0 training set $\mathcal{D}_{train}$ while data from $\mathcal{S}_2$ form level-0 test set $\mathcal{D}_{test}$. In level-0 step, multiple level-0 classifiers are learned from $\mathcal{D}_{train}$ then make predictions for data in both $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$. Since level-0 classifiers trained from some training data can not be used to make predictions for the same data, we split $\mathcal{D}_{train}$ into two subsets of the same size– $\mathcal{D}_{train1}$ and $\mathcal{D}_{train2}$–by splitting each individual subject's dataset into two parts $\mathcal{T}_s^{train1}$ and $\mathcal{T}_s^{train2}$, where $s \in \mathcal{S}_1$. Therefore, $\mathcal{D}_{train1} = \bigcup_{s\in\mathcal{S}_1} \mathcal{T}_s^{train1}$, $\mathcal{D}_{train2} = \bigcup_{s\in\mathcal{S}_1} \mathcal{T}_s^{train2}$. In order to create level-1 representations for $\mathcal{D}_{train2}$, we choose M random sets of K subjects from $\mathcal{D}_{train1}$. For each set of K subjects, data are pooled together to learn a level-0 classifier, which is used to generate predictive labels for samples in $\mathcal{D}_{train2}$. Therefore every sample in $\mathcal{D}_{train2}$ obtains M predictions, which are stacked to form the level-1 representation–a M-dimensional vector–for the given sample. To be clear, we denote level-1 dataset of $\mathcal{D}_{train2}$ as $\hat{\mathcal{D}}_{train2}$. With the same procedure, $\hat{\mathcal{D}}_{train1}$ is constructed by another set of M level-0 classifiers which are learned from $\mathcal{D}_{train2}$. The union of $\hat{\mathcal{D}}_{train1}$ and $\hat{\mathcal{D}}_{train2}$ forms level-1 training set, which is denoted as $\hat{\mathcal{D}}_{train}$. Since two sets of M level-0 classifiers are learned from $\mathcal{D}_{train}$, each set constructs one level-1 test set, which is denoted as $\hat{\mathcal{D}}_{test}^1$ and $\hat{\mathcal{D}}_{test}^2$, respectively. We use the average of $\hat{\mathcal{D}}_{test}^1$ and $\hat{\mathcal{D}}_{test}^2$ to construct final level-1 test set, which is denoted as $\hat{\mathcal{D}}_{test}$. In level-1 step, level-1 classifier is trained on $\hat{\mathcal{D}}_{train}$ and then generalizes to $\hat{\mathcal{D}}_{test}$. The flowchart of stacked generalization is illustrated in Figure 4.5.

4.6.2 Experiment setting

We use the same cross validation splits as the ones adopted in Study 1, which is, 50 splits using leave-ten-subjects-out cross validation scheme for fMRI data and 120 splits using leave-two-subjects-out scheme for MEG data. The baseline performance is obtained by the same procedure of multi-source standardization in Study 1, which includes first rescaling individual subject's features with multi-source standardization, then implementing logistic regression on standardized data with predefined cross vali-

dation scheme. Subspace alignment and stacked generalization are applied on the same standardized datasets to examine how these two strategies impact the generalization performance of the learning algorithm.

For subspace alignment, the value of d is chosen with two methods. One is based on the strategy detailed in [44] where the value of d is selected automatically. The other is manually selecting the value within {100, 200, 300, 400, 500, 1000, 2000, 3000, 4000, max} where max represents the size of feature, i.e. 5400 for fMRI data and 4896 for MEG data. Then logistic regression is applied on the d-dimensional representations.

For stacked generalization, since fMRI data adopted leave-ten-subjects-out cross validation, $S^1 = 90$, $S^2 = 10$. The number of level-0 classifiers, M, is selected within {40, 80, 120, 160, 200} and each level-0 classifier is learned from data drawn from K subjects, where K is chosen from {20,40,60,80}. With the MEG data, we employed leave-two-subject-out cross validation, therefore $S^1 = 14$, $S^2 = 2$. M level-0 classifiers are learned and each classifier is trained from K subjects' level-0 data, where M is selected from {40, 80, 120, 160, 200} and K is chosen from {3,5,7,9}. Both level-0 and level-1 classifiers are implemented by logistic regression.

4.6.3 Results

We first present generalization performances of ISPA which are obtained from subspace alignment with different values of d, then we show results of stacked generalization. In order to examine how well these two strategies improve the generalization performance, we chose the parameters which produce the highest mean generalization performance over all splits for each strategy, then we compare each strategy's best performance to the baseline with paired t-test.

Performance obtained with subspace alignment

We examine the generalization performance, separately for each value of d. Results are displayed in Figure 4.6 which shows how generalization performance changes when the value of d varies. For all three datasets, the value of d chosen automatically is smaller than 100. The mean generalization performance increases as the value of d increases, until reaching a plateau. In fMRI_1 and fMRI_2, the plateau is reached when the value of d is equal to 400, while in MEG dataset, the value of d is 2000.

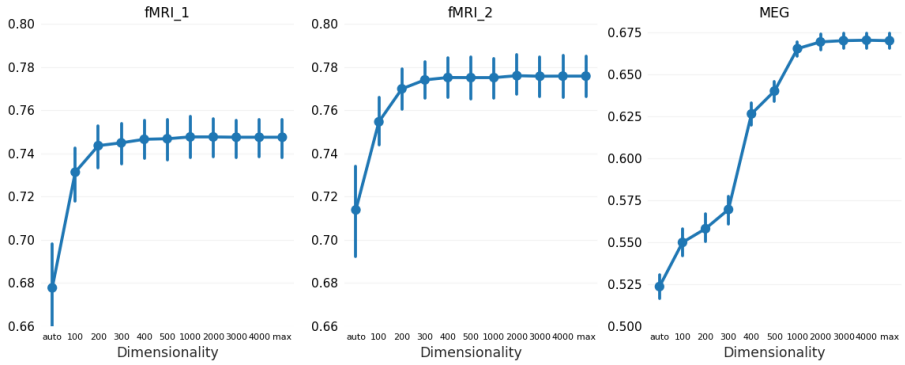


Figure 4.6: Generalization performance obtained on three datasets with subspace alignment. Subspace alignment extracts the first d eigenvectors from the training and test set. The value of d is automatically chosen (auto) or manually selected from 100 to the size of the features in each sample (max, 5400 features in fMRI data, 4896 features in MEG data). For each value of d, generalization performance of logistic regression is evaluated with 50 splits (fMRI data) or 120 splits (MEG). In general, generalization performance increases as the value of d increases, until reaching a plateau.

Performance obtained with stacked generalization

Figure 4.7 displays how mean generalization performance over all cross validation splits changes as the number of level-0 classifiers changes, each classifier is learned from data drawn from multiple subjects in the training set. In general, given the number of level-0 classifiers, as the number of subjects whose data are pooled to train level-0 classifiers increases, level-1 classifier's generalization performance first increases, then drops. Particularly, in fMRI_1 dataset, when multiple level-0 classifiers are trained from 20 or 40 subjects, best level-1 performances are obtained. In fMRI_2, situations where level-0 classifiers are learned from 40 subjects always have the highest average level-1 accuracy. While in MEG dataset, best performances are obtained when the number of subjects is five. In all three datasets, situations where level-0 classifiers are trained from the largest number of subjects have poor performance compared to others (80 subjects for fMRI, 9 subjects for MEG).

However, for each dataset the gap between every two curves in Figure 4.7 is small. Therefore we also analyzed level-1 performances obtained from all cross validation splits using a two-way ANOVA, for which the two factors are the number of subjects and the number of level-0 classifiers. Results showed that neither of these two main factors had significant effect for all three datasets.

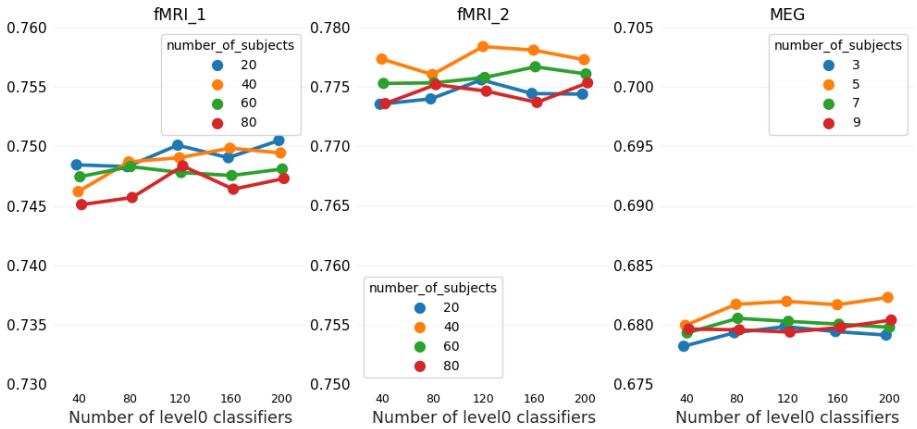


Figure 4.7: Generalization performance obtained on three datasets with stacked generalization. Each colored line corresponds to a given number of subjects from which data are drawn to train level-0 classifiers. The horizontal axis shows the influence of the number of level-0 classifiers.

Comparison to baseline

In this part, we compare the performance of subspace alignment and stacked generalization to the baseline. For each strategy, we choose the parameters which generate the best average generalization performance, i.e. with subspace alignment maximal number of eigenvectors is selected, with stacked generalization, for fMRI_1 200 level-0 classifiers are trained and each classifier is learned from data drawn from 20 subjects, for fMRI_2 we train 120 level-0 classifiers and each classifier is trained with data of 40 subjects, for MEG 200 classifiers are trained and each is trained with data of 5 subjects .

We used paired t-tests to assess the difference between mean performances of every two strategies. The results are shown in Figure 4.8. We first compare the performance of stacked generalization to baseline, it shows that stacked generalization significantly improves performance for fMRI_2 and MEG data ($t = 2.290$, $p < 0.05$ for fMRI_2, $t = 5.90$, $p < 10^{-7}$ for MEG), while not for fMRI_1. For all three datasets, subspace alignment does not outperform the baseline. Additionally, the performance of subspace alignment degrades significantly for MEG ($t = 7.19$, $p < 10^{-10}$). Furthermore, stacked generalization significantly outperforms subspace alignment in all cases ($t = 2.09$, $p < 0.05$ for fMRI_1, $t = 2.36$, $p < 0.05$ for fMRI_2, $t = 11.07$, $p < 10^{-19}$ for MEG)

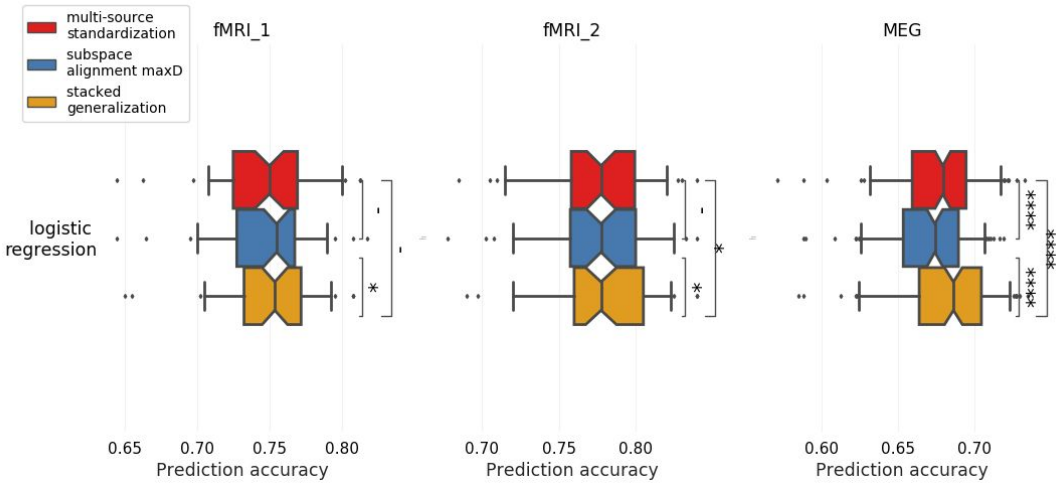


Figure 4.8: Generalization performance of subspace alignment and stacked-generalization. Compared to the baseline (red), stacked generalization (yellow) yields better generalization performances with fMRI_2 and MEG data. Subspace alignment (blue) does not outperform baseline in all cases, in addition, subspace alignment underperform baseline significantly with MEG data. In all cases, stacked generalization outperforms subspace alignment significantly. * denotes a significant difference ($p < 0.05$), ** denotes a significant difference ($p < 0.01$), **** denotes a highly significant difference ($p < 10^{-4}$), while - designates comparison for which no significant difference was found ($p > 0.05$).

4.7 Discussion

4.7.1 The multi-source transductive transfer setting is valuable

In Study 1, the improvement of generalization performance under multi-source transductive setting demonstrates that there exist distribution shifts between subjects. Furthermore, compared to classical standardization, multi-source standardization improves the performance of a classifier from two points of view: i) multi-source standardization takes into account the multi-subject nature of the data by performing feature construction subject by subject; ii) the transductive nature of ISPA allows access to the test set during the training phase, which allows multi-source standardization to be performed on both the training and test set. Therefore, in Study 1 inter-subject variability is reduced by exploiting the multi-source transductive nature of ISPA, which shows that ISPA is a multi-source transductive question. Study 2 shows that after common space construction and multi-subject feature designing, it is still challenging for a classifier to generalize to a test set formed by data from unseen subjects, which demonstrates that difference between the training and test set still exists. Therefore, ISPA is a transfer learning question as well. Overall, Study 1 and 2 show the well-foundedness of our multi-source

transductive transfer formalization of ISPA.

4.7.2 Guidelines and suggestions for ISPA studies.

We now attempt to capitalize on the ISPA formalization introduced in Chapter 3 and on the results of the experiments described in this chapter to provide a set of recommendations that could help researchers in designing an experimental paradigm for which ISPA can prove effective for group-level multivariate analysis.

A first element – that has recently been discussed in the neuroimaging literature (see for instance [134]) – is the importance of the sample size available for training. While training a model on data from multiple subjects allows increasing the sample size by at least an order of magnitude, it adds heterogeneity in the training set, thus potentially making the learning task more challenging. Study 2 has shown that when new subjects are added to the training set, which mechanically increases the sample size, the predictive power on new subjects is improved. This means that even with a simple pooling of the data, the added heterogeneity seems to be counter-balanced by the positive diversity and the increased sample size so that it is overall beneficial. Trying to decipher the influence of these two factors, Study 3 has shown that the most beneficial factor is in fact the within-subject sample size. Therefore, the researcher should favor experimental paradigms for which the number of samples / trials will be as high as possible for each subject, which in practice means choosing event-related designs over block designs, and maximizing the number of trials using fast designs. In conclusion, our advice is to favor (rapid) event-related designs and to maximize the number of subjects that will be scanned.

