Multivariate methods for the joint analysis of neuroimaging and genetic data

Edith Le Floch Le Floch

Université Paris Sud XI, Faculté des Sciences
d'Orsay

École Doctorale STITS
(Sciences et Technologies de l'Information et de la Communication)

Thèse

en vue d'obtenir le grade de

Docteur, spécialité Traitement du Signal

présentée et soutenue par

Edith LE FLOCH

Méthodes multivariées pour l'analyse jointe de données de neuroimagerie et de génétique

Thèse soutenue le 28 Septembre 2012

Composition du jury :

Rapporteurs	Dr Thomas Nichols	Univ. of Warwick, Dept of Statistics, Coventry, UK
	Pr Hervé Abdi	Univ. of Texas, School of Behavioural and Brain Sciences, Dallas, USA
Examinateurs	Dr Jean-Philippe Vert	Mines-ParisTech/Institut Curie, U900, Paris, France
	Dr Stéphane Jamain	Institut Mondor, INSERM U955, Créteil, France
	Pr Jean-Michel Poggi	Univ. Paris-Sud XI, Dép. de Mathématiques, Orsay, France
Encadrant	Dr Edouard Duchesnay	CEA, Neurospin, Saclay, France
Directeur	Vincent Frouin	CEA, Neurospin, Saclay, France
Co-directrice	Dr Monica Zilbovicius	Hôpital Necker, INSERM-CEA U1000, Paris, France

Acknowledgements

Je voudrais d'abord remercier Edouard Duchesnay et Vincent Frouin, mes deux encadrants, pour leur disponibilité, leur patience, leurs précieux conseils, leur confiance et leur soutien tout au long de ces quatre années de thèse. J'espère avoir la chance de continuer à collaborer encore longtemps avec eux... Je voudrais également remercier Monica Zilbovicius ma co-directrice de thèse pour son aide, ses conseils et son regard de médecin toujours très pertinent.

Je voudrais ensuite remercier les membres de mon jury: mes deux rapporteurs Hervé Abdi et Thomas Nichols, ainsi que mes examinateurs Jean-Philippe Vert, Stéphane Jamain et Jean-Michel Poggi.

Je remercie également toutes les personnes avec lesquelles j'ai eu la chance de travailler et d'échanger sur les problématiques qui nous sont chères, et qui m'ont beaucoup apporté: Vincent Guillemot, Arthur Tenenhaus, Laura Trinchera, Jean-Baptiste Poline, Philippe Pinel, Antonio Moreno, Christophe Lalanne, Bertrand Thirion, Thomas Bourgeron, Roberto Toro, Cathy Philippe, Anne-Laure Fouque et Cécilia Damon.

Je voudrais maintenant remercier tous mes collègues et amis de Neurospin avec qui j'ai partagé beaucoup de choses et qui m'ont aidée, écoutée et soutenue pendant ces quatres années, notamment dans les moments un peu plus difficiles. Vous allez me manquer...

Enfin, je voudrais remercier mes amis plus anciens, mes parents, mon frère et Bruno pour avoir cru en moi et m'avoir soutenue (et supportée ;-)) pendant ma thèse et déjà bien avant...

Contents

Résumé

Contexte

L'imagerie cérébrale connaît un intérêt grandissant, en tant que phénotype intermédiaire (ou endophénotype), dans la compréhension du chemin complexe qui relie les gènes à un phénotype comportemental ou clinique. En effet, l'imagerie cérébrale fournit un phénotype quantitatif, que l'on espère plus riche et plus près de la génétique que ne peut l'être un phénotype final tel qu'un statut clinique par exemple.

Ainsi, par analogie avec les études de neuroimagerie qui se composent généralement de quelques dizaines de sujets, on peut espérer que les études d'imagerie génétique puissent découvrir des effets plus importants et nécessitent une taille d'échantillon moins grande que les études génétiques classiques d'association, qui exigent souvent une taille d'échantillon de plusieurs milliers de sujets. Néanmoins, dans ce domaine de l'imagerie génétique, un premier objectif est de proposer des méthodes capables d'identifier la part de variabilité génétique qui explique une certaine part de la variabilité observée en neuroimagerie.

Les études d'imagerie génétique, qui comprennent une grande quantité de données tant du point de vue de l'imagerie que de la génétique, sont en effet confrontées à des défis pour lesquels la communauté de neuroimagerie n'a pas de réponse définitive à ce jour. Les études d'imagerie génétique actuelles sont souvent soit limitées à l'étude de quelques endophénotypes candidats de neuroimagerie versus un grand nombre de variables génétiques telles que les *Single Nucleotide Polymorphisms* (SNP) (par exemple, Furlanello et al. 2003), soit limitées à quelques SNPs candidats versus le cerveau entier (par exemple, McAllister et al. 2006, Roffman et al. 2006, Glahn et al. 2007). En l'absence d'a priori fort sur les régions du cerveau ou du génome impliquées, on peut se diriger vers des méthodes exploratoires. Toutefois, lorsque l'on est confronté à la fois à un grand nombre de variables génétiques et un grand nombre de variables de neuroimagerie, il faut concevoir une stratégie d'analyse appropriée aussi sensible et spécifique que possible.

L'approche exploratoire la plus simple en imagerie génétique est d'appliquer une analyse univariée massive (Stein et al. 2010), que l'on peut appeler modélisation linéaire massivement univariée ou *Massive Univariate Linear Modelling* (MULM). Cependant, si les techniques univariées sont plus simples, elles rencontrent un problème de comparaisons multiples de l'ordre de 10^{12} lorsqu'elles sont appliquées au niveau du génome entier et du cerveau entier. En outre, elles ne tiennent pas compte du fait que le lien entre les données de génétique et d'imagerie est susceptible d'être en partie multivarié. En effet, sur le plan génétique, l'interaction

entre plusieurs loci génétiques (épistasie) ou l'accumulation de plusieurs petits effets sont des phénomènes très probables dans les maladies ou les traits communs (Frazer et al. 2009, Yang et al. 2010). Ainsi, les endophénotypes d'imagerie sont probablement influencés par les effets combinés de plusieurs variants génétiques. De même, du côté de l'imagerie, le phénomène de pléiotropie peut se produire également, ce qui signifie que différentes régions du cerveau peuvent être influencées par le(s) même(s) variant(s) génétique(s). En outre, il peut être intéressant de rendre compte de la nature multivariée des données d'imagerie, telle que la connectivité anatomique ou la co-activation de régions cérébrales, lors de la recherche d'associations d'imagerie génétique.

Une première manière de prendre partiellement en compte la nature multivariée des données génétiques peut être de tester l'effet conjoint de plusieurs SNPs au sein d'un même gène sur les différents voxels du cerveau (Hibar et al. 2011, Kohannim et al. 2011).

D'autres stratégies multivariées ont été également proposées récemment en imagerie génétique, telles que la *Reduced Rank Regression* parcimonieuse (Vounou et al. 2010; 2012), la régression multi-tâche paricmonieuse au niveau du groupe (Wang et al. 2012a;b) ou l'Analyse en Composantes Independantes parallèle (Liu et al. 2009), qui analysent conjointement les deux blocs de données, pour prendre en compte les effets conjoints potentiels qui peuvent exister entre les SNPs ou les covariations possibles entre différentes régions du cerveau. Cependant, dans de grandes dimensions telles que celles des études d'imagerie génétique, les méthodes multivariées peuvent être sujettes à des problèmes de sur-apprentissage, même dans leur version régularisée/parcimonieuse.

Objectifs

Le but de ce travail est de trouver des méthodes qui tiennent compte de la nature potentiellement multivariée des données, localement ou à plus longue échelle, tout en gérant la très grande dimensionnalité du problème afin d'augmenter la puissance de détection.

Nous cherchons tout d'abord à améliorer la sensibilité de l'approche univariée en exploitant la nature multivariée des données de SNPs d'une manière locale, en recherchant soit des clusters de SNPs adjacents associés à un même phénotype soit des clusters 4D dans l'espace $voxel \times SNP$.

Nous essayons ensuite d'aller plus loin et d'identifier un réseau cérébral qui covarie avec un ensemble de polymorphismes génétiques, en utilisant deux méthodes multivariées conçues pour l'analyse conjointe de deux blocs de données: la régression *Partial Least Squares* (moindres carrés partiels) pour deux blocs de données (PLS 2) et l'analyse canonique ou *Canonical Correlation Analysis* (CCA). En outre, nous comparons différentes stratégies de régularisation et de réduction de dimension, combinées avec la PLS 2 ou la CCA, pour faire face à la très grande dimension des données d'imagerie génétique. Nous proposons une étude comparative des différentes stratégies sur un jeu de données simulées tout d'abord, puis sur un jeu de données réelles d'imagerie par résonance magnétique fonctionnelle (IRMf) et de SNPs.

Chapitre 1

Dans ce chapitre, nous présentons le contexte et le but des études d'imagerie génétique. Ensuite, nous décrivons brièvement les différentes modalités de neuroimagerie IRM qui peuvent être utilisées pour de telles études, en mettant l'accent sur les données d'IRM fonctionnelle. Enfin, les données génétiques utilisées dans les études d'imagerie génétique peuvent être de différentes natures, et nous introduisons quelques notions sur l'ADN et la nature de ces données.

La neuroimagerie: un phénotype intéressant

Les études d'imagerie génétique reposent sur l'idée que les données de neuroimagerie peuvent être considérées comme un phénotype intermédiaire intéressant (ou endophénotype) afin de comprendre le chemin complexe entre la génétique et un phénotype comportemental ou clinique.

Un nombre croissant d'études tendent à montrer que certains phénotypes de neuroimagerie sont en effet héréditaires, tels que la quantité de matière grise dans certaines régions cérébrales ou l'activation cérébrale lors d'une tâche particulière. Le premier type d'études sont les études de jumeaux, qui tentent d'évaluer l'héritabilité du phénotype, qui est la part de variabilité phénotypique expliquée par une certaine variabilité génétique. En effet, ils comparent des jumeaux homozygotes avec des jumeaux hétérozygotes, ce qui tient compte par nature des effets environnementaux car les jumeaux sont censés vivre dans le même environnement.

Le deuxième type d'études sont les études de population, où l'on recherche des associations entre des polymorphismes génétiques et un phénotype à travers une population d'individus non apparentés.

Les différentes modalités d'IRM

Les données de neuroimagerie utilisées dans les études d'imagerie génétique peuvent être de différents types. Parmi eux, l'imagerie par résonance magnétique (IRM) a l'avantage d'être non invasive et d'avoir une bonne résolution spatiale. Il existe trois principales modalités IRM. Tout d'abord, l'IRM anatomique peut être utilisée pour distinguer les différents types de tissus: la matière grise faite de corps cellulaires de neurones (le cortex et les noyaux gris centraux), les fibres de la substance blanche faites d'axones reliant les différentes régions du cerveau, et le liquide céphalorachidien qui sert de support et protège les cellules nerveuses.

Deuxièmement, l'IRM de diffusion reflète le mouvement des molécules d'eau dans le cerveau, et est en particulier intéressante pour la reconstruction de fibres de matière blanche.

Enfin, l'IRM fonctionnelle (IRMf) permet une mesure indirecte de l'activation cérébrale au cours d'une tâche spécifique ou au repos. Dans cette thèse, nous nous concentrons sur l'IRMf.

Les données génétiques

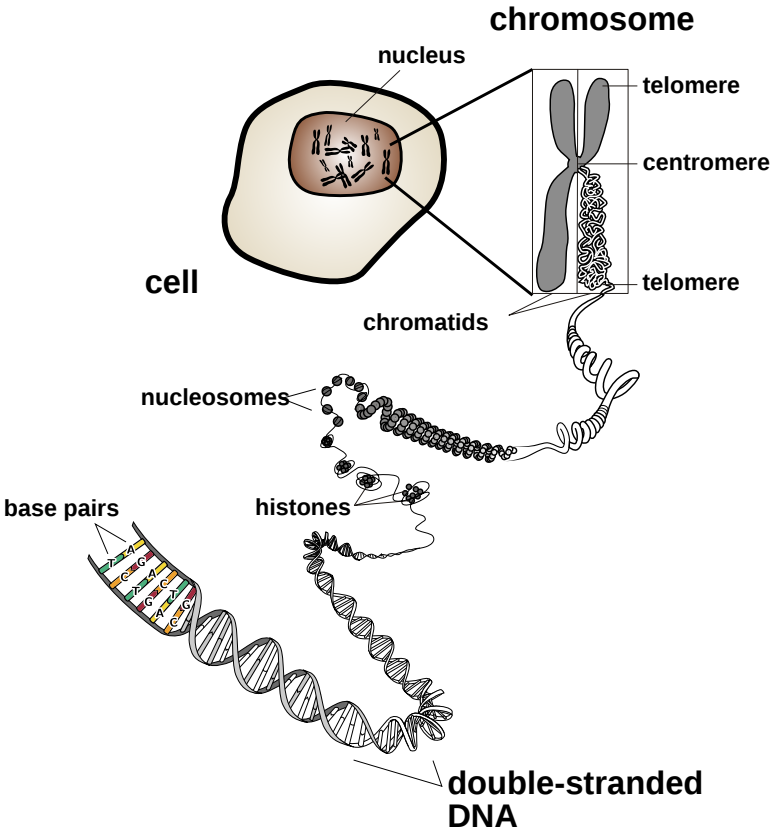
Les données génétiques utilisées dans les études d'imagerie génétique peuvent être de différents types, qui nous allons brièvement décrire. Nous

allons d'abord introduire quelques notions sur l'ADN et décrire la nature de ces données génétiques.

Définitions

L'ADN (acide désoxyribonucléique) est une molécule présente dans toutes les cellules des organismes vivants qui contient le code génétique de l'organisme. Il s'agit d'une macromolécule composée de deux brins complémentaires de nucléotides, chaque brin étant constitué d'environ 3 milliards de nucléotides chez l'homme. Un nucléotide est constitué de trois éléments: un groupe phosphate, un sucre à cinq carbones appelé désoxyribose et une nucléobase. Il existe quatre types de bases nucléiques: l'adénine (A), la thymine (T), la guanine (G) et la cytosine (C).

Figure 1 – *ADN.*



Chez les eucaryotes, qui sont des organismes tels que l'homme, caractérisés par la présence d'un noyau et de mitochondries dans leurs cellules, la molécule d'ADN est divisée en plusieurs segments, que sont les chromosomes. Les humains ont 23 paires de chromosomes, 22 paires de chromosomes autosomiques et une paire de chromosomes sexuels (XX ou XY). Après réplication de l'ADN, chaque chromosome est constitué de deux chromatides identiques reliées par le centromère (voir figure 1 tirée de http://www.genome.gov/Pages/Hyperion//DIR/VIP/Glossary/Illustra-/ chromosome.shtml).

Un gène est un morceau de séquence d'ADN qui code pour un acide ribonucléique (ARN) et qui est ensuite en général traduit en acides aminés pour former une protéine. Cependant, certains gènes codent pour

des ARN fonctionnels qui ne sont pas traduits en protéines. Les humains ont entre 20000 et 25000 gènes, ce qui représente moins de 30% de leur génome. À l'intérieur d'un gène, des séquences appelées introns sont transcrites en ARN précurseur mais sont ensuite éliminées lors de la formation de l'ARN mature par le phénomène d'épissage. Les exons sont les séquences restantes qui sont jointes ensemble pour former un ARN mature et coder la protéine dans le cas d'un gène codant une protéine. Au final, la portion d'ADN codant des protéines représente moins de 2% du génome.

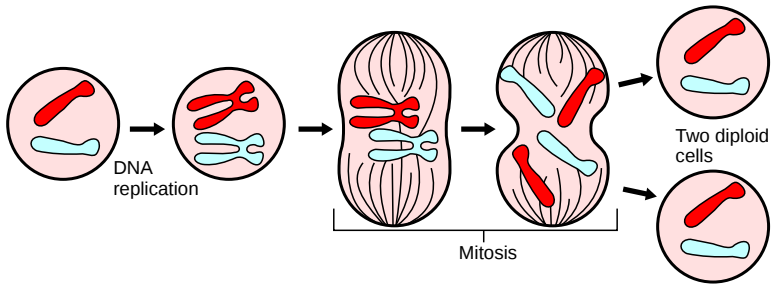
Variabilité de l'ADN

Causes Il existe différentes sources de variabilité de l'ADN humain. La première est la mutation. Les mutations peuvent être causées par un événement externe, tel que les radiations, les virus, les produits chimiques mutagènes. Elles peuvent aussi être la conséquence d'erreurs lors de la réplication de l'ADN, avant la division cellulaire. Elles peuvent être héritées si elles se produisent dans une cellule qui deviendra une cellule sexuelle et sera fécondée par la suite.

Une autre source de variabilité de l'ADN humain peut être liée aux événements chromosomiques qui se produisent lors de la division cellulaire. Il existe deux types de divisions cellulaires: la mitose et la méiose.

La mitose est un ensemble d'événements chromosomiques survenant pendant la duplication cellulaire classique d'une cellule mère à deux cellules filles génétiquement identiques. Lors de la mitose, les deux chromatides de chaque chromosome de la cellule mère se séparent pour former le matériel génétique des deux nouvelles cellules. Après la mitose, l'ADN sera dupliqué dans chaque cellule fille afin de reformer les deux chromatides de chaque chromosome, rendant alors une nouvelle division cellulaire possible. Toutefois, certaines erreurs peuvent se produire lors de la mitose, ce qui entraîne le gain ou la perte d'un chromosome ou d'un segment chromosomique chez les cellules filles (voir figure 2 provenant de http://www.ncbi.nlm.nih.gov/About/primer/genetics_cell.html).

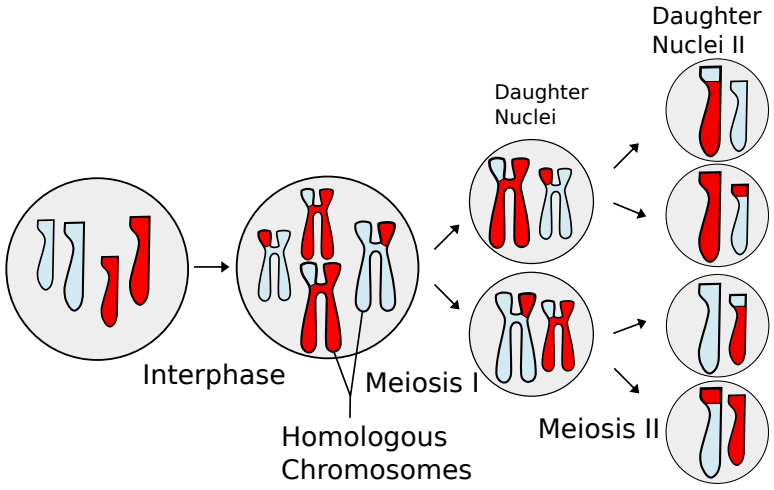
Figure 2 – *Mitose.*



La méiose est le deuxième type de division cellulaire conduisant à la production des cellules sexuelles ou gamètes. Elle est très similaire à la mitose et peut conduire aux mêmes erreurs. Cependant, elle diffère de la mitose à deux égards. Tout d'abord, une recombinaison peut se produire car des segments de chromatides peuvent être échangés entre chromosomes homologues. Ce phénomène est appelé *crossing-over*. En outre, la méiose conduit à quatre cellules filles hap-

loïdes (avec un seul chromosome de chaque paire) appelées gamètes, qui ne pourront pas se diviser à nouveau (voir figure 3 provenant de http://www.ncbi.nlm.nih.gov/About/primer/genetics_cell.html).

Figure 3 – *Méiose et crossing-over.*



Types de variabilité La plupart des variations au niveau de l'ADN sont des polymorphismes nucléotidiques ou *Single Nucleotide Polymorphisms* (SNPs), des petites insertions ou suppressions, ou des insertions et des suppressions plus grandes appelées variations du nombre de copies ou *Copy Number Variants* (CNV).

Un SNP est un variant concernant un seul nucléotide. Selon les connaissances actuelles, il s'agit de la forme la plus courante de la variation de l'ADN. Il y a environ 20 millions de SNPs fréquents et encore plus de rares. Les SNP communs sont généralement bi-alléliques avec deux versions ou allèles, l'allèle le plus fréquent étant appelé l'allèle majeur et le plus rare l'allèle mineur. Ils peuvent être situés dans une région intronique ou exonique, au sein d'un gène qui code pour une protéine ou non, ou même en dehors de tout gène. Même s'il se trouve dans une zone exonique d'un gène codant pour une protéine, un SNP ne modifie pas nécessairement la protéine résultante car il peut être synonyme en raison de la redondance du code génétique. Cependant, il peut aussi mener à un faux-sens (se traduisant par un autre acide aminé) ou un non-sens (conduisant à un arrêt prématuré de la protéine). Cependant, les SNPs synonymes ou ceux qui se trouvent dans des régions non-codantes peuvent encore affecter l'épissage, la régulation transcriptionnelle ou traductionnelle et la stabilité de l'ARN.

Les CNV sont moins nombreux que les SNPs. Cependant, ils comportent beaucoup plus de nucléotides. En effet, un CNV est un polymorphisme du nombre de copies d'un segment d'ADN d'au moins une kilobase (1000 nucléotides), sous forme de délétions ou de duplications.

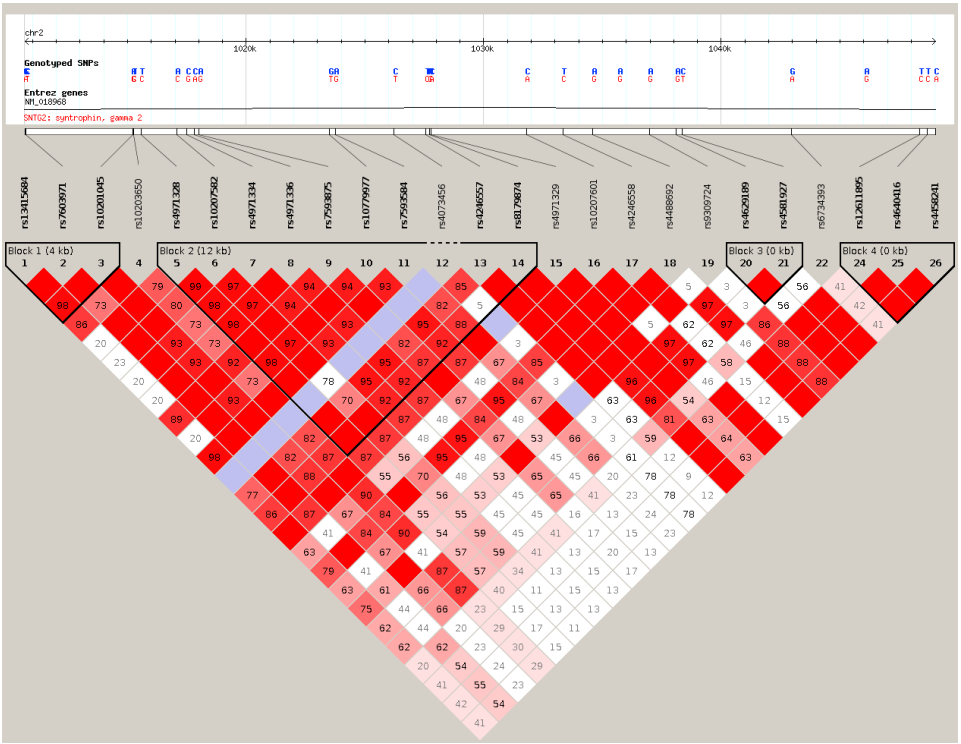
Génotypage des SNP et des CNV Les plateformes actuelles de génotypage à haut débit permettent le génotypage d'un million de SNPs en utilisant la technologie des puces à ADN. L'identification de CNVs peut

également être réalisée en utilisant indirectement les plateformes de génotypage de SNPs, mais la façon la plus courante consiste à utiliser des puces avec des sondes spécifiques aux CNVs. La technologie de séquençage de nouvelle génération sera de plus en plus utilisée à l'avenir, ce qui permettra le génotypage de tous les variants rares.

Déséquilibre de liaison

Il existe une structure de corrélation dans les données génétiques, appelée déséquilibre de liaison ou *linkage disequilibrium* (LD), comme on peut le remarquer sur la figure 4. Le déséquilibre de liaison signifie l'association non aléatoire entre les allèles de deux polymorphismes génétiques, à deux positions distinctes (loci). Cela signifie que certaines paires d'allèles correspondant aux deux loci sont observées plus souvent ensemble sur le même chromosome que ce ne serait prévu par le hasard. Ce phénomène est dû à la liaison physique existant entre les loci voisins sur le même chromosome, appelée liaison génétique, et au fait que leurs allèles sont transmis ensemble d'une génération à l'autre.

Figure 4 – *Déséquilibre de liaison calculé pour le chromosome 2 entre 1010kb et 1050kb en utilisant la base de données HapMap (V3 version 2), et illustré à l'aide du logiciel Haploview.*



L'une des causes de la diminution du LD est la recombinaison, survenant au cours de la méiose. Ainsi, on s'attend à ce que plus la distance physique (en termes de nombre de bases) entre deux polymorphismes est grande, plus la probabilité de recombinaison est importante et donc plus le LD est faible. Toutefois, ce n'est pas toujours le cas. En effet, l'étendue du LD varie généralement entre quelques kilobases et quelques dizaines de kilobases chez l'homme, mais elle peut atteindre quelques centaines de

kilobases dans certaines régions de l'ADN. Ceci est dû au fait que le taux de recombinaison n'est pas constant le long du génome. Il existe certaines régions avec un taux élévé de recombinaison, appelées points chauds, et d'autres avec un faible taux de recombinaison, appelées blocs haplotypiques, de 10 à 20 kilobases en moyenne jusqu'à quelques centaines de kilobases. Certaines régions sont même presque non-recombinantes telles que le chromosome Y, certaines parties du chromosome X et les régions proches des centromères des chromosomes autosomiques. Cela conduit à un autre type de mesure de distance, appelée distance génétique, qui dépend de la fréquence de recombinaison entre deux loci et dont l'unité est la centimorgan (cM). Un centimorgan correspond à une fréquence de recombinaison de 0.01 d'une génération à l'autre. Par définition, la distance physique correspondant à un centimorgan varie le long du génome en fonction du LD, mais en moyenne un centimorgan représente environ 1000 kilobases chez l'homme.

Un aspect intéressant du LD peut être que cela réduit le nombre de polymorphismes génétiques nécessaires pour capturer la plupart de la variabilité génétique, ce qui implique que l'on ne doit génotyper qu'un nombre réduit de SNPs indépendants et très informatifs, appelés tagSNPs. Par exemple, 99% des SNPs communs (avec une fréquence de l'allèle mineur> 5%) sont liés avec un LD $r^2 > 0.8$ à un million de tagSNPS. La plupart des puces commerciales utilisent ces tagSNPs.

Une conséquence est que dans la plupart des cas, quand un SNP se trouve être associé à un phénotype donné, il ne s'agit pas du SNP causal lui-même mais d'un SNP en LD avec le SNP causal non génotypé.

Chapitre 2

Dans le chapitre précédent, nous avons brièvement détaillé les différentes modalités d'IRM et les différents types de données génétiques qui sont couramment utilisés dans les études d'imagerie génétique. Avant de passer à l'analyse conjointe de ces deux types de données, nous faisons dans ce chapitre une revue des méthodes d'analyse statistique qui sont classiquement utilisées dans chaque domaine, respectivement. En fait, nous montrons que les méthodes d'analyse statistique sont très similaires en neuroimagerie et en génétique. Nous présentons d'abord l'approche classique univariée et son application dans les deux domaines. Ensuite, nous présentons quelques approches multivariées communes aux deux domaines. Enfin, nous montrons que les deux types de données sont en général de très grande dimension et nécessitent une étape préliminaire de réduction de dimension.

L'approche classique: l'analyse massivement univariée

Lors de la recherche d'associations entre une première variable d'intérêt (comme une variable comportementale ou d'un phénotype donné) et plusieurs autres variables d'un autre type (comme des données de neuroimagerie ou des données génétiques), l'approche la plus simple et la plus classique est d'analyser cette première variable par rapport à cha-

cune des autres variables de façon indépendante. Une telle approche est appelée une analyse massivement univariée.

En ne considérant qu'une seule paire de variables à la fois, on peut d'abord distinguer deux cas: lorsque les deux variables sont quantitatives et quand au moins une variable est qualitative. Dans le premier cas, une régression linéaire est classiquement utilisée, tandis que dans le second cas on effectue plutôt des tests t, des tests F, ou des tests du chi-deux. Dans cette thèse, nous nous concentrons sur le premier cas où toutes les variables sont considérées comme quantitatives.

Application en neuroimagerie: l'inférence classique

L'approche massivement univariée est le moyen le plus classique pour analyser des données de neuroimagerie, où l'on effectue un test statistique pour chaque voxel de l'image ou pour chaque caractéristique d'intérêt extraite de l'image, en recherchant des associations avec la variable comportementale d'intérêt. Cette approche est appelée inférence classique ou *Voxel-Based Analysis* lorsque l'analyse est effectuée au niveau du voxel.

Les données de neuroimagerie sont classiquement des données quantitatives, tandis que la variable comportementale ou phénotypique peut être quantitative ou qualitative. Lorsque cette variable comportementale est quantitative, une régression linéaire sera utilisée, en ajustant un modèle linéaire pour chaque voxel/caractéristique d'intérêt (la variable cible à expliquer) avec la variable de comportement comme prédicteur. A l'inverse, si la variable de comportement se trouve être catégorielle, on préfèrera un test t à deux échantillons ou un test F.

Application en génétique

L'approche massivement univariée est également l'approche la plus classique dans les études génétiques. Elle peut s'appliquer sur des données de familles, avec une couverture relativement peu dense des polymorphismes génétiques, en utilisant l'analyse de liaison, qui analyse la co-ségrégation de chaque polymorphisme génétique avec un caractère d'intérêt et est basée sur une statistique univariée appelée *LOD score* (*logarithm of odds*). Avec le développement des nouvelles techniques de génotypage à haut débit, on a évolué vers des études d'association sur des individus non apparentés avec une couverture plus dense de la variabilité du génome, où l'association entre le phénotype et chaque polymorphisme génétique (chaque SNP par exemple) est testé indépendamment. Dans le cas d'une couverture des polymorphismes génétiques de l'ensemble du génome, cette approche est appelée *Genome-Wide Association Study* (GWAS).

Les données de SNPs peuvent être considérées comme des données quantitatives ou qualitatives. En effet, pour chaque SNP bi-allélique (par exemple avec un allèle majeur A et un allèle mineur T), les génotypes possibles peuvent être codés soit AA, AT et TT, soit 0, 1 et 2 lorsqu'on utilise un codage génétique additif où le génotype représente le nombre d'allèles mineurs. Lorsqu'ils sont confrontés à un phénotype quantitatif, les SNPs sont souvent considérés comme des variables quantitatives (en supposant des effets génétiques additifs) et sont successivement testés

comme prédicteurs potentiels du phénotype (la variable cible), en utilisant un modèle de régression linéaire simple. A l'inverse, dans le cas d'un phénotype catégoriel, ils sont plutôt considérés comme des variables catégorielles et l'analyse génétique est classiquement effectuée à l'aide de tests du chi-deux. Dans cette thèse, nous considérons les SNPs comme des variables quantitatives, en supposant un modèle génétique additif. Néanmoins, différents modèles génétiques, tels que les modèles dominant, récessif ou génotypique, pourraient également être étudiés dans des travaux ultérieurs.

Limitations

Bien que les techniques univariées soient relativement simples, elles rencontrent un problème de comparaisons multiples. En effet, la p-valeur obtenue pour chaque test doit être corrigée pour le nombre de tests effectués qui peut être très élevé tant en neuroimagerie qu'en génétique. Par exemple, la correction de Bonferroni est la correction la plus commune et contrôle le taux d'erreur en divisant le seuil de significativité par le nombre de tests effectués. Cette approche est très stricte et suppose que les tests sont indépendants. Une correction moins sévère telle que le *False Discovery Rate* peut également être utilisée. Enfin on peut aussi utiliser une correction basée sur des permutations lorsque la distribution des variables n'est pas connue, mais cela nécessite des calculs plus intensifs.

En outre, le lien entre la variable comportementale (phénotype) et les données de neuroimagerie ou de génétique est susceptible d'être en partie multivarié. Par exemple, l'épistasie est un phénomène fréquent dans les traits ou maladies communs, ce qui signifie que le phénotype est susceptible d'être influencé par les effets combinés de plusieurs SNPs ou gènes. De même, les différentes régions du cerveau peuvent également être corrélées et associées à une même variable comportementale.

Approches multivariées

Afin de remédier à ces limitations de l'approche univariée, les approches multivariées sont également devenues populaires en neuroimagerie et en génétique. Le but de ces méthodes est de prédire une variable cible (une variable comportementale ou un phénotype donné, par exemple) à partir de plusieurs variables prédictives quantitatives telles que des données de neuroimagerie ou de génétique. Nous notons $\mathbf{x}_1$, $\mathbf{x}_2$, ... , $\mathbf{x}_p$ les p variables prédictives et $\mathbf{y}$ la variable cible à expliquer. $\mathbf{X}$ sera la matrice des prédicteurs de taille $n \times p$ et $\mathbf{y}$ la cible de taille $n \times 1$, n étant le nombre d'observations.

Les méthodes multivariées prédictives peuvent être divisées en plusieurs catégories. Par exemple, on peut distinguer les méthodes de classification dans le cas d'une variable comportementale (phénotype) qualitative et les méthodes de régression dans le cas d'une variable quantitative.

Modèles de régression linéaire

Comme pour l'approche univariée, nous nous concentrons sur le cas où toutes les variables sont considérées comme quantitatives, c'est-à-dire sur les méthodes de régression, et plus particulièrement sur les modèles linéaires.

Le modèle de régression linéaire multiple recherche le vecteur des coefficients β de taille $p \times 1$ qui minimise l'erreur quadratique entre $\mathbf{y}$ et son estimation $\mathbf{X}\beta$:

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \tag{1}$$

Cependant, quand le nombre d'observations n est trop faible par rapport au nombre de prédicteurs p, l'estimation de β est sujette au problème de sur-apprentissage. Afin de limiter ce phénomène, on peut alors utiliser des techniques de régularisation en appliquant une contrainte sur β, telles que la régularisation Ridge (Hoerl and Kennard 1970), la régularistion Lasso (Tibshirani 1996) ou la régularisation Elastic Net (Zou and Hastie 2005).

Application en neuroimagerie: l'inférence inverse En neuroimagerie, de telles approches multivariées sont appelées inférence inverse. En effet, au lieu d'essayer d'expliquer le signal dans chaque voxel en utilisant par exemple une variable comportementale, les différents voxels (les prédicteurs) sont utilisés simultanément pour prédire la variable comportementale (la cible).

L'inférence inverse doit faire face au problème du sur-apprentissage en raison de la très grande dimension des données d'imagerie, avec jusqu'à un million de voxels par image. On se trouve dans le cas typique où $n \ll p$ et où la régularisation est nécessaire. Récemment, la régression régularisée comme la régression Elastic Net a été utilisée avec succès dans un but de prédiction de données comportementales à partir de données d'IRMf (Carroll et al. 2009).

Application en génétique De même en génétique, on peut essayer d'expliquer un phénotype donné (la cible) en utilisant les effets conjoints de plusieurs SNPs (les prédicteurs). Les puces actuelles contiennent jusqu'à un million de SNPs sur l'ensemble du génome, ce qui conduit au même phénomène de sur-apprentissage qu'en neuroimagerie. La régression régularisée, comme la régression Lasso, a également été appliquée avec succès dans le cas des études d'association, afin d'identifier les polymorphismes génétiques associés à un phénotype donné (Shi et al. 2011).

Méthodes d'analyse pour deux blocs de données

Dans cette section, nous considérons plusieurs variables cibles formant une matrice $\mathbf{Y}$ de taille $n \times q$, où q est le nombre de variables cibles considérées.

Régression PLS et analyse canonique Afin d'analyser conjointement deux blocs de données, plusieurs techniques ont été développées, telles

que la régression PLS à deux blocs (PLS2) et l'analyse canonique, qui sont décrites dans le chapitre 5. Elles ont été appliquées avec succès dans le domaine de la génétique, afin d'analyser conjointement l'expression de gènes et des données de CNVs par exemple.

PLS-SVD Une variante de la PLS2, appelée analyse inter-batterie de Tucker (Tucker 1958) ou PLS-SVD (McIntosh et al. 1996), a également été beaucoup appliquée dans le domaine de la neuroimagerie. Cette variante est symétrique et consiste à calculer la décomposition en valeurs singulières (SVD) de $\mathbf{XY}$ en une seule fois. Toutes les paires de vecteurs singuliers à gauche et à droite forment les vecteurs de poids $\mathbf{a}_h$ et $\mathbf{b}_h$ pour les variables de $\mathbf{X}$ et de $\mathbf{Y}$ respectivement. Les combinaisons linéaires ainsi obtenues, $\mathbf{Xa_h}$ et $\mathbf{Yb_h}$, sont appelées variables latentes. La PLS-SVD donne les mêmes résultats que la régression PLS sur la première paire de variables latentes, mais diffère sur les paires suivantes en raison d'une contrainte d'orthogonalité différente. Tandis que la régression PLS force les variables latentes successives de chaque bloc à être orthogonales, la PLS-SVD force les vecteurs de poids successifs de chaque bloc à être orthogonaux.

Réduction de Dimension

Le principal problème rencontré par les méthodes multivariées est le phénomène de sur-apprentissage, qui se produit en grandes dimensions. Lorsque la régularisation n'est pas suffisante pour faire face au sur-apprentissage, on peut la combiner avec une étape préliminaire de réduction de dimension, comme c'est couramment le cas en génétique et en neuroimagerie.

La réduction de dimension est basée essentiellement sur deux paradigmes: l'extraction de caractéristiques et la sélection de variables. Les méthodes d'extraction de caractéristiques telles que l'analyse en composantes principales (ACP) ou l'analyse en composantes indépendantes recherche une représentation des données de faible dimension, tandis que les méthodes de sélection de variables telles que le filtrage univarié vise à éliminer les variables inutiles.

Chapitre 3

Dans le chapitre précédent, nous avons montré la similitude entre les méthodes d'analyse statistique qui sont classiquement utilisées en neuroimagerie et la génétique, respectivement. Dans ce chapitre, nous allons maintenant passer aux méthodes existantes pour l'analyse conjointe de ces deux types de données de grande dimension, qui sont essentiellement des extensions des méthodes du chapitre précédent au cas de plusieurs variables cibles. Nous allons d'abord décrire l'approche classique univariée puis des approches multivariées, à savoir la régression sur composantes principales, la régression pénalisée et les méthodes multivariées à deux blocs.

Désormais, nous notons $\mathbf{Y}$ la matrice de données de neuroimagerie de taille $n \times q$ et $\mathbf{X}$ la matrice de SNPs de taille $n \times p$, n étant le nombre de

sujets, q le nombre de phénotypes de neuroimagerie et p le nombre de SNPs. Les phénotypes d'imagerie sont des variables quantitatives. De même, les SNPs sont considérés comme des variables quantitatives, en supposant un modèle génétique additif.

L'approche classique: analyse massivement univariée

Voxelwise Genome-Wide Association Studies

Lors de la recherche d'associations entre deux ensembles de variables, telles que des données de neuroimagerie et de génétique, l'approche la plus simple est à nouveau d'effectuer une analyse univariée massive, où l'on teste chaque variable du premier groupe par rapport à chaque variable du second groupe de façon indépendante.

On utilise classiquement un modèle de régression linéaire simple pour régresser chaque voxel (ou tout autre phénotype extrait de l'image) sur chaque SNP, en supposant un modèle génétique additif (Stein et al. 2010). Toutefois, les SNPs peuvent également être considérés comme des données catégorielles, et on peut utiliser un test F sans aucune hypothèse sur le modèle génétique ou un test t si l'on suppose un modèle dominant ou récessif.

Limitations

Les limites de l'approche univariée sont les mêmes que celles mentionnées dans le chapitre précédent, lorsque l'on avait une seule variable cible.

La première est la question de comparaisons multiples, qui est encore plus importante en imagerie génétique qu'en neuroimagerie ou en génétique. En effet, il faut corriger pour le nombre de tests effectués ($p \times q$) dans le cas des études d'associations cerveau entier et génome entier, qui peut atteindre 10^{12}. La correction de Bonferroni est la procédure la plus courante, mais est très conservatrice, en particulier lorsque les variables sont corrélées, comme c'est le cas pour les SNPs en raison du déséquilibre de liaison. Ainsi, (Stein et al. 2010) a proposé une correction moins stricte basée sur une ACP des données de SNPs, afin de déterminer le nombre effectif de tests indépendants en fonction du nombre de composantes principales nécessaire pour expliquer 99.5% de la variance des SNPs. Cependant, cela n'a pas suffi dans leur exemple pour que les SNPs passent le seuil de significativité corrigée.

La seconde limitation est qu'une telle approche ne tient pas compte de l'aspect potentiellement multivarié du lien entre les données de génétique et d'imagerie, alors que l'épistasie ou la pléiotropie sont des phénomènes fréquents dans les traits communs. En effet, les endophénotypes d'imagerie cérébrale sont probablement influencés par les effets combinés de plusieurs SNPs, et différentes régions du cerveau peuvent aussi être corrélées et influencées par le(s) même(s) SNP(s).

Approches multivariées

Afin de répondre aux limites de l'analyse univariée, certaines méthodes multivariées ont été récemment utilisées dans les études d'imagerie géné-

tique. Nous présentons d'abord deux exemples classiques de régression, la régression sur composantes principales et la régression régularisée, où la nature multivariée des données génétiques a été prise en compte. Ensuite, nous passons aux méthodes multivariées à deux blocs, qui sont conçues pour l'analyse conjointe de deux blocs de données et tirent profit de l'aspect multivarié au sein de chaque bloc.

Méthodes multivariées classiques avec un seul bloc de variables

Les méthodes multivariées classiques, comme celles présentées dans le chapitre 2, sont généralement conçues pour l'analyse d'un bloc de variables explicatives par rapport à une seule variable cible. Néanmoins, ces méthodes peuvent aussi être intéressantes pour l'analyse de deux ensembles de variables $\mathbf{Y}$ et $\mathbf{X}$, telles que des données de neuroimagerie et de génétique, car elles peuvent être appliquées pour chaque variable cible de $\mathbf{Y}$ (versus les variables du bloc $\mathbf{X}$) de façon indépendante. Par exemple, la régression sur composantes principales (Hibar et al. 2011) et la régression régularisée (Kohannim et al. 2011) ont été utilisées en imagerie génétique, afin d'étudier les effets combinés des différents SNPs d'un même gène ou d'une même région génomique sur un phénotype d'imagerie.

Limites La première limite de ces deux approches est qu'elles ne tiennent pas compte des effets multivariés à longue distance entre les SNPs. En effet, seules les interactions entre les SNPs au sein du même gène ou dans la même région sont prises en compte.

De plus, elles n'exploitent pas la nature multivariée des données de neuroimagerie, puisque chaque phénotype d'imagerie est analysé de façon indépendante.

Enfin, la question des comparaisons multiples reste critique, car il faut toujours tenir compte du nombre de gènes (ou régions génomiques) et de phénotypes d'imagerie testés.

Méthodes multivariées à deux blocs

Pour bien prendre en compte de l'aspect multivarié au sein de chaque bloc de données, on peut alors utiliser des méthodes multivariées à deux blocs. Nous présentons ici trois méthodes qui ont montré des résultats prometteurs sur des données d'imagerie génétique: l'analyse en composantes indépendantes parallèle, la régression *Reduced Rank* parcimonieuse et la régression multi-tâches parcimonieuse en termes de groupes.

Régression *Reduced Rank* parcimonieuse Nous rappelons d'abord les notations $\mathbf{Y}$ (de taille $n \times q$) pour les q phénotypes d'imagerie et $\mathbf{X}$ (de taille $n \times p$) pour les p SNPs. Nous allons d'abord décrire le modèle de régression *Reduced Rank* (RRR) introduit par Anderson (1951), puis sa version parcimonieuse introduite par Vounou et al. (2010).

Afin de mieux comprendre la régression *Reduced Rank*, rappelons d'abord le modèle de régression multiple multivarié classique:

$$\mathbf{Y} = \mathbf{XC} + \mathbf{E} \tag{2}$$

où **C** est la matrice des coefficients de régression à estimer de taille $p \times q$, et **E** la matrice des erreurs résiduelles de taille $n \times q$. La solution classique des moindres carrés équivaut à réaliser q régressions linéaires multiples, pour chaque phénotype indépendamment.

L'idée de la RRR repose sur le fait d'imposer que la matrice des coefficients **C** soit de rang réduit, ce qui revient à considérer la nature multivariée des phénotypes.

Vounou et al. (2010) propose une version régularisée de la RRR qui mène à une solution parcimonieuse. Ils appliquent alors la RRR parcimonieuse sur des données simulées d'imagerie génétique et montrent des performances supérieures à l'approche univariée classique. Cependant ils ne proposent pas de validation des résultats obtenus par cette méthode sur des données réelles.

Régression multi-tâches parcimonieuse en termes de groupes Une autre façon de prendre en compte la nature multivariée des variables cibles (les phénotypes d'imagerie) dans un cadre de régression multiple multivariée, tout en imposant une certaine parcimonie sur les prédicteurs (SNPs), est de forcer un parcimonie structurée de la matrice des coefficients **C**. En effet, on peut utiliser une forme de régularisation qui couple la sélection d'un prédicteur au travers de toutes les cibles (ou tâches) (Argyriou et al. 2007, Obozinski et al. 2010). On parle alors de régression multi-tâches.

De plus, Wang et al. (2012a) proposent de rajouter une autre forme de régularisation à la régression multi-tâches, en imposant de la parcimonie en termes de groupes de prédicteurs. Ils appliquent cette méthode sur des données de SNPs et d'IRM anatomique, et montrent des performances améliorées par rapport aux méthodes multivariées classiques telles que la régression multiple mutivariée.

Analyse en composantes indépendantes parallèle Liu et al. (2009) proposent un nouvel algorithme appelé parallèle ICA, qui cherche des composantes indépendantes pour chaque bloc de données, tout en essayant de maximiser la corrélation entre chaque paire de composantes. Ils appliquent cette méthode sur des données de SNP et d'IRM fonctionnelle mais ne proposent pas de validation de leurs résultats.

Chapitre 4

Comme mentionné précédemment, les études classiques en imagerie génétique sont souvent basées sur un test massivement univarié de chaque voxel de l'image versus chaque donnée génétique, ce qui conduit à un énorme problème de comparaisons multiples. Dans le domaine de la neuroimagerie, des stratégies visant à limiter le nombre de comparaisons multiples ont été proposées en se basant sur la détection des groupes de voxels contigus activés, ce qui peut augmenter la sensibilité au dépend de la localisation anatomique. Dans ce chapitre, nous décrivons d'abord brièvement l'inférence au niveau du cluster en neuroimagerie. Ensuite, nous présentons une stratégie similaire sur les données génétiques, en

recherchant des groupes de SNPs associés à un même phénotype, et montrons des résultats très préliminaires. Enfin, nous montrons que cette idée pourrait être étendue à la détection de clusters 4D dans l'espace voxel $\times$ SNP.

Inférence au niveau du cluster 3D en neuroimagerie

Concept

En neuroimagerie, l'inférence au niveau du cluster a été introduite par Poline and Mazoyer (1993) et Roland et al. (1993), et se fonde sur l'idée qu'il peut être plus puissant de détecter des groupes de voxels contigus actifs plutôt que des voxels actifs isolés. En effet, en effectuant une inférence au niveau du cluster, on peut réduire le nombre de tests, car le nombre de clusters est beaucoup plus faible que le nombre de voxels. Par ailleurs, on peut penser qu'un groupe de voxels contigus actifs reflète la structure anatomique sous-jacente du cerveau et est donc moins susceptible d'être du bruit qu'un signal à l'échelle du voxel. Enfin, des groupes de voxels contigus actifs peuvent être plus faciles à interpréter en termes de régions anatomiques par rapport à de voxels isolés.

Méthode

La première étape de l'inférence au niveau du groupe consiste à effectuer un test massivement univarié du signal de chaque voxel par rapport à la variable cible, comme dans l'inférence classique.

Les cartes statistiques sont ensuite seuillées à une valeur donnée u, qui correspond classiquement à une p-valeur de 10^{-3}, et on peut alors identifier des groupes de voxels contigus avec une statistique au-dessus de ce seuil.

Enfin, la taille des clusters de voxels devient la nouvelle statistique d'intérêt afin de détecter des activations.

Le degré de significativité des tailles de cluster peut alors être évalué en calculant la distribution sous l'hypothèse nulle de la taille de cluster, en utilisant plusieurs techniques.

La théorie des champs aléatoires est la méthode la plus largement utilisée (Friston et al. 1993), elle nécessite quelques hypothèses telles qu'une image lisse, un lissage uniforme, et un seuil statistique suffisamment élevé (Worsley et al. 1992). De plus, afin de corriger pour le nombre de clusters testés, on peut utiliser une approche maxT pour estimer la distribution sous l'hypothèse nulle de la statistique maximale sur les clusters (au lieu de la distribution sous l'hypothèse nulle de la statistique proprement dite) et comparer les tailles de clusters avec cette distribution afin d'obtenir des p-valeurs corrigées.

Une autre approche basée sur des permutations a été introduite par Holmes et al. (1996), ce qui ne nécessite pas d'hypothèses sur la distribution sous l'hypothèse nulle de la taille du cluster. Pour obtenir des p-valeurs corrigées pour le nombre de clusters testés, l'approche maxT peut être effectuée par permutations, en dérivant des permutations la distribution sous l'hypothèse nulle de la statistique maximale sur les clusters.

Cas génétique: clusters 1D

Concept

Comme pour l'inférence au niveau du cluster en neuroimagerie, nous avons essayé d'identifier des groupes de SNPs successifs le long du génome qui seraient corrélés à un même phénotype d'imagerie. En effet comme mentionné précédemment, il existe une structure de corrélation sous-jacente entre SNPs voisins, similaire aux corrélations entre voxels, appelée déséquilibre de liaison.

En outre, il existe déjà des techniques qui tentent de combiner les p-valeurs obtenues par des SNPs adjacents, basées sur l'idée qu'une telle combinaison sera plus importante et plus pertinente biologiquement que le fait de considérer les SNPs indépendamment. Par exemple, la méthode de Tippett (la p-valeur minimale parmi plusieurs SNPs), la méthode de Fisher (qui calcule le produit des différentes p-valeurs) et la méthode de Stouffer (qui moyenne les valeurs en Z correspondantes) sont actuellement utilisées pour combiner la significativité de plusieurs SNPs. Certaines contributions plus récentes de Liang and Kelemen (2008), Neale and Sham (2004), Hoh and Ott (2000) proposent également une série de tests basés sur l'agrégation de p-valeurs.

Méthode

La première étape de notre méthode consiste à ajuster un modèle linéaire additif pour chaque SNP k par rapport au phénotype d'intérêt, et à calculer une p-valeur (p_k) pour chaque SNP k .

Nous appliquons ensuite une opération morphologique de fermeture en niveaux de gris à $-log_{10}(p_k)$ le long de l'axe des SNPs, qui est en quelque sorte similaire à l'étape de lissage des données d'IRM.

L'étape suivante consiste à détecter des groupes de SNPs successifs avec des p-valeurs inférieures à un seuil donné α.

Afin de prendre en compte la non-stationnarité le long de la région génétique due à la non-homogénéité du LD, nous avons décidé d'utiliser une statistique locale en chaque SNP: la taille du cluster auquel il appartient.

La significativité de cette statistique a ensuite été estimée par permutations pour chaque SNP. Nous avons effectué 100000 permutations afin d'obtenir des p-valeurs empiriques précises.

Si l'utilisation de la statistique de taille de cluster reportée en chaque SNP permet d'obtenir par permutations une p-valeur non-corrigée localement pour chaque SNP, la non-stationnarité reste un problème lorsque l'on souhaite corriger pour les comparaisons multiples. En effet, la correction maxT ne gère pas la non-stationnarité. Elle donne de bons résultats lorsque la distribution sous l'hypothèse nulle de la statistique est la même pour toutes les variables mais favorise les régions lisses sinon. Ainsi, nous avons utilisé à la place une autre procédure de ré-échantillonnage afin de corriger pour les comparaisons multiples, appelée correction minP (Westfall and Young (1993), Dudoit et al. (2004)), qui gère à la fois la dépendance entre les variables et de l'inhomogénéité de la statistique. Elle consiste à déduire à partir des permutations la distribution sous l'hypothèse nulle de

la p-valeur minimale (au travers des SNPs) de la taille de cluster reportée en chaque SNP, puis à en dériver les p-valeurs corrigées.

Résultats sur un jeu de données réelles

Après avoir obtenu des résultats prometteurs sur données simulées, nous avons appliqué cette méthode sur un jeu de données réelles avec comme phénotype un indexe de latéralisation issu de données d'IRMf lors d'une tâche de lecture, et les SNPs du gène candidat DYX2 pour la dyslexie sur le chromosome 6. Comme pour le jeu de données simulées, nous avons d'abord comparé en chaque SNP les p-valeurs de taille de cluster non-corrigées avec les p-valeurs univariées non-corrigées, pour deux tailles différentes de la structure de fermeture (2 et 4) et pour un seuil de p-valeur de formation des clusters $\alpha = 0.01$. Les p-valeurs de taille de cluster semblent similaires aux p-valeurs univariées, ce que l'on peut voir sur les figures 5 et 6 où la ligne bleue représente les p-valeurs univariées et la rouge les p-valeurs de taille de cluster. Aucune d'entre elles n'est significative après une correction de Bonferroni.

Nous avons ensuite comparé les effets de la correction minP sur les p-valeurs univariées et sur les p-valeurs de taille de cluster, pour les deux tailles différentes de la structure de fermeture (2 et 4). La correction minP n'a pas donné de résultats significatifs sur les p-valeurs univariées. Cependant, la combinaison du test sur la taille de cluster et de la correction minP conduit cette fois-ci à la détection d'un cluster significatif, pour une structure de fermeture de taille 2. En effet, dans ce cas, la meilleure association est significative avec une p-valeur corrigée de 0.013, comme on peut le voir sur la figure 7 où la ligne bleue représente les p-valeurs univariées corrigées par minP et la rouge les p-valeurs de taille de clusters corrigées par minP.

Avec un élément de fermeture de taille de 4, les résultats sont légèrement inférieurs, mais la meilleure association a presque atteint le niveau de significativité avec une p-valeur corrigée de 0.059, comme on peut le voir sur la figure 8 où la ligne bleue représente les p-valeurs univariées corrigées par minP et la rouge les p-valeurs de taille de cluster corrigées par minP.

Figure 5 – *p-valeurs non-corrigées univariées (en bleu) et de taille de cluster (en rouge) pour un élément de taille* 2*, sur les données réelles. La matrice de corrélation en-dessous illustre le LD.*

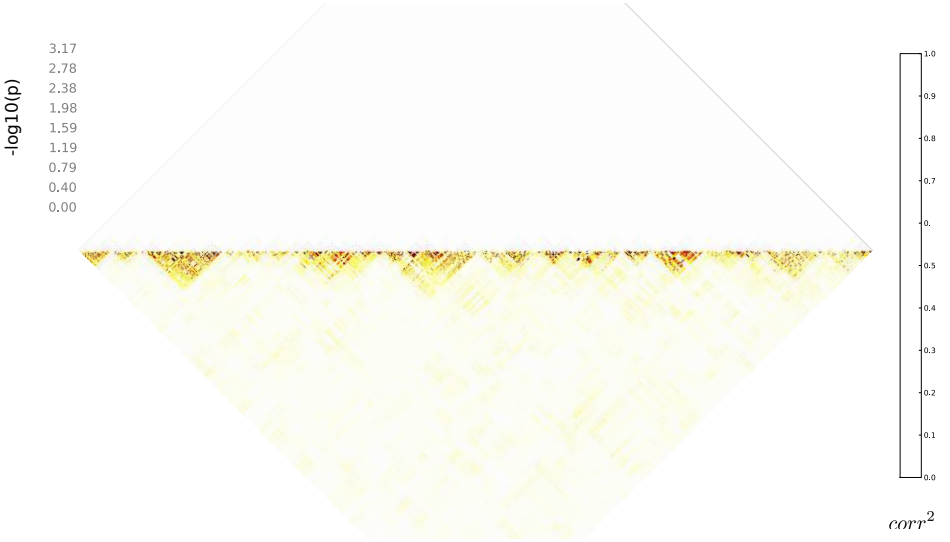


Figure 6 – *p-valeurs non-corrigées univariées (en bleu) et de taille de cluster (en rouge) pour un élément de taille* 4*, sur les données réelles. La matrice de corrélation en-dessous illustre le LD.*

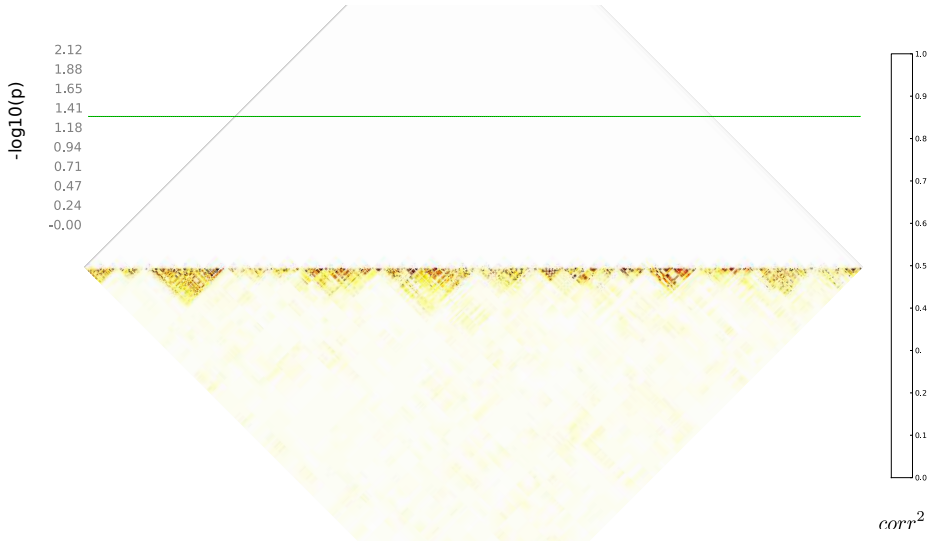


Figure 7 – *p-valeurs univariées (en bleu) et de taille de cluster pour un élément de taille* 2 *(en rouge), corrigées par minP, sur les données réelles. La matrice de corrélation en-dessous illustre le LD. La ligne verte représente le niveau de significativité de* 0.05*.*

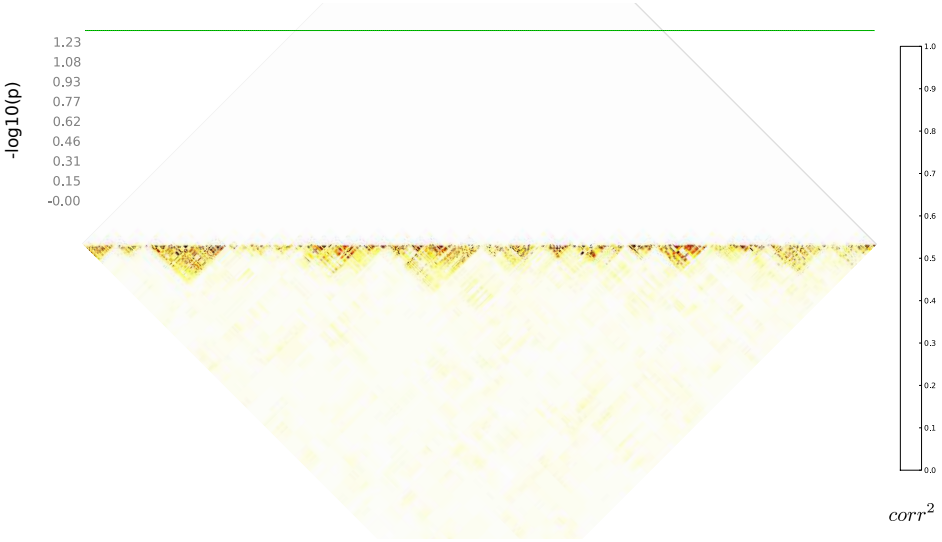


Figure 8 – *p-valeurs univariées (en bleu) et de taille de cluster pour un élément de taille* 4 *(en rouge), corrigées par minP, sur les données réelles. La matrice de corrélation en-dessous illustre le LD. La ligne verte représente le niveau de significativité de* 0.05*.*

Clusters 4D en imagerie génétique

Ensuite, nous avons essayé d'étendre le concept des clusters en combinant les clusters de voxels et de SNPs, et en utilisant un test de cluster 4D qui détecte conjointement les régions du cerveau et du génome avec des associations élevées.

La méthode est similaire à celle des clusters 1D, si ce n'est que la correction pour les test multiples est effectuée par l'approche maxT comme en neuroimagerie, en raison du coup calculatoire trop important de l'approche minP pour une étude cerveau entier (ou génome entier).

Nous avons obtenu 21 clusters significatifs après correction (voir figure 9).

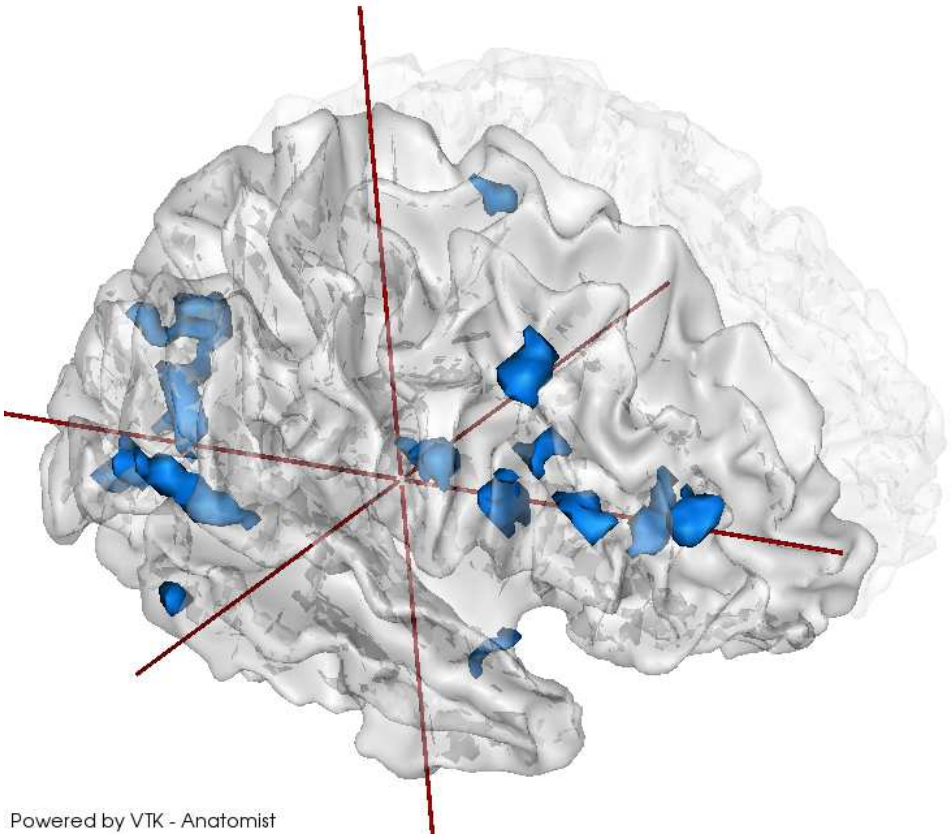


Figure 9 – *Régions cérébrales impliquées dans les clusters 4D détectés*

Chapitre 5

Dans ce chapitre, nous décrivons les approches multivariées à variables latentes que nous avons étudiées, l'analyse canonique (CCA) et la régression Partial Least Squares à deux blocs (PLS2), afin de rechercher des associations entre des données d'IRMf et de SNPs. Nous évoquons également les modalités de régularisation et les stratégies de réduction de dimension que nous avons utilisées pour résoudre le problème sur-apprentissage, qui se produit avec les méthodes multivariées en grandes dimensions. Enfin, nous présentons les procédures de validation classiques afin d'évaluer et de comparer les performances des différentes stratégies étudiées.

Méthodes multivariées basées sur des variables latentes

Notons $\mathbf{Y}$ (de taille $n \times q$) la matrice des q phénotypes d'imagerie et $\mathbf{X}$ (de taille $n \times p$) la matrice des p SNPs pour les n individus observés.

Régression Partial Least Squares

La régression Partial Least Squares est utilisée pour modéliser les associations entre deux blocs de variables en supposant qu'ils sont liés par le biais

de variables latentes non observées. Une variable latente (ou composante) correspondant à un bloc est une combinaison linéaire des variables observées de ce bloc. Une illustration de ces méthodes multivariées basées sur l'extraction des variables latentes est donnée Figure 10.

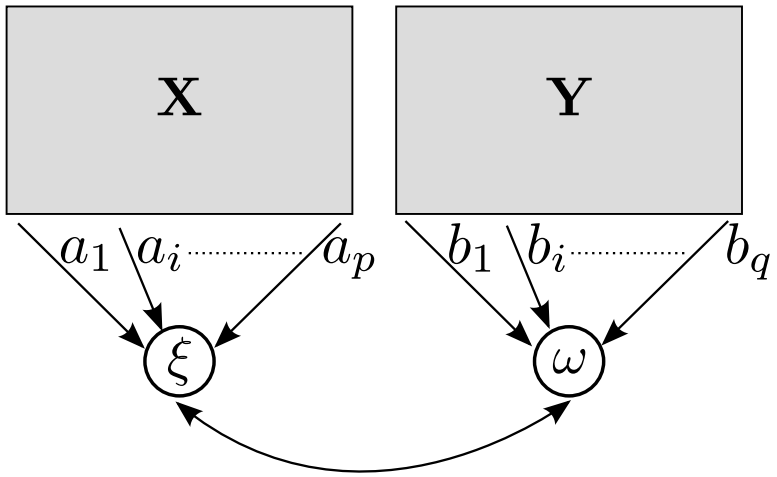


Figure 10 – *Illustration d'une méthode multivariée basée sur le concept de variables latentes.*

Plus précisément, la régression PLS recherche pour chaque bloc des variables latentes successives et orthogonales, de telle sorte qu'à chaque étape la covariance entre les deux variables latentes soit maximale.

A chaque étape h, la PLS optimise le critère suivant:

$$\max_{\|\mathbf{a}_h\|_2=\|\mathbf{b}_h\|_2=1} \operatorname{cov}(\mathbf{X}_{h-1}\mathbf{a}_h, \mathbf{Y}_{h-1}\mathbf{b}_h) \tag{3}$$

$$= \max_{\|\mathbf{a}_h\|_2=\|\mathbf{b}_h\|_2=1} \operatorname{corr}(\mathbf{X}_{h-1}\mathbf{a}_h, \mathbf{Y}_{h-1}\mathbf{b}_h) \sqrt{\operatorname{var}(\mathbf{X}_{h-1}\mathbf{a}_h)} \sqrt{\operatorname{var}(\mathbf{Y}_{h-1}\mathbf{b}_h)}$$

où $\mathbf{a}_h$ et $\mathbf{b}_h$ sont les vecteurs de poids pour les combinaisons linéaires des variables des blocs de $\mathbf{X}$ et $\mathbf{Y}$ respectivement. $\mathbf{X}_{h-1}$ et $\mathbf{Y}_{h-1}$ sont les résidus des matrices $\mathbf{X}$ et $\mathbf{Y}$ après leur régression sur les $h-1$ paires précédentes de variables latentes, en commençant par $\mathbf{X}_0 = \mathbf{X}$ et $\mathbf{Y}_0 = \mathbf{Y}$.

Ce problème d'optimisation est résolu à l'aide d'un algorithme itératif appelé NIPALS (Wold 1966).

Analyse Canonique

L'analyse canonique, ou *Canonical Correlation Analysis* (CCA), est une méthode similaire à la PLS, où l'on cherche désormais à maximiser la corrélation entre les variables latentes au lieu de la covariance.