Secondly, we wish to come back to the choice of the cross-validation scheme. Because in ISPA, population-wise inference is enabled through the estimation of the generalization power of a decoder on new subject's data, an appropriate cross-validation scheme should leave out all the samples of one or more subjects for the test set (leave-P-subjects-out). We wish to insist that contrarily to the leave-one-sample-out scheme that was recently shown to be unstable and biased in neuroimaging problems [134], the leave-one-subject-out (i.e working with $P = 1$) scheme provides unbiased estimates of the decoding performances because all samples from a subject are left-out. However, we recommend to increase the number of data splits used in the cross-validation in order to obtain a larger number of measurements for the statistical assesment of the results at the second level (see Figure 2.1). With $P > 1$, the number of possible splits jumps to $\frac{S!}{P!(S-P)!}$, and choosing the ratio $\frac{P}{S}$ between $\frac{1}{10}$ and $\frac{1}{5}$ is usually admitted to offer accurate

performance estimates. For instance, with $S = 25$ subjects, switching from $P = 1$ to $P = 2$ makes the number of available splits go from 25 to 300; randomly drawing 50 to 100 splits amongst these 300 is a good solution to obtain reliable estimates of the performances.

Thirdly, we would like to put in perspective the cases of ISPA with calibration datasets, i.e. ISPA with anatomical and functional calibration datasets introduced in section 3.4.1. Comparisons of the performances offered by these two types of calibration datasets have been tackled in several papers (e.g [58, 148]), giving a clear advantage for the functional one where the common space is constructed through functional alignment rather than standard spatial registration based solely on the anatomy. We therefore advise to favor functional alignment when feasible, i.e. when it is possible to add a calibration scan in the scanning protocol. A question then naturally arises: which type of calibration scan? Most functional alignment studies use passive movie watching as calibration, arguing that such task recruits a large variety of cognitive capabilities, offers a large diversity and maintains the attention level. But these scans are usually very long, which can be prohibitive in many cases, and their effectiveness for higher-level cognitive functions remains to be evaluated.

In parallel, it has been shown that resting-state data is capable to be exploited as calibration data [97]. Furthermore, the connectivity at rest allows explaining inter-individual functional variability [125]. The authors actually suggested in their conclusion that *Resting-state data may provide a means for "calibration" of the BOLD signal.* In addition, exploiting the information offered by other MR acquisition types is also an interesting possibility. In particular, diffusion MRI has been shown to carry information that can explain the functional variability observed across subjects [118]. It has also been shown in [45] that diffusion MRI provides better performance as calibration data than standard spatial normalization. Given these elements, it seems clear that it would be valuable for the community to evaluate and compare the interest of these different scans to provide calibration data for ISPA.

4.7.3 The dimensionality of data affects generalization performance

All three datasets suffer from the "curse of dimensionality", i.e. in fMRI dataset the training set has 3600 samples (90 subjects, 40 samples per subject) and each has 5400 features, while in MEG dataset 8120 samples in the training set (14 subjects, 580 samples per subject) and each sample has 4896 features. High-dimensionality features and training set of small size provide challenges for machine learning algorithms, especially

when there exists inter-subject variability across samples. In the third panel of Figure 4.3, the green curve shows that learning from data drawn from larger number of subjects does not improve classifier's performances. Besides the influence of heterogeneity in the training set, this could also imply that high dimensionality hinders classifiers from picking up important features when more subjects are involved. High dimensionality also impacts the performance of subspace alignment in Study 4. Subspace alignment extracts first d eigenvectors from both training and test set. However, d eigenvectors do not cover enough information because of the small size of the test set but high dimensionality, we believe that is the reason why curves in Figure 4.6 reach a plateau for $d > 400$ and $d > 1000$, with fMRI and MEG dataset, respectively (fMRI and MEG datasets have 400 and 1160 samples in the test set, respectively).

In order to perform dimensionality reduction and reduce the difference between the training and test sets at the same time, there are several machine learning methods other than subspace alignment and stacked generalization in the literature. Zhang et al. proposed to learn a low-dimensional latent space with transfer component analysis [151]. However, similar to subspace alignment, the difference between subjects in the training set is ignored. Chen et al. proposed to learn a SRM which is a low-dimensional template shared by both training and test subjects [24]. [132] and [68] perform feature selection with regularization to select features that are common across subjects. These studies are rational candidates for performing dimensionality reduction in the context of ISPA. Therefore further methodological and experimental work are needed to exploit the existing methods and explore new methods.

4.7.4 Subspace alignment and stacked generalization provide different results

In Study 4 we have performed ISPA with two multi-source transductive transfer machine learning strategies: subspace alignment and stacked generalization. Our results show that with all three datasets, subspace alignment does not outperform baseline, while stacked generalization does not underperform baseline.

Figure 4.8 shows that in fMRI_1 neither of these two machine learning strategies outperforms baseline, in fMRI_2 only stacked generalization outperform baseline, additionally in MEG performance of stacked generalization has been significantly improved, on the contrary performance of subspace alignment has a significant drop compared to baseline. These results could be explained by two factors. The first is the respective levels of heterogeneity of the data, both within the training set and between the training and test set.

In Figure 4.3 the width of the gap between two curves is different in each panel, which implies that the amount of heterogeneity is different in each dataset. This point is again proved in Study 3, that curves in the first panel of Figure 4.4 are smoother than curves in the other two panels, which implies that fixing the sample size but involving more subjects in the training set has less influence on fMRI_1 than on the other two datasets. However, this point is not tested with a criteria which could measure the amount of heterogeneity in the dataset. The second point is the machine learning strategy used for performing ISPA. Subspace alignment transfers knowledge from the training set to the test set while stacked generalization focuses on constructing an invariant feature space between the training and test set. Both of them are attempts of reducing the difference between the training and test set. However, when the amount of heterogeneity in the dataset is little, it is challenging for both of them to improve performance (compared to baseline, neither of them significantly improves performance in fMRI_1). When the amount of heterogeneity is large, inter-subject variability and high dimensionality are obstacles for subspace alignment to extract invariant features through eigenvectors, while stacked generalization constructs feature space shared by the training and test set through hierarchical strategy. Furthermore, in order to examine how the distribution shifts are reduced by implementing different transfer learning strategies, we suggest to use a divergence measure to quantify the differences between training and test sets, e.g. $H\Delta H$ divergence [11] which measures divergence between distributions. However, the small size of individual dataset may make it difficult to compute such divergence.

4.8 Conclusion

In this chapter we have examined the well-foundedness of our formalization of ISPA as a multi-source transductive transfer learning problem. Four experimental studies have been implemented and we have demonstrated that several factors affect how well a classifier generalizes to new individuals, i.e. feature designing strategies, the size of the training set, the amount of heterogeneity across subjects and transfer learning strategies. Therefore, it remains challenging to overcome inter-individual variability in the context of ISPA.

Chapter 5

Multivariate group-level analysis using L^p distance-based optimal transport

Publication associated with this chapter: Wang, Q., Redko, I. and Takerkart, S., 2018, June. Population Averaging of Neuroimaging Data Using L^p Distance-based Optimal Transport. In 2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI) (pp. 1-4). IEEE.

5.1 Introduction

In previous chapters, we have introduced several widely used group level analysis methods, i.e. univariate group analysis with the GLM, multivariate group analysis such as G-MVPA and ISPA. However, in order to implement the conventional group analysis methods with the GLM on fMRI datasets, usually a spatial smoothing operation is applied to the data to overcome the potential differences that exist in the locations of activation foci across individuals. This allows improving the spatial overlap across subjects. But it results in an amplitude dump which harms the detection power of standard group analysis methods based on euclidean averaging computed independently at each brain location. Additionally, for multivariate group analysis methods, searchlight strategy which exploits local multivariate patterns is usually involved to detect significant regions in the population level. However, searchlight only adopts patterns from voxels in the searchlight sphere, which ignores the global information and the geometric

properties of the brain.

Motivated by these limitations, we aim to provide a new multivariate group analysis method which exploits patterns of all voxels and the geometric information contained in fMRI data.

Recently, a new method was introduced to compute the barycenter of a set of empirical probability measures with respect to a distance from the optimal transportation theory, called the Wasserstein distance [4]. The Wasserstein barycenter given by a probability measure is defined to minimize the sum of its Wasserstein distances to each element in that set. An important advantage offered by the Wasserstein distance is its capacity of taking into account the geometry of the data underlying the probability distributions by the means of the cost-matrix associated to it.

In the context of neuroimaging-based population studies, the Wasserstein barycenter therefore sounds appealing because its geometrical grounding should facilitate handling individual differences on the one hand, and because it intrinsically takes into account the multivariate nature of brain patterns on the other hand. However, the challenge lies in the specific nature of neuroimaging data which cannot be naturally represented as empirical probability measures, i.e histograms. Indeed, the natural transformation that maps arbitrary data into a histogram is the affine function that ensures non-negativity and normalizes the mass of the data to having the value of one. Applying such transformation separately on each subject would discard the relative amplitude differences that exist both across the images and across individuals, which are critical to correctly assess the informative content of brain patterns. New optimal transport algorithms are therefore needed to handle unnormalized data such as offered in neuroimaging.

In order to compute barycenters from unnormalized measures which are non-negative with mass smaller than one, [51] proposed to add a virtual point to the considered probability measure and then to minimize the Wasserstein distance by exploiting the dual optima as the gradient of barycenter [29]. On the other hand, the method in [9] does not transform data to fulfill the properties of a probability measure, but proposes to use the discrepancy of the functional features as a cost function between voxels. Furthermore, [127] proposes to define a transportation L^p distance (TL^p), a modification of the Wasserstein distance that integrates the amplitudes and the geometric information of the data in the cost matrix.

Therefore we propose a new multivariate group analysis which includes computing barycenter for unnormalized measures by minimizing the sum of their TL^p distances. Our method is implemented on both artificial and real fMRI data. In our method,

data from multiple subjects are first transformed into a common space, i.e. by performing spacial normalization on fMRI data or by generating artificial data in gaussian distribution. Then data are carried forward into our proposed algorithm to compute their barycenter, followed by statistical model evaluated on the barycenter to detect the significant regions at the group level.

In this chapter, in section 5.2 we first introduce some background on optimal transport and the method proposed in [51]. Then in section 5.3 we introduce our new algorithm that computes a barycenter from a set of unnormalized measures based on the TL^p distance. We provide an empirical study of its effectiveness, which we present in section 5.4. Our evaluation includes a direct comparison with the algorithm of [51], using artificial data, which allows studying the robustness of the methods when confronted with different levels of noise, as well as an application on a real fMRI dataset.

5.2 Background of optimal transport

In this section we first introduce the background for Wasserstein distance of probability measures, then we move to distance between non-negative unnormalized measures.

Notations. For any vector $u \in \mathbb{R}_+^d$, we denote $|u|_1$ the mass of u, $|u|_1 = \sum_{j=1}^d |u_j|$. There are two types of measures involved in this chapter, one is probability measures, $u \in \Sigma_d$, $\Sigma_d = \{u \in \mathbb{R}_+^d, \sum_{j=1}^d u_j = 1\}$. The other is the unnormalized measure h, whose mass is smaller than or equal to 1, $h \in S_d$, where $S_d = \{h \in \mathbb{R}_+^d, \sum_{j=1}^d h_j \leq 1\}$. Let $C \in \mathbb{R}_+^{d \times d}$, C is the cost matrix of pairwise distances between the locations of elements in d-dimensional measure, i.e. C_{ij} is the Euclidean distance between locations of i-th and j-th elements.

5.2.1 Wasserstein barycenter of probability measures

For two histograms a and b, $a, b \in \Sigma_d$, when $p \geq 1$, the p- Wasserstein distance between a and b is the p^{th} root of the optimum of a Optimal Transport (OT) problem, which is defined as:

$$W_p(a,b) = \mathbf{OT}(a,b,C^p)^{\frac{1}{p}} = \min_{T \in U(a,b)} \langle T, C^p \rangle^{\frac{1}{p}} \tag{5.1}$$

where C^p is C raised to the power of p, $U(a,b)$ is a set of matrices whose row and column marginals are equal to a and b, respectively:

$$U(a,b) = \{T \in \mathbb{R}_+^{d \times d} | T\mathbb{1}_d = a, T^T\mathbb{1}_d = b\} \tag{5.2}$$

To make it clear, we fix $p = 1$ in the rest of this section. When $p = 1$, the distance obtained from 5.1 is also known as Earth Mover's Distance [114].

Since solving 5.1 is computational expensive, especially when d gets larger, in [28], Cuturi introduced an efficient algorithm with an entropic regularization term to compute the Sinkhorn distance which approximates to the optimal transport distance:

$$\mathbf{OT}^{\lambda}(a, b, C) = \langle T^{\lambda}, C \rangle, \ where \ T^{\lambda} = \underset{T \in U(a,b)}{\operatorname{argmin}} \langle T, C \rangle - \lambda h(T) \tag{5.3}$$

where $h(T)$ is the entropy of T, $h(T) = -\sum T_{ij} \log T_{ij}$. This regularized transport problem $\mathbf{OT}^{\lambda}$ not only smoothes the original problem with an entropic regularization term, but also obtains a unique transportation plan T^{λ} several orders of magnitude faster.

Then Cuturi et al. (2014) introduced a way to compute the mean of a set of empirical probability measures, known as Wasserstein barycenter [4], by minimizing the sum of barycenter's distance to each measure in that set [29].

$$a = \underset{u \in \Sigma_d}{\operatorname{argmin}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{OT}^{\lambda}(u, b^i, C) \tag{5.4}$$

where a is the Wasserstein barycenter of N probability measures b^i. In order to compute the barycenter, [29] proposed to compute the dual optimal of N transportation problems to form the subgradient of a. Followed the approach proposed in [28], the gradients were computed efficiently which lead to the convergence of a.

5.2.2 Barycenter of unnormalized measures

Above we introduced the background necessary for computing the Wasserstein barycenter of multiple probability measures. However, the approach proposed above is not able to be applied on unnormalized measures with mass different than 1. In order to handle this situation, Gramfort et al. (2015) proposed to compute Kantorovich's distance between unnormalized measures [51]. Computing Kantorovich's distance can be cast as a regular optimal transport problem by adding a virtual point [53]. The Kantorovich's distance between two unnormalized measures is defined as follows:

$$K(g, q) = \mathbf{OT}(\begin{bmatrix} g \\ 1 - |g|_1 \end{bmatrix}, \begin{bmatrix} q \\ 1 - |q|_1 \end{bmatrix}, \bar{C}) \tag{5.5}$$

where $g, q \in S_d$, $\bar{C} = \begin{bmatrix} C & \Delta \\ \Delta^T & 0 \end{bmatrix}$, $\bar{C} \in \mathbb{R}_+^{d+1 \times d+1}$, Δ contains the distances from the virtual point to elements of the unnormalized measure, $\Delta \in \mathbb{R}_+^d$.