La CCA construit donc des variables latentes successives et orthogonales pour chaque bloc, telles qu'à chaque étape h le critère suivant soit optimisé:

$$\max \operatorname{corr}(\mathbf{X}\mathbf{a}_h, \mathbf{Y}\mathbf{b}_h) \tag{4}$$

où $\mathbf{a_h}$ et $\mathbf{b_h}$ sont toujours des vecteurs de poids.

La solution peut être obtenue en calculant la décomposition en valeurs singulières de $\mathbf{D} = (\mathbf{X}'\mathbf{X})^{-1/2}\,\mathbf{X}'\mathbf{Y}\,(\mathbf{Y}'\mathbf{Y})^{-1/2}$.

Méthodes de régularisation

Régularisation L2 de la CCA

Afin de résoudre les problèmes de sur-apprentissage de la CCA, on peut utiliser une régularisation basée sur une pénalisation L2, en remplaçant les matrices $\mathbf{X}'\mathbf{X}$ et $\mathbf{Y}'\mathbf{Y}$ par $\mathbf{X}'\mathbf{X} + \lambda_{2X}\mathbf{I}$ et $\mathbf{Y}'\mathbf{Y} + \lambda_{2Y}\mathbf{I}$ respectivement (Vinod 1976, Leurgans et al. 1993). Cependant, dans de telles dimensions, on utilise souvent l'approximation qui consiste à remplacer les matrices $\mathbf{X}'\mathbf{X}$ et $\mathbf{Y}'\mathbf{Y}$ par des matrices identité, ce qui correspond à une régularisation extrême et rend la CCA équivalente à la PLS-SVD (et donc à la régression PLS pour la première composante).

Régularisation L1 de la PLS

En poussant plus loin la régularisation, on peut également envisager une pénalisation L1. Lê Cao et al. (2008) a récemment proposé une approche qui inclut la sélection de variables dans la régression PLS, basée sur une pénalisation L1 (Tibshirani 1996) et conduisant à une solution parcimonieuse. Dans la régression PLS parcimonieuse, ou *sparse PLS* (sPLS), le critère de la régression PLS est modifié par l'ajout d'une pénalisation L1 sur les vecteurs de poids $\mathbf{a}$ et $\mathbf{b}$ pour chaque paire de composantes:

$$\min_{\|\mathbf{a}\|_2=\|\mathbf{b}\|_2=1} -\mathbf{a}'\mathbf{X}'\mathbf{Y}\mathbf{b} + \lambda_{1X} \parallel \mathbf{a} \parallel_1 + \lambda_{1Y} \parallel \mathbf{b} \parallel_1 \tag{5}$$

où λ_{1X} et λ_{1Y} sont les paramètres de pénalisation L1 pour les vecteurs de poids des blocs $\mathbf{X}$ et $\mathbf{Y}$ respectivement.

Toutefois, si la PLS parcimonieuse permet de lutter contre le problème de sur-apprentissage, son efficacité dans le cas de données de très grandes dimensions reste une question ouverte. C'est la raison pour laquelle nous avons décidé de la combiner avec une première étape de réduction de dimension sur les SNPs.

Méthodes de réduction de dimension

Analyse en composantes principales

Une première manière de réaliser la réduction de dimension est d'ajouter une étape préliminaire d'analyse en composantes principales, ou *Principal Component Analysis* (PCA), sur chaque bloc de données avant d'appliquer la PLS ou la CCA. La régularisation n'est plus nécessaire dans ce cas, étant donné que la dimension a été considérablement réduite.

Filtrage univarié sur les SNPs

Une autre façon de procéder afin de réduire la dimension est d'ajouter une première étape de filtrage univarié avant la sPLS ou la CCA régularisée. Cette étape se compose de 1) $p \times q$ régressions linéaires basées sur un modèle génétique additif, 2) le classement des SNPs en fonction de la p-valeur minimale obtenue par chaque SNP au travers des phénotypes, et 3) le seuillage afin de conserver les SNPs avec les plus basses p-valeurs.

Même si le filtrage univarié peut sembler en contradiction avec la nature multivariée de méthodes telles que la PLS ou la CCA, il leur permet de continuer à extraire des modèles multivariés parmi les variables restantes et peut même être nécessaire pour surmonter le problème de sur-apprentissage très important dans ces dimensions.

L'évaluation des performances

Capacité de généralisation

Une façon classique d'évaluer les performances des différentes méthodes multivariées peut être d'évaluer si le lien obtenu sur un échantillon donné entre les deux blocs de données peut être généralisé à un nouvel échantillon.

Lorsque le nombre d'observations est trop faible, ce qui se trouve être notre cas comme nous le verrons dans le chapitre suivant, une autre façon d'évaluer la généralisabilité peut être aussi d'utiliser une validation croisée ou *cross-validation* (CV).

Ainsi, nous avons choisi d'utiliser une procédure de validation croisée afin de comparer les performances des différentes stratégies mentionnées ci-dessus, c'est-à-dire la combinaison de la PLS ou la CCA avec des techniques de régularisation et de réduction de dimension (voir Tableau 1 pour un résumé des différentes stratégies étudiées).

Méthode	Acronyme
Mass Univariate Linear Modelling	MULM
Partial Least Squares	PLS
Kernel Canonical Correlation Analysis	KCCA
sparse PLS	sPLS
regularised KCCA	rKCCA
Principal Component Analysis + PLS	PCPLS
Principal Component Analysis + KCCA	PCKCCA
Filtrage + (sparse) PLS	f(s)PLS
Filtrage + (regularised) KCCA	f(r)KCCA

Table 1 – *Résumé des différentes stratégies étudiées*

Pour chaque méthode et à chaque *fold* de la validation croisée, l'estimation du modèle (les vecteurs de poids) est effectuée sur l'échantillon d'apprentissage et testée sur l'échantillon de test (voir figure 11 pour les méthodes basées sur le filtrage et figure 12 pour les méthodes basées sur la PCA). En effet, à chaque fold i, les poids obtenus sur l'échantillon d'apprentissage sont utilisés pour construire les scores de l'échantillon de test (l'échantillon mis de côté), $\mathbf{X}^i\mathbf{a}^{-i}$ et $\mathbf{Y}^i\mathbf{b}^{-i}$, et le coefficient de corrélation entre ces scores est calculé. Cela donne un coefficient de corrélation de test moyenné au travers des folds qui reflète la reproductibilité du lien entre les deux blocs sur de nouveaux sujets.

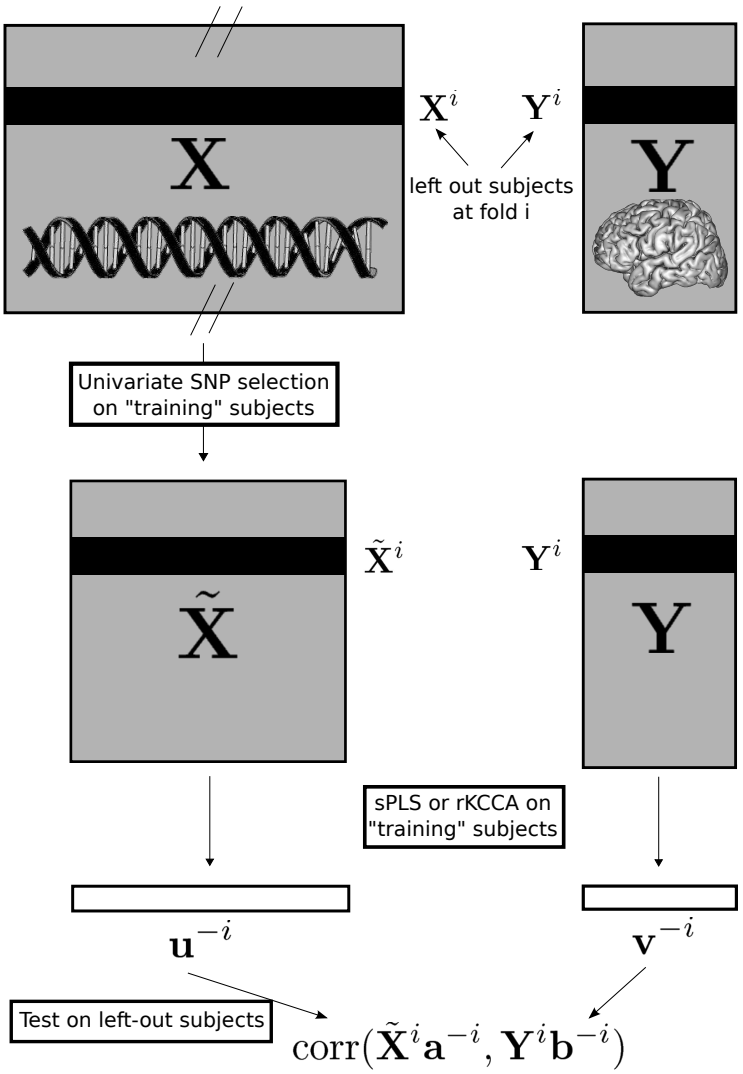


Figure 11 – *Illustration du schéma de validation croisée pour les méthodes basées sur le filtrage univarié. A chaque fold i, une sélection univariée de k SNPs est effectuée sur l'échantillon d'apprentissage* $\mathbf{X}^{-i}$ *et* $\mathbf{Y}^{-i}$*; des vecteurs de poids,* $\mathbf{a}^{-i}$ *et* $\mathbf{b}^{-i}$*, sont alors estimés par la sPLS ou la CCA régularisée sur cet échantillon d'apprentissage et finalement les scores des sujets de l'échantillon de test correspondant à ce fold sont calculés à partir de leurs variables observées,* $\tilde{\mathbf{X}}^i$ *et* $\mathbf{Y}^i$*, et des vecteurs de poids.*

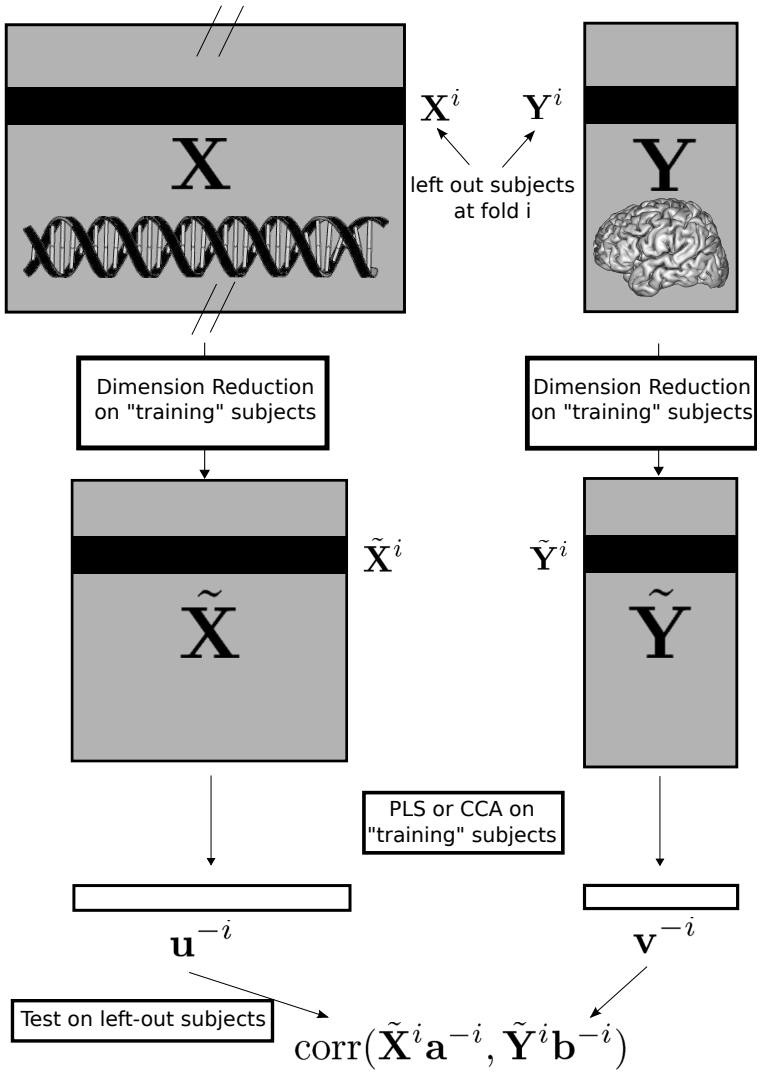


Figure 12 – *Illustration du schéma de validation croisée pour les méthodes basées sur la PCA. A chaque fold i, deux PCAs sont effectuées sur les SNPs et sur les phénotypes de l'échantillon d'apprentissage* $\mathbf{X}^{-i}$ *et* $\mathbf{Y}^{-i}$*; des vecteurs de poids,* $\mathbf{a}^{-i}$ *et* $\mathbf{b}^{-i}$*, sont alors estimés par la PLS ou la CCA sur cet échantillon d'apprentissage et finalement les scores des sujets de l'échantillon de test correspondant à ce fold sont calculés à partir de la projection de leurs variables observées sur les composantes principales,* $\tilde{\mathbf{X}}^i$ *and* $\tilde{\mathbf{Y}}^i$*, et des vecteurs de poids.*

Chapitre 6

Dans ce chapitre, nous présentons l'application, sous la forme d'une étude comparative, des différentes méthodes décrites dans le chapitre précédent.

Nous utilisons d'abord un jeu de données simulées reproduisant des données d'IRMf et des données de SNPs afin de comparer les performances des différentes méthodes, en évaluant leur capacité de détection, ainsi que leur capacité à généraliser le lien qui se trouve entre les deux blocs avec une procédure de validation croisée. En effet, nous avons d'abord comparé la PLS et la CCA, puis nous avons étudié l'influence de la régularisation L2 sur la CCA et de la régularisation L1 sur la PLS.

Enfin, nous avons essayé d'ajouter une première étape de réduction de dimension telle que la PCA ou le filtrage univarié.

Enfin, nous appliquons ces différentes méthodes avec la même procédure de validation croisée sur des données réelles d'IRMf et des données de SNPs, et la significativité statistique du lien obtenu sur les sujets de test est évaluée par le biais de permutations.

Résultats sur données simulées

Dans un premier temps, 34 phénotypes d'imagerie ont été simulés pour 500 individus à partir d'une distribution normale multivariée avec des paramètres estimés sur des données expérimentales d'IRMf.

Afin de simuler des données de génotypage avec une structure génétique similaire à celle des données réelles, nous avons considéré une méthode de simulation qui utilise la base de données HapMap. Nous avons généré 85772 SNPs pour nos 500 individus.

Nous avons ensuite sélectionné aléatoirement 10 SNPs et 8 phénotypes d'imagerie et induit deux modèles causaux indépendants en ajoutant à chaque fois l'effet cumulé de 5 SNPs sur 4 phénotypes d'imagerie.

Les SNPs ayant un coefficient r^2 d'au moins 0.2 avec l'un des 10 SNPs causaux ont été isolés et considérés comme informatifs.

Evaluation

Pour les différentes méthodes décrites précédemment, nous avons évalué la capacité à généraliser la corrélation entre variables latentes, en utilisant une validation croisée *5-fold* avec des échantillons d'apprentissage de 100 sujets et des échantillons de test de 400 sujets.

De plus, la vérité terrain étant connue, nous avons également pu comparer les performances des différentes méthodes en termes de valeur prédictive positive (VPP) lorsque 50 SNP sont choisis par chaque méthode. C'est presque l'équivalent dans notre cas de la sensibilité de chaque méthode lorsque 50 SNP sont sélectionnés, car il y a 56 SNPs causaux dans notre jeu de données simulées.

Influence de la régularisation

Nous nous sommes d'abord intéressés à la comparaison des performances de la PLS et de la CCA lorsque le nombre de SNPs p augmente, en partant de 200 SNPs (dont les 198 SNPs informatifs) jusqu'à 85772 SNPs (pour la plupart du bruit), ainsi qu'à l'influence de la régularisation L1 sur la PLS et de la régularisation L2 sur la CCA.

En faibles dimensions, sur la première paire de composantes, la CCA légèrement régularisée donne les meilleurs résultats.

Néanmoins, l'augmentation du nombre de SNPs (avec du bruit) met clairement en évidence la supériorité de la PLS et plus encore de la sPLS en grandes dimensions: la performance de la CCA régularisée diminue rapidement, tandis que la sPLS tolère une augmentation de la dimension jusqu'à 1000 SNPs avant que sa performance ne commence à diminuer.

Sur la deuxième paire de composantes, les résultats sont moins clairement interprétables. Cependant les performances de la sPLS sont au-dessus de celles de la CCA régularisée. De plus, tandis que le premier modèle causal semble avoir été capturé par la première paire de composantes, la seconde paire semble avoir détecté le second modèle.

Influence de la réduction de dimension

Ensuite, nous avons étudié l'influence d'une première étape de réduction de dimension.

Les résultats sur la première parie de composantes montrent que toutes les méthodes basées sur la PCA n'ont pas réussi à identifier de covariations généralisables lorsque le nombre de SNPs non pertinents augmente.

La réduction de dimension basée sur le filtrage, quant à elle, a légèrement amélioré la performance de la CCA et grandement amélioré la performance de la PLS: la fPLS obtenant la deuxième meilleure performance dans notre étude comparative.

Enfin, la meilleure performance est obtenue en combinant la régularisation L1 de la PLS et le filtrage (fsPLS), lorsque 100 SNPs sont conservés après filtrage et 50% d'entre eux sélectionnés par sPLS. Cependant, la stratégie purement univariée montre une faible capacité de généralisation, ce qui suggère que même si le filtrage apparaît nécessaire afin d'enlever des éléments non pertinents, il n'est pas en mesure d'identifier seul le lien imagerie/génétique et doit être combiné avec une étape multivariée pour tirer profit des effets cumulatifs de plusieurs SNPs.

Résultats sur données réelles

Dans un deuxième temps, nous avons donc appliqué ces différentes méthodes sur un jeu de données réelles avec 622534 SNPs et de 34 phénotypes issus de données d'IRMf correspondant à des indexes de latéralisation dans des régions d'intérêts pour des contrastes de lecture et de compréhension orale.

Comme sur les données simulées, nous avons évalué la capacité de généralisation des différentes méthodes, en utilisant cette fois-ci une validation croisée *10-fold*, avec des échantillons d'apprentissage de 85 sujets et des échantillons de test de 9 sujets.

Le tableau 2 montre les corrélations sur les échantillons d'apprentissage et de test pour les deux premières paires de variables latentes. On peut voir que, pour la première paire de composantes, la régularisation L1 de la PLS ne peut pas résoudre seule le problème de sur-apprentissage. En effet, comme la PLS, la sPLS n'a pas réussi à extraire un lien généralisable dans de telles dimensions et n'a capturé que du bruit. Dans ces grandes dimensions, la CCA nécessite une régularisation L2 tellement extrême qu'elle devient équivalente à la PLS.

Par conséquent, une première étape de réduction de dimension semble être nécessaire pour surmonter le problème de sur-apprentissage. En effet, même si les méthodes basées sur la PCA n'y parviennent pas non plus, les méthodes basées sur le filtrage univarié fonctionnent beaucoup mieux.

Parmi ces méthodes basées sur le filtrage, fsPLS donne le coefficient de corrélation le plus élevé sur les échantillons de test: 0.43 lorsque 1000 SNPs sont gardés après le filtre univarié et 5% de ces SNPs sont conservés par la sPLS. La deuxième meilleure performance sur la première paire de composantes est obtenue avec frKCCA avec un coefficient de corrélation sur les échantillons de test de 0.24 ($k = 1000$ et $s_{\lambda 2} = 1,000,000$). Cependant, avec une régularisation L2 tellement extrême, elle devient presque équivalente à fPLS, comme on peut le voir sur le tableau 2.

En ce qui concerne la seconde paire de composantes, le coefficient de corrélation obtenu sur les échantillons de test par fsPLS est plus faible que sur la première paire de composantes. Cependant, pour les autres méthodes utilisant la PLS, la corrélation semble être légèrement plus élevée sur la deuxième paire de composantes que sur la première. Cela peut s'expliquer par le fait qu'une fois que le bruit menant au sur-apprentissage sur la première paire de composantes a été soustrait, certains effets peuvent alors être observés sur d'autres composantes. Enfin, MULM et PCKCCA ne semblent en mesure de capter d'effets généralisables sur aucune des paires de composantes.

	ρ^1_{test}	ρ^2_{test}	$\rho^1_{training}$	$\rho^2_{training}$
MULM	0.036	-0.104	-0.458	-0.451
PLS	-0.092	0.218	0.990	0.984
sPLS ($s_{\lambda_{1X}} = 0.1\%$)	0.008	0.201	0.938	0.922
PCKCCA	0.010	0.008	1.000	1.000
PCPLS	-0.088	0.217	0.990	0.984
frKCCA ($k = 1000$,$\lambda_2 = 1,000,000$)	0.245	0.324	0.963	0.954
fPLS ($k = 1000$)	0.236	0.268	0.962	0.953
fsPLS ($k = 1000$,$s_{\lambda_{1X}} = 5\%$)	0.432	0.210	0.772	0.788

Table 2 – *Les deux premiers coefficients de corrélation obtenus en moyenne sur les échantillons de "test" et les échantillons d'"apprentissage".*

Afin d'obtenir les SNPs et les phénotypes d'imagerie impliqués dans la liaison entre les deux blocs, nous avons ensuite appliqué fsPLS sur tous les sujets en même temps pour le couple de paramètres donnant les résultats les plus significatifs sur la première paire de composantes: 1000 SNPs sélectionnés avec le filtre univarié et un taux de sélection de la PLS parcimonieuse de 5% ($k = 1000$, $s_{\lambda_{1X}} = 0.05$). Il faut noter à ce stade que la significativité statistique de l'analyse multivariée doit être considérée à l'échelle de l'ensemble de SNPs et non SNP par SNP, nous devons donc rester très prudents quant à l'interprétation des résultats.

La figure 13 illustre la localisation des SNPs sélectionnés par fsPLS.

La distribution des 1000 SNPs ayant passé le filtre univarié est illustrée par des traits noirs le long des 22 autosomes. Les 5% de ces SNPs qui ont été ensuite sélectionnés par sPLS sont indiqués en rouge. Comme on peut le voir, ils sont répartis sur tous les autosomes et certains d'entre eux semblent en déséquilibre de liaison. Parmi les 50 SNPs sélectionnés par fsPLS, certains d'entre eux se trouvent au sein d'un gène. Dix-huit

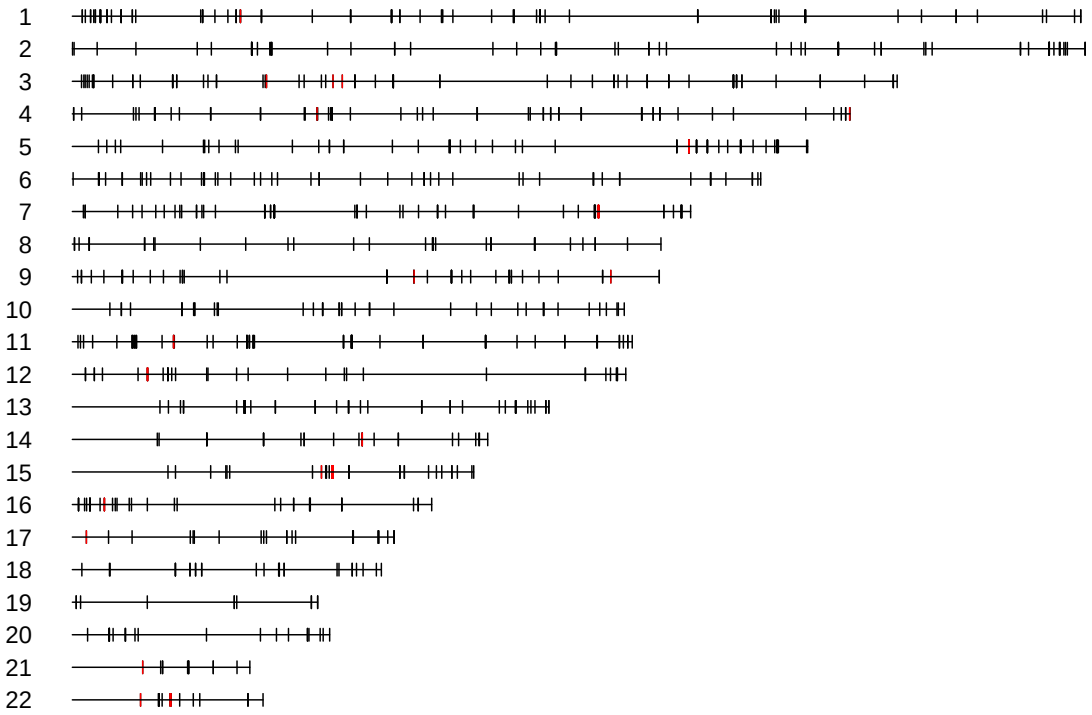


Figure 13 – *Distribution des* 1000 *SNPs les plus significatifs (après le test univarié) le long du génome. Les* 50 *SNPs sélectionnés ensuite par la sPLS sont surlignés en rouge.*

gènes ont ainsi été identifiés, tel que PPP2R2B et RBFOX1, qui ont été signalés comme étant liés à l'ataxie et à une mauvaise coordination des mouvements de la parole et du corps, ou encore PDE4B qui a été associé avec la schizophrénie et les troubles bipolaires.

La figure 14 montre la localisation des indexes de latéralisation, pour les cartes de contraste de "lecture" et de "compréhension orale". Les pondérations attribuées par la sPLS aux différents phénotypes d'imagerie sont illustrées à l'aide du dégradé de couleur. Les phénotypes ayant obtenu les plus grands poids (en valeur absolue) proviennent principalement du contraste de "lecture", en particulier du lobe temporal.

Pris ensemble, nos résultats montrent que fsPLS semble établir un lien significatif entre un sous-ensemble de SNPs répartis le long du génome et un réseau cérébral fonctionnel activé lors d'une tâche de lecture, certains de ces SNPs étant probablement indirectement liés aux phénotypes de neuroimagerie par le biais du déséquilibre de liaison. Ceci suggère que la variabilité individuelle du génome contient des prédicteurs de la variabilité observée dans l'activation du cerveau lors de tâches de langage. On peut remarquer que tous les phénotypes ne semblent pas contribuer de la même façon à la première composante et il semble y avoir une plus forte implication des phénotypes obtenus à partir du "contraste" de lecture.

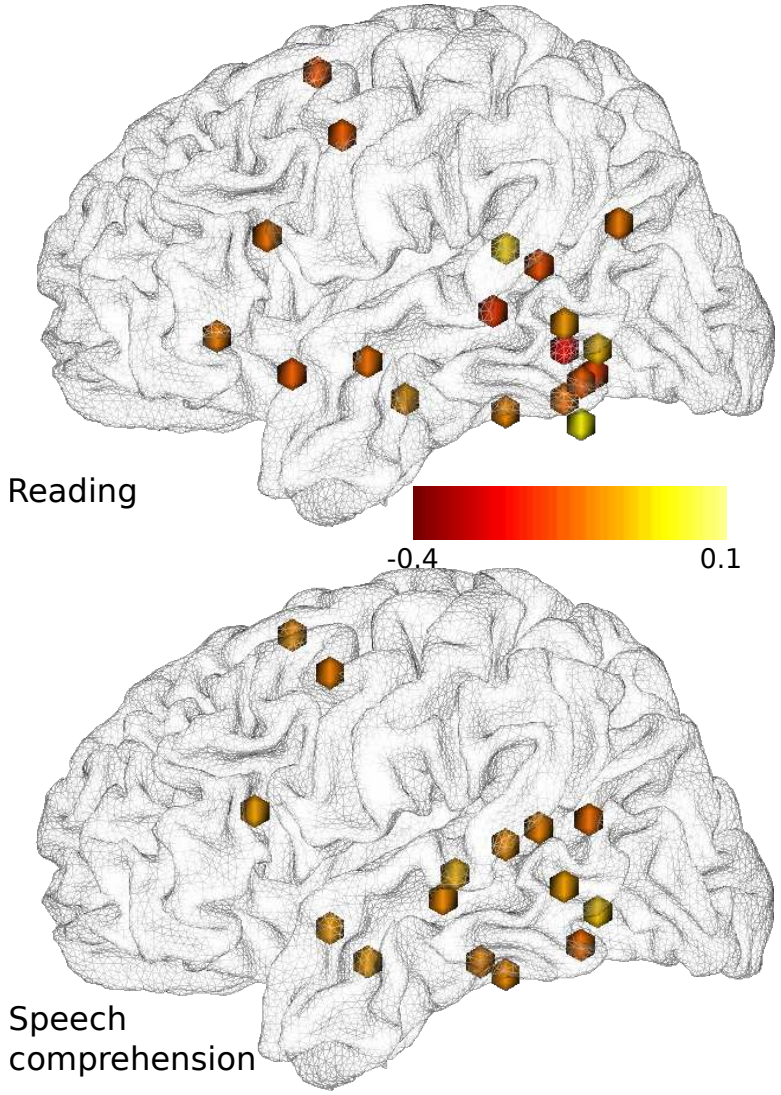


Figure 14 – *Localisation des 19 phénotypes extraits à partir de la carte de contraste de "lecture" et des 15 phénotypes extraits de la carte de contraste de "compréhension orale". Les poids attribués par la sPLS aux phénotypes sont illustrés à l'aide de barre de couleur. (Le signal qui apparait à l'extérieur de la surface corticale fait partie du cervelet.)*

Conclusion

Le but de ce travail a été de trouver des méthodes qui tiennent compte de la nature multivariée potentielle des données de génétique et d'imagerie, localement ou à plus longue distance, tout en faisant face à la très grande dimensionnalité du problème, afin d'augmenter la puissance de détection par rapport à l'approche univariée classique.

Clusters de SNPs 1D et clusters 4D

Notre première contribution a été d'améliorer la sensibilité de l'approche univariée en tirant profit de la nature multivariée des données de SNPs d'une manière locale, en recherchant des clusters 1D de SNPs adjacents associés à un même phénotype d'imagerie.

L'originalité de l'approche des clusters 1D a été d'adapter les tech-

niques d'inférence au niveau du cluster existant en neuroimagerie à des données de SNPs, en utilisant une statistique en chaque SNP combinée avec une correction minP pour faire face à la non-stationnarité du LD le long du génome. Nous avons appliqué cette approche à la fois sur des données réelles et simulées. Les résultats sont très préliminaires, mais la méthode proposée semble prometteuse afin d'améliorer la sensibilité des études d'association génétique tout en contrôlant l'erreur de type I.

Ensuite, nous avons poussé plus loin le concept de clusters et nous avons combiné les clusters de SNPs et de voxels, en utilisant un test simple de cluster 4D qui détecte conjointement les régions du cerveau et du génome associées. Nous avons calibré le test par permutations. Sur des données réelles, ce test a montré une plus grande sensibilité que le test d'association en chaque SNP et chaque voxel, ou que le test de cluster 3D. Nous espérons que cela permettra d'améliorer la capacité de détection lors d'études d'imagerie génétique avec des jeux de données encore plus grands.

Limites et extensions

Cependant, ces deux approches rencontrent quelques limitations, qui sont semblables à celles qui se produisent avec l'inférence au niveau du cluster en neuroimagerie.

Choix du seuil et localisation de l'activation Tout d'abord, le choix du seuil pour la formation des clusters reste critique et peut conduire à des résultats relativement différents.

Deuxièmement, le test cluster 4D fournit une statistique au niveau du cluster, de sorte que même si un cluster est important, on ne peut rejeter l'hypothèse nulle pour un voxel ou un SNP spécifique. L'interprétation devrait être que dans cette région du cerveau et parmi cet ensemble de SNPs contigus, il y a une association imagerie/génétique anormalement forte.

Une extension possible de ce travail serait d'étudier la stratégie récemment proposée par Smith and Nichols (2009) qui traite de ces deux questions, en utilisant une procédure sans seuil qui augmente les statistiques des voxels au sein d'un cluster, mais il peut y avoir des problèmes de pivotalité liés à cette technique, qui devraient encore être étudiés afin d'assurer un contrôle précis de la spécificité (Ge et al. 2003).

Taille de l'élément morphologique De plus, le choix de la taille de l'élément morphologique pour les clusters 1D est essentielle, comme le degré de lissage en neuroimagerie. Il convient également de noter que la taille de l'élément morphologique est exprimée en nombre de SNPs et non pas comme une distance biologique réelle telle que le nombre de paires de bases ou de la distance génétique (en cM). Il serait justement intéressant de tester l'utilisation d'un élément morphologique exprimé en cM ou en nombre de paires de bases.

Non-stationnarité et pivotalité Une autre limite de l'inférence sur la taille des clusters 4D est liée à la non-stationnarité de l'image et du LD

le long du génome. En effet, s'il est déjà connu que les régions cérébrales lisses auront une plus grande sensibilité que les régions non-lisses, nous avons remarqué également dans nos résultats que beaucoup de clusters sont détectés dans les zones de fort LD.

A l'inverse, l'approche des clusters 1D gère la non-stationnarité du LD, en utilisant une statistique de taille de cluster reportée au niveau de chaque SNP ainsi qu'une correction minP pour les comparaisons multiples. Malheureusement, la gestion de la non-stationnarité induit souvent des problèmes de pivotalité, puisque la statistique en chaque SNP dépend des statistiques des SNPs voisins. Par conséquent, la correction minP n'est plus valide dans ce cas.

Coût calculatoire Enfin, ces deux approches sont coûteuses en calculs, en particulier l'approche cluster 1D en raison de la correction minP, qui nécessite plus de permutations que la stratégie maxT afin d'évaluer avec précision les p-valeurs non-corrigées.

Autres extensions possibles D'autres extensions de ces tests seraient également intéressantes à étudier. Par exemple, on pourrait essayer de combiner l'intensité de l'association avec la taille du cluster. Des stratégies équivalentes à celles déjà proposées dans la littérature de neuroimagerie testant à la fois l'étendue spatiale et l'intensité moyenne ou la valeur maximale (Poline et al. 1997) au sein du cluster pourraient être également mises au point dans le contexte des études d'imagerie génétique.

La correction family-wise error rate peut conduire à des seuils très élevés et à une sensibilité très faible dans les jeux de données avec beaucoup de voxels et de SNPs. Ainsi il serait intéressant d'étendre ce travail à d'autres corrections moins sévères telles que la correction False Discovery Rate (FDR), qui peut également être facilement appliquée sur des données 4D d'imagerie génétique (Nichols 2006). Il faut cependant noter que même avec une procédure moins rigoureuse telle que FDR, la détection d'associations sur des jeux de données d'un million de SNPs est susceptible d'avoir une faible sensibilité.