The Kantorovich mean of N unnormalized measures is defined as:

$$g = \underset{h \in S_d}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} K(h, q^i) = \underset{h \in S_d}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \mathbf{OT}(\begin{bmatrix} h \\ 1 - |h|_1 \end{bmatrix}, \begin{bmatrix} q^i \\ 1 - |q^i|_1 \end{bmatrix}, \bar{C}) \tag{5.6}$$

Where g is the barycenter of N unnormalized measures q^i.

In order to solve 5.6, [51] proposed to add a constrain on the mass of the g, where $|g|_1 = \frac{1}{N} \sum_{i=1}^{N} |q^i|_1$. By implementing the approach proposed in [29], the gradient of barycenter was computed for updating g in each iteration. The barycenter of N unnormalized measures proposed in [51] is defined as follows:

$$g = \underset{h \in S_d,\ |h|_1 = \frac{1}{N} \sum_{i=1}^{N} |q^i|_1}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \mathbf{OT}^{\lambda}(\begin{bmatrix} h \\ 1 - |h|_1 \end{bmatrix}, \begin{bmatrix} q^i \\ 1 - |q^i|_1 \end{bmatrix}, \bar{C}) \tag{5.7}$$

We denote this algorithm (Kantorovich Barycenter with Constrained Mass [51]) as KBCM.

5.3 Methods

In this chapter, we focus on computing a barycenter from unnormalized measures. We assume that there are N subjects, and denote by $q^i \in S_d$ the unnormalized measure for i^{th} subject, with $i \in \{1, ..., N\}$.

Our objective is to find a barycenter g which minimizes the sum of its distance to q^i. Given the unnormalized measures, we propose to construct a cost matrix with locations and intensities of bins in q^i, which is related to the cost matrix in TL^p distance [127]. Then, the barycenter can be efficiently computed using the entropic regularized optimal transport and Iterative Bregman projections [12]. Specifically, we aim at comparing our method, referred to as TL^p-BI, to KBCM. We use both artificially constructed and real fMRI datasets where inter-individual differences are strong to test whether such algorithms can handle such variability.

5.3.1 TL^p distance

When measures considered above represent images, the cost matrix C in the definition of the Wasserstein distance allows to take into account geometric information but does not include the information regarding the intensities of pixels. This drawback was addressed by the TL^p distance [127] defined as follows:

$$TL^p_\lambda(\mu,\nu;g,q) = \min_{\hat{T}\in U(\mu,\nu)} \langle \hat{T}, \hat{C}\rangle \tag{5.8}$$

$$\hat{C} = \lambda C + \tilde{C} \tag{5.9}$$

where μ, ν are probability measures, C is the cost matrix as in the Wasserstein distance and $\tilde{C}$ is the cost matrix of pairwise distances between intensities of g and q, i.e. $\tilde{C}_{m,n} = |g_m - q_n|^p_p$. Contrary to the Wasserstein distance, the TL^p distance makes no assumptions regarding the non-negativity of measures g and q and allows them to have different mass.

5.3.2 Barycenter with the TL^p distance

As activation intensities are important for group analysis, we assume that taking into account the intensity in the cost matrix may help to preserve the amplitudes of unnormalized data. Under this assumption, we propose to find a barycenter g that represents the mean of unnormalized measures by minimizing the sum of its TL^p distances to $\{q^1, ..., q^N\}$, where p is set to be 2.

To proceed, let us denote by $\hat{C}^i$ the cost matrix between g and q^i. This cost can be calculated as the combination of squared Euclidean distances between the locations and intensities of pixels, respectively, i.e. $\hat{C}^i_{m,n} = |\mathrm{loc}(g_m) - \mathrm{loc}(q^i_n)|^2_2 + \eta|g_m - q^i_n|^2_2$, where η is the trade-off parameter between two distances and loc() is the coordinate of a pixel on the discrete grid. The considered optimization problem thus reads:

$$g = \underset{h\in S_d}{\mathrm{argmin}} \frac{1}{N}\sum_{i=1}^{N} TL^2(\hat{h}, \hat{q}^i; h, q^i) \tag{5.10}$$

where $q^i \in S_d$, $\hat{q}^i = q^i/|q^i|_1$, $\hat{h} = h/|h|_1$.

As mentioned in the introduction, entropic regularization of the barycenter leads to an optimization problem that can be solved efficiently with the Iterative Bregman projections algorithm which provides strong convergence guarantees. To benefit from it, we

propose a two-stage strategy that first computes a barycenter g with Iterative Bregman projections for given cost matrices $\{\hat{C^1}, ..., \hat{C^N}\}$, and then updates cost matrices using the obtained barycenter. This process is continued until the barycenter becomes stable. The proposed algorithm is summarized in Algorithm 1.

Algorithm 1 Barycenter with TL^p distance

Input: $\{q^1, ..., q^N\}$, $\{\hat{q}^1, ..., \hat{q}^N\}$, geometry cost matrix C, $\eta > 0$, $\lambda > 0$, weights $\{\beta^1, ..., \beta^N\}$, $\sum \beta^i = 1$
Output: barycenter g

mean mass $\rho = \frac{1}{N} \sum_i |q^j|_1$
$\hat{g} = \mathbb{1}_d/d$, $g = \mathbb{1}_d/d \times \rho$
$u = ones(d, N)$
$v = ones(d, N)$
while g changes **do**
    for each q^i **do**
        $\hat{C}^i_{m,n} = C_{m,n} + \eta |g_m - q^i_n|^2_2$
        $K^i = \exp(-\frac{1}{\lambda} \hat{C}^i)$
    end for
    while $\hat{g}$ changes **do**
        for each $\hat{q}^i$ **do**
            $u^i = \hat{q}^i / (K^i v^i)$
        end for
        $\hat{g} = \prod_{i=1}^N (v^i \times (K^i u^i))^{\beta^i}$
        for each $\hat{q}^i$ **do**
            $v^i = \hat{g} / ({K^i}^T u^i)$
        end for
    end while
    $g = \rho \times \hat{g}$
end while

5.3.3 Artificial fMRI data

First, we generate artificial fMRI datasets by simulating one contrast map per individual on a square of size 50×50. Each datasets consists in 20 subjects, i.e. 20 contrast maps.

We assume that each subject presents a unique activated region of gaussian shape, with a small size ($\sigma = 1$ pixel) compared to the size of the image. We would therefore expect a summary representation of the population to present the same property. The inter-individual variability is induced by allowing both the location of the activated region and its amplitude to vary across subjects: its center is randomly chosen (uniform distribution) within a circle of radius 15 centered in the middle of the image and its amplitude is drawn from a Gaussian distribution $\mathcal{N}(5, 1)$. In order to evaluate the robustness of

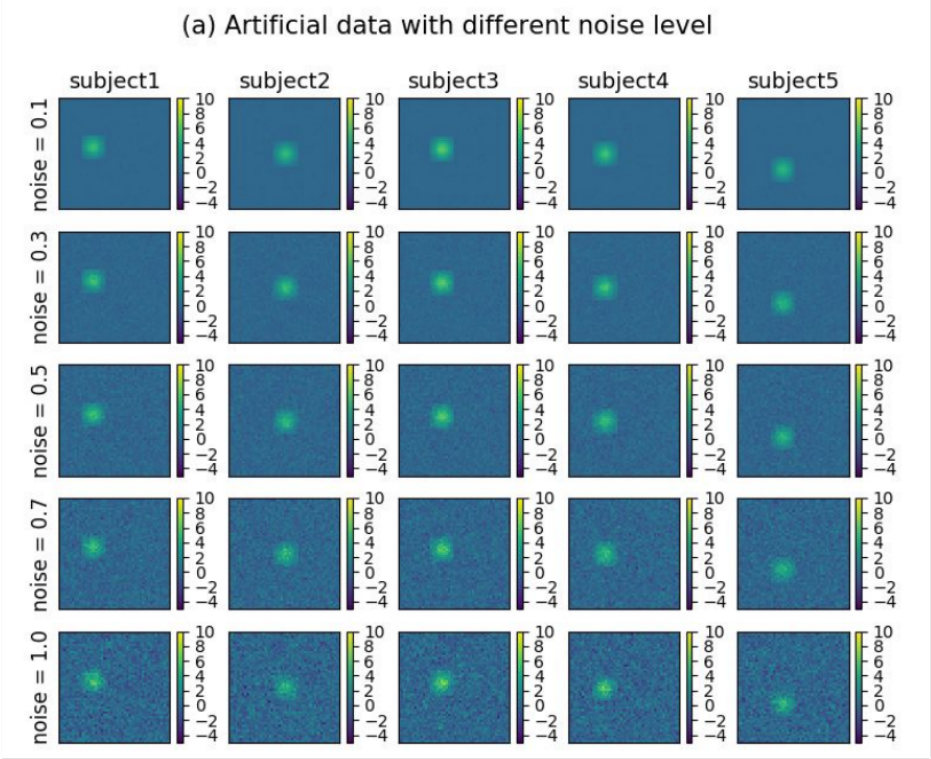


Figure 5.1: Illustration of the artificial datasets. Each line is a subpart of a single dataset (5 subjects shown amongst 20 in (a), 10 subjects shown in (b). (a): influence of the noise level (increasing noise from top to bottom). (b): five datasets obtained with the same noise level (noise=0.1)

these algorithms in the presence of noise, we add noise onto each image, generated from a Gaussian distribution $\mathcal{N}(0, \sigma)$. The noise level is parametrically controlled by using

$\sigma \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$. Furthermore, for each noise level, we generated 10 datasets. Figure 5.1a illustrates the influence of the noise level, Figure 5.1b shows different datasets generated with the same noise level.

5.3.4 Real fMRI data

The fMRI dataset exploited in this chapter is fMRI_1 used in Chapter 4, which was acquired as each subject passively listened to a fixed set of vocal and non vocal stimuli. A within-subject general linear model analysis was set-up to estimate the contrast between the perception of vocal and non vocal sounds for each subject, which ends up with 100 contrast maps in the fMRI dataset, one contrast map per subject. These contrast maps served as inputs to compute the population average maps, within a large region of interest that included the auditory cortex as well as the voice sensitive regions (illustrated on Figure 5.9). The study in [109] showed that large differences exist in the organization of the temporal voice areas, but that a structure comprising three voice patches can be identified at the group level using advanced processing techniques, not with standard group analysis. This data is therefore perfectly suited to challenge the potential gains offered by optimal transport techniques.

5.3.5 Experimental setting

We randomly define 10 sets of 20 subjects to simulate a group analysis obtained with a group of standard size, both for the artificial data – by construction – and for the real data – by drawing subsets from the full set of available subjects. We compute a barycenter image separately for each of these sets. For this, we first vectorize each contrast map to obtain a vector $v^i \in \mathbb{R}^d$. We then transform these vectors with $q_n^i = \frac{v_n^i - \alpha}{max_i \sum_{n=1}^{d}(v_n^i - \alpha)}$, where α is the minimal intensity of the full set of data $\{v^1, ..., v^{20}\}$ – which allows preserving the relative amplitude differences across subjects – so that the elements of q^i become all non-negative, $max_i \sum_{n=1}^{d}(v_n^i - \alpha)$ is the maximal mass of $\{v^1 - \alpha, ..., v^{20} - \alpha\}$ – which allows the total mass of q^i smaller than or equal to 1 – so that the necessary properties for KBCM is ensured.

The cost matrix used for KBCM, C^{KBCM}, contains the pairwise squared Euclidean distances of locations on the grid, $C_{m,n}^{\text{KBCM}} = |\text{loc}(g_m) - \text{loc}(q_n^i)|_2^2$. In TL^p-BI, the cost matrix takes into account both the locations and intensities of pixels, $\hat{C}_{m,n}^{\text{TL}^p-\text{BI}} = C^{\text{KBCM}} + \eta|g_m - q_n^i|_2^2$, where $\eta \in \{10.0, 1.0, 0.1, 0.01, 0.001\}$. Additionally, since both KBCM and TL^p-BI adopt entropic regularization term for computing the barycenter, λ

is chosen from $\{0.1, 0.01, 0.001, 0.0001, 0.00001, 0.000001\}$ for both KBCM and TL^p-BI.

For each λ or each combination of λ and η, KBCM and TL^p-BI compute one barycenter respectively. The obtained barycenters are examined both qualitatively and quantitatively. First, we assess the properties of the barycenters visually. Then the amplitude of barycenter is considered as an important criterion for selecting parameters, because this criterion was used for comparing barycenters in [51]. The barycenters of maximum amplitude are selected, which are carried forward to the statistical assessment for detecting significant regions.

5.3.6 Statistical assessment

In order to localize regions where there exist non null group level effects, we use the same permutation test described in section 2.5 for both KBCM and TL^p-BI on artificial and real fMRI datasets. Given a set of 20 subjects, the barycenter of maximal amplitude is selected. Since the null hypothesis is that there is no group level effects in the barycenter, shuffling signs of intensities of pixels in the input maps will not influence obtaining the barycenter. We implement 500 permutations across subjects, each computes a non-effect barycenter with the same parameters which generate the barycenter of maximal amplitude with true contrast maps. With the given significance threshold ($p < 0.05$), the significance level is assessed and assigned to each pixel in the barycenter map.

In order to compare the results provided by KBCM and TL^p-BI, for both artificial and real fMRI data, we examine the thresholded statistical map and compare the size of significant regions obtained from each strategy.

5.4 Results

In this section, we present barycenters and the results obtained from statistical assessment on both artificial and real fMRI datasets. With the artificial datasets, because we added noise of different levels to data, our focus is therefore on how the noise level influences the barycenter. Then for both artificial and real fMRI datasets, we describe the differences between results produced by KBCM and TL^p-BI.

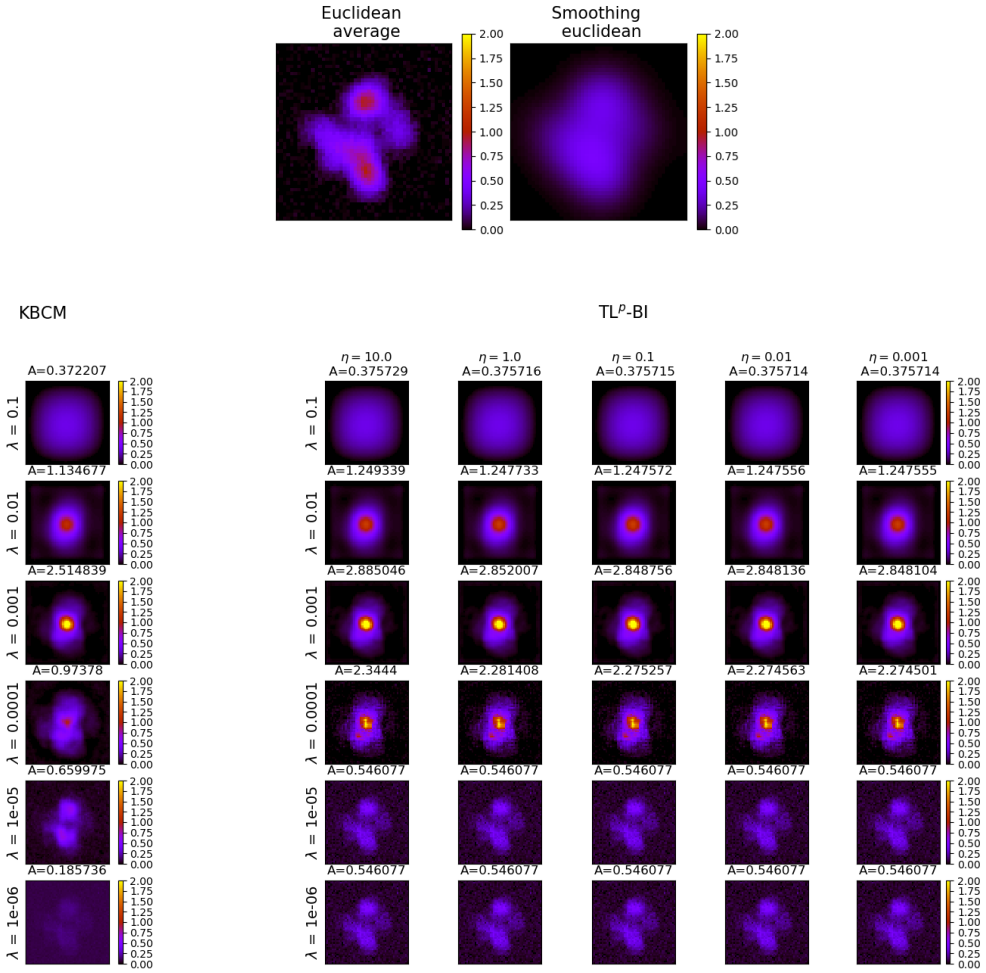


Figure 5.2: Barycenters of artificial data, noise level is 0.1

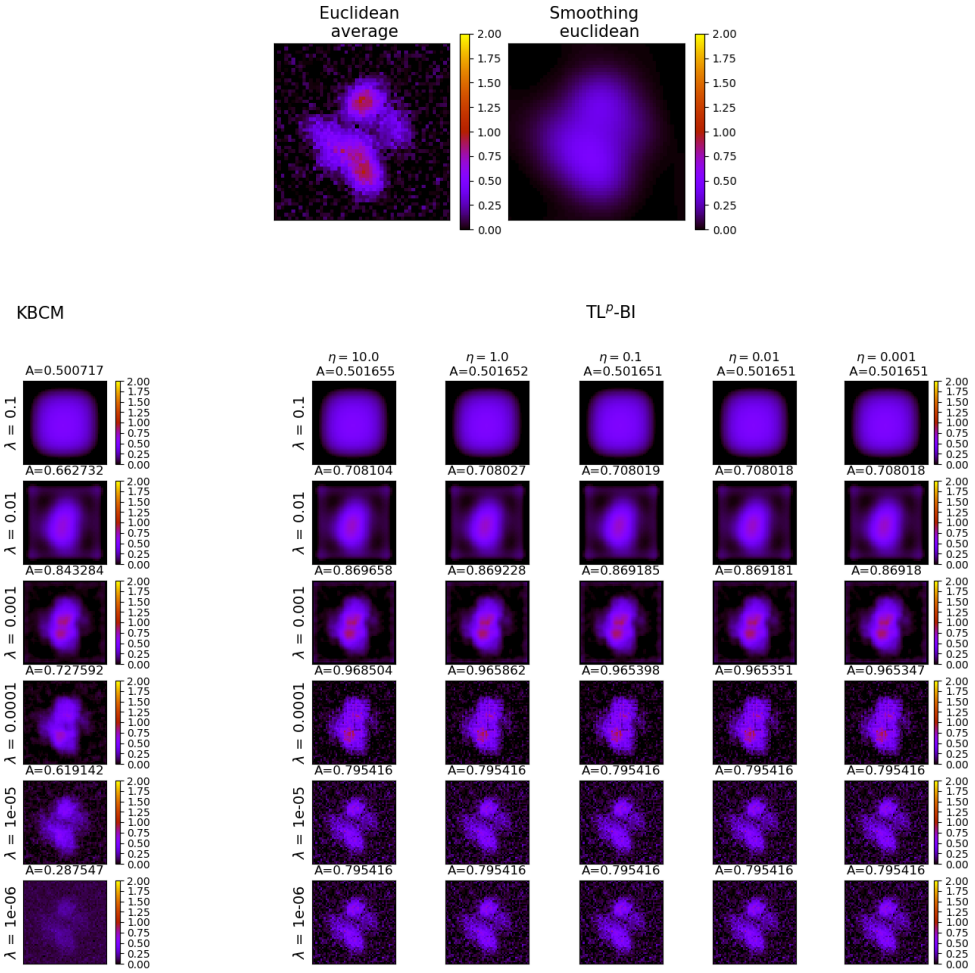


Figure 5.3: Barycenters of artificial data, noise level is 0.3

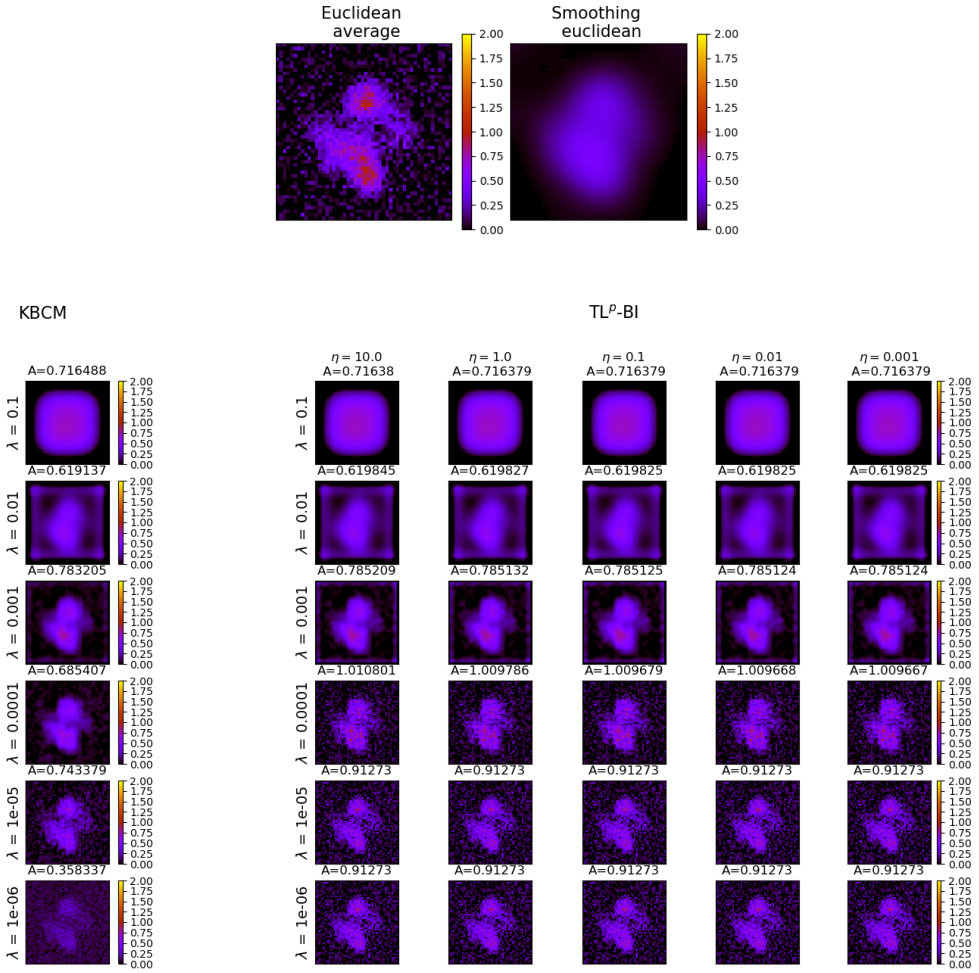


Figure 5.4: Barycenters of artificial data, noise level is 0.5

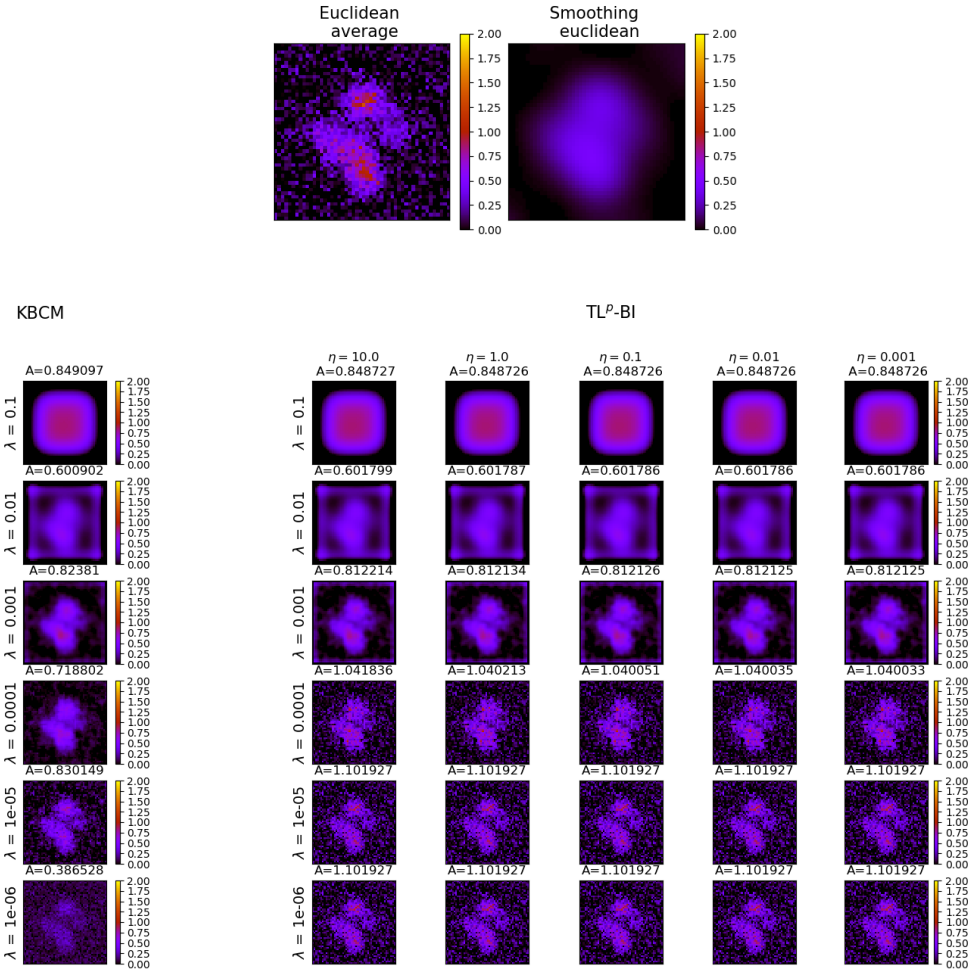


Figure 5.5: Barycenters of artificial data, noise level is 0.7

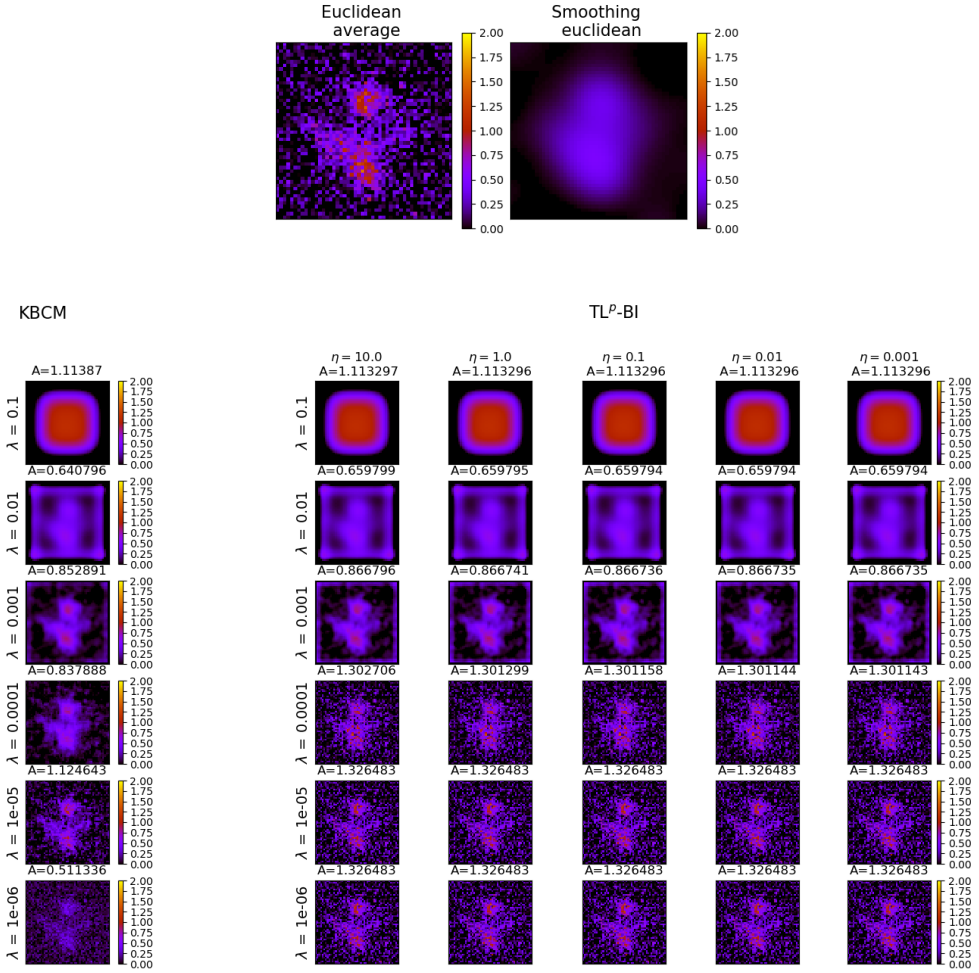


Figure 5.6: Barycenters of artificial data, noise level is 1.0

5.4.1 Results on artificial data

Figure 5.2 – Figure 5.6 illustrate barycenters generated from artificial data with noise level from 0.1 to 1.0. On each figure, barycenters computed from different parameters are displayed for both KBCM and TL^p-BI. In general, λ influences the main shape of the barycenter for both two strategies. When $\lambda > 0.0001$, KBCM and TL^p-BI generate similar images, however, when $\lambda \leq 0.0001$, there is difference between images obtained from these two strategies. When λ is 0.1, the obtained image is smooth, low-intensity or high-intensity pixels cover the majority of the image. As λ decreases from 0.01 to 0.001, images become sparse and multiple high-intensity pixels are grouped in the center of the image. When $\lambda \leq 0.0001$, pixels spreads out into several regions, images obtained from TL^p-BI are similar to those obtained from Euclidean averaging used in the standard GLM, while images computed from KBCM has less high-intensity pixels and smaller amplitude. For TL^p-BI, η does not impact the shape of the barycenter but has slight influences on the amplitude of the barycenter. When λ is smaller than 0.0001, η has little influence no matter the noise is weak or strong. When λ is larger than 0.0001, in most cases, the amplitude increases as η increases. Furthermore, the noise level impacts the obtaining of the barycenter. In the cases where the noise level is smaller than 0.7, the maximal amplitude is obtained when λ is 0.001 or 0.0001. However, with strong noise, when the noise level is larger, the maximal amplitude is obtained with smaller λ.