Enfin, ces méthodes pourraient être appliquées à tout type de phénotype quantitatif de neuroimagerie et même facilement étendues à des phénotypes qualitatifs.

Interprétation des résultats Dans ce travail, nous n'avons pas étudié les implications de nos résultats du point de vue des neurosciences cognitives. Cependant, on sait que l'asymétrie du langage est un phénotype héréditaire, et que les régions génétiques DYX2 et DYX5 sont susceptibles d'être impliquées dans le phénotype que l'on observe, si bien que les résultats obtenus ici sont susceptibles d'être interprétables.

Cependant, nous devons rester prudents quant à l'interprétation des résultats, car ces méthodes peuvent détecter des associations directes (peut-être dues à un ou plusieurs SNPs causaux), mais aussi des associations indirectes dues à des corrélations entre SNPs au sein d'un bloc LD.

Régression Partial Least Squares parcimonieuse associée à une réduction de dimension préliminaire

Notre deuxième contribution a été d'étudier l'utilisation de méthodes multivariées à deux blocs de données, à savoir la régression Partial Least Squares et l'analyse canonique, pour augmenter la puissance de détection des études d'imagerie génétique, en tenant compte de la nature multivariée, potentiellement à longue échelle, de l'association tant sur le plan de l'imagerie que de la génétique. Pour faire face à la très forte dimensionnalité du problème, l'originalité de ce travail a été de comparer différentes stratégies combinant à la fois une régularisation et une réduction de dimension préliminaire, associées à la PLS ou la CCA, sur des jeux de données simulées et réelles d'imagerie génétique.

Performance de la méthode fsPLS

Nous avons montré que l'approche en deux étapes appelée fsPLS, combinant filtrage univarié et PLS parcimonieuse, montre de bien meilleurs résultats que les autres stratégies multivariées sur les données simulées et réelles.

En effet sur le jeu de données simulées, même si la PLS parcimonieuse montre de meilleurs résultats que la PLS et la CCA (régularisée) quand la dimension augmente, elle n'est pas parvenue à surmonter seule le problème de sur-apprentissage, ce qui suggère qu'une première étape de réduction de dimension est également nécessaire. Le filtrage univarié semble être la meilleure solution, surtout lorsqu'il est combiné avec la PLS parcimonieuse, alors que les méthodes basées sur l'ACP ont échoué.

Par ailleurs, nos résultats sur le jeu de données expérimentales ont montré que la fsPLS était suffisamment sensible pour détecter un lien multivarié généralisable et significatif entre des données de génétique et de neuroimagerie, tandis que le test univarié seul n'a pas pu détecter aucune association phénotype/SNP significative après correction.

Influence des paramètres de filtrage univarié et de régularisation L1

Nous avons également étudié l'influence des paramètres de filtrage univarié et de régularisation L1 sur la capacité à généraliser le lien trouvé par la fsPLS entre les deux blocs de données.

Nous avons observé sur le jeu de données expérimentales que la fsPLS extrait le lien imagerie/génétique le plus significatif et généralisable lorsque 1000 SNPs sont conservés après filtrage et 50 de ces SNPs sélectionnés par la sPLS. Ces résultats ainsi que ceux obtenus sur les données simulées soulèvent la question de la contribution relative du filtrage univarié et de la contrainte de parcimonie dans la sélection de caractéristiques pertinentes. Un nombre relativement important de SNPs conservés après filtrage semble être nécessaire, jusqu'à un compromis entre le nombre de vrais et de faux positifs, afin de permettre à la sPLS d'extraire une association robuste entre un ensemble de SNPs et un réseau de neuroimagerie. Cependant, les résultats sur les jeux de données simulées et expérimentales ont démontré qu'un seuil trop lâche sur le filtrage conduit à une

augmentation importante du sur-apprentissage de la PLS quelque soit la parcimonie.

Une autre raison d'effectuer un filtrage univarié est liée au fait que la PLS et même la PLS parcimonieuse essayent d'expliquer la variance de chaque bloc de données, tandis qu'elles cherchent à maximiser la corrélation entre les deux blocs. Par conséquent, dans les très grandes dimensions telles qu'en imagerie génétique, les deux termes d'écart-type intra-bloc pèsent trop par rapport au terme de corrélation inter-blocs. En outre, la PLS et la PLS parcimonieuse peuvent être influencées par la non-stationnarité du LD et ont tendance à donner des poids plus élevés (et à sélectionner dans le cas de la PLS parcimonieuse) dans les régions de fort LD. Le filtrage univarié permet de résoudre ces problèmes en réduisant le nombre de SNPs et en cassant partiellement la structure du LD, car il ne sélectionne que les SNPs qui sont les plus corrélés avec le phénotype d'imagerie. Le lien entre le seuil optimal et la densité d'échantillonnage de LD devrait être étudiée plus longuement.

Ainsi, même si il peut sembler en contradiction avec la nature multivariée des méthodes telles que la PLS, le filtrage univarié semble nécessaire pour permettre à la PLS parcimonieuse de surmonter le problème de sur-apprentissage et d'extraire un lien imagerie/génétique généralisables et significatif. En fait, il serait intéressant d'étudier l'influence du filtrage univarié sur d'autres approches multivariées. Par exemple, on pourrait essayer de combiner le filtrage et la régression multitâche ou l'ICA parallèle, et de les comparer avec notre approche fsPLS.

Limitations potentielles de fsPLS

Cependant, le filtrage univarié n'est peut-être pas la meilleure technique de réduction de dimension et il pourrait être comparé à d'autres travaux avec des filtres multivariés qui tiennent compte des interactions potentielles entre les prédicteurs (par exemple, pour une revue voir Díaz-Uriarte and Alvarez de Andrés 2006).

Le choix du type de régularisation est aussi essentiel. Nous avons vu que la PLS parcimonieuse est semblable à une régularisation Elastic Net. On pourrait étudier des pénalisations plus sophistiquées qui intègrent les connaissances a priori sur la structure de corrélation des données (par exemple, le déséquilibre de liaison ou la connectivité fonctionnelle) ou sur des groupes de variables biologiquement significatifs, tels que des ensembles de gènes appartenant à la même voie ou des régions cérébrales anatomiquement connectées. Par exemple, il serait intéressant de comparer la PLS parcimonieuse avec la régression multitâche parcimonieuse au niveau du groupe (Wang et al. 2012a).

Une autre limitation de ces méthodes multivariées est qu'elles ne fournissent pas de degré de significativité pour chaque variable ou de contrôle explicite des faux positifs. Il serait toutefois intéressant de regarder plutôt la robustesse de la sélection, en utilisant des techniques de *bootstrap* par exemple.

De plus, le codage génétique additif que nous avons utilisé n'est pas nécessairement le plus approprié pour capturer certains effets non-linéaires du nombre d'allèles mineurs. Un codage génétique différent, tel

qu'un codage dominant/récessif ou génotypique, devrait être étudié dans des travaux ultérieurs.

Interprétation des résultats

Sur le jeu de données expérimentales, la fsPLS nous a permis de détecter un lien significatif entre un ensemble de SNPs et un réseau cérébral fonctionnel activé lors d'une tâche de lecture, dans un cadre d'analyse génome entier. Ceci suggère que la variabilité individuelle dans le génome contient des prédicteurs de la variabilité observée dans l'activation du cerveau lors de tâches de langage. Nous avons montré que nous pouvions généraliser notre modèle sur des nouveaux sujets. Cependant, il serait très intéressant de reproduire ces résultats sur un jeux de données indépendant et plus grand.

De plus, l'interprétation des résultats reste un problème très difficile et la pertinence de ces résultats neuroscientifiques devrait être étudiée dans des recherches ultérieures. Quant à la méthode fsPLS elle-même, des règles de filtrage plus élaborées et d'autres formes sophistiquées de pénalisation devraient également être étudiées, ce qui pourrait peut-être aider à l'interprétation des résultats.

Enfin, l'objectif ultime de ce travail serait de relier les résultats d'imagerie génétique à des phénotypes finaux tels que des scores cognitifs ou des phénotypes cliniques. Par exemple, les méthodes conçues pour l'analyse multi-blocs, telles que l'analyse canonique généralisée régularisée (Tenenhaus and Tenenhaus 2011), pourrait être intéressante pour étudier les relations entre ces trois types de données.

Introduction

Context

Brain imaging is increasingly recognised as an interesting intermediate phenotype (or endophenotype) to understand the complex path between genetics and behavioural or clinical phenotypes. Indeed, brain imaging is a quantitative phenotype, hopefully richer and closer to genetics than a final phenotype such as a patient *vs* control status for instance. Thus, by analogy with neuroimaging studies which usually consist of a few tens of subjects, one might hope that imaging genetics studies will uncover stronger effects and require a smaller sample size than classical genetic association studies, which often require a sample size of several thousands of subjects. Nevertheless, in this *imaging genetics* field, a first goal is to propose methods to identify the part of genetic variability that explains some neuroimaging variability.

Imaging genetics studies, which include a large amount of data in both the imaging and the genetic components, are indeed facing challenges for which the neuroimaging community has no definitive answer so far. Current imaging genetics studies are often either limiting the brain imaging endophenotypes studied to a few candidate variables but testing their relationship with a large number of genetic variables such as Single Nucleotide Polymorphisms (SNPs), as one usually proceeds during gene screening (e.g., Furlanello et al. 2003), or limiting the number of candidate SNPs or genes to be tested on the whole brain or some large portion of it (e.g., McAllister et al. 2006, Roffman et al. 2006, Glahn et al. 2007). Without any strong priors on genetic or brain regions involved, one may investigate exploratory methods. However, when faced with both a large number of genetic variables and a large number of neuroimaging variables, one has to design an appropriate analysis strategy that should be as sensitive and specific as possible.

The simplest approach to exploratory imaging genetics studies is clearly to apply a massive univariate analysis on both genetic and imaging data (Stein et al. 2010), which may be called Mass-Univariate Linear Modelling (MULM). However, while univariate techniques are simpler, they may reach a multiple comparison problem in the order of 10^{12} when applied at the genome-wide and brain-wide level. Moreover, they do not account for the fact that the link between genetic and imaging data is likely to be in part multivariate. Indeed, on the genetic side, the interaction between several genetic loci (epistasis) or the accumulation of several small effects are very likely phenomena in common traits or diseases (Frazer et al. 2009, Yang et al. 2010). Thus, brain imaging endophenotypes are probably influenced by the combined effects of several genetic variants. Similarly, on the imaging side, pleiotropy may occur as well, which means

that different brain regions may be influenced by the same genetic variant(s). Moreover, it might be interesting to account for the multivariate nature of imaging data, such as anatomical connectivity or the co-activation of brain regions, when searching for imaging genetics associations.

A first way to partially take into account the multivariate nature of the genetic data may be to use a gene-based method for associating the joint effect of the different SNPs within each gene across the voxels of the whole brain (Hibar et al. 2011, Kohannim et al. 2011).

Some other multivariate strategies have also recently been proposed for imaging genetics, such as sparse Reduced Rank Regression (Vounou et al. 2010; 2012), Group-sparse Multitask Regression (Wang et al. 2012a;b) or parallel Independent Component Analysis (Liu et al. 2009), which analyse jointly the two blocks of data and take into account the potential joint effects that may exist between SNPs or the potential covariations between brain regions. However, in high-dimensional settings such as imaging genetics studies, multivariate methods may be prone to overfitting issues, even in their regularised/sparse version.

Objectives

The purpose of this work is to find methods that account for the potential multivariate nature of the data, either locally or at a longer range, while facing the very high dimensionality of the problem and increasing the detection power.

We first try to improve sensitivity of the univariate approach by taking advantage of the multivariate nature of SNP data, in a local way, by looking either at clusters of adjacent SNPs associated with the same phenotype or even at 4D clusters in the voxel $\times$ SNP space.

Then, we try to go further and to identify a brain network covarying with a set of genetic polymorphisms, by investigating the use of two multivariate methods designed for the joint analysis two blocks of data: two-block Partial Least Squares (PLS2) regression and Canonical Correlation Analysis (CCA). Moreover, we compare different strategies of regularisation and dimension reduction, combined with PLS2 or CCA, to face the very high dimensionality of imaging genetics data. We propose a comparison study of the different strategies on a simulated dataset first and then on a real fMRI and SNP dataset.

Organisation and contributions

Part I: Overview of Imaging Genetics Studies

In the first part, we will make an overview of Imaging Genetics studies, starting with a description of the context and the different types of data, and then moving to the statistical methods that have been proposed to analyse such data.

Chapter 1: Imaging Genetics data

In this chapter, we will first introduce the context and purpose of imaging genetics studies. Then we will briefly describe several types of imaging and genetic data that may be of interest for such studies.

Chapter 2: Statistical analysis in Neuroimaging or in Genetics: a similar problem

Before moving to the joint analysis of the two types of data, we will try in this chapter to make a review of the statistical analysis methods that are classically used in each field respectively. In fact, we will see that statistical analysis methods are very similar in neuroimaging and in genetics. We will first present the classical univariate approach and its application in both fields. Then, we will introduce some multivariate approaches common to both fields. Finally, we will show that both types of data are usually very-high dimensional, which implies the regularisation of those multivariate methods, possibly combined with an initial step of dimension reduction.

Chapter 3: Statistical analysis in Imaging Genetics: a joint analysis

In this chapter, we will move to the existing methods for the joint analysis of these two types of high-dimensional data, which are basically extensions of the methods from Chapter 2 to the case of multiple variables within both blocks of data. We will first describe the classical univariate approach and then detail multivariate approaches, namely principal components regression, penalised regression and two-block multivariate methods.

Part II: Contributions

In the second part, we will present the two approaches that we investigated and that account for the potential multivariate nature of the data, while facing the very high dimensionality of the problem and increasing the detection power.

Chapter 4: Clusters of SNPs

In this chapter, we will present the first approach that we developed, accounting for the potential multivariate nature of the data, in a local way. In the neuroimaging field, strategies to limit the multiple comparisons issue of the univariate approach have been proposed, based on the detection of clusters of contiguous activated voxels, which may increase sensitivity by trading off anatomical specificity. In this chapter, we will first briefly describe cluster-level inference in neuroimaging. Then, we will investigate a similar strategy on the genetic data, by searching for SNP clusters, and show preliminary results. Finally, we will show that this idea could be extended to the detection of 4D clusters of voxels and SNPs.

Chapter 5: Dimension reduction and regularisation combined with Partial Least Squares

In this chapter, we will describe the two-block multivariate approaches that we investigated, namely Canonical Correlation Analysis (CCA) and two-block Partial Least Squares Regression (PLS2), in order to search for association between fMRI and SNP data. We will also detail the regularisation and dimension reduction strategies that we used to solve the overfitting issue, which occurs with multivariate methods in such high-dimensional settings. Finally, we will present classical validation procedures in order to assess and to compare the performances of the different strategies that we investigated.

Chapter 6: Application and assessment on simulated and real datasets

In this chapter, we will present the application and a comparison study of the different methods described in the previous chapter.

We first use a simulated dataset mimicking fMRI and genome-wide SNP data and compare the performances of the different methods, by assessing their detection power, as well as their capacity to generalise the link found between the two blocks with a cross-validation procedure. Indeed, we first compared PLS and CCA, then we investigated the influence of L2 regularisation on CCA and L1 regularisation on PLS, and finally we tried to add a first step of dimension reduction such as PCA or filtering.

Finally, we apply these different methods with the same cross-validation procedure on a real dataset made of fMRI and genome-wide SNP data and the statistical significance of the link obtained on "test" subjects is assessed with randomisation techniques.

Part I

Overview of Imaging Genetics Studies

Imaging Genetics data

1

In this chapter, we will first introduce the context and purpose of imaging genetics studies. Then we will briefly present the different types of MRI neuroimaging techniques that may be of interest for such studies, with a focus on the functional MRI data. Finally, we will see that the genetic data used in imaging genetics studies may be of different kinds, and introduce a few notions about DNA and the nature of such data. We will also see how DNA is transcribed into RNA and then translated into proteins, which may be of interest as well in some cases.

Contents

1.1 Context

1.1.1 Neuroimaging: an interesting endophenotype

Imaging genetics studies rely on the idea that neuroimaging data may be considered as an interesting intermediate phenotype (or endophenotype) to understand the complex path between genetics and behavioural or clinical phenotypes. Compared to the final phenotypes, this endophenotype is hoped to be richer and closer to genetics. In this context, a first goal is to propose methods to identify the part of genetic variability that explains some neuroimaging variability.

1.1.2 Heritability of neuroimaging phenotypes

An increasing number of studies tend to prove that some neuroimaging phenotypes are indeed heritable, that is to say some proportion of the phenotype variability between individuals is due to genetic differences.

The first kind of studies are twin studies, which try to assess the heritability of the phenotype, that is the part of phenotypic variability explained by some genetic variability. Indeed they compare monozygotic twins with dizygotic twins, accounting by nature for environmental effects, since twins are supposed to live in the same environment. For instance, Thompson et al. (2001) showed that grey matter quantity was heritable in Broca's and Wernicke's areas as well as in frontal regions. Similarly for twin studies with fMRI data, Koten Jr. et al. (2009) found a significant genetic influence on brain activation in neural networks supporting digit working memory tasks.

The second type of studies are population-based studies, searching for associations between genetic polymorphisms and phenotypes across a population of unrelated individuals. For instance, Hariri et al. (2002) showed that individuals with one or two copies of the short allele of the serotonin transporter (5-HTT) promoter polymorphism, which has been associated with reduced 5-HTT expression and function and increased fear and anxiety-related behaviours, exhibit greater amygdala neuronal activity, as assessed by fMRI, in response to fearful stimuli compared with individuals homozygous for the long allele. As for anatomy, Pezawas et al. (2004) found bilateral reductions of hippocampal gray matter volumes in met-BDNF carriers compared with val/val-BDNF subjects, which seems to be consistent with the fact that the met-BDNF allele predicts variation in human memory, and is associated with several neurological and psychiatric disorders.

1.2 Neuroimaging data

Neuroimaging data used in imaging genetics studies may be of different types. Among them, Magnetic Resonance Imaging (MRI) has the advantage of being non invasive and of having a good spatial resolution. There are three main MRI modalities. First, anatomical MRI may be used to distinguish between the different types of tissues: grey matter made of cell bodies of neurons (the cortex and basal ganglia), white matter fibres made

of axons connecting the different regions of the brain, and cerebrospinal fluid which supports and protects neural cells. Second, diffusion MRI reflects the movement of water molecules in the brain, and is in particular interesting to reconstruct white matter fibres. Finally, functional MRI (fMRI) allows an indirect measurement of brain activation during a specific tasks or at rest. In this thesis, we will focus on fMRI. We will also very briefly describe the features of interest that can be extracted from MRI images.

1.2.1 Functional MRI

1.2.1.1 The BOLD contrast

Functional MRI aims at identifying the brain regions that are activated during specific conditions, such as specific tasks or at rest. It is based on the Blood Oxygenation Level Dependent (BOLD) contrast, which allows visualisation of blood flow variations. Indeed, the activation of a specific brain region is associated with a local increase of blood flow and oxygen consumption. But the increase in oxygen consumption is less important, which leads to an increase in the oxyhemoglobin level, and thus a detectable decrease in the relative level of desoxyhemoglobin due to its paramagnetic properties.

1.2.1.2 Single subject data analysis

fMRI data are submitted to a few pre-processing steps for each individual, before group-level analysis. Classically, fMRI data consists in temporal sequences of $3D$ images, every 2 to 3 seconds. Spatial resolution is classically $3mm^3$ on standard 3 Tesla (T) scanners, this volume unit being called a voxel. First, slice timing correction temporally realigns the different slices within each $3D$ image. Motion correction is also applied to spatially realign the different $3D$ images, to compensate for the subject's movements. Each $3D$ fMRI image is then registered onto a $3D$ anatomical image of the subject.

In the case of a study with several subjects, the images are also registered into a common referential, a reference brain called template, in order to be able to compare the different subjects afterwards. This step, called spatial normalisation, is usually followed by a smoothing step using a Gaussian kernel, in order to compensate for registration errors between subjects. The spread of the kernel is usually two or three times the voxel size.

The pre-processed fMRI data of each subject are then analysed using a regression framework, called the General Linear Model (GLM), as proposed by (Friston et al. 1994). For each voxel independently, the time series of the signal in that voxel is regressed onto several regressors (called the design matrix) corresponding to the presence or not over time of each of the different conditions. The linear regression coefficient associated with each condition is thus estimated in each voxel, leading to different coefficient maps (or activation maps) corresponding to the different conditions (see Figure 2.1). Coefficient maps may then be used for statistical inference. Indeed, one may test for the presence (in each voxel separately) of

an effect of interest, called a contrast, such as the difference between the coefficient maps of two conditions. This approach is called classical inference in neuroimaging and consists in computing a test statistic in each voxel, resulting in a statistical map. The significance degree of the contrast may then be derived from this test statistic in each voxel. We will see in the next chapter that this leads to huge multiple comparison problem.

Coefficient maps may also be used for group analysis, where the goal is to test whether an effect of interest is reproducible across subjects.

1.2.2 Extraction of features of interest

Instead of looking at each voxel of the image, one may prefer to extract features of interest from the image. For instance in the case of anatomical MRI, the volumes of grey and white matters may be computed, sulci and gyri of the cortex may also be extracted. Similarly, white matter fibres can be reconstructed using diffusion MRI, as mentioned earlier. Finally, in fMRI, the signal may be averaged in specific regions of interest.

1.3 Genetic data

The genetic data used in imaging genetics studies may be of different kinds, which will be described in this section. We will first introduce a few notions about DNA and then describe the nature of these genetic data. Finally, we will see how DNA is transcribed into RNA and then translated into proteins, which may also be of interest in some cases.

1.3.1 DNA

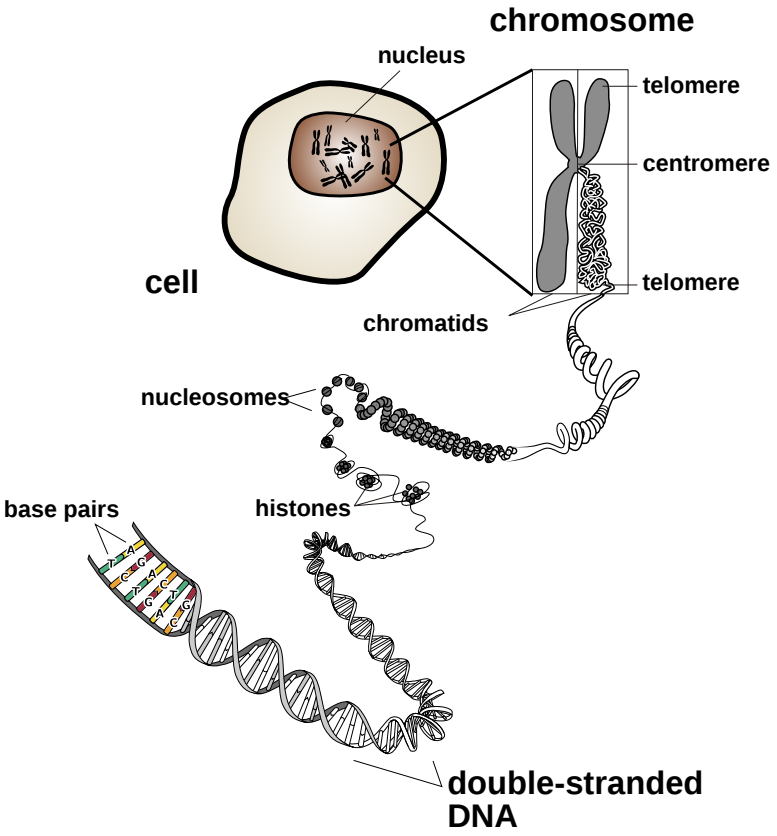
1.3.1.1 Definitions

DNA (Deoxyribonucleic acid) is a molecule present in every cell of living organisms that contains the genetic code of the organism. It is a macromolecule consisting in two complementary strands of nucleotides and it represents about 3 billions of nucleotide pairs in humans. A nucleotide is made of three elements: a phosphate group, a five-carbon sugar called deoxyribose and nucleobase. There are four types of nucleobase: Adenine (A), Thymine (T), Guanine (G) and Cytosine (C).

In eukaryotes, which are organisms such as humans characterized by the presence of a nucleus and mitochondria in their cells, the DNA molecule is divided into several segments, called chromosomes. Humans have 23 pairs of chromosomes, 22 pairs of autosomal chromosomes and one pair of sex chromosomes (XX or XY). After DNA replication, each chromosome is made of two identical chromatids attached by the centromere (see Figure 1.1 from http://www.genome.gov/Pages/Hyperion//DIR/VIP/Glossary/Illustration/chromosome.shtml).

A gene is a DNA segment that codes for a ribonucleic acid (RNA) that is then usually translated into amino acids to form a protein. However some genes code for functional RNAs that are not translated into proteins. Humans have between 20000 and 25000 genes, which represents less than

Figure 1.1 – *DNA.*



30% of their genome. Within a gene, some sequences called introns are transcribed into precursor RNA but are then removed during the formation of mature RNA by the splicing phenomenon. Exons are the remaining sequences that are joined together to form mature RNA and to code for the protein in the case of a protein-coding gene. Finally protein-coding DNA represents less than 2% of the genome.

1.3.1.2 DNA variability

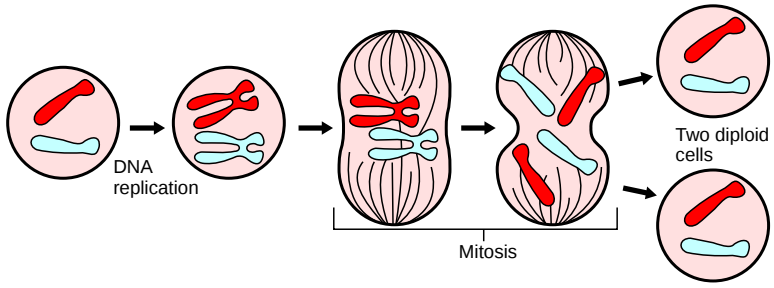
Causes There exist different sources of variability of human DNA. The first one is mutation. Mutations may be caused by an external event such as radiations, viruses, mutagenic chemicals. They may also be the consequence of errors during DNA replication, before cell division. They may be inherited if they occur in a cell that will become a sex cell and will be fertilized afterwards.

Another source of variability of human DNA may be the chromosomal events occurring during cell division. There are two types of cell divisions: mitosis and meiosis.

Mitosis is the chromosomal events occurring during classical cell duplication from one mother cell to two genetically identical daughter cells. During mitosis, the two chromatids of each chromosome of the mother cell split to form the genetic material of the two new cells. After mitosis, DNA will be duplicated in each daughter cell to re-form the two chromatids of each chromosome, making a new cell division possible. However some errors may occur during mitosis, resulting in the gain or the loss of a

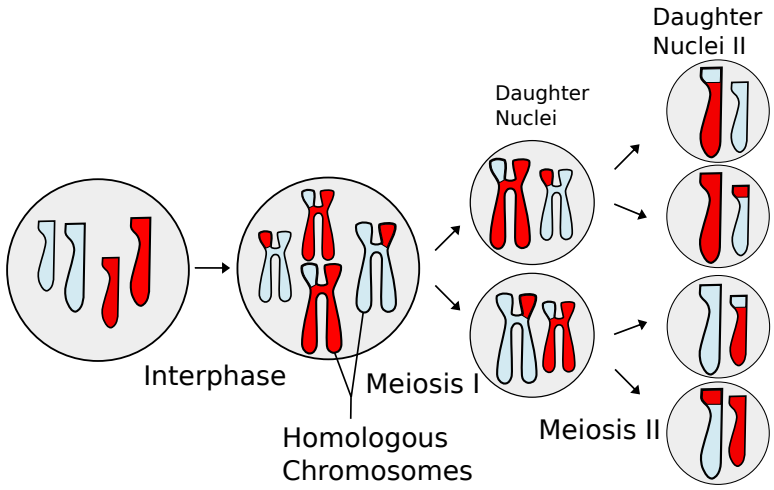
chromosome or a chromosomal segment, for the daughter cells (Figure 1.2 from http://www.ncbi.nlm.nih.gov/About/primer/genetics_cell.html).

Figure 1.2 – *Mitosis.*



Meiosis is the second type of cell division resulting in the production of sex cells or gametes. It is very similar to mitosis and may results in the same errors. However it differs from mitosis in two respects. First, recombination may occur since segments of chromatids may be exchanged between homologous chromosomes. This phenomenon is called crossing-over. Moreover, meiosis leads to four haploid (with only one chromosome of each pair) daughter cells called gametes, which will not divide again (Figure 1.3 from http://www.ncbi.nlm.nih.gov/About/primer/genetics_cell.html).

Figure 1.3 – *Meiosis and crossing-over.*



Types of variability Most variation at the DNA level consists in Single Nucleotide Polymorphisms (SNPs), small insertions and deletions, and some larger insertions and deletions called copy number variations (CNVs).

A SNP is a variation at a single nucleotide. According to current knowledge, this is the most common form of DNA variation. There are about 20 millions frequent SNPs and even more rare SNPs. Common SNPs are usually bi-allelic with two versions or alleles, the more frequent allele being called the major allele and the rarer the minor allele. They may be located within an exonic or intronic region of a gene that codes for a protein or not, or even outside any gene. Even though it lies within an exonic region of protein-coding gene, a SNP does not necessarily modify

the resulting protein and may be synonymous because of the redundancy of the genetic code. However it may also be missense (resulting in another amino acid) or nonsense (leading to a premature stop). Moreover synonymous and non-coding SNPs may still affect splicing, transcriptional or translational regulation and RNA stability.

CNVs are less numerous than SNPs. However they involve many more nucleotides. Indeed, a CNV is a polymorphism in the number of copies of a DNA segment at least one kilobase (1000 nucleotides) long, including deletions or duplications.

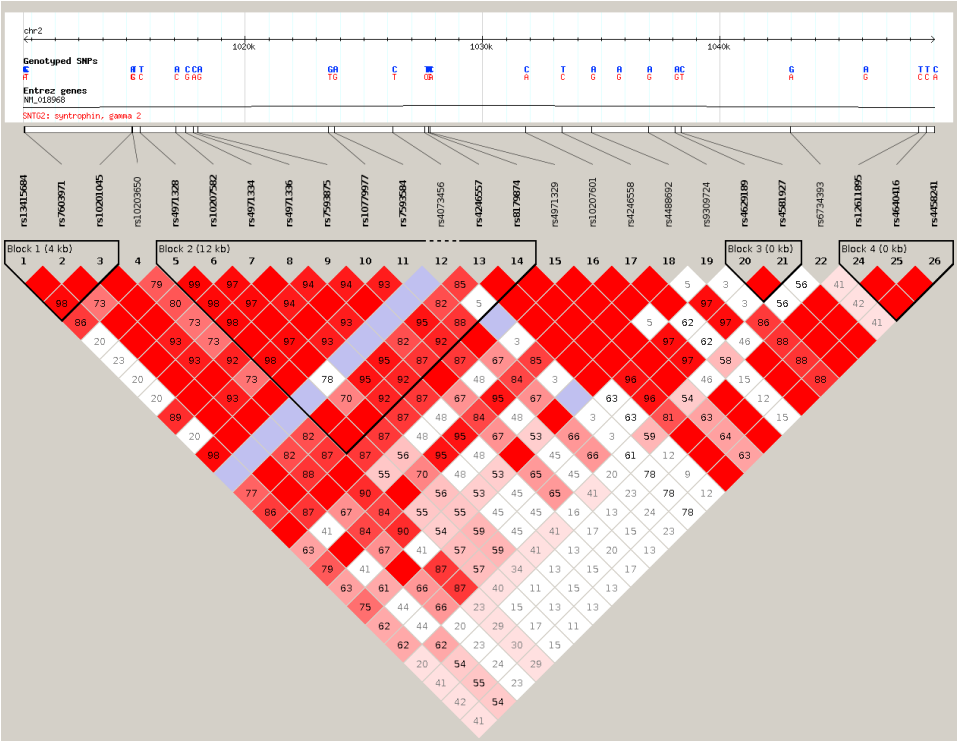
SNP and CNV Genotyping Current high-throughput genotyping platforms allows the genotyping of one million of SNPs using DNA microarray technology. CNV identification may also be indirectly performed using SNP genotyping platforms, but the most common way is to use arrays with specific CNV probes. Next-generation sequencing technologies will be more and more used in the future, allowing the genotyping of all rare variants.

1.3.1.3 Linkage Disequilibrium

There exists a correlation structure within genetic data, called linkage disequilibrium (LD), as can be observed in Figure 1.4. LD refers to the non-random association between the alleles of two genetic polymorphisms, at two distinct positions (loci). This means that certain pairs of alleles corresponding to the two loci are seen more often together on the same chromosome than expected by chance. This phenomenon is due to the physical link existing between neighbouring loci on the same chromosome, called genetic linkage, and the fact that their alleles are inherited together from one generation to another.

One cause of decay of LD is recombination, occurring during the crossing over step of meiosis. Thus one expects that the larger the physical distance (in terms of number of bases) between two polymorphisms, the higher the probability of recombination and thus the lower LD. However this is not always the case. Indeed, the extent of LD usually varies between a few kilobases and a few tens of kilobases in humans, but it may reach a few hundreds of kilobases in some DNA regions. This is due to the fact that the recombination rate is not constant over the genome. There exist some regions with a higher recombination rate, called hot spots, and some others with a low recombination rate, called haplotype blocks, of 10 to 20 kilobases on average up to a few hundreds of kilobases. Some regions are even almost non-recombining such as the Y chromosome, some parts of the X chromosome and the centromere-proximal regions of autosomal chromosomes. This leads to a different type of distance measure, called genetic distance, which depends on the recombination frequency between two loci and whose unit is the centimorgan (cM). A centimorgan corresponds to a recombination frequency of 0.01 in a single generation. By definition, the corresponding physical distance varies along the genome according to LD but on average one centimorgan represents about 1000 kilobases in humans.