Figure 5.7 illustrates the effects of the variability on barycenters obtained from optimal transport strategies and from the Euclidean averaging used in the standard GLM (with or without spatial smoothing of the data). For this, we show the results for two groups of subjects (set1 and set2), for different noise levels (from weak to strong from top to bottom rows). Only optimal transport methods offer a summary representation of the population which corresponds to the model that was used to generate the data, i.e that show one small activated region. But this is the case only with a weak noise level (noise level is 0.1). In all other cases, the obtained images show high-intensity pixels that are more spread out throughout the images, eventually divided into several regions, and that depend on the content of the group of subjects itself which are different between set1 and set2.

Given the noise level, for each set of 20 subjects we select parameters which generate barycenter of the maximal amplitude (λ for KBCM, λ and η for TL^p-BI), then with the same parameters we obtain one statistical map through permutation test for each case. Figure 5.8 presents the quantitative comparisons of results obtained from KBCM and TL^p-BI algorithms. Figure 5.8a show that both algorithms yield equivalent results: one

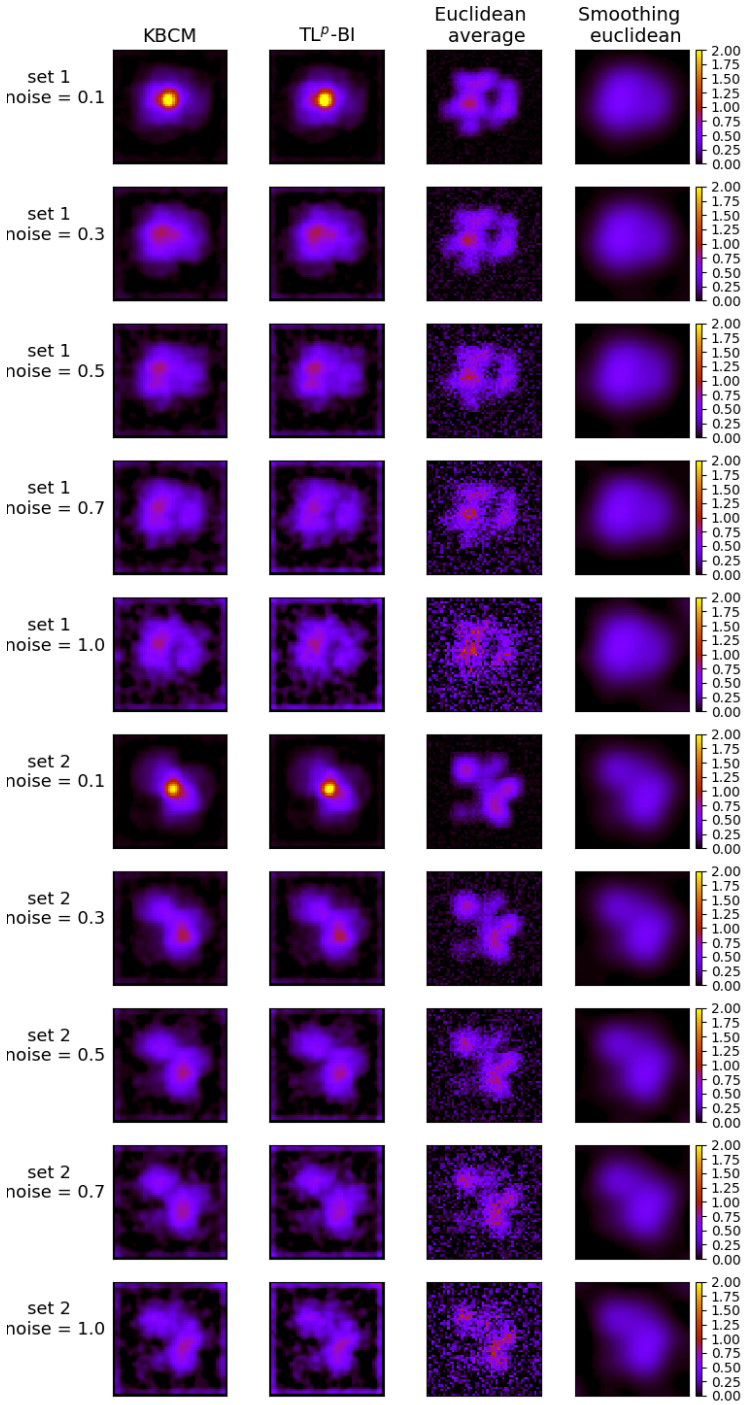


Figure 5.7: Barycenters computed from artificial data with noise of different levels. Two groups of subjects are used (set1 and set2), for each group, barycenters generated from artificial data with different noise levels (from 0.1 to 1.0 from top to bottom) are shown.

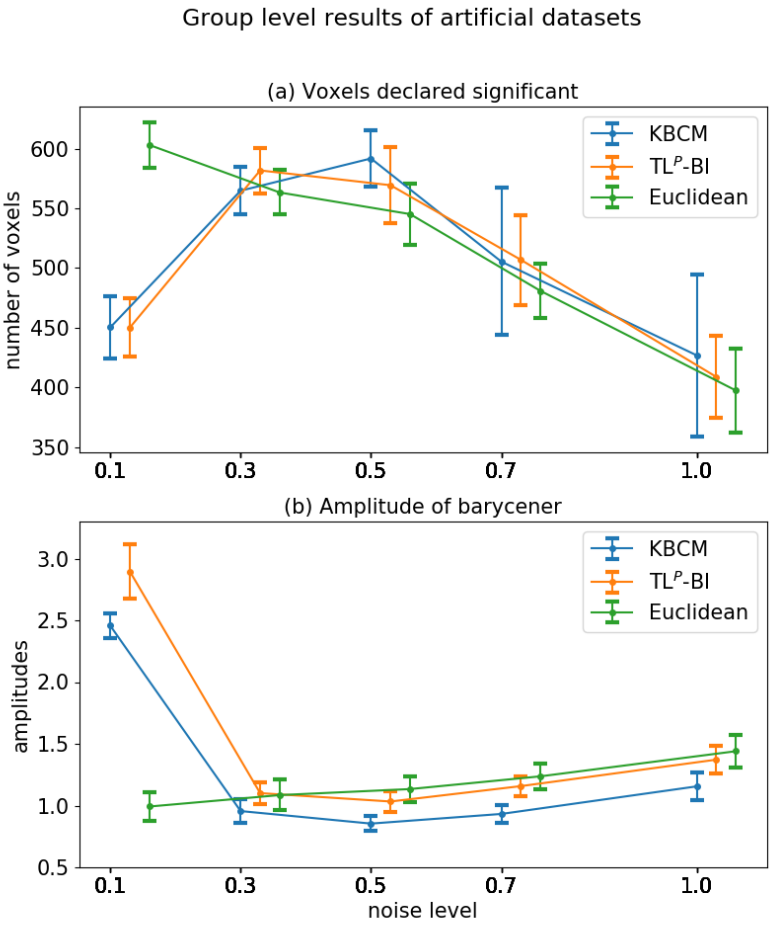


Figure 5.8: Comparisons of KBCM and TL^p-BI on artificial datasets

small focus of activation when the noise is weak, as expected from the generative model – note that this cannot be obtained with standard group analysis; and a deviation from this (increasing number of significant voxels) when the noise gets stronger. However, as the noise level is larger than 0.5, the number of significant voxels decreases, which is similar to the results generated from standard group analysis with Euclidean averaging. Figure 5.8b shows that when the noise level increases, the peak amplitude of the barycenters decreases. For each noise level, TL^p-BI yields barycenters with higher peak amplitude than those of KBCM ($p < 0.01$ for 0.1, $p < 0.001$ for 0.3, $p < 0.0001$ for 0.5 and 0.7, $p < 0.001$ for 1.0). When noise level is stronger than 0.1, standard group analysis with Euclidean averaging generates similar results as optimal transport methods. However, for noise is 0.1, standard group analysis has lower amplitude than two optimal transport methods.

5.4.2 Results on real fMRI data

Figure 5.9 displays barycenters computed from real fMRI dataset recorded from one group of 20 subjects. As with the artificial datasets, we also exploit several values of λ and η for KBCM and TL^p-BI, barycenters computed with different parameters are shown. Figure 5.9 shows that both strategies produce similar barycenters when $\lambda \geq 0.001$: high-intensity pixels cover the majority of the image when λ is 0.1, high-intensity pixels in the middle of the image for λ of 0.01 and 0.001. When λ is smaller than 0.001, differences exist between images generated by the two strategies. When λ is 0.0001, from a qualitative point of view, the map produced by KBCM appears smoother while the barycenter obtained with TL^p-BI provides finer spatial details. When λ is smaller, the map produced by KBCM has less high-intensity pixels, but map from TL^p-BI stays consistent. For both KBCM and TL^p-BI, the barycenter obtains the maximal amplitude when $\lambda < 0.001$.

Figure 5.10 shows barycenters computed from three groups of 20 subjects. It shows that images computed from both KBCM and TL^p-BI display large variability across groups. In general, The map generated from real fMRI dataset are very similar to maps computed form artificial data with strong noises.

After performing the statistical test, we compare the peak amplitudes of the barycenters and the number of significant voxels obtained with the two algorithms (see Figure 5.11). Results show that our algorithm provides larger number of significant voxels (paired t-test, $p < 0.01$) and higher amplitude than KBCM (paired t-test, $p < 0.001$).

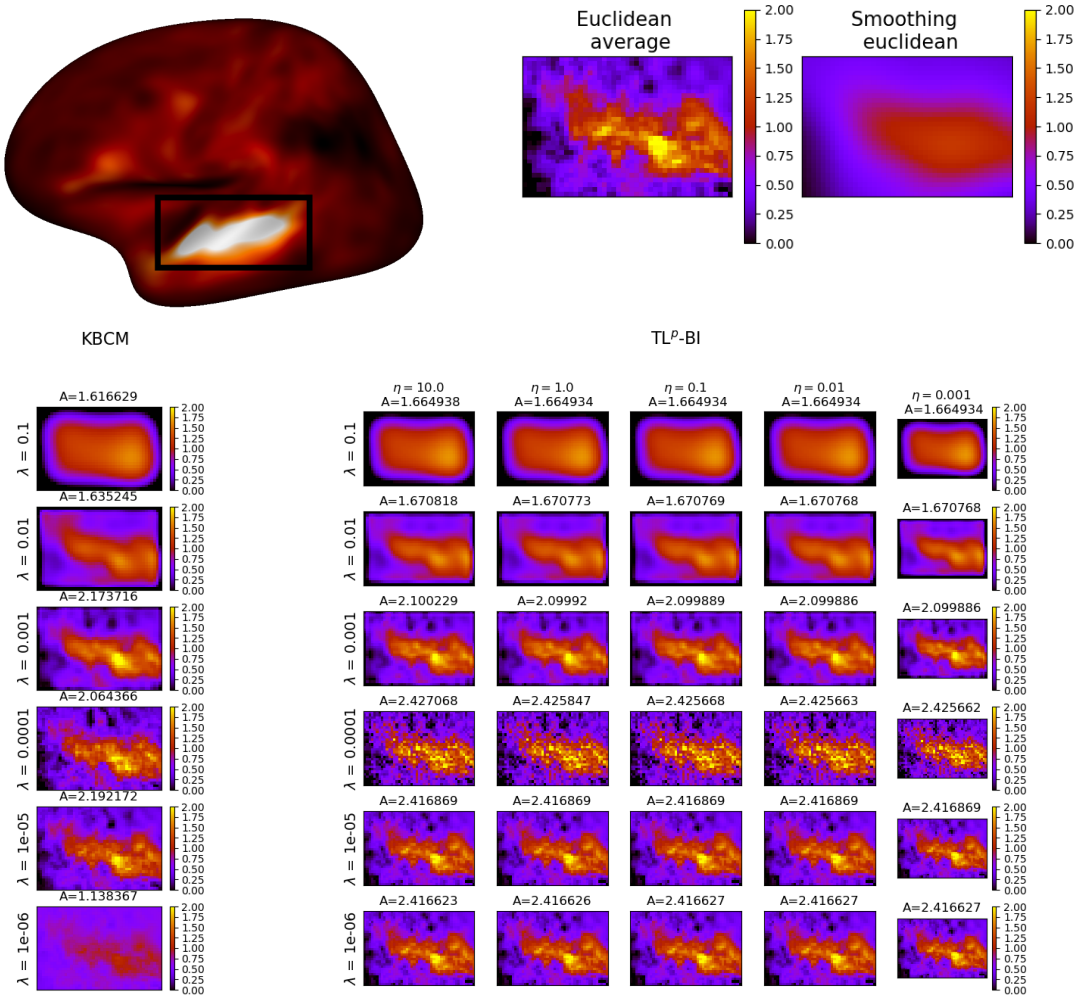


Figure 5.9: Barycenters of fMRI data

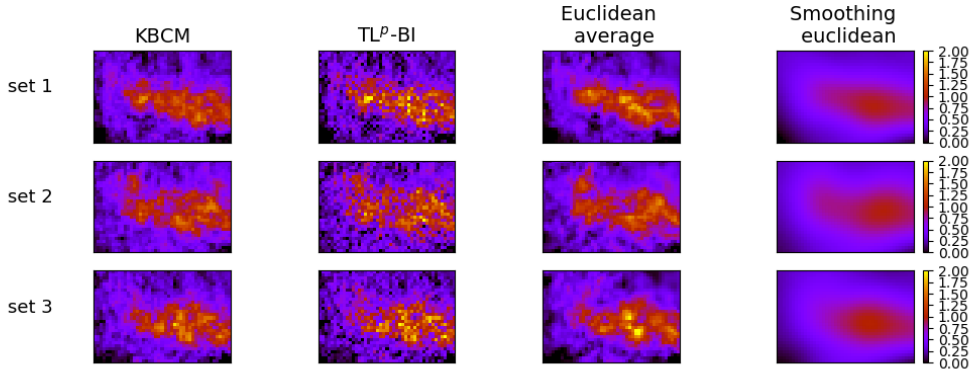


Figure 5.10: Barycenters computed from real fMRI data. Three groups of subjects are shown (set1, set2 and set3), for each group, barycenters generated by four strategies are shown.

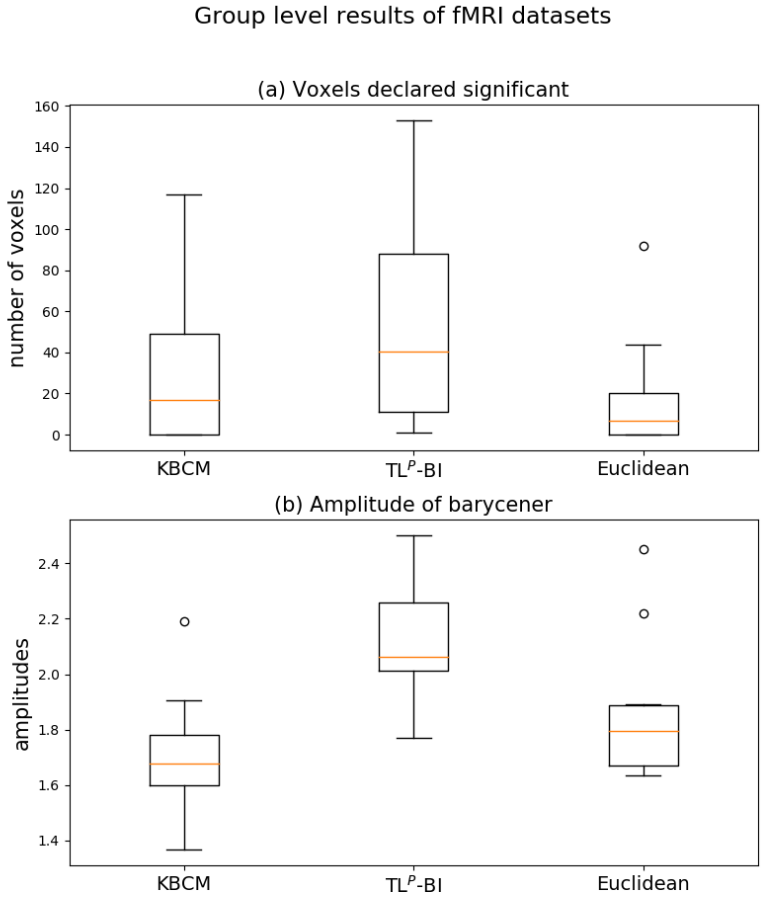


Figure 5.11: Comparisons of KBCM and TL^p-BI on real fMRI datasets

5.5 Discussion

5.5.1 Influence of the noise on the optimal transport methods

In this study we have first performed experiments on artificial data which consist of one unique region of activation generated by a gaussian model. In order to understand how noise level affects the computation of barycenter for optimal transport methods, we added noise of different levels to artificial data. Our results show that noise level has large impact on the estimated barycenters. When the noise level is weak (noise = 0.1), both optimal transport methods hold the property that they are capable to overcome the inter-subject variability and generate barycenters that correspond to the generative model, regardless of the set of subjects. However, as the noise level increases the property is no longer held, it is challenging for both optimal transport methods to obtain such barycenters. Additionally, results from statistical assessment show this trend as well. Figure 5.8 displays that when noise level is 0.1, optimal transport methods produce results that are better than results of Euclidean average on both the number of significant voxels and the amplitude of barycenter. However, when the noise level is stronger than 0.1, optimal transport and Euclidean average methods generate similar results. Therefore, optimal transport methods outperform Euclidean averaging when working on data with weak noise, but strong noise hinders optimal transport methods from holding this advantage.

5.5.2 From artificial to real fMRI data

In this study we applied group level analysis on both artificial and real fMRI datasets. Results computed from fMRI data present similarity to those obtained from artificial data with strong noise level, e.g. both Figure 5.6 and Figure 5.9 show that the maps generated with $\lambda = 0.1$ are covered by large majority of high-intensity pixels, as the value of λ decreases, high-intensity pixels spread out through maps, additionally, the map of maximal amplitude is obtained when λ is smaller than 0.001. We demonstrated that with an increased level of noise in the input data, optimal transport methods do not obtain barycenters that correspond to the generative model, it implies that these methods are not capable to obtain reasonable barycenters for fMRI data either. Furthermore, we use the same fMRI dataset as the one adopted in [109], where three voice patches are identified at the group level using an advanced clustering technique applied after the univariate GLM. Our hope was that optimal transport methods could detect these three regions, which were not directly distinguished by a standard group-level Euclidean

averaging. Compared to the artificial data that only have one activated region, real fMRI data are more complex. Therefore, whether optimal transport methods are better than the GLM on detecting several sub regions and how to handle the noise in real fMRI data is still a open question. In the future, we will for instance explore the effect of a light smoothing of the data, which allows reducing the noise level, on the performances of our optimal transport algorithm.

5.5.3 Criterion for choosing parameters

In this study, for each of two optimal transport methods we compute barycenters with several sets of parameters (λ for KBCM, λ and η for TL^p-BI). Each set of parameters generates one map of baryceter, therefore a criterion is required to select the parameters for further analysis. In the literature, [28] proves that when $\lambda \Rightarrow 0$ Sinkhorn distance approximates to the original optimal transport distance. However, in practice, when computing barycenter with Sinkhorn regularization, λ is usually set between 0.01 and 0.001. For instance, [29] set λ to $\frac{\text{median}(C)}{60}$, where C is the squared-Euclidean distance matrix on the grid. Similarly, in [51] λ is set to be $\frac{\text{median}(C)}{100}$, where C is the Euclidean distance matrix between pixels in the grid, and $\lambda = \frac{2}{N}, N = 256$ in [12]. In our study, we do not fix the value for parameters but select them arbitrarily. With the artificial data, we have access to the generative model – one small activated region per subject, what we expect is to obtain barycenters that not only correspond to the generative model but also overcome the inter-subject variability. When the noise is weak, barycenters corresponding to our expectation are observed with the maximal amplitude. Additionally, when the maximal amplitude is obtained, the value of λ is 0.001, which corresponds to the setting in the literature. Therefore we set amplitude of barycenter as the criterion for selecting parameters. However, in our further study on artificial data with stronger noise levels, this criterion does not seem to fully hold. For instance, Figure 5.6 show that maps have the maximal amplitude when $\lambda = 0.00001$, but we believe maps obtained by $\lambda = 0.001$ makes more sense. Therefore further work are needed for selecting parameters.

5.5.4 KBCM and TL^p-BI

In this chapter, we introduced a new algorithm to compute a barycenter from a set of images using techniques from the optimal transportation theory. First, we confirm the encouraging results presented in [51] – that were obtained with a different algorithm – i.e that Wasserstein-like barycenters can allow overcoming inter-individual differences in a

population. We also showed that overall, both algorithms, ours and the one of [51], offer equivalent results. It is however to be noted that during our experiment, our algorithm generate barycenters of significantly higher amplitude compared to the ones produced by KBCM, which takes the value of amplitude as an important criterion in [51]. It is however to be noted that during our experiments, our algorithm was on average 12 times faster, making it a good candidate to perform optimal transport-based inference at the population level with statistical tools such as permutation tests. For this, we provide a Python implementation of our TL^p-BI algorithm at https://github.com/SylvainTakerkart/PRNI2018_TLp_bary.

5.6 Conclusion

In this chapter, we introduced a new algorithm to compute a barycenter of a group of unnormalized measures by minimizing the sum of optimal transport distances. We have compared our algorithm, named TL^p-BI, to the one, KBCM, that adds a virtual point to compensate the deficiency of mass of unnormalized measures. Both strategies offer equivalent results. TL^p-BI offers barycenter with higher amplitude and less computational cost, making it a versatile scheme for group level multivariate analyses.

Chapter 6

Conclusion

This thesis focuses on multivariate group analysis that aims at identifying invariant traits in functional neuroimaging data within the population.

We explored multivariate group analysis in two directions. First, unlike univariate method which encodes task-related information in the experimental design, multivariate pattern analysis decodes categories of task from functional patterns using machine learning models. In Chapter 2, we compared two decoding-based multivariate group analysis strategies: the first is the standard method that aggregates within-subject decoding results and a second one that directly seeks to decode neural patterns at the group level in an inter-subject scheme, which we denote as inter-subject pattern analysis (ISPA). The comparison was performed on both artificial and real fMRI datasets, inter-subject pattern analysis offered a higher detection power to detect weak distributed effects and facilitated the interpretation. Therefore, in Chapter 3 we focus on inter-subject pattern analysis. Motivated by the lack of unifying definition of ISPA framework and the lack of ISPA review, in Chapter 3 we introduced a complete formalization of ISPA as a multi-source transductive transfer learning problem. Then we provided a survey of the methods that have been proposed to improve the decoding performance of inter-subject pattern analysis. In Chapter 4 we examined the well-foundedness of our formalization of ISPA through four experimental studies. In each study, decoding performances obtained from different strategies which focused on specific aspect of ISPA were compared. The results demonstrated that there existed large inter-subject variability between individual datasets and our formalization of ISPA as a multi-source transductive transfer learning problem is valuable for future ISPA methods that aim at learning group level decoding principle.

In the other direction, we explored multivariate group analysis through learning an averaging brain image across subjects. Analyzing neuroimaging data at the population level relies on averaging images that have been acquired on a group of individuals drawn from this population. However, traditional univariate group analysis is based on euclidean averaging computed independently at each brain location, which is largely impacted by inter-individual differences. In order to overcome inter-subject variability, we proposed to compute the averaging image of a group of subjects by using optimal transport, which leverages the geometrical properties of multivariate brain patterns. Although when applied to fMRI dataset, whether our approach is better than the univariate method on detecting several sub regions is still a open question, our approach proved effective on artificial datasets to obtain a barycenter overcoming inter-subject variability. Therefore, it is possible to detect regions in fMRI datasets by handling the noise in real fMRI data.

Moreover, these two directions of our research yield to an extension of our work that could implement optimal transport for inter-subject pattern analysis. As reviewed in Section 3.4.1, several studies construct a common space across subjects to align multiple datasets, which further helps improve the decoding performance in inter-subject pattern analysis. Since optimal transport is capable to learn an average image of subjects, which could be taken as the common template across subjects for alignment of individual datasets. Additionally, optimal transport has been applied for domain adaptation which reduced the distribution shifts between training and test sets and improved generalization ability of machine learning models. Therefore it seems that it is possible to adopt optimal transport for multi-source domain adaptation which would reduce the heterogeneity of multiple datasets in ISPA.

Bibliography

[1] Mojtaba Khomami Abadi, Jacopo Staiano, Alessandro Cappelletti, Massimo Zancanaro, and Nicu Sebe. Multimodal engagement classification for affective cinema. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 411–416. IEEE, 2013.

[2] I. V. Accamma, H. N. Suma, and M. Dakshayini. Decoding Multiple Subject fMRI Data using Manifold based Representation of Cognitive State Neural Signatures . *International Journal of Computer Applications*, 115(15), 2015.

[3] Virginia Aglieri, Bastien Cagna, Pascal Belin, and Sylvain Takerkart. InterTVA. A multimodal MRI dataset for the study of inter-individual differences in voice perception and identification. *OpenNeuro*, February 2019.

[4] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[5] Carsten Allefeld, Kai Görgen, and John-Dylan Haynes. Valid population inference for information-based imaging: From the second-level t -test to prevalence inference. *NeuroImage*, 141:378–392, November 2016.

[6] Francis Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *The Journal of Machine Learning Research*, 15(1):595–627, 2014.

[7] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Classification of covariance matrices using a Riemannian-based kernel for BCI applications. *Neurocomputing*, 112:172–178, July 2013.

[8] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. Learning representations from eeg with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*, 2015.

[9] Thomas Bazeille, Hugo Richard, Hicham Janati, and Bertrand Thirion. Local optimal transport for functional brain template estimation. In *International Conference on Information Processing in Medical Imaging*, pages 237–248. Springer, 2019.

[10] Pascal Belin, Robert J. Zatorre, Philippe Lafaille, Pierre Ahad, and Bruce Pike. Voice-selective areas in human auditory cortex. *Nature*, 403(6767):309–312, January 2000. bibtex: belin_voiceselective_2000.

[11] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007.

[12] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[13] Léon Bottou and Chih-Jen Lin. Support vector machine solvers. *Large scale kernel machines*, 3(1):301–320, 2007.

[14] Kay H. Brodersen, Jean Daunizeau, Christoph Mathys, Justin R. Chumbley, Joachim M. Buhmann, and Klaas E. Stephan. Variational Bayesian mixed-effects inference for classification studies. *NeuroImage*, 76:345–361, August 2013.

[15] Kay H Brodersen, Jean Daunizeau, Christoph Mathys, Justin R Chumbley, Joachim M Buhmann, and Klaas E Stephan. Variational bayesian mixed-effects inference for classification studies. *Neuroimage*, 76:345–361, 2013.

[16] Kay H Brodersen, Christoph Mathys, Justin R Chumbley, Jean Daunizeau, Cheng Soon Ong, Joachim M Buhmann, and Klaas E Stephan. Bayesian mixed-effects inference on classification performance in hierarchical data sets. *Journal of Machine Learning Research*, 13(Nov):3133–3176, 2012.

[17] Carlos Cabral, Margarida Silveira, and Patricia Figueiredo. Decoding visual brain states from fMRI using an ensemble of classifiers. *Pattern Recognition*, 45(6):2064–2074, June 2012.