Figure 1.4 – *Linkage disequilibrium computed for each SNP pair of chromosome 2 between 1010 kb and 1050kb using the HapMap database (V3 Release 2) and displayed using the Haploview software. The more red (and the higher the coefficient), the higher the LD.*



One interesting aspect of LD may be that it reduces the number of genetic polymorphisms necessary to capture most of the genetic variability, which implies that one only needs to genotype a reduced number of independent and highly informative SNPs, called tagSNPs. For instance 99% of common SNPs (with a minor allele frequency > 5%) are tagged with a LD of $r^2 > 0.8$ by one million of tagSNPS. Most of commercial chips use these tagSNPs.

A consequence is that in most cases when a genotyped SNP is found to be associated with a given phenotype, it is not the causal SNP itself but in LD with the causal SNP.

1.3.2 Other sources of genomic data

We will now present other sources of genomic data, which might be interestingly integrated in imaging genetics studies and whose understanding is anyway very useful in order to interpret the association results of genetic variants.

1.3.2.1 RNA

Definitions Ribonucleic acid is obtained from DNA after transcription. A RNA is a copy of a DNA segment, it only differs from DNA in a few respects: the desoxyribose sugar is replaced by a ribose, Thymine bases

are replaced by Uracil bases, it has a single strand, and finally it is shorter (between a few tens and a few thousands bases long).

Transcription involves the action of enzymes called RNA polymerase and of specific proteins called transcription factors. It is often followed by a maturation step where RNA modifications may occur such as splicing.

There exist different types of RNA. The first one is messenger RNA (mRNA) which carries the genetic information of a gene from DNA to the ribosome in order to produce the protein. There are also some other types of RNAs corresponding to non-protein coding genes, such as ribosomal RNAs, which play an enzymatic role in translation, or transfer RNAs (tRNAs), which bring amino-acids during translation.

Gene expression studies Gene expressions studies look at the expression level of a given gene in a particular cell or tissue type under specific conditions. Ideally one would like to measure the final product of a gene, that is the protein for protein-coding genes. However it is easier to measure messenger RNA levels than protein levels, which explains why gene expression analysis consists in measuring levels of RNA transcripts instead. The level of expression of a gene is of interest to study for instance the reaction of cell to a viral infection, to assess individual susceptibility to cancer, or to predict response to a drug. It may be also be used to distinguish between different types of cells that express different parts of the genetic code (for instance neuron tissues versus blood).

Different technologies can be used such as DNA microarrays, in which case one first needs to perform reverse transcription of the target mRNA to produce cDNA (complementary DNA), which can then be tested on a DNA microarray. However, this does not lead to an absolute quantification of mRNA levels and may only be used to compare mRNA levels between genes or between samples. Next-generation RNA sequencing techniques will solve this issue.

1.3.2.2 Proteins

Definitions Proteins are the final products of protein-coding genes. They are obtained after translation of mRNAs into amino-acid chains by a ribosome, with the help of tRNAs (which transport amino-acids to the ribosome) and enzymes. The genetic code of mRNAs is "read" by triplets of nucleotides, called codons, and each codon corresponds to a specific amino acid. Because of the redundancy of the genetic code, only 20 amino acids are encoded by codons; there are also a START codon and some STOP codons. The variation of a single nuceotide within a codon may be redundant or lead to a different amino acid or to a premature stop and thus to a different protein.

Proteomics After genotyping and transcriptomics, measuring the expression level of the different proteins may be of interest. Indeed mRNA level is not be directly proportional to protein level, and small changes in mRNA level may cause large changes in the level of the corresponding protein.

Proteomics may be very useful to develop effective diagnostic techniques, by using specific protein biomarkers, and to develop personalised disease treatments drugs that are more effective for the individual.

The most common techniques used to measure protein expression levels are protein electrophoresis, mass spectrometry and enzyme-linked immunosorbent assay.

1.3.2.3 Gene categorisation and pathways

After the detection of associations between genetic variants within some genes and a given phenotype, or the identification of genes that are differentially expressed in different conditions, it may be interesting to look for patterns among these genes and to determine whether some important biological relationships can emerge from them. There exist some categorisations of gene products, such as Gene Ontology, according to the type of cell component they belong to, their molecular function or the biological process they are involved in. One may also consider gene sets corresponding to molecular interaction networks, called pathways, such as in the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway database. Thus it may be interesting to test if the set of genes corresponding to a specific category or pathway is over-represented among the differentially expressed genes. However, it should be noted that these databases often derive from automatic annotations by analogy with other species. Other databases grouping specific genes frequently associated together with a disease may represent a more reliable source. Such databases are commercially available (e.g. the BioCarta database).

2 Statistical analysis in Neuroimaging or in Genetics: a similar problem

In the previous chapter, we have briefly detailed the different MRI modalities and the different types of genetic data that are commonly used in imaging genetics studies. Before moving to the joint analysis of these two types of data, we will try in this chapter to make a review of the statistical analysis methods that are classically used in each field respectively. In fact, we will see that statistical analysis methods are very similar in neuroimaging and in genetics. We will first present the conventional univariate approach and its application in both fields. Then, we will introduce some multivariate approaches common to both fields. Finally, we will show that both types of data are usually very-high dimensional and need an initial step of dimension reduction.

Contents

2.1 Conventional approach: massive univariate analysis

When searching for associations between a first variable of interest (such as a behavioural variable or a given phenotype) and several other variables of a different type (such as neuroimaging or genetic data), the simplest and most classical approach is to analyse this first variable versus each of the other variables independently. Such an approach is called a massive univariate analysis.

2.1.1 Univariate approach

We will now describe the univariate approach by considering only one pair of variables at once. One may first distinguish between two cases: when the two variables are quantitative and when at least one variable is qualitative. In the first case, a regression approach is classically used, while in the second case one rather performs t-tests, F-tests or Chi-squared tests. In this thesis, we will focus on the first case where all the variables are considered as quantitative. Nevertheless, we will also briefly describe the second case with categorical variables.

2.1.1.1 Simple Linear Regression

In this paragraph, where the two variables are supposed to be quantitative, we will note $\mathbf{x}$ the predictor variable and $\mathbf{y}$ the target variable to explain. $\mathbf{x}$ and $\mathbf{y}$ will be of size $n \times 1$, n being the number of samples. Moreover, without loss of generality, we will assume that both the predictor and the target are centred, that is to say they have zero mean.

The simple linear regression model searches for the coefficient β that minimises the quadratic error between $\mathbf{y}$ and its estimate $\mathbf{x}\beta$:

$$\min_{\beta} \|\mathbf{y} - \mathbf{x}\beta\|_2^2 \tag{2.1}$$

where the L2 (or Euclidean) norm of a vector $\mathbf{x}$ is given by $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$.

This is solved by computing the Ordinary Least Squares (OLS) estimate for β:

$$\hat{\beta}^{\text{OLS}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} \tag{2.2}$$

A statistical test may then be performed to test if β is significantly different from zero, by computing the corresponding t-statistic:

$$t = \frac{\hat{\beta}^{\text{OLS}}}{Std(\hat{\beta}^{OLS})} \tag{2.3}$$

where the standard deviation of $\hat{\beta}^{OLS}$ is $Std(\hat{\beta}^{OLS}) = \frac{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\sum_{i=1}^{n} x_i^2}$.

The degree of significance of this statistic may be obtained by deriving the corresponding p-value.

2.1.1.2 Two-sample t-test, F-test or Chi-squared test

When one of the two variables happens to be categorical, one may test if the mean of the quantitative variable differs across the different categories of the categorical one, using a two-sample t-test or an F-test, depending on whether the categorical variable has two or more categories. Finally, when the two variables are categorical, one may simply perform a Pearson's Chi-squared test to test for the independence between the two categorical variables. p-values can also be derived from these different statistical tests.

2.1.2 Application to the neuroimaging case: Classical Inference

The massive univariate approach is the most classical way to analyse neuroimaging data, where one performs a statistical test for each voxel of the image or for each feature of interest extracted from the image, searching for associations with the behavioural variable of interest (after having removed the effects of potential nuisance variables). This approach is called Classical Inference or Voxel-Based Analysis when the analysis is performed voxel-wise.

Neuroimaging data are classically quantitative data, while the behavioural/phenotypical variable may be either quantitative or qualitative. When this behavioural variable is quantitative, a linear regression approach will be used, by fitting a linear model for each voxel/feature of interest (the target variable to explain) with the behavioural variable as predictor. On the opposite, if the behavioural variable happens to be categorical, a two-sample t-test or an F-test will be preferred.

2.1.3 Application to the genetic case

2.1.3.1 Genome Wide Association Studies

Massive univariate analysis is also the most classical approach in genetic studies. It used to be performed on pedigrees with a sparse coverage of genetic polymorphisms, using linkage analysis, which tests for co-segregation (from parents to children) of each genetic polymorphism with a trait of interest and is based on a univariate statistic called a LOD (Logarithm Of Odds) score. With the development of new high-throughput genotyping techniques, one has moved towards association studies on unrelated individuals with a dense coverage of genome variability, where the association between the phenotype and each genetic polymorphism, each SNP for instance, is tested independently. In the case of genome-wide coverage of genetic polymorphisms, this approach is called a Genome-Wide Association Study (GWAS).

SNP data may be considered either as quantitative or categorical data. Indeed for each bi-allelic SNP (e.g. allele A and allele T), the possible genotypes may either be coded as AA, AT and TT, or 0, 1 and 2 when using the additive genetic coding where the genotype represents the number of minor alleles (T alleles in our example). When faced with a quantitative phenotype, SNPs are often considered as quantitative variables (assuming additive genetic effects) and are successively tested as potential

predictors of the phenotype (the target variable), using a simple linear regression model. On the opposite, in the case of a categorical phenotype, they are rather seen as categorical variables and the genetic analysis is classically performed using Chi-squared tests. Please note that in the rest of this thesis, we will consider SNPs as quantitative variables, assuming the additive genetic model. Nevertheless, different genetic models, such as dominant/recessive or genotypic models, could also be investigated in further work.

2.1.3.2 Differential analysis of gene expression data

Similarly, in the case of gene expression data, the massive univariate approach is classically used to identify the genes that are differentially expressed between two conditions, such as cancerous versus non-cancerous cells for instance. Since gene expression data are typically quantitative, a two-sample t-test can be used for each gene to compare the mean expression between the two conditions.

2.1.4 Limitations

While univariate techniques are relatively simple, they encounter a multiple comparison issue. Indeed the p-value obtained for each test has to be corrected for the number of tests performed, which may be very high both in neuroimaging and in genetics. For instance, each MRI image usually represents between a few 10s of 1000s and a few millions of voxels and there are about one million of SNPs on common SNP microarrays. Bonferroni correction is the most common correction and controls the family-wise error rate by dividing the p-value significance threshold by the number of tests performed. This approach is very stringent and assumes that the tests are independent. A less conservative correction such as False Discovery Rate may also be preferred. Finally a permutation-based correction may also be used when nothing is known about the distribution of the variables, but it is computationally intensive.

Moreover, the link between the behavioural variable (or phenotype) and the neuroimaging (or genetic) data is likely to be in part multivariate. For instance in common traits or diseases, the phenotype is very likely to be influenced by the interactions of several SNPs or genes (this phenomenon is called epistasis), or at least by some cumulative genetic effects. Similarly, different brain regions may also be correlated and associated to the same behavioural variable.

2.2 Multivariate approaches

In order to address the limitations of the univariate approach mentioned in the previous section, multivariate approaches have also become popular both in neuroimaging and in genetics. The purpose of such methods is to predict a *target* variable (a behavioural variable or a given phenotype for instance) from several *predictor* variables such as neuroimaging or genetic data. In the rest of this chapter, we will note $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ the p quantitative predictor variables and $\mathbf{y}$ the target variable to explain. $\mathbf{X}$ will be the

matrix of predictors of size $n \times p$ and $\mathbf{y}$ the target of size $n \times 1$, n being the number of samples.

Multivariate predictive methods may be divided into several categories. For instance, one may distinguish classification methods in the case of a qualitative behavioural variable/phenotype and regression methods in the case of a quantitative variable.

Similarly to the previous section 2.1 on the univariate approach, we will focus on the case where all the variables are considered as quantitative, that is to say on regression methods, and more specifically on linear models. Nevertheless, we will briefly mention a few classical classification methods. Finally, we will see that the multivariate approach may be extented to the case of several target variables, using some multivariate methods designed for the analysis of two blocks of data.

2.2.1 Linear regression models

In this section, we consider that all variables are quantitative. Moreover, without loss of generality, we will assume that both the predictors and the target are centred, which means they have zero mean.

2.2.1.1 Multiple linear regression

The multiple linear regression model searches for the vector of coefficients β of size $p \times 1$ that minimises the quadratic error between $\mathbf{y}$ and its estimate $\mathbf{X}\beta$:

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \tag{2.4}$$

When the problem is well-posed, that is when $\mathbf{X}$ is full rank and thus $\mathbf{X}'\mathbf{X}$ is invertible, the solution is obtained by computing the unbiased Ordinary Least Squares (OLS) estimate:

$$\hat{\beta}^{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{2.5}$$

However, this estimate is very sensitive to the conditioning of $\mathbf{X}$ and may have a large variance for small changes of $\mathbf{X}$ in the case of highly correlated predictors (multicollinearity) for instance. Moreover, in high-dimensional settings where $n \ll p$, the inverse of $\mathbf{X}'\mathbf{X}$ is not even defined anymore. To solve these issues, regularisation techniques based on some constraint on β can be used, which introduces some bias in the estimation of β but reduces its variance, leading to a better estimation at the end. We will now present three classical forms of regularised regression: Ridge Regression, Lasso Regression and Elastic Net Regression.

2.2.1.2 Ridge Regression: L2-regularisation

Ridge regression was first presented by Hoerl and Kennard (1970) and is based on adding to the OLS criterion a constraint on the L2-norm of the regression coefficients:

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_2 \|\beta\|_2 \tag{2.6}$$

with $\lambda_2 \geq 0$ and $\|\beta\|_2 = \sqrt{\sum_{i=1}^{p} \beta_i^2}$.

It may be solved by replacing $\mathbf{X}'\mathbf{X}$ by $\mathbf{X}'\mathbf{X} + \lambda_2\mathbf{I}$ in the OLS estimate of β, which gives more weight to the diagonal elements of the scatter matrix and facilitates its inversion. Thus, the solution becomes:

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \lambda_2\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \tag{2.7}$$

One interesting property of Ridge Regression is that it gives similar coefficients to correlated predictors, reducing the variability of the coefficients that occurs with multiple regression in high dimensions. An extreme case of regularisation would lead to the replacement of $\mathbf{X}'\mathbf{X}$ by $\lambda_2\mathbf{I}$, ignoring the correlations within predictors and becoming equivalent to several independent simple linear regressions, up to a factor $\frac{1}{\lambda_2}$.

However, Ridge regression does not assign exactly zero coefficients to predictors, while in high-dimensional settings many variables are expected to be irrelevant and should be removed from the model. We will see now that such variable selection may be achieved using Lasso regression.

2.2.1.3 Lasso Regression: L1-regularisation

Lasso (Least Absolute Shrinkage and Selection Operator) regression (Tibshirani 1996) is based on applying a penalty on the L1-norm of the coefficient vector β. The following criterion is optimised:

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1 \tag{2.8}$$

with $\lambda_1 \geq 0$ and $\|\beta\|_1 = \sum_{i=1}^{p} |\beta_i|$.

It can be solved by different algorithms such as Least Angle Regression (LARS) (Efron et al. 2004) or coordinate descent (Friedman et al. 2007). Morevoer, if $\mathbf{X}'\mathbf{X}$ is diagonal, i.e. if the predictor variables are uncorrelated, the solution may be obtained more easily by soft-thresholding for each predictor $\mathbf{x}_i$:

$$\hat{\beta}_i^{\text{lasso}} = g_{\lambda_1/2}((\mathbf{x}_i'\mathbf{x}_i)^{-1}\mathbf{x}_i'\mathbf{y}) \tag{2.9}$$

where $g_\lambda(y) = \text{sign}(y)(|y| - \lambda)_+ = \text{sign}(y)\max(0, |y| - \lambda)$ is the soft-thresholding function.

One interesting property of Lasso regression is that, contrary to Ridge regression, it leads to a sparse solution for β, by selecting at most n non-null coefficients when $n \ll p$. This may be of interest when interpreting the results. However, because it tends to select only one variable among a set of correlated variables, the selection may be unstable and the interpretation of the results should remain very careful.

2.2.1.4 Elastic Net Regression: L1+L2 regularisation

Elastic Net regression was introduced by Zou and Hastie (2005) and is a combination of L1 and L2 regularisations. Indeed, the following criterion is optimised:

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2 \tag{2.10}$$

with $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$.

Like Lasso regression, it can be solved by different algorithms such as Least Angle Regression (LARS) (Efron et al. 2004) or coordinate descent (Friedman et al. 2007).

Elastic Net regression has the advantages of both Ridge and Lasso regressions, in that it yields a sparse and interpretable solution while it gives similar and relatively stable coefficients to correlated predictors.

2.2.1.5 Partial Least Squares Regression

Partial Least Squares (PLS) regression, in the univariate case with only one target variable to predict (called PLS1), is used to model the association between the target variable and the predictor variables, hypothesising that the predictors may be summarised by unobserved latent variables. A latent variable (or component) extracted from a block of variables is a linear combination of the observed variables of this block.

More precisely, PLS1 builds successive and orthogonal latent variables from the predictor block, such that at each step the covariance between the latent variable and the target variable is maximal. For each step h in $1..H$, where H is the maximal number of components and equals the rank of $\mathbf{X}$, it optimises the following criterion:

$$\max_{\|\mathbf{u}_h\|_2=1} \operatorname{cov}(\mathbf{X}_{h-1}\mathbf{u}_h, \mathbf{y}) \tag{2.11}$$

$$= \max_{\|\mathbf{u}_h\|_2=1} \operatorname{corr}(\mathbf{X}_{h-1}\mathbf{u}_h, \mathbf{y}) \sqrt{\operatorname{var}(\mathbf{X}_{h-1}\mathbf{u}_h)} \sqrt{\operatorname{var}(\mathbf{y})}$$

where $\mathbf{u}_h$ is the weight vector for the linear combination of the predictors of block $\mathbf{X}$. $\mathbf{X}_{h-1}$ is the residual (deflated) $\mathbf{X}$ after its regression on the $h-1$ previous latent variables , starting with $\mathbf{X}_0 = \mathbf{X}$.

Once the variables are standardised, the previous criterion for each new pair of components is equivalent to optimising:

$$\max_{\|\mathbf{u}_h\|_2=1} \mathbf{u}_h' \mathbf{X}_{h-1}' \mathbf{y} \tag{2.12}$$

It should be noted that the criterion tends to maximise the relative value of the covariance, which implies that the covariance is forced to be null or positive. In the case of a negative association between a predictor variable from block $\mathbf{X}$ and the target variable $\mathbf{y}$, a negative weight will thus be assigned to this predictor in order to obtain a positive covariance.

Since the variables are standardised, the optimal weight vector $\mathbf{u}_h$ (before its normalisation) consists in fact of the univariate correlation coefficients between each variable of block $\mathbf{X}_{h-1}$ and the target variable $\mathbf{y}$. Thus, on the first component, it is equivalent to an extreme regularisation of Ridge regression, up to a factor as mentioned before.

If $\mathbf{X}$ is full-rank, the final model obtained with H components corresponds to the OLS solution.

2.2.1.6 Application in neuroimaging: Inverse Inference

In neuroimaging, such multivariate approaches are called inverse inference. Indeed, instead of trying to explain the signal within each voxel

using for instance a behavioural variable, the different voxels (the predictors) are used simultaneously to predict the behavioural variable (the target).

Inverse inference has to face the overfitting issue due to the very high dimensionality of the imaging data, with up to one million of voxels per image. This is the typical case where $n \ll p$ and regularisation is needed. Recently, regularised regression such as Elastic Net regression has been successfully used for the purpose of prediction of behavioural data from fMRI data (Carroll et al. 2009). PLS regression has also been applied by Giessing et al. (2007) to predict a behavioural variable from fMRI data.

2.2.1.7 Application to the genetic case

Similarly in genetics, one may try to explain a given phenotype of interest (the target) using the joint effects of several SNPs (the predictors). Up to one million of SNPs can be genotyped on common genome-wide chips, which leads to the exact same overfitting phenomenon as in neuroimaging. Regularised regression, such as Lasso regression, has also been successfully applied in the case of association studies, in order to identify the genetic polymorphisms associated with a given phenotype (Shi et al. 2011). PLS regression has also been used to predict a phenotype from SNP data (Sarkis et al. 2006).

2.2.2 Classification methods

Even though this is not the subject of this thesis as mentioned earlier, we will now give a brief overview of the classification methods that are classically used in both the neuroimaging and the genetics fields. Indeed, it may be interesting to notice the strong similarities between the statistical methods commonly used within each field, like in the regression case.

Within classification methods, one may distinguish between discriminative and generative approaches.

Discriminative approaches

A discriminative approach computes directly the probability *a posteriori* of the label of the target given the value of the predictors $p(\mathbf{y} = k|\mathbf{X})$, such as Support Vectors Machines or Logistic Regression. The advantage of such approaches is that they perform well when the distribution of the data is unknown or difficult to model. Moreover they are often relatively simple and computationally efficient. However, the interpretation of the results may be delicate sometimes, since they aim at predicting the target rather than modelling the data.

Generative approaches

On the other hand, a generative approach will estimate the probability of the different labels for the target $p(\mathbf{y} = k)$ and the distribution of the predictors given the label of the target $p(\mathbf{X}|\mathbf{y} = k)$, allowing the generation of new data. Generative approaches yield interpretable results but require more assumptions and are more computationally expensive. The

generative approaches with the most simple assumptions, such as Naive Bayes classifiers or Linear Discriminant Analysis (LDA), are also the most common and efficient ones.

Applications to inverse inference in neuroimaging and to classification of genetic data

Concerning discriminative approaches, Support Vectors Machines are classically used both for the classification of neuroimaging patterns (Cox and Savoy 2003) and for the classification of gene expression data (Speed 2003). Logistic Regression is also commonly used in both fields, usually in a regularised form to deal with the high dimensionality of the data, such as sparse or elastic net logistic regression in neuroimaging (Yamashita et al. 2008, Ryali et al. 2010) and penalised logistic regression for the classification of microarray data (Zhu and Hastie 2004).

As for simple generative approaches, Naive Bayes classifiers have also been applied in neuroimaging (Mitchell et al. 2004) and in genetics to predict multi-genic diseases from SNP data (Moore et al. 2006). Similarly, LDA has been applied in neuroimaging (Cox and Savoy 2003) and for the classification of gene expression data (Hakak et al. 2001).

2.2.3 Two-block methods

In this section, we will consider several quantitative target variables forming a matrix $\mathbf{Y}$ of size $n \times q$, where q is the number of target variables considered.

PLS Regression and Canonical Correlation Analysis

In order to analyse jointly two blocks of data, several techniques have been developed such as two-block PLS Regression (PLS2) and Canonical Correlation Analysis, which will be described in Chapter 5. They have been successfully applied in the genetics field, in order to analyse jointly gene expression and CNV data for instance.

PLS-SVD

A variant of PLS2, called Tucker Inter-battery Analysis (Tucker 1958) or PLS-SVD (McIntosh et al. 1996), has also been widely applied in the neuroimaging field. This variant is symmetric and consists in computing the complete Singular Value Decomposition (SVD) of $\mathbf{X}'\mathbf{Y}$ at once. All the pairs of left and right singular vectors will form weight vectors $\mathbf{a}_h$ and $\mathbf{b}_h$ for $\mathbf{X}$ and $\mathbf{Y}$ variables respectively. The linear combinations thus obtained, $\mathbf{Xa_h}$ and $\mathbf{Yb_h}$, are called latent variables. PLS-SVD gives the same results as PLS regression on the first pair of latent variables, but differs on further pairs due to a different orthogonality constraint. While PLS regression forces successive latent variables of each block to be orthogonal, PLS-SVD forces successive weight vectors of each block to be orthogonal, which leads to the orthogonality between each latent variable $\mathbf{Xa_h}$ of block $\mathbf{X}$ and each latent variable $\mathbf{Yb_j}$ of block $\mathbf{Y}$, as long as they are of different order ($h \neq j$).

2.3 Dimension Reduction

The critical issue encountered by multivariate methods is the overfitting issue, which occurs in the case of very high-dimensional data as mentioned earlier. When regularisation is not sufficient to face the overfitting issue, they may be combined with a preliminary dimension reduction methods.

Dimension reduction is essentially based on two paradigms: feature extraction and feature selection. Feature extraction looks for a low-dimensional representation of the data, while feature selection aims at removing irrelevant features.

2.3.1 Feature extraction

We will now briefly describe two classical feature extraction methods: Principal Components Analysis (PCA) and Independent Component Analysis (ICA). It should be noted that these two techniques are unsupervised methods, since they are performed on the predictor block $\mathbf{X}$ to reduce its dimensionality without taking into account the target $\mathbf{y}$. However, one may also use supervised feature extraction methods, such as PLS regression, since the identification of PLS components may also be seen as feature extraction.

2.3.1.1 Principal Component Analysis

Principal Components Analysis (PCA) transforms a set a correlated variables into a set of linearly uncorrelated components, called principal components, that explain most of the variability of the data. The components are ordered according to the amount of variance of the original data that they explain.

Assuming that the variables of block $\mathbf{X}$ are centred, principal components are obtained by computing the Singular Value Decomposition (SVD) of $\mathbf{X}$: $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}'$, where $\mathbf{U}$ is an $n \times n$ orthogonal matrix, $\mathbf{\Lambda}$ an $n \times p$ diagonal matrix and $\mathbf{V}'$ an $p \times p$ orthogonal matrix. The principal components are the columns of $\mathbf{U}\mathbf{\Lambda}$, while the different columns of $\mathbf{V}$ are the weights of the original variables for the different principal components.

The number of components necessary to explain the variability of $\mathbf{X}$ depends on the rank of $\mathbf{X}$ and is smaller than or equal to $\min(n, p)$. Thus, PCA may be used to reduce the dimensionality of the data without loosing too much information, by keeping the k first principal components that explain a certain percentage of the variance.

2.3.1.2 Independent Component Analysis

A classical Independent Component Analysis (ICA) model assumes that there exist some non-observable, non-gaussian and independent sources, and that the observed data are a linear mixing of those sources:

$$\mathbf{X} = \mathbf{A}\mathbf{S} \tag{2.13}$$

where $\mathbf{S} = [\mathbf{s_1}, ..., \mathbf{s_m}]^T$ is an $m \times n$ matrix made of m random variables that correspond to m independent sources, $\mathbf{X} = [\mathbf{x_1}, ..., \mathbf{x_p}]^T$ is a $p \times n$ matrix

representing the p observed random variables, and $\mathbf{A}$ is a $p \times m$ mixing matrix. Typically, we have $m < p$, which implies that $\mathbf{A}$ is full-rank.

The goal of ICA is to find an approximation $\mathbf{Z}$ of the sources:

$$\mathbf{Z} = \mathbf{WX} \tag{2.14}$$

where $\mathbf{W}$ is an unmixing matrix of size $m \times p$.

Different algorithms may be used to solve this problem and we will only very briefly detail one of them, called the Infomax algorithm (Bell and Sejnowski 1995). The Infomax algorithm looks for a $\mathbf{W}$ matrix that maximises an entropy function:

$$\max_{\mathbf{W}} \{H(Z) = -E[\ln(f(\mathbf{Z}))]\} \tag{2.15}$$

where H is the entropy function, E the expected value, f the probability density function and $\mathbf{Z}$ the logistic function:

$$\mathbf{Z} = \frac{1}{1 + \exp^{-\mathbf{U}}}, \text{ with } \mathbf{U} = \mathbf{WX} + \mathbf{W}_0 \tag{2.16}$$

This optimisation problem may then be solved using a gradient-based approach.

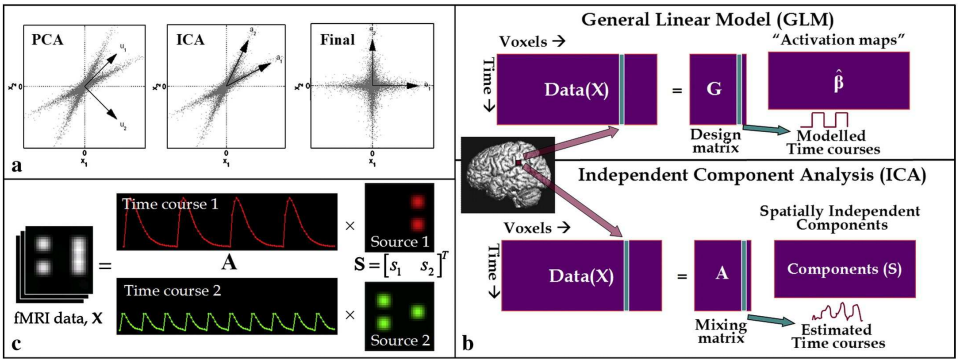
2.3.1.3 Applications in neuroimaging and in genetics

Dimension reduction techniques based on feature extraction are widely used both in the neuroimaging and genetics fields.

For instance, PCA may be used to reduce the dimensionality of neuroimaging data before using regression or classification techniques, as in the work by Strother et al. (2002) in the case of fMRI data. As for gene expression data, PCA is also commonly applied to identify dominant patterns of gene expression and to reduce dimensionality (Hastie et al. 2000, Khan et al. 2001). Similarly, PCA may be applied to SNP data at the gene level (Horne and Camp 2004), or to CNV data (Soneson et al. 2010). PCA may also be used as a quality control procedure on genetic data (such as SNPs) to identify and remove a hidden sampling effect such as ethnicity (Novembre et al. 2008), which could lead to confounding effects and biased results.

ICA has become a common tool as well in neuroimaging, especially for the analysis functional MRI data. It should be noted that most applications of ICA to fMRI seek spatially independent components, called brain modes, considering time points (and subjects in the case of a group study) as the random variables and voxels as the observations of these random variables (see figure 2.1). Calhoun et al. (2008) proposed to use spatial ICA at the group level (Calhoun et al. 2001) in order to extract two brain modes, before classification of the subjects relatively to a phenotype of interest. However, it should be noted that ICA often requires a first dimension reduction step performed by PCA, so that it can be applied afterwards. Similarly, Liu et al. (2004) performed ICA on gene expression data before applying classifiers.

Figure 2.1 – *a) Comparison of the results of PCA and ICA. b) Comparison of GLM and spatial ICA for fMRI data. c) Illustration from spatial ICA for fMRI data. Taken from Calhoun et al. (2009).*



2.3.2 Feature selection

Feature selection methods may be divided into two categories: some univariate methods, which select relevant features independently from each other, and some multivariate methods, which consider feature interrelations to select a subset of variables (Guyon et al. 2006). These methods are supervised methods since they take into account the target $\mathbf{y}$.

2.3.2.1 Univariate filters

The univariate methods based on statistical tests that have been presented in section 2.1, such as simple linear regression, two-sample t-test, F-test or Chi-squared test, may be used as a first step of dimension reduction, called filtering, by ranking the different predictor variables and selecting only the best ranked variables. A multivariate method may then be used on selected variables.

Univariate filters are relatively simple but ignore the potential interactions between predictors.

2.3.2.2 Multivariate feature selection

In order to deal with the limitations of univariate filters, most of classification and regression multivariate methods mentioned in section 2.2, such as SVM, regularised regression or PLS regression, may also be used as a preliminary step of feature selection. Indeed the different predictor variables may be ranked according to the absolute value of the weights they have been assigned by the predictive function, and one may select the best ranked variables. Then, a second multivariate predictive method (potentially the same one) may be used on selected variables.

The selection process may also be performed iteratively, such as in the Recursive Feature Elimination (RFE) approach (Guyon et al. 2002), where the predictive function is first computed on all variables, then the worst ranked are removed and a new predictive function is computed on the remaining variables, and so on, up to a given number of remaining variables. With such an approach, one does not need to apply a second multivariate predictive method, since the final predictive function computed at the last

iteration plays that role. In that case, feature selection is included in the learning process of the predictive function, and this is called an embedded method.

It should be noted that in the case of Lasso and Elastic Net regressions, feature selection is already embedded in the learning process of the predictive function.

The main drawbacks of these multivariate feature selection techniques are that they are more computationally intensive than univariate filters and prone to overfitting as mentioned in section 2.2.

2.3.2.3 Applications in neuroimaging and in genetics

Univariate filters have been successfully applied on fMRI data by (Cox and Savoy 2003) before using multivariate classifiers such as SVM or LDA. Multivariate feature selection has also provided good results as well, such as RFE-SVM (De Martino et al. 2008) or Elastic Net logistic regression(Ryali et al. 2010) for the classification of fMRI data, especially when combined with some initial univariate filtering.

Similarly, univariate filters are widely used as an initial dimension reduction step for classification of gene expression data and seem to perform even better than multivariate approaches such as RFE-SVM, Lasso or Elastic Net (Haury et al. 2011).

3 Statistical analysis in Imaging Genetics: a joint analysis

In the previous chapter, we have shown the similarity between the statistical analysis methods that are classically used in neuroimaging and in genetics respectively. In this chapter, we are now moving to the existing methods for the joint analysis of these two types of high-dimensional data, which are basically extensions of the methods from the previous chapter to the case of multiple variables within both blocks of data. We will first describe the conventional univariate approach and then multivariate approaches, namely principal components regression, penalised regression and two-block multivariate methods.

Contents

In the rest of this thesis, we will note $\mathbf{Y}$ the neuroimaging data matrix of size $n \times q$ and $\mathbf{X}$ the SNP matrix of size $n \times p$, n being the number of subjects, q the number of neuroimaging phenotypes and p the number of SNPs. Imaging phenotypes are classically quantitative variables. Thus, unless otherwise specified, SNPs will be considered as quantitative variables as well, assuming an additive genetic model.

3.1 Conventional approach: massive univariate analysis

3.1.1 Voxelwise Genome-Wide Association Studies

When searching for associations between two sets of variables, such as neuroimaging and genetic data, the simplest approach is again to perform a massive univariate analysis, where one tests each variable from the first set versus each variable of the second set independently.