[18] Almudena Capilla, Pascal Belin, and Joachim Gross. The Early Spatio-Temporal Correlates and Task Independence of Cerebral Voice Processing Studied with MEG. *Cerebral Cortex*, 23(6):1388–1395, 05 2012.

[19] Xin Chai, Qisong Wang, Yongping Zhao, Xin Liu, Dan Liu, and Ou Bai. Multi-subject subspace alignment for non-stationary EEG-based emotion recognition. *Technology and Health Care*, 26(S1):327–335, January 2018.

[20] Kai-min Kevin Chang, Tom Mitchell, and Marcel Adam Just. Quantitative modeling of the neural representation of objects: How semantic feature norms can account for fMRI activation. *NeuroImage*, 56(2):716–727, May 2011.

[21] Ian Charest, Rogier A. Kievit, Taylor W. Schmitz, Diana Deca, and Nikolaus Kriegeskorte. Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences*, 111(40):14565–14570, October 2014.

[22] Po-Hsuan Chen, J Swaroop Guntupalli, James V Haxby, and Peter J Ramadge. Joint svd-hyperalignment for multi-subject fmri data alignment. In *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2014.

[23] Po-Hsuan Chen, Xia Zhu, Hejia Zhang, Javier S Turek, Janice Chen, Theodore L Willke, Uri Hasson, and Peter J Ramadge. A convolutional autoencoder for multi-subject fmri data aggregation. *arXiv preprint arXiv:1608.04846*, 2016.

[24] Po-Hsuan (Cameron) Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James Haxby, and Peter J Ramadge. A Reduced-Dimension fMRI Shared Response Model. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 460–468. Curran Associates, Inc., 2015. bibtex: chen_reduced-dimension_2015.

[25] John A. Clithero, David V. Smith, R. McKell Carter, and Scott A. Huettel. Within- and cross-participant classifiers reveal different neural coding of information. *NeuroImage*, 56(2):699–708, May 2011.

[26] Juan Abdon Miranda Correa, Mojtaba Khomami Abadi, Niculae Sebe, and Ioannis Patras. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 2018.

[27] Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. *The Journal of Machine Learning Research*, 9:1757–1774, 2008. bibtex: crammer_learning_2008.

[28] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[29] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.

[30] Saudamini Roy Damarla, Vladimir L Cherkassky, and Marcel Adam Just. Modality-independent representations of small quantities based on brain activation patterns. *Human brain mapping*, 37(4):1296–1307, 2016.

[31] Richard [dataset]Henson, Daniel Wakeman, Vladimir Litvak, and Karl Friston. A Parametric Empirical Bayesian Framework for the EEG/MEG Inverse Problem: Generative Models for Multi-Subject and Multi-Modal Integration. *Frontiers in Human Neuroscience*, 5:76, 2011.

[32] Cyril R. [dataset]Pernet, Phil McAleer, Marianne Latinus, Krzysztof J. Gorgolewski, Ian Charest, Patricia E.G. Bestelmeyer, Rebecca H. Watson, David Fleming, Frances Crabbe, Mitchell Valdes-Sosa, and Pascal Belin. The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage*, 119:164–174, October 2015. bibtex: pernet_human_2015.

[33] 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. *NeuroImage*, 28(3):663 – 668, 2005.

[34] Pierre A Devijver and Josef Kittler. *Pattern recognition: A statistical approach*. Prentice hall, 1982.

[35] J. Dubois, A. O. de Berker, and D. Y. Tsao. Single-Unit Recordings in the Macaque Face Patch System Reveal Limit ations of fMRI MVPA. *Journal of Neuroscience*, 35(6):2791–2802, February 2015.

[36] Eugene P Duff, Aaron J Trachtenberg, Clare E Mackay, Matt A Howard, Frederick Wilson, Stephen M Smith, and Mark W Woolrich. Task-driven ica feature generation for accurate and interpretable prediction using fmri. *Neuroimage*, 60(1):189–203, 2012.

[37] Hedwig Eisenbarth, Luke J Chang, and Tor D Wager. Multivariate brain prediction of heart rate and skin conductance responses to social threat. *Journal of Neuroscience*, 36(47):11987–11998, 2016.

[38] J. A. Etzel. MVPA Permutation Schemes: Permutation Testing for the Group Level. In *2015 International Workshop on Pattern Recognition in NeuroImaging*, pages 65–68, June 2015.

[39] Joset A Etzel, Nikola Valchev, Valeria Gazzola, and Christian Keysers. Is brain activity during action observation modulated by the perceived fairness of the actor? *PLoS One*, 11(1):e0145350, 2016.

[40] Joset A. Etzel, Nikola Valchev, and Christian Keysers. The impact of certain methodological choices on multivariate analysis of fMRI data with support vector machines. *NeuroImage*, 54(2):1159–1167, January 2011.

[41] Fatemeh Fahimi, Zhuo Zhang, Wooi Boon Goh, Tih-Shi Lee, Kai Keng Ang, and Cuntai Guan. Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI. *Journal of Neural Engineering*, 16(2):026007, January 2019.

[42] Yong Fan, Dinggang Shen, and Christos Davatzikos. Detecting cognitive states from fMRI images by machine learning and multivariate classification. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, pages 89–89. IEEE, 2006.

[43] S. Fatima and A. M. Kamboh. Decoding brain cognitive activity across subjects using multimodal M/EEG neuroimaging. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3224–3227, July 2017.

[44] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pages 2960–2967, 2013.

[45] Takuya Fuchigami, Yumi Shikauchi, Ken Nakae, Manabu Shikauchi, Takeshi Ogawa, and Shin Ishii. Zero-shot fMRI decoding with three-dimensional registration based on diffusion tensor imaging. *Scientific Reports*, 8(1):1–11, August 2018.

[46] John D.E. Gabrieli, Satrajit S. Ghosh, and Susan Whitfield-Gabrieli. Prediction as a Humanitarian and Pragmatic Contribution from Human Cogniti ve Neuroscience. *Neuron*, 85(1):11–26, January 2015.

[47] M Gajdos and M Mikl. 23. statistical characteristics of event related and block design datasets. *Clinical Neurophysiology*, 125(5):e32, 2014.

[48] A. Gammerman, V. Vovk, and V. Vapnik. Learning by Transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pages 148–155, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[49] Roee Gilron, Jonathan Rosenblatt, Oluwasanmi Koyejo, Russell A Poldrack, and Roy Mukamel. What's in a pattern? examining the type of signal multivariate analysis uncovers at the group level. *NeuroImage*, 146:113–120, 2017.

[50] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 513–520, 2011. bibtex: glorot_domain_2011.

[51] Alexandre Gramfort, Gabriel Peyré, and Marco Cuturi. Fast optimal transport averaging of neuroimaging data. In *International Conference on Information Processing in Medical Imaging*, pages 261–272. Springer, 2015.

[52] Logan Grosenick, Brad Klingenberg, Kiefer Katovich, Brian Knutson, and Jonathan E. Taylor. Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage*, 72:304–321, May 2013.

[53] Kevin Guittet. Extended kantorovich norms: a tool for optimization. 2002.

[54] J. Swaroop Guntupalli, Michael Hanke, Yaroslav O. Halchenko, Andrew C. Connolly, Peter J. Ramadge, and James V. Haxby. A Model of Representational Spaces in Human Cortex. *Cerebral Cortex*, page bhw068, March 2016. bibtex: guntupalli_model_2016.

[55] 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. *Science*, 293(5539):2425–2430, 2001.

[56] James V. Haxby, Andrew C. Connolly, and J. Swaroop Guntupalli. Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of Neuroscience*, 37(1):435–456, July 2014.

[57] James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

[58] James V. Haxby, J. Swaroop Guntupalli, Andrew C. Connolly, Yaroslav O. Halchenko, Bryan R. Conroy, M. Ida Gobbini, Michael Hanke, and Peter J. Ra-

madge. A Common, High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex. *Neuron*, 72(2):404–416, October 2011.

[59] John-Dylan Haynes. A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, a nd Perspectives. *Neuron*, 87(2):257–270, July 2015. bibtex: haynes_primer_2015.

[60] He He and Dongrui Wu. Transfer Learning for Brain-Computer Interfaces: A Euclidean Space Data Alignment Approach. August 2018.

[61] S. M. Helfinstein, T. Schonberg, E. Congdon, K. H. Karlsgodt, J. A. Mumford, F. W. Sabb, T. D. Cannon, E. D. London, R. M. Bilder, and R. A. Poldrack. Predicting risky choices from brain activity patterns. *Proceedings of the National Academy of Sciences*, 111(7):2470–2475, February 2014.

[62] Sarah M Helfinstein, Tom Schonberg, Eliza Congdon, Katherine H Karlsgodt, Jeanette A Mumford, Fred W Sabb, Tyrone D Cannon, Edythe D London, Robert M Bilder, and Russell A Poldrack. Predicting risky choices from brain activity patterns. *Proceedings of the National Academy of Sciences*, 111(7):2470–2475, 2014.

[63] Pipei Huang, Gang Wang, and Shiyin Qin. Boosting for transfer learning from multiple data sources. *Pattern Recognition Letters*, 33(5):568–579, April 2012. bibtex: huang_boosting_2012.

[64] Keise Izuma, Kazuhisa Shibata, Kenji Matsumoto, and Ralph Adolphs. Neural predictors of evaluative attitudes toward celebrities. *Social cognitive and affective neuroscience*, 12(3):382–390, 2017.

[65] Hojin Jang, Sergey M Plis, Vince D Calhoun, and Jong-Hwan Lee. Task-specific feature extraction and classification of fmri volumes using a deep neural network initialized with a deep belief network: Evaluation using sensorimotor tasks. *NeuroImage*, 145:314–328, 2017.

[66] Jiefeng Jiang, Christopher Summerfield, and Tobias Egner. Visual Prediction Error Spreads Across Object Features in Human Visual Cortex. *The Journal of Neuroscience*, 36(50):12746–12763, December 2016.

[67] Marcel Adam Just, Vladimir L. Cherkassky, Sandesh Aryal, and Tom M. Mitchell. A Neurosemantic Theory of Concrete Noun Representation Based on the Underlying Brain Codes. *PLoS ONE*, 5(1):e8622, January 2010.

[68] Seyed Mostafa Kia, Fabian Pedregosa, Anna Blumenthal, and Andrea Passerini. Group-level spatio-temporal pattern recovery in MEG decoding using multi-task joint feature learning. *Journal of Neuroscience Methods*, 285:97–108, June 2017.

[69] Eunwoo Kim and HyunWook Park. Pairwise Classifier Ensemble with Adaptive Sub-Classifiers for fMRI Pa ttern Analysis. *Neuroscience Bulletin*, 33(1):41–52, February 2017.

[70] Jongwan Kim, Jing Wang, Douglas H Wedell, and Svetlana V Shinkareva. Identifying core affect in individuals from fmri responses to dynamic naturalistic audio-visual stimuli. *PloS one*, 11(9):e0161589, 2016.

[71] Etienne Koechlin and Thomas Jubault. Broca's Area and the Hierarchical Organization of Human Behavior. *Neuron*, 50(6):963–974, June 2006.

[72] Sotetsu Koyamada, Yumi Shikauchi, Ken Nakae, Masanori Koyama, and Shin Ishii. Deep learning of fMRI big data: a novel approach to subject-transfer decoding. *arXiv preprint arXiv:1502.00093*, 2015.

[73] Philip A. Kragel, Leonie Koban, Lisa Feldman Barrett, and Tor D. Wager. Representation, Pattern Information, and Brain Signatures: From Neurons to Neuroimaging. *Neuron*, 99(2):257–273, July 2018.

[74] Nikolaus Kriegeskorte. Pattern-information analysis: from stimulus decoding to computational-model testing. *Neuroimage*, 56(2):411–421, 2011.

[75] Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, 103(10):3863–3868, 2006.

[76] Nikolaus Kriegeskorte and Rogier A. Kievit. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8):401–412, August 2013.

[77] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008.

[78] Eleni Kroupi, Jean-Marc Vesin, and Touradj Ebrahimi. Subject-independent odor pleasantness classification using brain and peripheral signals. *IEEE Transactions on Affective Computing*, 7(4):422–434, 2016.

[79] H. Li, W. Zheng, and B. Lu. Multimodal Vigilance Estimation with Adversarial Domain Adaptation Networks. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, July 2018.

[80] He Li, Yi-Ming Jin, Wei-Long Zheng, and Bao-Liang Lu. Cross-subject emotion recognition using deep adaptation networks. In *International Conference on Neural Information Processing*, pages 403–413. Springer, 2018.

[81] Linling Li, Gan Huang, Qianqian Lin, Jia Liu, Shengli Zhang, and Zhiguo Zhang. Magnitude and Temporal Variability of Inter-stimulus EEG Modulate the Linear Relationship Between Laser-Evoked Potentials and Fast-Pain Perception. *Frontiers in Neuroscience*, 12, 2018.

[82] Martin A. Lindquist, Anjali Krishnan, Marina López-Solà, Mariek e Jepma, Choong-Wan Woo, Leonie Koban, Mathieu Roy, Lauren Y. Atlas, Liane Schmidt, Luke J. Chang, Elizabeth A. Reynolds Losin, Hedwig Eisenbarth, Yoni K. Ashar, Elizabeth Delk, and Tor D. Wager. Group-regularized individual prediction: theory and application to pain. *NeuroImage*, 145:274–287, January 2017.

[83] Alexander Lorbert and Peter J. Ramadge. Kernel hyperalignment. In *Advances in Neural Information Processing Systems*, pages 1790–1798, 2012.

[84] Na Lu, Tengfei Li, Xiaodong Ren, and Hongyu Miao. A deep learning scheme for motor imagery classification based on restricted boltzmann machines. *IEEE transactions on neural systems and rehabilitation engineering*, 25(6):566–576, 2016.

[85] Yun Luo, Si-Yang Zhang, Wei-Long Zheng, and Bao-Liang Lu. WGAN Domain Adaptation for EEG-Based Emotion Recognition. In *Neural Information Processing*, pages 275–286. Springer, Cham, December 2018.

[86] Abdelhak Mahmoudi, Sylvain Takerkart, Fakhita Regragui, Driss Boussaoud, and Andrea Brovelli. Multivoxel Pattern Analysis for fMRI Data: A Review. *Computational and Mathematical Methods in Medicine*, 2012:1–14, 2012. bibtex: mahmoudi_multivoxel_2012.

[87] Loizos Markides and Duncan Fyfe Gillies. Towards identification and characterisation of selective fmri feature sets using independent component analysis. In *Pattern Recognition in NeuroImaging (PRNI), 2012 International Workshop on*, pages 17–20. IEEE, 2012.

[88] KV MARolA, JT KBNT, and JM Bibly. *Multivariate analysis*. AcadeInic Press, Londres, 1979.

[89] Andre F. Marquand, Michael Brammer, Steven C.R. Williams, and Orla M. Doyle. Bayesian multi-task learning for decoding multi-subject neuroimaging data. *NeuroImage*, 92:298–311, May 2014.

[90] Robert A. Mason and Marcel Adam Just. Neural Representations of Physics Concepts:. *Psychological Science*, April 2016.

[91] V. Michel, A. Gramfort, G. Varoquaux, E. Eger, C. Keribin, and B. Thirion. A supervised clustering approach for fMRI-based inference of brain states. *Pattern Recognition*, 45(6):2041 – 2049, 2012.

[92] Vincent Michel, Evelyn Eger, Christine Keribin, and Bertrand Thirion. Multi-class sparse bayesian regression for fmri-based prediction. *Journal of Biomedical Imaging*, 2011:2, 2011.

[93] Vincent Michel, Alexandre Gramfort, Gaël Varoquaux, Evelyn Eger, and Bertrand Thirion. Total variation regularization for fmri-based prediction of behavior. *IEEE transactions on medical imaging*, 30(7):1328–1340, 2011.

[94] Rupert G Miller. Normal univariate techniques. In *Simultaneous statistical inference*, pages 37–108. Springer, 1981.

[95] Tatsuya Mima, Norihiro Sadato, Shogo Yazawa, Takashi Hanakawa, Hidenao Fukuyama, Yoshiharu Yonekura, and Hiroshi Shibasaki. Brain structures related to active and passive finger movements in man. *Brain*, 122(10):1989–1997, October 1999.

[96] Tom M. Mitchell, Rebecca Hutchinson, Radu S. Niculescu, Francisco Pereira, Xuerui Wang, Marcel Just, and Sharlene Newman. Learning to decode cognitive states from brain images. *Machine learning*, 57(1):145–175, 2004.

[97] Hiroshi Morioka, Atsunori Kanemura, Jun-ichiro Hirayama, Manabu Shikauchi, Takeshi Ogawa, Shigeyuki Ikeda, Motoaki Kawanabe, and Shin Ishii. Learning a common dictionary for subject-transfer decoding with resting calibration. *NeuroImage*, 111:167–178, 2015.

[98] Janaina Mourao-Miranda, Emanuelle Reynaud, Francis McGlone, Gemma Calvert, and Michael Brammer. The impact of temporal compression and space selection on svm analysis of single-subject and multi-subject fmri data. *Neuroimage*, 33(4):1055–1065, 2006.

[99] Janaina Mourão-Miranda, Arun L.W. Bokde, Christine Born, Harald Hampel, and Martin Stetter. Classifying brain states and determining the discriminating acti-

vation patterns: Support Vector Machine on functional MRI data. *NeuroImage*, 28(4):980–995, December 2005.

[100] Jeanette A. Mumford, Benjamin O. Turner, F. Gregory Ashby, and Russell A. Poldrack. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, 59(3):2636–2643, February 2012.

[101] Brian Murphy, Massimo Poesio, Francesca Bovolo, Lorenzo Bruzzone, Michele Dalponte, and Heba Lakany. Eeg decoding of semantic category reveals distributed representations for single concepts. *Brain and language*, 117(1):12–22, 2011.

[102] Thomas E. Nichols, Samir Das, Simon B. Eickhoff, Alan C. Evans, Tristan Glatard, Michael Hanke, Nikolaus Kriegeskorte, Michael P. Milham, Russell A. Poldrack, Jean-Baptiste Poline, Erika Proal, Bertrand Thirion, David C. Van Essen, Tonya White, and B. T. Thomas Yeo. Best practices in data analysis and sharing in neuroimaging using mri. *bioRxiv*, 2016.

[103] Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25, 2002.

[104] Kenneth A Norman, Sean M Polyn, Greg J Detre, and James V Haxby. Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in cognitive sciences*, 10(9):424–430, 2006.

[105] Emanuele Olivetti, Seved Mostafa Kia, and Paolo Avesani. MEG decoding across subjects. In *Pattern Recognition in Neuroimaging, 2014 International Workshop on*, pages 1–4. IEEE, 2014.

[106] Emanuele Olivetti, Sriharsha Veeramachaneni, and Ewa Nowakowska. Bayesian hypothesis testing for pattern discrimination in brain decoding. *Pattern Recognition*, 45(6):2075–2084, June 2012.

[107] Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010. bibtex: pan_survey_2010.

[108] Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, 45(1):S199–S209, March 2009.

[109] Cyril R Pernet, Phil McAleer, Marianne Latinus, Krzysztof J Gorgolewski, Ian Charest, Patricia EG Bestelmeyer, Rebecca H Watson, David Fleming, Frances Crabbe, Mitchell Valdes-Sosa, et al. The human voice areas: Spatial organization

and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage*, 119:164–174, 2015.

[110] Russell A Poldrack, Jeanette A Mumford, and Thomas E Nichols. *Handbook of functional MRI data analysis*. Cambridge University Press, 2011.

[111] Rajeev DS Raizada and Andrew C. Connolly. What makes different people's representations alike: neural similarity space solves the problem of across-subject fMRI decodi ng. *Journal of cognitive neuroscience*, 24(4):868–877, 2012.

[112] Christoph Reichert, Robert Fendrich, Johannes Bernarding, Claus Tempelmann, Hermann Hinrichs, and Jochem W. Rieger. Online tracking of the contents of conscious perception using real-time fMRI. *Frontiers in Neuroscience*, 8, 2014.

[113] Jesse Rissman, Henry T Greely, and Anthony D Wagner. Detecting individual memories through the neural decoding of memory states and past experience. *Proceedings of the National Academy of Sciences*, 107(21):9849–9854, 2010.

[114] Yossi Rubner, Leonidas J Guibas, and Carlo Tomasi. The earth mover's distance, multi-dimensional scaling, and color-based image retrieval. In *Proceedings of the ARPA image understanding workshop*, volume 661, page 668, 1997.

[115] Raif M. Rustamov and Leonidas Guibas. Hyperalignment of Multi-subject fMRI Data by Synchronized Projections. In Irina Rish, Georg Langs, Leila Wehbe, Guillermo Cecchi, Kai-min Kevin Chang, and Brian Murphy, editors, *Machine Learning and Interpretation in Neuroimaging: 4th International Workshop, MLINI 2014, Held at NIPS 2014, Montreal, QC, Canada, December 13, 2014, Revised Selected Papers*, pages 115–121. Springer International Publishing, Cham, 2016. DOI: 10.1007/978-3-319-45174-9_12.

[116] Srikanth Ryali, Kaustubh Supekar, Daniel A Abrams, and Vinod Menon. Sparse logistic regression for whole-brain classification of fmri data. *NeuroImage*, 51(2):752–764, 2010.

[117] Srikanth Ryali, Kaustubh Supekar, Daniel A. Abrams, and Vinod Menon. Sparse logistic regression for whole-brain classification of fMRI data. *NeuroImage*, 51(2):752–764, June 2010.

[118] Zeynep M Saygin, David E Osher, Kami Koldewyn, Gretchen Reynolds, John D E Gabrieli, and Rebecca R Saxe. Anatomical connectivity patterns predict face selectivity in the fusiform gyrus. *Nature Neuroscience*, 15(2):321–327, December 2011.

[119] Svetlana V Shinkareva, Vicente L Malave, Robert A Mason, Tom M Mitchell, and Marcel Adam Just. Commonality of neural representations of words and pictures. *Neuroimage*, 54(3):2418–2425, 2011.

[120] Svetlana V. Shinkareva, Robert A. Mason, Vicente L. Malave, Wei Wang, Tom M. Mitchell, and Marcel Adam Just. Using fMRI Brain Activation to Identify Cognitive States Associated with Perception of Tools and Dwellings. *PLoS ONE*, 3(1):e1394, January 2008.

[121] Vishwajeet Singh, K. P. Miyapuram, and Raju S. Bapi. Detection of Cognitive States from fMRI Data Using Machine Learning Techniques. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 587–592, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[122] Johannes Stelzer, Yi Chen, and Robert Turner. Statistical inference and multiple testing correction in classification-based multi -voxel pattern analysis (MVPA): Random permutations and cluster size control. *NeuroImage*, 65:69–82, January 2013.

[123] Johannes Stelzer, Yi Chen, and Robert Turner. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (mvpa): random permutations and cluster size control. *Neuroimage*, 65:69–82, 2013.

[124] Sylvain Takerkart, Guillaume Auzias, Bertrand Thirion, and Liva Ralaivola. Graph-based inter-subject pattern analysis of fMRI data. *PloS one*, 9(8):e104586, 2014.

[125] Tavor, Smith, and Jbabdi. Task-free MRI predicts individual differences in brain activity during task performance. *Science*, 352(6282):213–216, April 2016.

[126] Bertrand Thirion, Guillaume Flandin, Philippe Pinel, Alexis Roche, Philippe Ciuciu, and Jean-Baptiste Poline. Dealing with the shortcomings of spatial normalization: Multi-subject parcellatio n of fMRI datasets. *Human Brain Mapping*, 27(8):678–693, August 2006.

[127] Matthew Thorpe, Serim Park, Soheil Kolouri, Gustavo K Rohde, and Dejan Slepčev. A transportation lp distance for signal analysis. *Journal of mathematical imaging and vision*, 59(2):187–210, 2017.

[128] Michael T. Todd, Leigh E. Nystrom, and Jonathan D. Cohen. Confounds in multivariate pattern analysis: Theory and rule representation case study. *NeuroImage*, 77:157–165, August 2013. bibtex: todd_confounds_2013.

[129] Jia-Jun Tong, Yun Luo, Bo-Qun Ma, Wei-Long Zheng, Bao-Liang Lu, Xiao-Qi Song, and Shi-Wei Ma. Sleep quality estimation with adversarial domain adaptation: From laboratory to real scenario. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[130] Javier S Turek, Cameron T Ellis, Lena J Skalaban, Nicholas B Turk-Browne, and Theodore L Willke. Capturing shared and individual information in fmri data. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 826–830. IEEE, 2018.

[131] Javier S Turek, Theodore L Willke, Po-Hsuan Chen, and Peter J Ramadge. A semi-supervised method for multi-subject fmri functional alignment. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1098–1102. IEEE, 2017.

[132] Marcel van Gerven, Christian Hesse, Ole Jensen, and Tom Heskes. Interpreting single trial data using groupwise regularisation. *NeuroImage*, 46(3):665–676, July 2009.

[133] Cara E Van Uden, Samuel A Nastase, Andrew C Connolly, Ma Feilong, Isabella Hansen, M Ida Gobbini, and James V Haxby. Modeling semantic encoding in a common neural representational space. *Frontiers in neuroscience*, 12:437, 2018.

[134] Gaël Varoquaux. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, 180:68–77, October 2018.

[135] Gaël Varoquaux, Pradeep Reddy Raamana, Denis A. Engemann, Andrés Hoyos-Idrobo, Yannick Schwartz, and Bertrand Thirion. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, 145:166–179, January 2017.

[136] Sandro Vega-Pons, Paolo Avesani, Michael Andric, and Uri Hasson. Classification of inter-subject fMRI data based on graph kernels. In *Pattern Recognition in Neuroimaging, 2014 International Workshop on*, pages 1–4. IEEE, 2014.

[137] Carina Walter, Philipp Wolter, Wolfgang Rosenstiel, Martin Bogdan, and Martin Spüler. Towards cross-subject workload prediction. In *Proceedings of the 6th International Brain-Computer Interface Conference*, 2014.

[138] Jing Wang, Vladimir L Cherkassky, and Marcel Adam Just. Predicting the brain activation pattern associated with the propositional content of a sentence:

modeling neural representations of events and states. *Human brain mapping*, 38(10):4865–4881, 2017.

[139] Qi Wang, Bastien Cagna, Thierry Chaminade, and Sylvain Takerkart. Inter-subject pattern analysis: a straightforward and powerful scheme for group-level mvpa. *bioRxiv*, page 587899, 2019.

[140] Xuerui Wang, Rebecca A. Hutchinson, and Tom M. Mitchell. Training fMRI Classifiers to Discriminate Cognitive States across Multiple Subjects. In *NIPS*, pages 709–716, 2003.

[141] David M Watson, Andrew W Young, and Timothy J Andrews. Spatial properties of objects predict patterns of neural response in the ventral visual pathway. *NeuroImage*, 126:173–183, 2016.

[142] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[143] Keith J Worsley, Alan C Evans, Sean Marrett, and P Neelin. A three-dimensional statistical analysis for cbf activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism*, 12(6):900–918, 1992.

[144] H. Xu, A. Lorbert, P. J. Ramadge, J. S. Guntupalli, and J. V. Haxby. Regularized hyperalignment of multi-set fMRI data. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 229–232, August 2012.

[145] Tonglin Xu, Muhammad Yousefnezhad, and Daoqiang Zhang. Gradient hyperalignment for multi-subject fmri data alignment. In *Pacific Rim International Conference on Artificial Intelligence*, pages 1058–1068. Springer, 2018.

[146] Kentaro Yamada, Yoichi Miyawaki, and Yukiyasu Kamitani. Inter-subject neural code converter for visual image representation. *NeuroImage*, 113:289–297, June 2015. bibtex: yamada_inter-subject_2015.

[147] Muhammad Yousefnezhad and Daoqiang Zhang. Local Discriminant Hyperalignment for multi-subject fMRI data alignment. *arXiv preprint arXiv:1611.08366*, 2016.

[148] Muhammad Yousefnezhad and Daoqiang Zhang. Deep Hyperalignment. In *Advances in Neural Information Processing Systems*, pages 1603–1611, 2017.

[149] Muhammad Yousefnezhad and Daoqiang Zhang. Multi-objective cognitive model: a supervised approach for multi-subject fmri analysis. *Neuroinformatics*, 17(2):197–210, 2019.

[150] Hejia Zhang, Po-Hsuan Chen, Janice Chen, Xia Zhu, Javier S Turek, Theodore L Willke, Uri Hasson, and Peter J Ramadge. A searchlight factor model approach for locating shared information in multi-subject fmri analysis. *arXiv preprint arXiv:1609.09432*, 2016.

[151] Yong-Qi Zhang, Wei-Long Zheng, and Bao-Liang Lu. Transfer components between subjects for eeg-based driving fatigue detection. In *International Conference on Neural Information Processing*, pages 61–68. Springer, 2015.

[152] Wei-Long Zheng and Bao-Liang Lu. Personalizing EEG-based affective models with transfer learning. pages 2732–2738. AAAI Press, July 2016.