A simple linear regression model is classically used to regress each voxel (or each imaging phenotype extracted from the image) on each SNP, assuming the additive genetic model (Stein et al. 2010). However, SNPs may also be considered as categorical data, and one may use an F-test without any assumption on the genetic model or a t-test if one assumes a dominant or recessive model.

3.1.2 Limitations

The limitations of the univariate approach are the same as the ones mentioned in the previous chapter, in the case of one set of predictor variables versus a target variable.

The first one is the multiple comparisons issue, which is even more important in imaging genetics than in neuroimaging or in genetics. Indeed, one has to correct for the $p \times q$ tests performed in the case of a voxel-wise genome-wide association studies, which may reach 10^{12}. Bonferroni correction is the most common procedure but is very conservative, especially when variables are correlated as it is the case for SNPs due to linkage disequilibrium. Thus, (Stein et al. 2010) proposed a less stringent correction based on PCA of the SNP data, in order to determine the effective number of independent tests by computing the number of principal components necessary to explain 99.5% of the SNP variance. However, it was not sufficient in this example to allow SNPs to survive the corrected significance threshold.

The second limitation is that it does not take into account the multivariate part of the link between genetic and imaging data, while for instance epistasis or pleiotropy are likely phenomena in common traits. Indeed, brain imaging endophenotypes are probably influenced by the combined effects of several SNPs and different brain regions may also be correlated and influenced by the same SNP(s).

3.2 Multivariate approaches

In order to address the limitations of univariate analysis, multivariate methods have recently been used in imaging genetics studies. We will first present two classical examples, principal components regression and penalised regression, where the multivariate nature of the genetic data has been taken into account. Then we will move to two-block multivariate methods, which are designed for the joint analysis of two blocks of data and take advantage of the multivariate aspect within each block.

3.2.1 Classical one-block multivariate methods

Classical multivariate methods, as the ones mentioned in Chapter 2, are usually designed for the analysis of one block of predictor variables versus a single target variable. Nevertheless, these methods may also be of interest for the analysis of two sets of variables $\mathbf{Y}$ and $\mathbf{X}$, such as neuroimaging and genetic data, since they may be applied on each target variable from block $\mathbf{Y}$ independently. We give two examples that have been used in imaging genetics studies: principal components regression and penalised regression.

3.2.1.1 Principal Component Regression

A first way to partially take into account epistasis may be to perform a gene-based analysis for testing the joint effect of the different SNPs within each gene across the voxels of the whole brain (Hibar et al. 2011). The authors propose to perform voxel-wise and gene-wide Principal Component Regression (PCR), by applying PCA on the SNPs of each gene, keeping the first principal components that explain 95% of the SNP variance for each gene, and fitting a multiple regression model of the signal at each voxel onto the selected principal components of each gene. An overall p-value may thus be derived for each gene at each voxel. Moreover they compute the effective number of independent tests (to correct for) by permutations and simulations, accounting for LD and spatial smoothness within the image. In the case of small genes with a diluted effect, gene-wide PCR appears to be more powerful to detect associations than SNP-based regression.

3.2.1.2 Regularised regression

Another way to partially account for the potential interactions between SNPs may be to perform regularised regression, such as Ridge regression, of each imaging phenotype onto a set of neighbouring SNPs (Kohannim et al. 2011). The authors proposed to select the SNPs that passed a univariate GWAS p-value threshold (p-value ≤ 0.1) with a structural MRI-based measure, such as hippocampal or temporal lobe volume. Then, they performed Ridge regression of the structural measure onto different genomic windows centred on selected SNPs and of fixed sizes (50 Kbp, 100 Kbp, 500 Kbp or 1 Mbp), and a p-value could be derived for each SNP of the window from regression coefficients. p-values were corrected for the effective number of independent univariate tests performed along the genome,

using PCA on SNPs like in Stein et al. (2010). The results they obtained tend to show that the power of GWAS may be increased by incorporating multi-SNP dependencies in the model. However, it should be noted that if p-values were corrected for the number of univariate tests, there was no correction for the number of multiple regressions performed.

3.2.1.3 Limitations

The first limitation of these two approaches is that they do not account for long-range multivariate effects among SNPs. Indeed, only the interactions between SNPs within the same gene or in the same neighbourhood are taken into account.

Moreover, they do no take advantage of the multivariate nature of neuroimaging data since each imaging phenotype is analysed independently.

Finally, the multiple comparisons issue remains critical, since one still has to correct for the number of genes (or genomic regions) and imaging phenotypes tested.

3.2.2 Two-block multivariate methods

In order to fully account for the multivariate aspect within each block of data, two-block multivariate methods may be used. We present here three methods that have shown promising results on imaging genetics data: Parallel Independent Component Analysis, Sparse Reduced Rank Regression and Group-Sparse Multi-Task Regression.

3.2.2.1 Sparse Reduced Rank Regression

We first recall the notations $\mathbf{Y}$ (of size $n \times q$) for the q imaging phenotypes and $\mathbf{X}$ (of size $n \times p$) for the p SNPs. Without loss of generality, we assume that all variables are centred and normalised. We will first describe the Reduced Rank Regression model introduced by Anderson (1951) and then its sparse version introduced by Vounou et al. (2010).

Reduced Rank Regression To better understand Reduced Rank Regression (RRR), let us first recall the standard multivariate multiple regression model:

$$\mathbf{Y} = \mathbf{X}\mathbf{C} + \mathbf{E} \tag{3.1}$$

where $\mathbf{C}$ is the matrix of regression coefficients of size $p \times q$ and $\mathbf{E}$ the matrix of residual errors of size $n \times q$.

If $n \geq p$ and $\mathbf{X}$ is full-rank, the least squares estimate of the matrix of regression coefficients $\mathbf{C}$ may be obtained by minimising:

$$\min_{\mathbf{C}} \|\mathbf{Y} - \mathbf{X}\mathbf{C}\|_F^2 \tag{3.2}$$

where the squared Frobenius norm of a matrix $\mathbf{M}$ is given by $\|\mathbf{M}\|_F^2 = \text{Trace}\left\{\mathbf{M}\mathbf{M}'\right\} = \sum_{i=1}\sum_{j=1} m_{ij}^2$.

If $\mathbf{C}$ is supposed to be full-rank, the ordinary least square solution (extended to several target variables) is the maximum likelihood estimate for $\mathbf{C}$:

$$\hat{\mathbf{C}}^{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{3.3}$$

In fact, this solution is equivalent to q independent multiple regressions of each variable of block $\mathbf{Y}$ onto $\mathbf{X}$, which means that the potential multivariate nature of $\mathbf{Y}$ is not accounted for. Moreover, in imaging genetics studies, we are typically in the case where the number of subjects is very low compared to the number of SNPs ($n \ll p$) and where there is multi-collinearity between SNPs. Thus, the OLS estimate for $\mathbf{C}$ is not even defined.

A way to solve theses issues is to impose a rank condition on $\mathbf{C}$, that is $\text{rank}(\mathbf{C}) = r < \min(p, q)$, as in the Reduced Rank Regression model by Reinsel and Velu (1998). If $\mathbf{C}$ is supposed to have a rank r, it may be decomposed into the product of two matrices of rank r and the regression model may be rewritten as follows:

$$\mathbf{Y} = \mathbf{XBA} + \mathbf{E} \tag{3.4}$$

where the matrices $\mathbf{A}$ and $\mathbf{B}$ are full-rank, of size $p \times r$ and $r \times q$ respectively.

The maximum likelihood estimate of the reduced-rank model can be obtained by minimising the following weighted least squares criterion:

$$\min_{\mathbf{A},\mathbf{B}} \text{Trace}\left\{(\mathbf{Y} - \mathbf{XBA})\mathbf{\Gamma}(\mathbf{Y} - \mathbf{XBA})'\right\} \tag{3.5}$$

where $\mathbf{\Gamma}$ is set to be $((\mathbf{Y} - \mathbf{X}\hat{\mathbf{C}}^{\text{OLS}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{C}}^{\text{OLS}}))^{-1}$. Alternatively, one may set $\mathbf{\Gamma}$ to $(\mathbf{Y}'\mathbf{Y})^{-1}$, which leads to the same solutions for $\mathbf{A}$ and $\mathbf{B}$, up to some scaling, and to the exact same solution for $\mathbf{C}$.

The solutions may be obtained by using the singular value decomposition of the matrix $\mathbf{D} = (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{Y}\mathbf{\Gamma}^{1/2}$:

$$\mathbf{D} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}' \tag{3.6}$$

where $\mathbf{U}$ is the $p \times p$ orthogonal matrix whose columns are the right singular vectors, $\mathbf{V}$ is the $q \times q$ orthogonal matrix whose columns are the left singular vectors and $\mathbf{\Lambda}$ is a $p \times q$ diagonal matrix whose diagonal elements are the singular values. The estimates $\hat{\mathbf{A}}_r$ and $\hat{\mathbf{B}}_r$ that minimise (3.5) are:

$$\hat{\mathbf{A}}_r = \mathbf{V}_r'\mathbf{\Gamma}^{-1/2} \tag{3.7}$$

$$\hat{\mathbf{B}}_r = (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{U}_r\mathbf{\Lambda}_r \tag{3.8}$$

where $\mathbf{V}_r$ is the $q \times r$ matrix, whose columns are the first r right singular vectors of $\mathbf{D}$ associated with the r largest singular values.

Since they depend on singular vectors, which are normalised, $\hat{\mathbf{A}}_r$ and $\hat{\mathbf{B}}_r$ satisfy:

$$\hat{\mathbf{A}}_r\mathbf{\Gamma}\hat{\mathbf{A}}_r' = \mathbf{I}_r \tag{3.9}$$

$$\hat{\mathbf{B}}_r'\mathbf{X}'\mathbf{X}\hat{\mathbf{B}}_r = \mathbf{\Lambda}_r^2 \tag{3.10}$$

where $\mathbf{\Lambda}_r^2$ is the $r \times r$ diagonal matrix of the r first squared singular values of $\mathbf{D}$.

Finally, we have:

$$\hat{\mathbf{C}}_r = \hat{\mathbf{B}}_r\hat{\mathbf{A}}_r = (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{U}_r\mathbf{\Lambda}_r\mathbf{V}_r'\mathbf{\Gamma}^{-1/2} \tag{3.11}$$

$$\hat{\mathbf{C}}_r = \hat{\mathbf{B}}_r \hat{\mathbf{A}}_r = (\mathbf{X}'\mathbf{X})^{-1/2} \hat{\mathbf{D}}_r \boldsymbol{\Gamma}^{-1/2} \tag{3.12}$$

where $\hat{\mathbf{D}}_r$ is the rank r approximation of $\mathbf{D}$. If $r = \text{rank}(\mathbf{D})$, then $\hat{\mathbf{D}}_r = \mathbf{D} = (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{Y}\boldsymbol{\Gamma}^{1/2}$ and $\hat{\mathbf{C}}_r = \hat{\mathbf{C}}^{\text{OLS}}$.

It should be noted that when $\boldsymbol{\Gamma}$ is set to $(\mathbf{Y}'\mathbf{Y})^{-1}$, the solutions of RRR for $\hat{\mathbf{B}}_r$ and $\hat{\mathbf{A}}_r$ are closely related to the solutions of Canonical Correlation Analysis, described in section 5.1.2.

Sparse Reduced Rank Regression In order to face the very high dimensionality of imaging genetics data and to reduce overfittig, Vounou et al. (2010) developed a sparse version of RRR.

Before applying sparsity, they first made the classical approximation when $n \ll p$ that the covariance matrix of each block may be replaced by its diagonal elements (Parkhomenko et al. 2007; 2009, Waaijenborg et al. 2008, Witten and Tibshirani 2009), that is $\mathbf{X}'\mathbf{X} = \mathbf{I}$ and $\boldsymbol{\Gamma} = (\mathbf{Y}'\mathbf{Y})^{-1} = \mathbf{I}$ in our case. It is equivalent to an extreme shrinkage of the covariance matrices or an extreme regularisation up to a factor, as mentioned before in sections 2.2.1.2 and 2.2.1.5. In fact, this approximation of RRR leads to the same solutions as PLS-SVD described in section 2.2.3 (see Appendix A). Equation 3.5 thus becomes:

$$\min_{\mathbf{A},\mathbf{B}} \text{Trace}\left\{\mathbf{Y}'\mathbf{Y}\right\} - 2\text{Trace}\left\{\mathbf{A}\mathbf{Y}'\mathbf{X}\mathbf{B}\right\} + \text{Trace}\left\{\mathbf{A}\mathbf{A}'\mathbf{B}'\mathbf{B}\right\} \tag{3.13}$$

One may note that the first term does not depend on $\mathbf{A}$ and $\mathbf{B}$. Thus, a sparse solution may be obtained for the first component/rank, by applying an L1-regularisation on $\mathbf{a}$ and $\mathbf{b}$:

$$\min_{\mathbf{a},\mathbf{b}} -2\mathbf{a}\mathbf{Y}'\mathbf{X}\mathbf{b} + \mathbf{a}\mathbf{a}'\mathbf{b}'\mathbf{b} + \lambda_{1Y}\|\mathbf{a}\|_1 + \lambda_{1X}\|\mathbf{b}\|_1 \tag{3.14}$$

Similarly to Lasso Regression, the higher λ_{1Y} (respectively λ_{1X}), the fewer imaging phenotypes (respectively SNPs) are selected.

This criterion is biconvex in $\mathbf{a}$ and $\mathbf{b}$ and may be solved by an iterative algorithm using soft-thresholding. For fixed and normalised $\mathbf{a}$:

$$\begin{aligned}\hat{\mathbf{b}} &= \arg\min_{\mathbf{b}} -2\mathbf{a}'\mathbf{Y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{b} + \lambda_{1X}\|\mathbf{b}\|_1 \\ &= g_{\lambda_{1X}/2}(\mathbf{X}'\mathbf{Y}\mathbf{a}')\end{aligned} \tag{3.15}$$

where $g_\lambda(y) = \text{sign}(y)(|y| - \lambda)_+$ is the soft-thresholding function. $\mathbf{b}$ is then normalised such that it satisfies $\mathbf{b}'\mathbf{b} = \Lambda_1^2$, where Λ_1 is the largest singular value of $\mathbf{D}$.

Then for fixed $\mathbf{b}$:

$$\begin{aligned}\hat{\mathbf{a}} &= \arg\min_{\mathbf{a}} -2\mathbf{a}'\mathbf{Y}'\mathbf{X}\mathbf{b} + \Lambda_1^2\mathbf{a}\mathbf{a}' + \lambda_{1Y}\|\mathbf{a}\|_1 \\ &= \frac{1}{\Lambda_{1Y}^2} g_{\lambda_1/2}(\mathbf{b}'\mathbf{X}\mathbf{Y})\end{aligned} \tag{3.16}$$

$\mathbf{a}$ is then normalised to have a unitary L2-norm.

Starting with arbitrary coefficients $\hat{\mathbf{a}}_0$ and $\hat{\mathbf{b}}_0$, the solutions are obtained by updating $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$, according to equations 3.15 and 3.16, until convergence. It should be noted that when there is only one target variable $\mathbf{y}$

and when $\mathbf{X}'\mathbf{X}$ is really diagonal, the solution of the sparse RRR algorithm on the first component is the same as the Lasso solution.

A second component may be obtained by optimising the same criterion on the residuals of the matrices $\mathbf{X}$ and $\mathbf{Y}$ after regression on their own latent variable $\mathbf{X}\hat{\mathbf{b}}$ and $\mathbf{Y}\hat{\mathbf{a}}$ respectively, and so on. However, this departs from the PLS-SVD framework and becomes equivalent to the sparse version of PLS regression in its canonical mode, described in section 5.2.2 (see Appendix A).

Application to imaging genetics Vounou et al. (2010) applied the sparse RRR algorithm to a simulated imaging genetics dataset of 1000 subjects, 111 imaging phenotypes simulated from real anatomical MRI data and up to $40,000$ SNPs. They showed an increased power to detect associations compared to Mass-Univariate Linear Modelling.

Vounou et al. (2012) then applied the sparse RRR algorithm to a sample from the Alzheimer's Disease Neuroimaging Initiave (ADNI) database, made of 475 subjects (Alzheimer's Disease patients, Mild Cognitive Impairment and controls), an Alzheimer's Disease (AD) signature of $11,000$ voxels derived from anatomical MRI and $437,000$ SNPs. They did not look at the generalisability of their model, but they ranked the SNPs according to their selection stability across resampling. Among the top-ranked SNPs, they found APOE and TOMM40, which are widely known to be associated with AD.

3.2.2.2 Group-Sparse Multi-Task Regression

Another way to account for the multivariate nature of the target variables (the imaging phenotypes) in a multivariate multiple regression framework, while imposing sparsity on predictors (SNPs), is to force a structured sparsity of the coefficient matrix $\mathbf{C}$.

Multi-Task Regression For instance, one may use multi-task regression with joint predictor selection (where each target variable is called a task), based on L1,2 regularisation (Argyriou et al. 2007, Obozinski et al. 2010). The criterion to be optimised then becomes:

$$\min_{\mathbf{C}} \|\mathbf{Y} - \mathbf{X}\mathbf{C}\|_F^2 + \lambda_{1,2}\|\mathbf{C}\|_{1,2} \tag{3.17}$$

with the L1,2 norm defined as $\|\mathbf{C}\|_{1,2} = \sum_{i=1}^{p} \|\mathbf{c}_i\|_2$, where the $\mathbf{c}_i$ are the rows of $\mathbf{C}$. Such a regularisation couples the predictor selection across tasks to make use of the multivariate nature of the target variables.

Group-Sparse Multi-Task Regression Wang et al. (2012a) used a new version of joint predictor selection for multi-task regression, where the L1,2 penalty is combined with a group-sparse penalty that incorporates the biological group structures among SNPs:

$$\min_{\mathbf{C}} \|\mathbf{Y} - \mathbf{X}\mathbf{C}\|_F^2 + \lambda_{1,2}\|\mathbf{C}\|_{1,2} + \lambda_{G_{1,2}}\|\mathbf{C}\|_{G_{1,2}} \tag{3.18}$$

with the group $L_{G_{1,2}}$ norm defined as $\|\mathbf{C}\|_{G_{1,2}} = \sum_{k=1}^{K} \sqrt{\sum_{i \in \pi_k} \sum_{j=1}^{q} c_{ij}^2}$, where the SNPs are partitioned into K groups $\{\pi_k\}_{k=1}^{K}$. Two types of groups were used: groups of SNPs belonging to or in the neighbourhood of the same gene, and groups of SNPs correlated with each other due to linkage disequilibrium ($r^2 \geq 0.2$). Such a regularisation considers the regression coefficients of all the SNPs in each group with respect to all the imaging phenotypes together and enforces sparsity at the group level.

Sparse Multimodal Multi-Task Regression They also developed sparse multimodal multitask learning (Wang et al. 2012b), which is similar to group-sparse multitask learning, but where the groups of variables correspond to different modalities, such as anatomical MRI data or genetic data, and where the target variables consist in cognitive scores and disease status. However, instead of the previous group penalty, they imposed the classical group-lasso penalty (Yuan and Lin 2006) extended to the multivariate case for $\mathbf{Y}$, which enforces sparsity at the group/modality level for each task independently:

$$\min_{\mathbf{C}} \|\mathbf{Y} - \mathbf{X}\mathbf{C}\|_F^2 + \lambda_{1,2}\|\mathbf{C}\|_{1,2} + \lambda_{G_1}\|\mathbf{C}\|_{G_1} \tag{3.19}$$

with the group lasso L_{G_1} norm defined as $\|\mathbf{C}\|_{G_1} = \sum_{j=1}^{q} \sum_{k=1}^{K} \sqrt{\sum_{i \in \pi_k} c_{ij}^2}$, where the K groups correspond to the different modalities.

Application to imaging genetics Wang et al. (2012a) applied Group-Sparse Multi-Task Regression to a sample of 733 subjects (AD or MCI) from the ADNI dataset, with 1224 SNPs from 37 top AD candidate genes, in order to predict 10 grey matter density measures and 12 volumetric and cortical thickness values from regions of interest. Group-sparse multi-task regression obtained significantly better prediction scores than multivariate regression and ridge regression, which do not account for the multivariate nature of the imaging variables, and it also outperformed the classical joint predictor selection for multi-task regression.

In Wang et al. (2012b), Sparse Multimodal Multi-Task Regression was applied to the same sample as in Wang et al. (2012a), and four different modalities (86 grey matter density measures, 54 volumetric and cortical thickness values, 26 PET measures and 1224 SNPs) were used to predict jointly 5 memory scores and the disease status, showing increased prediction scores compared to other classical methods.

3.2.2.3 Parallel Independent Component Analysis

Liu et al. (2009) proposed a novel algorithm called parallel ICA, which looks for independent components for each block of data, while trying to maximise the correlation between each pair of components. Their algorithm is based on the ICA algorithm called Infomax (see section 2.3.1.2). Similarly to the one-block case, it searches for independent components maximising an entropy function, for each block of data, while maximising the correlation between each pair of components.

Let us first recall the notations $\mathbf{Y}$ (of size $n \times q$) for the q imaging phenotypes and $\mathbf{X}$ (of size $n \times p$) for the p SNPs, n being the number of subjects. Using also notations similar to Liu et al. (2009), one assumes that the SNP data $\mathbf{X} = \mathbf{A}_x\mathbf{S}_x$ are a linear mixing of m_x sources $\mathbf{S}_x$ (of size $m_x \times p$) using a mixing matrix $\mathbf{A}_x$ (of size $n \times m_x$), and that the imaging data $\mathbf{Y} = \mathbf{A}_y\mathbf{S}_y$ are a linear mixing of m_y sources $\mathbf{S}_y$ (of size $m_y \times q$) with a mixing matrix $\mathbf{A}_y$ (of size $n \times m_y$). It should be noted that the subjects play the role of random variables in this case, while each voxel and each SNP are considered as an imaging observation and a genetic observation respectively.

Parallel ICA looks for independent components $\mathbf{W}_x\mathbf{X}$ and $\mathbf{W}_y\mathbf{Y}$ ($\mathbf{W}_x$ and $\mathbf{W}_y$ being the unmixing matrices) that maximise:

$$\max_{\mathbf{W}_x\mathbf{W}_y} H(\mathbf{Z}_x) + H(\mathbf{Z}_y) + \text{cor}(\mathbf{A}_x, \mathbf{A}_y), \text{ with } \mathbf{A}_x = \mathbf{W}_x^{-1} \text{ and } \mathbf{A}_y = \mathbf{W}_y^{-1} \tag{3.20}$$

where H is the entropy function, $\mathbf{Z}_x$ and $\mathbf{Z}_x$ being defined as follows:

$$\mathbf{Z}_x = \frac{1}{1 + \exp^{-\mathbf{U}_x}}, \text{ with } \mathbf{U}_x = \mathbf{W}_x\mathbf{X} + \mathbf{W}_{x0} \tag{3.21}$$

$$\mathbf{Z}_y = \frac{1}{1 + \exp^{-\mathbf{U}_y}}, \text{ with } \mathbf{U}_y = \mathbf{W}_y\mathbf{Y} + \mathbf{W}_{y0} \tag{3.22}$$

This optimisation problem may be solved using an iterative algorithm based on a gradient-approach, maximising each one of the three terms iteratively, up to convergence. It should be noted in their implementation that the third term $\text{cor}(\mathbf{A}_x, \mathbf{A}_y) = \sum_{i,j} \text{cor}(\mathbf{A}_x^i, \mathbf{A}_y^j)$, where $\mathbf{A}_x^i$ and $\mathbf{A}_y^j$ are the ith column/component of $\mathbf{A}_x$ and jth column/component of $\mathbf{A}_y$ respectively, does not necessarily include all the possible pairs of components and varies during the optimisation process. In practice, for each iteration, the algorithm only imposes a constraint on the pairs of components with a correlation higher than 0.3, leading to varying constraints.

In order to reduce overfitting of the correlation term, an adaptive regularisation parameter was also introduced before this third term. Moreover, the choice of the number of components is critical and may lead to overfitting issues as well, when overestimated. Thus, Liu et al. (2009) propose to use a modified version of the Akaike information criterion (AIC) described in Li et al. (2007b) to fix the number of fMRI components, and select the number of SNP components according to the consistency (or stability) of SNP components across resampling.

Application to imaging genetics Liu et al. (2009) applied parallel ICA to a sample of 63 subjects including 20 patients with schizophrenia and 43 healthy controls, 7060 fMRI voxels and 367 SNPs, and they showed that the component pair with the highest correlation could significantly discriminate patients from controls.

More recently, Meda et al. (2012) applied Parallel ICA on a sample of 818 subjects (Late Onset AD patients, MCI patients and controls) from the ADNI database, with 94 bilateral volumes of interest and cortical thickness values, and an AD signature of 27150 SNPs. They found four significantly correlated pairs of SNP and imaging components, after correction for the

number of component pairs tested. However, the generalisability of their model was not assessed. Moreover, it should be noted that sparsity is not enforced by parallel ICA and that, for each component, one has to select the variables with a weight larger than an arbitrary threshold (in absolute value), in order to interpret the results.

3.2.2.4 Limitations

The first limitation of these multivariate methods is probably the overfitting issue, which rapidly occurs when the number of variables increases, even though regularisation is enforced.

Moreover when the dimension increases, the computational cost of such complex multivariate methods may be very high, especially if one needs to tune the regularisation parameters and to validate the results, using cross-validation and permutation schemes.

Part II

Contributions

Clusters of SNPs and 4D clusters

4

As mentioned before, classical imaging genetics studies are often based on a massive univariate testing of the voxels from the whole brain image versus genome-wide data, leading to a huge multiple comparison issue. In the neuroimaging field, strategies to limit the number of multiple comparisons have been proposed based on the detection of clusters of contiguous activated voxels, which may increase sensitivity by trading off anatomical specificity. In this chapter, we will first briefly describe cluster-level inference in neuroimaging. Then, we will investigate a similar strategy on the genetic data, by searching for SNP clusters, and show very preliminary results. Finally, we will show that this idea could be extended to the detection of 4D clusters of voxels and SNPs.

Contents

4.1 Imaging case: 3D clusters

4.1.1 Concept

In neuroimaging, cluster-level inference was introduced by Poline and Mazoyer (1993) and Roland et al. (1993), and is based on the idea that it may be more powerful to detect clusters of contiguous active voxels than single active voxels. Indeed, by performing a cluster-level inference, one may reduce the number of multiple tests since the number of clusters is much lower than the number of voxels. Moreover, one may think that a cluster of contiguous active voxels reflects the underlying anatomical structure of the brain and is thus less likely to be due to noise than some voxel-wise signal. Finally, clusters of contiguous active voxels may be easier to interpret in terms of anatomical regions than single voxels.

4.1.2 Method

Voxel-wise univariate tests The first step of cluster-level inference is to perform a massive univariate test of the signal within each voxel versus the target variable, like in classical inference.

Thresholding The statistical maps are then thresholded at a given value u, classically corresponding to a p-value of 10^{-3}, and clusters of contiguous voxels with a statistic above that threshold are identified.

Size of voxel clusters Finally, the size of the clusters of voxels becomes the new statistic of interest in order to detect activations.

Assessment of the cluster-size significance The degree of significance of cluster sizes may then be assessed by computing the null distribution of the cluster-size statistic, using several techniques.

Random Field Theory is the most widely used approach (Friston et al. 1993), which requires a few assumptions such as smooth image, a uniform smoothness, and a sufficiently high statistical threshold (Worsley et al. 1992). Moreover, in order to correct for the number of clusters tested, one may use a maxT approach by estimating the null distribution of the maximal statistic over clusters (instead of the null distribution of the statistic itself) and comparing the cluster-size statistics with that distribution to derive corrected p-values.

Another approach based on permutations was introduced by Holmes et al. (1996), which does not require any assumptions on the null distribution of the cluster-size statistic. To obtain corrected p-values for the number of clusters tested, the maxT approach may be performed by permutations, by deriving the null distribution of the maximal statistic over clusters, from the permutations.

4.1.3 Limitations

Though largely used for the analysis of functional neuroimaging data, cluster-level inference has a few limitations.

Choice of the threshold First, the choice of the statistical threshold u is critical and arbitrary. A low statistical threshold will result in merging regions that are functionally distinct, while a too high statistical threshold will miss activations.

Localisation of the activation Moreover, even though one detects an activation at the cluster level, this only guarantees that at least one voxel of the cluster is active, without telling which voxels are active within that cluster.

Non-stationarity Finally, another important limitation of cluster-size inference may be related to the non-stationarity of the imaging data. Indeed, the smoothness of the image is not uniform, which may disadvantage the RF-based approach. Moreover, it will also influence the results of the permutation-based approach, by giving more sensitivity for the detection of smooth brain regions compared to rough regions.

4.1.4 Variants

A few variants of these two approaches have been proposed in order to face the limitations mentioned above.

RESEL In order to account for non-stationarity, the concept of RESEL (RESolution ELement) was introduced by Worsley et al. (1992), which corresponds to a volume sampling corrected for the local smoothness. It may be used to compute adjusted cluster sizes and can then be combined with both the RF (Worsley 2002) and the permutation approaches (Hayasaka et al. 2004). However, it should be noted that maxT correction at the cluster-level may not always guarantee strong control of family-wise Type I error in that case. Indeed, if two clusters happen to be adjacent, their cluster-size statistics in terms of RESELs are not completely independent, and thus subset pivotality does not apply anymore (Westfall and Young 1993, Ge et al. 2003).

Threshold-free cluster enhancement Smith and Nichols (2009) proposed an interesting method, called Threshold-Free Cluster Enhancement (TFCE), that tends to enhance the height of spatially extended signals. This leads to a voxel-wise statistic that accounts for clusters, without having to choose any cluster-size threshold. However, they introduced two new parameters that need to be tuned as well.

The results they obtained show that TFCE tends to be more sensitive than cluster-level inference. Nevertheless, as for RESELs, it should be noted that they used permutation-based maxT correction, which may not guarantee strong control in that case since the statistic at one voxel depends on the statistics of the neighbouring voxels.

4.2 Genetic case: 1D clusters

4.2.1 Concept

Similarly to cluster-inference in neuroimaging, we have tried to identify clusters of successive SNPs along the genome which would be correlated to the same phenotype extracted from imaging data. Indeed, there is an underlying structure of correlation between neighbouring SNPs, similar to voxel correlation, called linkage disequilibrium (LD).

Moreover, there already exist some techniques that try to combine the p-values obtained at adjacent SNPs, based on the idea that such a combination will be more significant and more biologically relevant than considering the SNPs independently. For instance, Tippett's method (the minimum p-value among several SNPs), Fisher's method (that takes the product of the different p-values), Stouffer's method (that averages the corresponding Z-values) are methods that are currently used to combine the significance of several SNPs. Some more recent contributions by Liang and Kelemen (2008), Neale and Sham (2004), Hoh and Ott (2000) propose also a set of tests based on p-values aggregation (with a sliding window along the sequence or scan statistics).

4.2.2 Method

Voxel-wise univariate tests The first step of our method was to compute univariate p-values (p_k) for each SNP k versus the phenotype of interest, using an additive linear model.

Morphology operation We then applied a greyscale morphology closing operation on $-log_{10}(p_k)$ along the SNP axis, which is in some way similar to the smoothing step of fMRI data.

Thresholding The next step was to detect clusters of successive SNPs with p-values below a given threshold α.

SNP-wise cluster size To be able to account for non-stationarity along the genetic region due to LD inhomogeneity, we decided to use a local statistic at each SNP: the size of the cluster to which it belongs.

Assessment of cluster-size significance The significance of the SNP-wise cluster-size statistic, was then estimated by permutations for each SNP. We performed 100000 permutations to get accurate empirical p-values.

If the use of a SNP-wise cluster-size statistic yields an uncorrected p-value derived locally for each SNP by permutations, non-stationarity remains an issue when trying to correct for multiple comparisons. Indeed, as mentioned in section 4.1.3, maxT correction does not handle non-stationarity, since it performs well when the null distribution of the statistic is the same across variables and favours smooth regions otherwise.

Thus, we used instead another resampling procedure to correct for multiple comparisons, called minP correction (Westfall and Young (1993),Dudoit et al. (2004)), accounting for both the dependence between the tests and the inhomogeneity of the statistic. It consists in inferring the null distribution across permutations of the minimal SNP-wise cluster-size p-value over all SNPs, and then deriving adjusted p-values.

4.2.3 Application to imaging genetics

4.2.3.1 Experimental data

This study is based on $n = 94$ subjects who were genotyped for $1,054,068$ SNPs and participated in a general cognitive assessment fMRI task described in Pinel et al. (2007). The study (both imaging and genetics components) was approved by the local ethics committee and all subjects gave informed consent. The task consisted of a short 5 min BOLD acquisition during which subjects were reading, listening to speech, asked to perform a motor response (button click), to subtract numbers, or were shown visual checkerboard.

Brain imaging data The functional images were acquired either on a 3T Bruker scanner or a 3T Siemens trio scanner using an EPI sequence (TR=2400 ms, TE=60 ms, matrix size=64×64, FOV=$19,2$ cm$\times$ $19,2$ cm). T1 anatomical images were acquired during the same acquisition session with a resolution of $(1.1 \times 1.1 \times 1.2)$ mm^3. Pre-processing classically comprised slice-timing correction, motion estimation, spatial normalisation (with a resampling of the functional images at 3mm resolution) and smoothing (FWHM=10 mm). The preprocessings and first level model analyses were performed with SPM5 (`www.fil.ucl.ac.uk/spm`).

The experimental protocol included speech comprehension and reading conditions, as well as motor and arithmetics instructions presented visually or verbally, that allowed to construct various contrasts of interest. In this section, we focused only on one activation contrast: *reading minus checkerboard viewing*. We used a first level, subject-specific, General Linear Model (GLM), to obtain parametric estimates of the BOLD activity at each voxel in each subject; the analysis was performed using SPM5, with standard parameters (frequency cut=128s, AR(1) temporal noise model). For each subject s in $\{1,\dots,n\}$ and each voxel v of the normalised volume, we obtained a map $\hat{\beta}_s(v)$ that represents the amount of BOLD signal associated with the contrast, normalised by the average signal. We defined a global brain mask for the group by considering all the voxels that belong to at least half of the individual brain masks (the individual masks were estimated using the standard SPM5 procedure).

Then, we selected a brain location of interest in the inferior temporal gyrus, at MNI coordinates $(-52,-60,-14)$ (see Figure 4.1), which had been reported to be atypically activated during reading in dyslexia (Paulesu et al. 2001). Each contrast map was locally averaged within a 4 voxel-radius sphere centred on this peak, keeping only active clusters of voxels (T$\geq$ 1 and cluster size$\geq$ 10 voxels) (Pinel and Dehaene 2009). This yielded an average value corresponding to this region of interest (ROI)

and we computed the average value for the mirror ROI by symmetry with respect to the inter-hemispheric plane. Finally, a lateralisation index was derived from this region. For each subject s, an index was computed as follows:

$$\text{Index}_s = \frac{\hat{\beta}_s^{\text{right}} - \hat{\beta}_s^{\text{left}}}{\sqrt{\left(\hat{\beta}_s^{\text{right}}\right)^2 + \left(\hat{\beta}_s^{\text{left}}\right)^2}}. \tag{4.1}$$

The distribution of this index spanned the range of $[-1.5; 1.5]$. The term "phenotype" will now refer to the lateralisation index thus obtained.

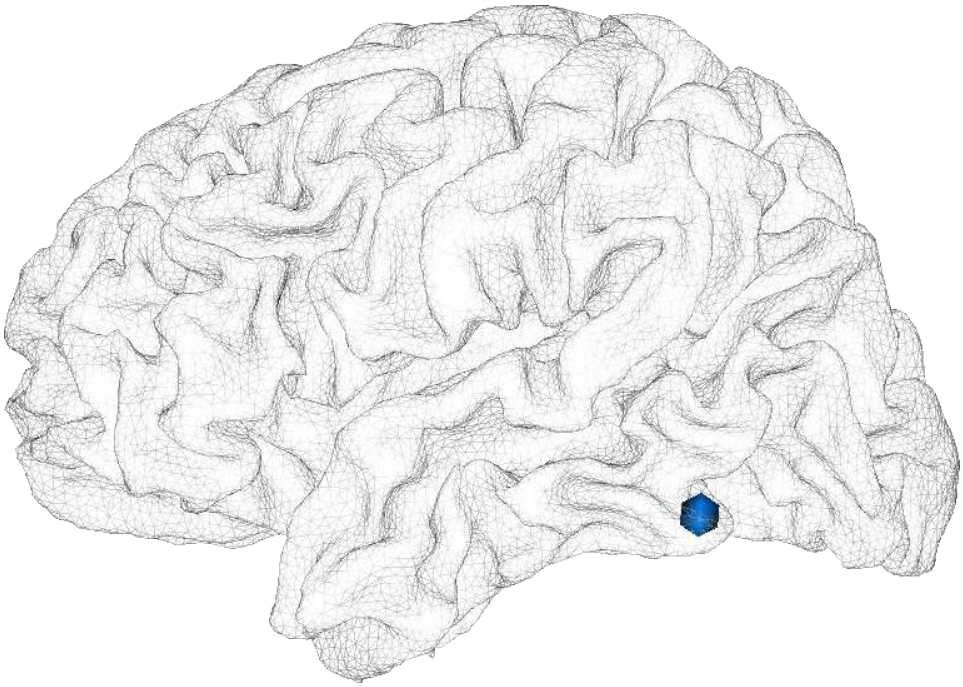


Figure 4.1 – *Location of the brain region of interest in the inferior temporal gyrus, at coordinates* $(-52, -60, -14)$.

SNP data For each subject, an Illumina HumanBeadQuad 1M platform was used to genotype $1,054,068$ SNPs and processed with the standard Illumina platform software. Considering all genotyped data available, we successively applied the following filters on all SNPs: (1) Minor Allele Frequency (MAF) at least 10%, (2) call rate at least 95%, and (3) Hardy-Weinberg test not significant at the 0.005 level. Genetic data were recoded as the number of minor alleles (denoted as A), $\{0, 1, 2\}$, hence a value of 0 means homozygous wild-type individuals (BB). The frequency of homozygous individuals for the minor allele (AA) was 0.03–0.13 in 75% of the cases. Missing SNP data were imputed with their corresponding median value across subjects[1]. These analyses were carried out using the open-source R software (R Development Core Team 2009) and the storage facilities for genetic data provided in the package `snpMatrix` (Clayton and Cheung 2007). After these preprocessing steps, $622,534$ SNPs were left for further analysis.

[1]Other imputation methods were tested, e.g. the Markov Chain based haplotyper proposed by Abecasis and coworkers (Willer et al. 2008, Sanna et al. 2008). All yield similar profiles of allele frequencies for our data set.

The database is dedicated to the study of the individual variability of some general cognitive processes, such as language in our case. Thus, the genotype of genes or genetic regions, with known implication in dysphasia or dyslexia, may be interestingly considered to study genetic associations with the imaging phenotype in the healthy volunteers database. Here, we focused on the so-called association site DYX2 (or KIAA0319), on chromosome 6 on the cytogenetic band 6p22.3, which had been reported to be associated with dyslexia (Cope et al. 2005). We selected 2738 contiguous SNPs from the HapMap database. From the selected genomic region, 568 SNP passed the filtering procedures described above.

4.2.3.2 Simulated data

The method was also tested on partially simulated data. We used the real genetic data from the 94 subjects from section 4.2.3.1. We arbitrarily selected a region of 1381 SNPs on chromosome 15 on the cytogenetic band 15q15, from HapMap. After the same filtering procedures as in section 4.2.3.1, 533 SNPs were kept for further analyses. Missing SNP data were imputed using their median value.

The quantitative phenotype was simulated, such that it was correlated to a specific SNP at a random position along the genetic region, using a linear model with p-value=0.002 and gaussian noise of variance 1.

4.2.3.3 Results on the simulated dataset

We first compared unadjusted cluster-size p-values with unadjusted SNP-wise p-values for two different sizes of the closing structure (2 and 4) and for a SNP-wise p-value threshold $\alpha = 0.01$. Cluster-size p-values were equivalent to SNP-wise p-values at the locations where clusters had been detected, as can be seen in figures 4.2 and 4.3 where the blue line represents SNP-wise p-values and the red one cluster-size p-values. None of them was significant after Bonferroni correction.

We then compared the effects of the minP correction on both SNP-wise p-values and cluster-size p-values for the two different sizes of the closing structure (2 and 4). The minP correction on SNP-wise p-values did not give any significant results, even though the correction appeared to be slightly less stringent than the Bonferroni correction, as it takes into account the dependence between SNPs. However, the combination of the cluster-size test and the minP correction led to the detection of two different significant clusters, for the two sizes of the closing structure tested.

In the case of a closing element of size 2, the most significant association reached a corrected p-value of 0.028, which is shown in figure 4.4 where the blue line represents SNP-wise p-values corrected by minP and the red one cluster-size p-values corrected by minP, the green line being the significance level of 0.05.

In the case of a closing element of size 4, the most significant association reached a corrected p-value of 0.018, which is shown in figure 4.5 where the blue line represents SNP-wise p-values corrected by minP and the red one cluster-size p-values corrected by minP. This peak happened to be located in the region where the association had been simulated, within a haplotype block (or LD block) of at least 4 SNPs with $r > 0.9$.

Figure 4.2 – *Uncorrected SNP-wise (in blue) and SNP-wise cluster-size (in red) p-values for a closing element of size* 2*, on the simulated data. The correlation matrix shown below illustrates LD.*

Figure 4.3 – *Uncorrected SNP-wise (in blue) and SNP-wise cluster-size (in red) p-values for a closing element of size* 4*, on the simulated data. The correlation matrix shown below illustrates LD.*

4.2.3.4 Results on the real dataset

Like for the simulated dataset, we first compared unadjusted cluster-size p-values with unadjusted SNP-wise p-values for two different sizes of the closing structure (2 and 4) and for a SNP-wise p-value threshold $\alpha = 0.01$. Cluster-size p-values were equivalent to SNP-wise p-values, which can be seen in figures 4.6 and 4.7 where the blue line represents SNP-wise p-values and the red one cluster-size p-values. None of them was significant after Bonferroni correction, like on the simulated dataset.

We then compared the effects of the minP correction on both SNP-wise p-values and cluster-size p-values for the two different sizes of the closing structure (2 and 4). The minP correction on SNP-wise p-values did not give any significant results. However, the combination of the cluster-size

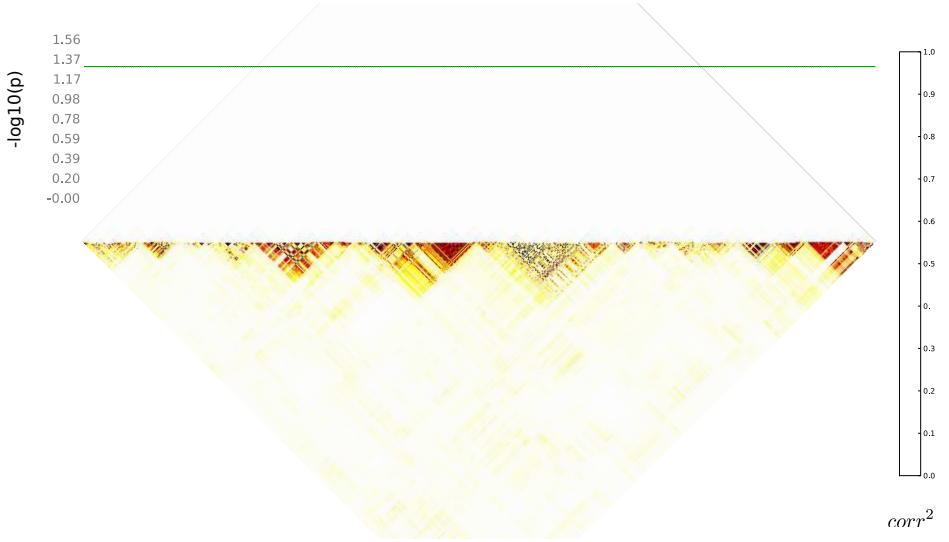


Figure 4.4 – *SNP-wise (in blue) and SNP-wise cluster-size (in red) p-values corrected by minP, for a closing element of size* 2*, on the simulated data. The green line represents the significance level of* 0.05*. The correlation matrix shown below illustrates LD.*

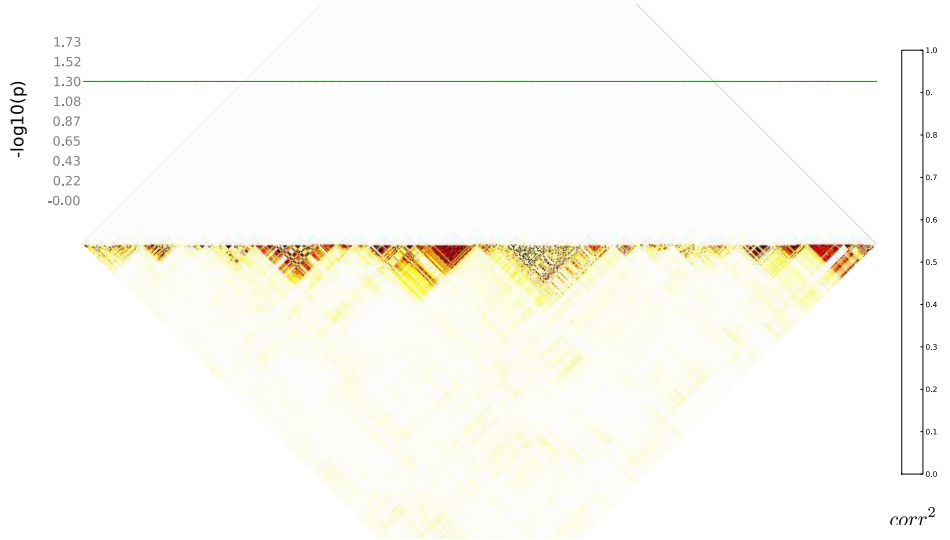


Figure 4.5 – *SNP-wise (in blue) and SNP-wise cluster-size (in red) p-values corrected by minP, for a closing element of size* 4*, on the simulated data. The green line represents the significance level of* 0.05*. The correlation matrix shown below illustrates LD.*

test and the minP correction led again to the detection of one significant cluster, for a closing structure of size 2. Indeed, in this case, the best association was significant with a corrected p-value of 0.013, which is shown in figure 4.8 where the blue line represents SNP-wise p-values corrected by minP and the red one cluster-size p-values corrected by minP.

With a closing element of size 4, the results were slightly lower but the best association almost reached the significance level with a corrected p-value of 0.059, which is shown in figure 4.9 where the blue line represents SNP-wise p-values corrected by minP and the red one cluster-size p-values corrected by minP.

Figure 4.6 – *Uncorrected SNP-wise (in blue) and SNP-wise cluster-size (in red) p-values for a closing element of size* 2*, on the real data. The correlation matrix shown below illustrates LD.*

Figure 4.7 – *Uncorrected SNP-wise (in blue) and SNP-wise cluster-size (in red) p-values for a closing element of size* 4*, on the real data. The correlation matrix shown below illustrates LD.*

4.2.4 Limitations

Even though this approach seems to improve sensitivity by taking into account the local multivariate nature of the genetic data and handles non-stationarity along the genome, it encounters a few limitations, which are similar to those mentioned above in the case of cluster-level inference in neuroimaging.

Choice of the threshold First, it should be noted that the choice of the p-value threshold may be critical, like for cluster-inference in neuroimaging.

Choice of the morphology element Moreover, the size of the morphology element is critical as well and may greatly influence the results, simi-

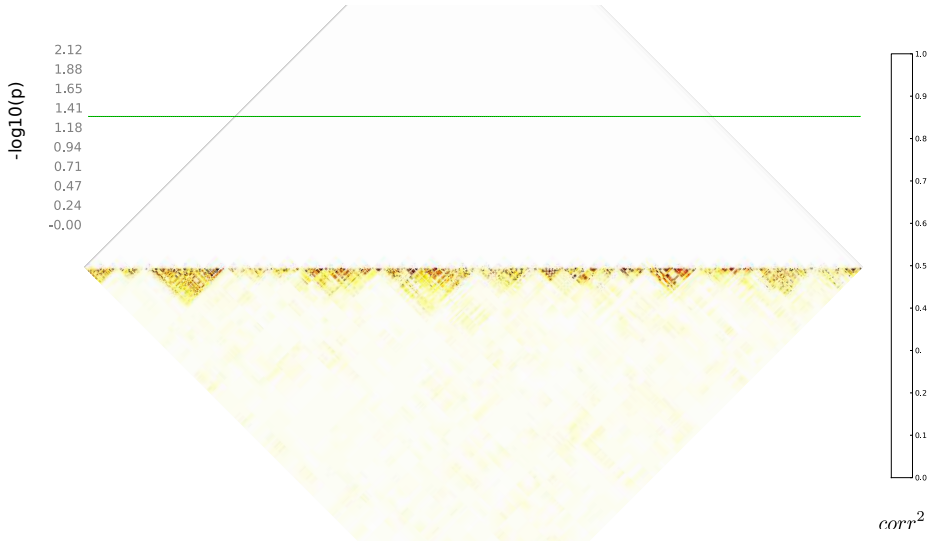


Figure 4.8 – *SNP-wise (in blue) and SNP-wise cluster-size (in red) p-values corrected by minP, for a closing element of size* 2*, on the real data. The green line represents the significance level of* 0.05*. The correlation matrix shown below illustrates LD.*

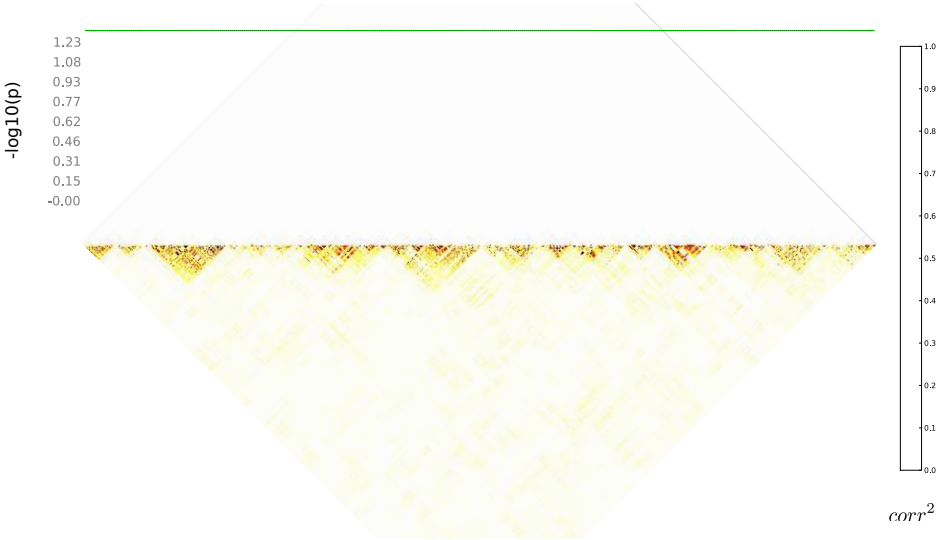


Figure 4.9 – *SNP-wise (in blue) and SNP-wise cluster-size (in red) p-values corrected by minP, for a closing element of size* 4*, on the real data. The green line represents the significance level of* 0.05*. The correlation matrix shown below illustrates LD.*

larly to the degree of smoothing in neuroimaging. Its optimal size seems to depend on the local structure of LD. It should be also noted that the size of the morphology element is expressed as a number of SNPs and not as a real biologic distance such as the number of base pairs or the genetic distance (in cM). However, common chips are usually designed using tagSNPs, such that the distance between two SNPs always roughly corresponds to the same genetic distance.

Subset pivotality Another limitation may be related to subset pivotality. Indeed, like with RESELs or the TFCE approach described in section 4.1.4, correction for the family-wise error rate such as minP may not be applicable in that case, since the statistic at one SNP depends on the statistics of the neighbouring SNPs.

Computational load Moreover, the minP procedure is very computationally expensive, since it requires a high precision (and thus a large number of permutations) for the computation of p-values, relatively to the number of variables to correct for.

4.2.5 Conclusion

The originality of our approach was to adapt cluster-inference techniques to SNP data and to handle non-stationarity along the genome axis, by using a SNP-wise statistic and combining it with a minP correction. This is a very preliminary study but the proposed method seems promising in order to improve the sensitivity of genetic association studies while controlling for type I error, even if subset pivotality issues and the choice of the threshold and of the morphology element should be further investigated.

Moreover, a significant association has been detected on healthy subjects between some loci of the DYX2 genetic region, involved in dyslexia, and a brain region reported to be atypically activated in dyslexia. This suggests that there might be a milder effect of this genetic region on the observed phenotype in healthy subjects. However, we should remain very cautious with the interpretation of the results. It should be noted as well that the proposed method may have detected direct associations (possibly due to one or several causal SNPs) or indirect ones due to correlations between SNPs within a haplotype block.

Finally, this method could be applied to any kind of quantitative phenotype such as neuroimaging data and even easily extended to qualitative phenotypes.

This work has been published in Le Floch et al. (2010).

4.3 Imaging Genetics case: 4D clusters

Then, we tried to extend the concept of clusters by combining voxel clusters and SNP clusters, using a 4D cluster test that jointly detects brain and genome regions with high associations.

4.3.1 Method

4D association maps The first step of our method was to apply a massive univariate approach, and to consider the association between the signal at each voxel and each SNP of interest, using an additive linear model. We constructed our 4D statistical map, with the corresponding p-values (one per SNP and voxel), which was then transformed into Z variables (4D-Z map). It should be noted that some SNP genotype values may be missing for some subjects, leading to small changes in the degrees of freedom of the test.

Thresholding We computed three types of statistics from our 4D-Z maps: SNP-wise voxel-wise Z value, 3D cluster sizes (6-connectivity) for

each SNP, defined with a threshold at $p = 0.005$, and 4D cluster sizes (8-connectivity: 6 in the 3D image space and 2 in the SNP direction) defined with the same $p = 0.005$ threshold.

Assessment of cluster-size significance The significance of the different statistics, was then estimated by permutations. This procedure preserves the covariance structure of our 4D data, while breaking the association between the imaging and the genetic data.

We recomputed the 4D-Z maps for 1600 permutations. The three types of statistics described above were computed on each permuted dataset, and we derived the distribution of the maximum of each statistic type under the null hypothesis (of no association). We then obtained maxT corrected p-values for family wise error rate. It should be noted that we did not use any variable-wise cluster-size statistic combined with minP correction in this 4D case, because of the computational load of the minP procedure when the number of tests increases (number of voxels $\times$ number of SNPs).

4.3.2 Application to imaging genetics data

4.3.2.1 Data

This study is based on the same dataset as in section 4.2.3.1, with $n = 94$ subjects who were genotyped for $1,054,068$ SNPs and participated in a general cognitive assessment fMRI task described in Pinel et al. (2007).

Brain imaging data This time, we focused on another activation contrast: *speech comprehension minus rest*. We performed the same preprocessing steps as in section 4.2.3.1 to obtain activation maps.

Finally, instead of considering a lateralisation index in a brain region of interest, a whole brain lateralisation index map was derived from the contrast maps. For each voxel in the normalised volume and in each subject, a symmetric voxel with respect to inter-hemispheric plane was selected and an index was computed as follows :

$$\hat{\beta}_s^{\text{index}} = \frac{\hat{\beta}_s^{\text{left}} - \hat{\beta}_s^{\text{right}}}{\sqrt{\left(\hat{\beta}_s^{\text{left}}\right)^2 + \left(\hat{\beta}_s^{\text{right}}\right)^2}} \tag{4.2}$$

This resulted in n maps with only one hemisphere and about $29,140$ voxels.

SNP data We performed the same preprocessing steps on genetic data as in section 4.2.3.1.

This time, we focused on the so-called association site DYX5 (Nopola-Hemmi et al. 2001), on chromosome 3, on the cytogenetic band 3p12.3 (from 74 M to 82 M bp). We selected 288 contiguous SNPs available on the Illumina chip at the beginning of this area: $74,882,614$ (*rs*9852358) to $76,247,020$ (*rs*7639460). From the selected genomic region, 175 SNPs passed the genotyping rate ($> 95\%$) and minor allele frequency (MAF $> 10\%$) cut-off.

4.3.2.2 Results

SNP-wise voxel-wise association value The estimated null distribution of the maximum SNP-voxel association was computed using the permutations. In our dataset, the maximum Z value was not significant at the 5% corrected level but reached a p-value of 11%.

3**D cluster-size** Then, we assessed the degree of significance of the maximum 3D cluster size per SNP (using the 175 distributions estimated by permutations for each SNP independently). Three SNPs were significantly associated with a 3D cluster at the 5% level, corrected for the number of clusters per SNP but uncorrected for the number of SNPs. None of them was significant after applying the Bonferroni correction for the number of SNPs.

4**D cluster-size** We compared the 4D cluster size statistics with the null distribution of their maximum. In our dataset, we detected 21 significant 4D clusters at the 5% level, corrected for the number of clusters.

Figure 4.10 shows the projection of the support of each detected cluster onto the 3D space (ie, summing the 4D support of the cluster along the dimension of the SNPs), after thresholding of each projection at its half maximum value.

In the genetic dimension, we projected the clusters onto the SNP axis and show in figure 4.11 the support of those projections. We show in this figure the linkage disequilibrium map obtained with Haploview (www.broad.mit.edu/mpg/haploview, (Barrett et al. 2005)).

We observed that the 4D cluster-size statistics showed greater sensitivity than the SNP-wise voxel-wise association test or the 3D cluster size test. The 3D representation of those clusters is distributed in a wide network in temporal, occipito-temporal and frontal areas. Although it is not the purpose of this work to discuss the interpretation of these regions, their localisation is not surprising in the context of the language asymmetry and the DYX5 genomic region.

In the genomic dimension, it is clear that 4D clusters are found more easily in regions with high LD (see figure 4.11). This is expected as the cluster sizes are more likely to reach higher values in genomic areas with stronger covariance. The 4D cluster size statistic should therefore be more sensitive in those areas, while overly conservative in areas with low LD, but the overall false positive rate is correct. We also note that some significant clusters are found even in places where LD is low, indicating that the test is indeed more sensitive than the SNP-voxel association and 3D cluster size alternatives. In the image direction, the smoothness of our data does not appear to have strong non-stationarity and the detected clusters positions are distributed in many brain regions.

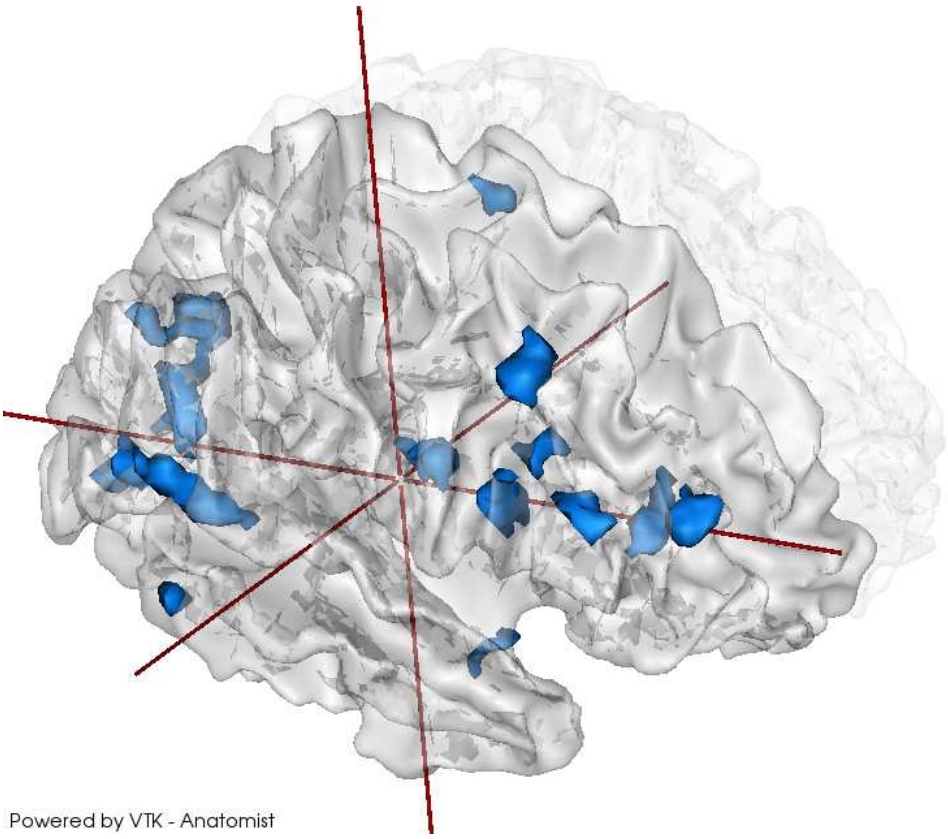


Figure 4.10 – *Three dimensional rendering of the* 21 *significant regions (corrected* 4*D cluster p-value* < 0.05*) superimposed on a left hemisphere template. The spatial anatomical support of these regions is obtained by first projecting the support of each* 4*D cluster onto the* 3*D space (ie, summing the cluster support along the dimension of the SNPs), and then thresholding the projection of the cluster support at its half maximum value.*

Interestingly, we observe that similar sets of SNPs may be linked with different brain regions. This may be an illustration of the general concept of pleiotropy in the context of the fMRI endophenotype. Conversely, we also observe that some of the brain regions detected are linked with several distinct clusters in the genomic region, indicating that the endophenotype could be due to a combination of the effects of the genotypes at several loci on the genome (epistasis).

4.3.3 Limitations

We have presented a simple procedure to search for brain regions that show a strong link with one or several contiguous SNPs. It has the advantages of sensitivity, simplicity and correction for family wise multiple comparison in the whole dataset. However, this test has clearly the same pitfalls or limitations as the usual 3D cluster test in neuroimaging and the proposed 1D cluster test in genetics.

Choice of the threshold First, it depends strongly on the threshold applied to the 4D data. A possible extension of the work would be to investigate the strategy recently proposed by Smith and Nichols (2009) using a

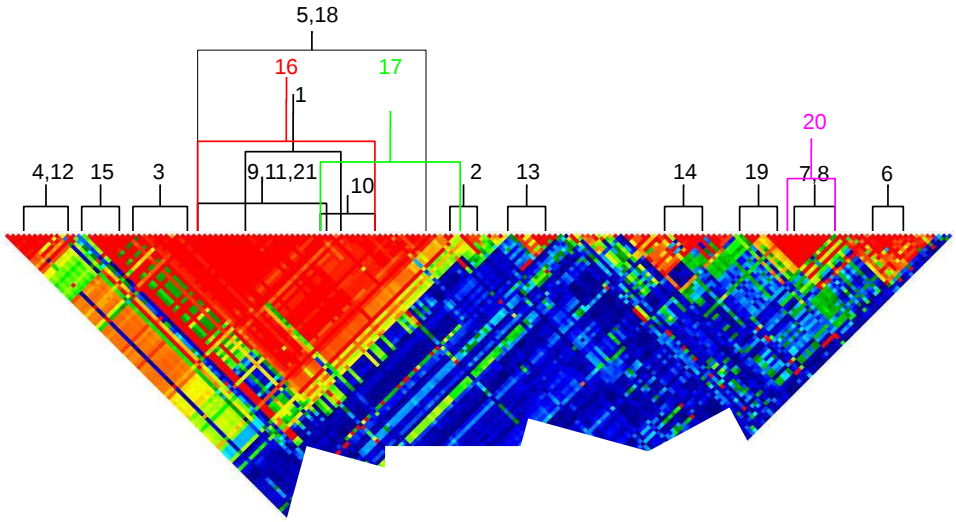


Figure 4.11 – *The run-lengths of the* 21 *significant regions (corrected* 4*D cluster p-value* < 0.05*) are registered to the linkage disequilibrium map computed from the* 175 *SNPs. The map was computed with Haploview (Barrett et al. 2005). Each cluster was assigned a number between* 1 *and* 21 *with an random order.*

threshold free procedure, but as mentioned earlier there may be pivotality issues associated with this technique that still have to be investigated to ensure an accurate specificity control.

Localisation of the activation Second, the test provides a statistic at the cluster level, such that even if a cluster is significant, one cannot reject the null hypothesis for any specific voxel and SNP. The interpretation should be that in this brain area, and this set of contiguous SNPs, there is an abnormally strong imaging genetics association. There is always the possibility that the definition of the cluster will aggregate noise voxels or SNPs. In our work, we limit this using a threshold at $p = 0.005$ for defining the clusters and a 6-nearest-neighbours connexity in the image space.

Non-stationarity Third, the procedure described may not be optimal as the spatial structure of the data is not necessarily stationary, in both the voxel or the SNP dimensions. The non-stationarity issue is important as we may have a test too liberal in the areas with higher covariance structures, and too conservative in other areas. In average, the test should provide the appropriate false detection rate (as shown on simulations in Hayasaka et al. (2004), even under strong non-stationarity the rate of false positive is controlled - see table 2 of their work). An extension of the 4D cluster test that would take into account local inhomogeneities of the (4D) spatial covariance structure would be a desirable. Again, we notice in our results that the detected clusters are distributed in the 3D space. In the genetic dimension, although we detect more clusters in the high LD area,

we also detect clusters in the low LD region. Moreover, the procedure we use at the moment may also not be ideal in the genetic domain, as any two adjacent SNPs may not be separated by an equivalent distance on the genome (either because of the sampling of the genotyping platform or because some SNPs are dropped in the preprocessing steps if they do not survive the MAF, or the calling threshold). This suggests extensions to account for non-stationarity in both the imaging and genetic dimensions.

Computational load Morover, the algorithm requires several days of computation on one processor (one core) for assessing the distribution under the null using permutations. Here we assessed the null distribution on a 175 (SNPs) * 29140 (voxels) * 94 (subjects) dimensional data-set, which required about 3 days of computation on a 8 cores Dell PC. Clearly, this task can be easily distributed on many processors, and for these numbers and for estimating a 5% threshold, the computation time can be reduced to the order of the hour with about 25 processors.

4.3.4 Conclusion and perspectives

To conclude, we have proposed a simple 4D cluster test that detects conjointly brain and genome regions with high associations, and we calibrate this test using permutations, accounting for multiple comparisons. This test shows promising preliminary results, with a greater sensitivity than SNP-wise voxel-wise association tests or 3D cluster tests. In our dataset, 21 regions were detected that showed some association between fMRI data asymmetry in a "listening to sentences" task and the DYX5 genomic region. We hope that this test will help assess the significance of the association in even larger imaging genetics datasets.

Other extensions of the test would be interesting to pursue. First, using only the size of the cluster in 4D, we discard the intensity of the association within the cluster. Strategies equivalent to those already proposed in the neuroimaging literature to test for both the spatial extent and the average intensity or maximum value (Poline et al. 1997) within the cluster may be also developed in the context of imaging genetic studies, or indeed the multiscale or multi filtering approaches (Poline et al. 1997, Worsley et al. 1996, Poline and Mazoyer 1994). The sensitivity of the test will clearly depend on the filtering applied both in the image and genetic domains. Filtering in the SNP dimension may be thought as an alternative to the combination of several p-values of neighbouring SNPs often proposed in association studies.

Because strict family wise correction may lead to very high thresholds and very low sensitivity in datasets with large voxel and SNP regions, it would be worth extending this work to other less stringent procedures such as False Discovery Rate (FDR) that may also easily be applied on the 4D imaging genetic data (see for instance, Nichols (2006)). Note that even with a less stringent FDR procedure, the detection across one million SNPs is likely to have a poor sensitivity.

In this work, we did not study the implications of our findings in the cognitive neuroscience field. However, it is known that language asymmetry is a heritable phenotype, and that the DYX5 region may be involved in

the phenotype we observe, such that the findings obtained here are likely to be interpretable (Nopola-Hemmi et al. 2001).

5 Dimension reduction and regularisation combined with Partial Least Squares

In this chapter, we will now describe the two-block multivariate approaches that we investigated, namely Canonical Correlation Analysis (CCA) and two-block Partial Least Squares Regression (PLS2), in order to search for association between fMRI and SNP data. We will also detail the regularisation and dimension reduction strategies that we used to solve the overfitting issue, which occurs with multivariate methods in such high-dimensional settings. Finally, we will present classical validation procedures in order to assess and to compare the performances of the different strategies that we investigated.

Contents

In this work we try to identify a functional brain network covarying with a set of genetic polymorphisms using multivariate methods, which take into account potential joint effects or covariations within each block of variables. Partial Least Squares (PLS) regression (Wold et al. 1983) and Canonical Correlation Analysis (CCA) (Hotelling 1936) appear to be good candidates in order to look for associations between two blocks of data, as they extract pairs of covarying/correlated latent variables (one linear combination of the variables for each block). Nevertheless, such multivariate methods encounter critical overfitting issues due to the very high dimensionality of the data.

To face these issues, strategies based on dimension reduction or regularisation can be used.

As for regularisation, a sparse (L1-regularised) version of Partial Least Squares (Parkhomenko et al. 2007; 2009, Waaijenborg et al. 2008, Lê Cao et al. 2008; 2009, Witten and Tibshirani 2009, Chun and Keleş 2010) and an L2-regularised version of CCA (Soneson et al. 2010) have recently been shown to provide good results in correlating two blocks of data such as transcriptomic and metabolomic data, gene expression levels and gene copy numbers, or gene expression levels and SNP data. One may note that such sparse multivariate methods based on L1 penalisation actually perform variable selection. The method proposed by Vounou et al. (2010), called sparse Reduced-Rank Regression (sRRR) and based on L1 penalisation, is very similar (see Appendix A) and was applied a simulated dataset made of 1000s of SNPs and brain imaging data. The implementation of the method becomes equivalent to sparse PLS in high dimensional settings, since they make the classical approximation that in this case the covariance matrix of each block may be replaced by its diagonal elements. This suggests that such multivariate methods may be appropriate to exploratory imaging genetics studies. However, whether these multivariate techniques can resist even higher dimensions remains an open question. In this thesis, we investigate this question by adding a first step of dimension reduction on SNPs, either by PCA or univariate filtering, before applying (sparse) PLS or (regularised) CCA.

In the next sections, we first introduce the multivariate methods that we used, and then detail the potential performance evaluation techniques to compare these different methods.

5.1 Multivariate methods based on latent variables

We first recall the notations $\mathbf{Y}$ (of size $n \times q$) for the q imaging phenotypes and $\mathbf{X}$ (of size $n \times p$) for the p SNPs. Without loss of generality, we assume that all variables are centred and normalised.

5.1.1 Partial Least Squares regression

Partial Least Squares regression is used to model the associations between two blocks of variables hypothesising that they are linked through unobserved latent variables. A latent variable (or component) corresponding to one block is a linear combination of the observed variables of this block.

An illustration of such multivariate methods based on the extraction of latent variables is given in Figure 5.1.

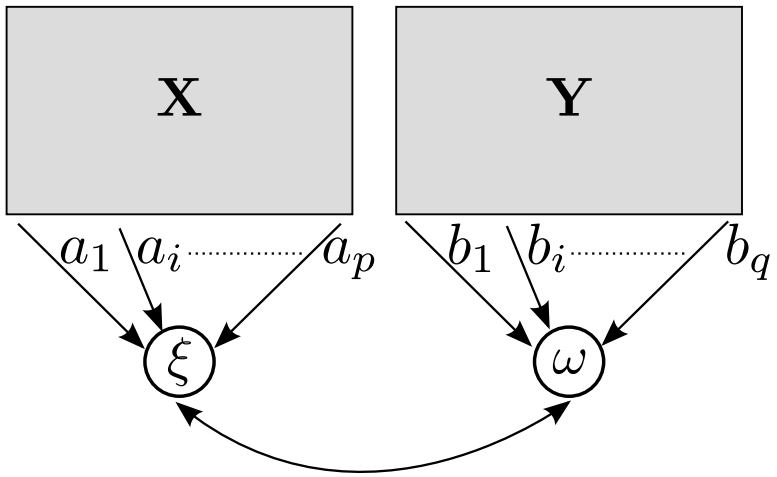


Figure 5.1 – *Illustration of a multivariate method based on the concept of latent variables.*

More precisely, PLS regression seeks successive and orthogonal latent variables for each block such that at each step the covariance between the pair of latent variables is maximal. For each step h in $1..H$, where H is the maximal number of pairs of components, it optimises the following criterion:

$$\max_{\|\mathbf{a}_h\|_2=\|\mathbf{b}_h\|_2=1} \operatorname{cov}(\mathbf{X}_{h-1}\mathbf{a}_h, \mathbf{Y}_{h-1}\mathbf{b}_h) \tag{5.1}$$

$$= \max_{\|\mathbf{a}_h\|_2=\|\mathbf{b}_h\|_2=1} \operatorname{corr}(\mathbf{X}_{h-1}\mathbf{a}_h, \mathbf{Y}_{h-1}\mathbf{b}_h) \sqrt{\operatorname{var}(\mathbf{X}_{h-1}\mathbf{a}_h)} \sqrt{\operatorname{var}(\mathbf{Y}_{h-1}\mathbf{b}_h)}$$

where $\mathbf{a}_h$ and $\mathbf{b}_h$ are the weight vectors for the linear combinations of the variables of blocks $\mathbf{X}$ and $\mathbf{Y}$ respectively. $\mathbf{X}_{h-1}$ and $\mathbf{Y}_{h-1}$ are the residual (deflated) $\mathbf{X}$ and $\mathbf{Y}$ matrices after their regression on the $h-1$ previous pairs of latent variables , starting with $\mathbf{X}_0 = \mathbf{X}$ and $\mathbf{Y}_0 = \mathbf{Y}$. There exist two ways of deflation: an asymmetric way (the original PLS regression) and a symmetric way (canonical-mode PLS). The difference is that in the first case both blocks are deflated on the latent variables of block $\mathbf{X}$ (which becomes the predictor block), while in the second case each bock is deflated on its own latent variables. In our case, we are more interested in canonical-mode PLS as we investigate exploratory methods trying to extract covarying networks among a huge amount of neuroimaging and SNP data, many of which are very likely to be irrelevant. Please note that, on the first pair of components, the original PLS regression and canonical-mode PLS give exactly the same results.

Since the variables are standardised, the previous criterion for each new pair of components is equivalent to optimising:

$$\max_{\|\mathbf{a}_h\|_2=\|\mathbf{b}_h\|_2=1} \mathbf{a}_h' \mathbf{X}' \mathbf{Y} \mathbf{b}_h \tag{5.2}$$

This optimisation problem is solved using the iterative algorithm called NIPALS (Wold 1966) and more precisely the NIPALS inner loop, the

NIPALS outer loop being the iteration over the number of pairs of components. Here is the NIPALS algorithm:

For h in 1..H:

1. Initialise $\mathbf{a}_h$ and $\mathbf{b}_h$ and normalise them.
2. Until convergence of $\mathbf{a}_h$ and $\mathbf{b}_h$:
   (a) For fixed $\mathbf{b}_h$:
   $$\widehat{\mathbf{a}_h} = \underset{\|\mathbf{a}_h\|_2=1}{\arg\min} \; -\mathbf{a}_h' \mathbf{X}_{h-1}' \mathbf{Y}_{h-1} \mathbf{b}_h = \mathbf{X}_{h-1}' \mathbf{Y}_{h-1} \mathbf{b}_h \tag{5.3}$$
   (b) Normalise $\mathbf{a}_h$: $\mathbf{a}_h \leftarrow \frac{\mathbf{a}_h}{\| \mathbf{a}_h \|_2}$.
   (c) For fixed $\mathbf{a}_h$:
   $$\widehat{\mathbf{b}_h} = \underset{\|\mathbf{b}_h\|_2=1}{\arg\min} \; -\mathbf{a}_h' \mathbf{X}_{h-1}' \mathbf{Y}_{h-1} \mathbf{b}_h = \mathbf{Y}_{h-1}' \mathbf{X}_{h-1} \mathbf{a}_h \tag{5.4}$$
   (d) Normalise $\mathbf{b}_h$: $\mathbf{b}_h \leftarrow \frac{\mathbf{b}_h}{\| \mathbf{b}_h \|_2}$.
3. Regress $\mathbf{X}_{h-1}$ and $\mathbf{Y}_{h-1}$ on the obtained latent variables to form $\mathbf{X}_h$ and $\mathbf{Y}_h$ respectively.

The optimal vectors $\mathbf{a}_h$ and $\mathbf{b}_h$ are in fact the first pair of singular vectors of the matrix $\mathbf{X}_{h-1}' \mathbf{Y}_{h-1}$. Please note that the criterion tends to maximise the relative value of the covariance, which implies that the covariance is forced to be null or positive. In the case of a negative association between a variable from block $\mathbf{X}$ and a variable from block $\mathbf{Y}$, a negative weight will thus be assigned to one of them in order to obtain a positive covariance.

One limitation of PLS regression is that such multivariate methods encounter overfitting issues in high-dimensional settings. For instance, (Chun and Keleş 2010) recently showed that asymptotic consistency of the PLS regression estimator does not hold when $p = \mathcal{O}(n)$, where p is the number of variables for blocks $\mathbf{X}$ and n the number of observations or individuals.

5.1.2 Canonical Correlation Analysis

A similar method is Canonical Correlation Analysis (CCA), which differs in that the correlation between the two latent variables, instead of the covariance, is maximised at each step. CCA builds successive and orthogonal latent variables for each block such as, at each step h in 1..H, they optimise the following the criterion:

$$\max \operatorname{corr}(\mathbf{X}\mathbf{a}_h, \mathbf{Y}\mathbf{b}_h) \tag{5.5}$$

where $\mathbf{a_h}$ and $\mathbf{b_h}$ are weight vectors.

Since the variables are standardised, it becomes equivalent to optimising:

$$\max \frac{\mathbf{a}_h' \mathbf{X}'\mathbf{Y}\mathbf{b}_h}{\sqrt{\mathbf{a}_h' \mathbf{X}'\mathbf{X}\mathbf{a}_h}\sqrt{\mathbf{b}_h' \mathbf{Y}'\mathbf{Y}\mathbf{b}_h}} \tag{5.6}$$

One may note that the correlation criterion does not depend on the norm of weight vectors, thus there exists an infinity of proportional solutions. Classically, the norm of weight vectors is chosen such that either:

$$\| \mathbf{a}_h \|_2 = \| \mathbf{b}_h \|_2 = 1$$

or

$$\| \mathbf{X}\mathbf{a}_h \|_2 = \| \mathbf{Y}\mathbf{b}_h \|_2 = 1$$

It should be noted that in the latter case, the criterion becomes equivalent to a covariance criterion as follows:

$$\max_{\|\mathbf{X}\mathbf{a}_h\|_2 = \|\mathbf{Y}\mathbf{b}_h\|_2 = 1} \operatorname{cov}(\mathbf{X}\mathbf{a}_h, \mathbf{Y}\mathbf{b}_h) \tag{5.7}$$

The solution may be obtained by computing the SVD of $\mathbf{D} = (\mathbf{X}'\mathbf{X})^{-1/2}\ \mathbf{X}'\mathbf{Y}\ (\mathbf{Y}'\mathbf{Y})^{-1/2}$. The successive pairs of weight vectors $\mathbf{a}_h$ and $\mathbf{b}_h$ are obtained by:

$$\mathbf{a}_h = \mathbf{X}'\mathbf{X}^{-1/2}\mathbf{U}_h \tag{5.8}$$

and

$$\mathbf{b}_h = \mathbf{Y}'\mathbf{Y}^{-1/2}\mathbf{V}_h$$

where the columns $\mathbf{U}_h$ and $\mathbf{V}_h$ of $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors respectively. The solution may also be obtained by an iterative algorithm, called PLS mode B (Wold 1985), which is a variant of the PLS algorithm presented in section 5.1.1.

It should be noted that the solutions for $\mathbf{a}_h$ and $\mathbf{b}_h$ are closely related to those obtained by RRR. Indeed both problems are solved by using the SVD of the same matrix $\mathbf{D} = (\mathbf{X}'\mathbf{X})^{-1/2}\ \mathbf{X}'\mathbf{Y}\ (\mathbf{Y}'\mathbf{Y})^{-1/2}$. $\mathbf{a}_h^{\text{CCA}}$ is in fact a scaled version of $\mathbf{b}_h^{\text{RRR}}$: $\mathbf{b}_h^{\text{RRR}} = \Lambda_h \mathbf{a}_h^{\text{CCA}}$, where Λ_h is the h largest singular value of $\mathbf{D}$. $\mathbf{b}_h^{\text{CCA}} = \mathbf{Y}'\mathbf{Y}^{-1/2}\mathbf{V}_h$ is the generalised inverse of $\mathbf{a}_h^{\text{RRR}} = \mathbf{V}_h'\mathbf{Y}'\mathbf{Y}^{1/2}$.

Like PLS regression, CCA has to face overfitting issues in high-dimensional settings. Moreover, CCA requires the inversion of the scatter matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{Y}'\mathbf{Y}$, which are ill-conditioned in our high-dimensional settings with very large p and q (numbers of variables for blocks $\mathbf{X}$ and $\mathbf{Y}$ respectively) and a small N (number of observations or individuals).

For numerical issues, we used the dual formulation of CCA based on a linear kernel: Kernel CCA (KCCA).

5.2 Regularisation techniques

5.2.1 L2 Regularisation of CCA

In order to first solve the overfitting issues of CCA, regularisation based on L2 penalisation may be used, by replacing the matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{Y}'\mathbf{Y}$ by $\mathbf{X}'\mathbf{X} + \lambda_2\mathbf{I}$ and $\mathbf{Y}'\mathbf{Y} + \lambda_2\mathbf{I}$ respectively (Vinod 1976, Leurgans et al. 1993). However, in such high-dimensional settings the approximation is often

made that the scatter matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{Y}'\mathbf{Y}$ may be replaced by identity matrices, which is an extreme case of shrinkage of the scatter matrices and makes CCA equivalent to PLS-SVD and thus to PLS regression as well on the first component. Shrinkage of the scatter matrices is similar to L2-regularisation, leading to proportional solutions for weight vectors (with a $1 + \lambda_2$ factor).

5.2.2 L1 Regularisation of PLS

In order to push further regularisation, one may use techniques based on L1 penalisation. Recently Lê Cao et al. (2008) proposed an approach that includes variable selection in PLS regression, based on L1 penalisation (Tibshirani 1996) and leading to a sparse solution. It should be noted that L1 penalisation could not have been easily implemented with PLS-SVD, described in section 2.2.3, without loosing the orthogonality constraint of PLS-SVD on weight vectors (Zou et al. 2006).

In the rest of this work, we have dropped the h index that stands for the number of pairs of components to make the notations simpler. In sparse PLS regression (sPLS), the PLS regression criterion for each new pair of components is modified by adding a L1 penalisation on weight vectors $\mathbf{a}$ and $\mathbf{b}$:

$$\min_{\|\mathbf{a}\|_2=\|\mathbf{b}\|_2=1} -\mathbf{a}'\mathbf{X}'\mathbf{Y}\mathbf{b} + \lambda_{1X} \parallel \mathbf{a} \parallel_1 + \lambda_{1Y} \parallel \mathbf{b} \parallel_1 \tag{5.9}$$

where λ_{1X} and λ_{1Y} are L1-penalisation parameters for the weight vectors of blocks $\mathbf{X}$ and $\mathbf{Y}$ respectively. The sPLS criterion is bi-convex in $\mathbf{a}$ and $\mathbf{b}$ and may be solved iteratively for $\mathbf{a}$ fixed or $\mathbf{b}$ fixed, using soft-thresholding of variable weights at each iteration of the NIPALS inner loop (see Appendix B). Weight vectors $\mathbf{a}$ and $\mathbf{b}$ are computed using the following algorithm:

1. Initialise $\mathbf{a}$ and $\mathbf{b}$ using for instance the first pair of singular vectors of the matrix $\mathbf{X}'\mathbf{Y}$ and normalise them.
2. Until convergence of $\mathbf{a}$ and $\mathbf{b}$:
   (a) For fixed $\mathbf{b}$:

$$\widehat{\mathbf{a}} = \underset{\|\mathbf{a}\|_2=1}{\arg\min} -\mathbf{a}'\mathbf{X}'\mathbf{Y}\mathbf{b} + \lambda_{1X} \parallel \mathbf{a} \parallel_1 = g_{\lambda_{1X}}(\mathbf{X}'\mathbf{Y}\mathbf{b}) \tag{5.10}$$

   where $g_\lambda(y) = \text{sign}(y)(|y| - \lambda)_+$ is the soft-thresholding function.

   (b) Normalise $\mathbf{a}$: $\mathbf{a} \leftarrow \dfrac{\mathbf{a}}{\parallel \mathbf{a} \parallel_2}$.

   (c) For fixed $\mathbf{a}$:

$$\widehat{\mathbf{b}} = \underset{\|\mathbf{b}\|_2=1}{\arg\min} -\mathbf{a}'\mathbf{X}'\mathbf{Y}\mathbf{b} + \lambda_{1Y} \parallel \mathbf{b} \parallel_1 = g_{\lambda_{1Y}}(\mathbf{Y}'\mathbf{X}\mathbf{a}) \tag{5.11}$$

   (d) Normalise $\mathbf{b}$: $\mathbf{b} \leftarrow \dfrac{\mathbf{b}}{\parallel \mathbf{b} \parallel_2}$.

In the version of sparse PLS that we used, L1 penalisation is performed by a variant of soft-thresholding, where instead of setting λ_{1X} and λ_{1Y} directly, the corresponding numbers of **X** and **Y** variables to be kept in the model are chosen. We then defined the sPLS selection rates, $s_{\lambda_{1X}}$ and $s_{\lambda_{1Y}}$, as the number of selected variables from each block out of the total number of variables of that block. In our case, we chose to apply sparsity on SNPs only and to set $s_{\lambda_{1Y}}$ to 1 for imaging phenotypes, as we had a very large number of SNPs and only a few imaging phenotypes.

It should be noted that when there is only one target variable **y**, the solution of sparse PLS on the first component is the same as the Lasso solution when $\mathbf{X}'\mathbf{X}$ is diagonal.

Sparse versions of CCA have also been proposed by (Parkhomenko et al. 2007; 2009, Waaijenborg et al. 2008, Witten and Tibshirani 2009). However, in order to solve the non-invertibility issue of $\mathbf{X}'\mathbf{X}$ and $\mathbf{Y}'\mathbf{Y}$, they make the approximation that the covariance matrices $\frac{1}{n-1}\mathbf{X}'\mathbf{X}$ and $\frac{1}{n-1}\mathbf{Y}'\mathbf{Y}$ may be replaced by their diagonal elements, which makes sparse CCA equivalent to sparse PLS regression.

However, whether sparse PLS can face overfitting issues by itself in the case of such high-dimensional data remains an open question. This is the reason why we decided to combine it with a first step of dimension reduction on SNPs.

5.3 Dimension reduction methods

5.3.1 PC-based dimension reduction

A first way to perform dimension reduction might be to add a first step of Principal Component Analysis on each block of data before applying PLS or CCA. Regularisation is not necessary anymore in that case, as the dimension has been dramatically reduced. For each block of data, we kept as many components as necessary to explain 99% of the variance of that block.

We also investigated the performance of Principal Component Regression (PCR) of the two first imaging principal components onto the genetic components explaining 99% of the genetic variance.

5.3.2 Univariate SNP filtering

Another way to perform dimension reduction is to add to sparse PLS or regularised CCA a first step of massive univariate filtering. This step consisted of 1) $p \times q$ pair-wise linear regressions based on an additive genetic model, 2) ranking the SNPs according to the minimal p-value each SNP gets across all phenotypes, and 3) keeping the set of SNPs with the lowest "minimal" p-values. Indeed, even though univariate filtering may seem to contradict the very nature of multivariate methods such as PLS or CCA, it still allows them to extract multivariate patterns among the remaining variables and may even be necessary to overcome the overfitting issue in very high dimensional settings.

5.4 Performance evaluation

5.4.1 Generalisation capacity

A classical way to evaluate the performances of the different multivariate methods may be to assess whether the link obtained on a given sample between the two blocks of data may be generalised to a new sample. Such an approach may be called a prediction approach in the sense that one tries to determine whether a model learned on a given sample may be generalised or replicated on a new sample.

One way to assess generalisability may be to split the sample into one training sample where the model is learned, and a testing sample where the generalisability of the learned model is tested. Cross-validation (CV) techniques may also be used instead when the number of observations is too low, which happens to be our case as we will see in the next chapter.

Thus, we chose to use a cross-validation procedure order to compare the performances of the different strategies mentioned above, combining regularisation and preliminary dimension reduction techniques with PLS or CCA. For each method, at each fold of the CV, the estimation of the model (weight vectors) was done on the training sample and tested on the left-out sample (Figure 5.2 for filter-based methods and Figure 5.3 for PC-based methods). Indeed, at each fold i, the weights thus obtained were used to build the scores of the "test" sample (the left-out sample), $\mathbf{X}^i\mathbf{a}^{-i}$ and $\mathbf{Y}^i\mathbf{b}^{-i}$, and the correlation coefficient between those scores was computed. This yielded an average "test" correlation coefficient over folds $\mathrm{E}_{CV}(\mathrm{corr}(\mathbf{X}^i\mathbf{a}^{-i}, \mathbf{Y}^i\mathbf{b}^{-i}))$, called the out-of-sample correlation coefficient, which reflects the replicability of the link between the two blocks on unseen subjects. This out-of-sample correlation coefficient might be also called the expected correlation by analogy with the expected risk in the prediction framework, while the average correlation on training samples would play the role of the empirical risk. Please note that at each fold, while the correlation coefficient obtained on the training samples is forced to be positive, the out-of-sample correlation coefficient may happen to be negative.

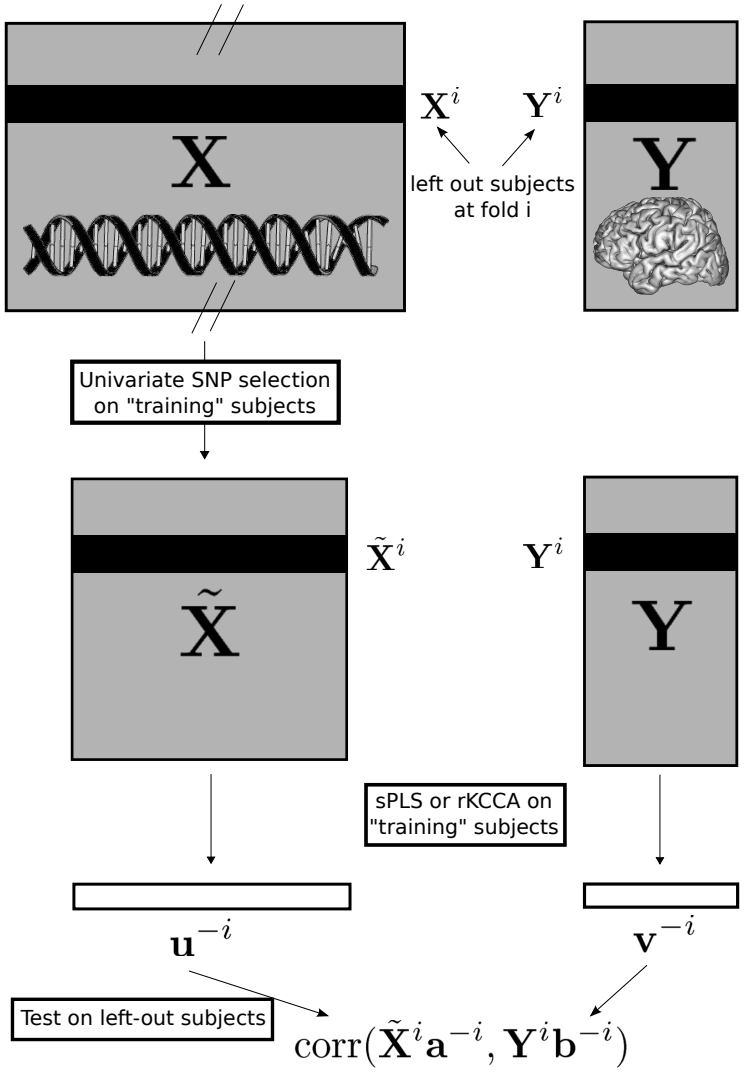


Figure 5.2 – *Illustration of the cross-validation scheme for filter-based methods. At each fold i, a univariate selection of k SNPs is performed on the data of "training" subjects* $\mathbf{X}^{-i}$ *and* $\mathbf{Y}^{-i}$*; the weight vectors,* a^{-i} *and* b^{-i}*, are then estimated by sPLS or rKCCA on the "training" subjects and finally the scores of the left out subjects corresponding to this* i^{th} *fold are computed using their observed responses* $\tilde{\mathbf{X}}^i$ *and* $\mathbf{Y}^i$ *and these weight vectors.*

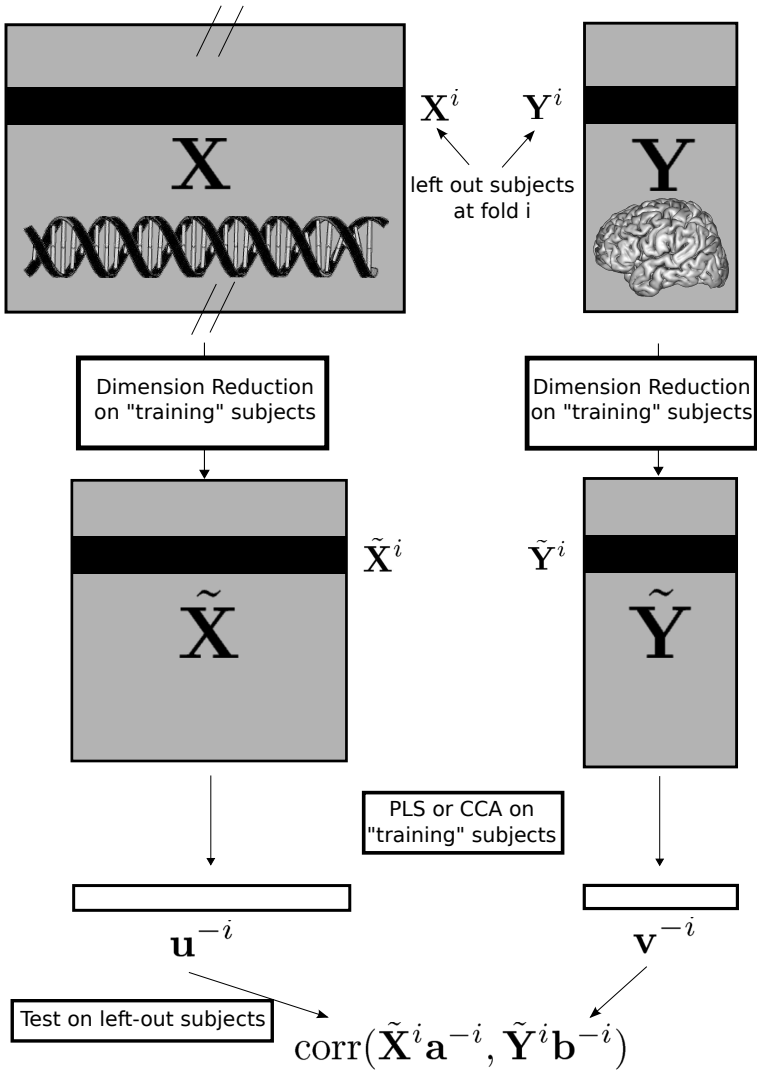


Figure 5.3 – *Illustration of the cross-validation scheme for PC-based methods. At each fold i, two PCAs are performed on SNPs and on phenotypes of "training" subjects* $\mathbf{X}^{-i}$ *and* $\mathbf{Y}^{-i}$*; the weight vectors,* a^{-i} *and* b^{-i}*, are then estimated by PLS or KCCA on the "training" subjects, and finally the scores of the left out subjects corresponding to this* i^{th} *fold are computed using the projection of their observed responses on the principal components,* $\tilde{\mathbf{X}}^i$ *and* $\tilde{\mathbf{Y}}^i$*, and these weight vectors.*

5.4.2 Statistical significance assessment

On the other hand, a detection approach, such as hypothesis testing, aims at detecting the causal variables. It is commonly used in the case of univariate methods to detect pair-wise associations, but it could also be performed in the case of multivariate methods to detect global effect, by looking at the significance of the empirical correlation using permutations for instance. However, multivariate methods suffer from overfitting in such high dimensions and hypothesis testing alone would be blind to this overfitting phenomenon, in the sense that it would only compare the overestimated or overfitted empirical correlation on the true data with its distribution under the null hypothesis obtained by permutations. Moreover even

if this empirical correlation was found to be significant, this would only guarantee that it is led by the true effect of at least one of the variables with non-null weights, without knowing which one(s).

This is the reason why it appeared to us that the prediction approach was more appropriate in order to obtain a generalisable link between two sets of variables, giving us more control over the overfitting phenomenon, even if it is more conservative. Nevertheless, we chose to add an outer permutation loop, in order to guarantee that this generalisable link is significant and led by at least one true effect.

In order to get a degree of significance for each variable, the weight assigned to a given variable on training samples could be compared with its distribution under the null hypothesis obtained by permutations, but this would ignore the fact that the variables with non-null weights need to be considered and assessed as a whole. Moreover in the case of L1 penalisation, it might suffer from some selection instability. Finally, this would lead to a major multiple comparisons issue, as in the univariate case.

5.4.3 Model selection

The performance of such multivariate methods may also be largely influenced by the choice of the parameters of dimension reduction and regularisation. The classical way to determine the best parameters is to choose the ones that show the best generalisation performance, using a cross-validation scheme.

In order to avoid overfitting due to the number of parameter combinations tested, this cross-validation procedure is often embedded in another validation procedure, such as an outer cross-validation loop or a replication on an independent dataset. However, we preferred to perform only one cross-validation loop and to correct the significance of the performance scores for the number of parameter combinations tested.

5.4.4 Bootstrap and stability selection

5.4.4.1 Bootstrap

As explained in section 5.4.2, individual variable significance cannot be easily assessed but, in the case of the multivariate methods that perform variable selection, one could try to assess the importance of each of the variables selected in the model by considering their stability of selection.

For instance, bootstrap techniques may be used to assess the robustness of selection of each variable across bootstrap samples (Efron 1979).

Similarly, since we chose a prediction approach using a cross-validation procedure, we looked at the robustness of selection of each variable across folds.

5.4.4.2 Stability selection

Another approach may be to include the stability condition into the variable selection process by integrating resampling techniques, such as boot-

strap in Bolasso (Bach 2008) or subsampling in stability selection (Meinshausen and Bühlmann 2010).

Stability selection is based on data subsampling without replacement, such that for each subsample the multivariate method selects some variables, and one only keeps at last the variables selected with a high probability across subsamples.

The results are supposed to be relatively insensitive to the amount of regularisation, which would allow us to avoid the delicate choice of the regularisation parameter. Unfortunately, this might not always be the case, especially in very high-dimensional settings for large ranges of potential regularisation values. Moreover, in order to control for false positives, the choice of the selection probability threshold remains critical.

Finally, stability selection does not yield any validation of the global multivariate model. Thus, it might need to be integrated into a validation procedure on an independent sample (or a cross-validation procedure), which may be very computationally expensive.

6 Application and assessment of the Partial Least Squares approach

In this chapter, we present the application and a comparison study of the different methods described in the previous chapter.

We first use a simulated dataset mimicking fMRI and genome-wide SNP data and compare the performances of the different methods, by assessing their detection power, as well as their capacity to generalise the link found between the two blocks with a cross-validation procedure. Indeed, we first compared PLS and CCA, then we investigated the influence of L2 regularisation on CCA and L1 regularisation on PLS, and finally we tried to add a first step of dimension reduction such as PCA or filtering.

Finally, we apply these different methods with the same cross-validation procedure on a real dataset made of fMRI and genome-wide SNP data and the statistical significance of the link obtained on "test" subjects is assessed with randomisation techniques.

In the next sections, we first introduce the simulated and real datasets, then detail the comparison study of the different multivariate methods presented in the previous chapter, and illustrate the results we obtained. Last we discuss the potential pitfalls and extensions of this work.

Contents

6.1 Data

6.1.1 Experimental dataset

This study is based on the same dataset as in section 4.2.3.1, with $n = 94$ subjects who were genotyped for $1,054,068$ SNPs and participated in a general cognitive assessment fMRI task described in Pinel et al. (2007).

fMRI data This time, we focused on two activation contrasts: *reading minus checkerboard viewing* and *speech comprehension minus rest*. We performed the same preprocessing steps as in section 4.2.3.1 to obtain activation maps.

This time, we selected thirty four brain locations of interest: 19 from the "reading" contrast (see Figure 6.1) and 15 from the "speech comprehension" contrast (see Figure 6.2). Most of them were the peaks of maximal activation, while the others had been reported to be atypically activated during reading in dyslexia (Paulesu et al. 2001). Each contrast map was locally averaged within 4 voxel-radius spheres centred on these peaks, keeping only active clusters of voxels (T$\geq$ 1 and cluster size$\geq$ 10 voxels) (Pinel and Dehaene 2009). This yielded 34 average values corresponding to 34 regions of interest (ROI) and we computed the average values for the 34 mirror ROIs by symmetry with respect to the inter-hemispheric plane. Finally, lateralisation indexes were derived from those regions. For each pair of ROIs in the normalised volume and in each subject s, an index was computed as follows:

$$\text{Index}_s = \frac{\hat{\beta}_s^{\text{right}} - \hat{\beta}_s^{\text{left}}}{\sqrt{\left(\hat{\beta}_s^{\text{right}}\right)^2 + \left(\hat{\beta}_s^{\text{left}}\right)^2}}. \tag{6.1}$$

The distribution of these indexes spanned the range of $[-1.5; 1.5]$, and variances were homogeneous across regions of interest. The term "phenotypes" will now refer to the lateralisation indexes thus obtained in the different regions.

SNP data We performed the same preprocessing steps on genetic data as in section 4.2.3.1. $622,534$ SNPs were left for further analysis. This we did not consider any specific candidate regions.

Our analyses were thus performed on two blocks of data $\mathbf{Y}$ (fMRI) and $\mathbf{X}$ (genetics) of size 94×34 and $94 \times 622,534$ respectively.

6.1.2 Simulated dataset

A simulated dataset mimicking the real dataset was also simulated in order to study the behaviour of the different methods of interest, while knowing ground truth. 500 samples of 34 imaging phenotypes were simulated from a multivariate normal distribution with parameters estimated from the experimental data.

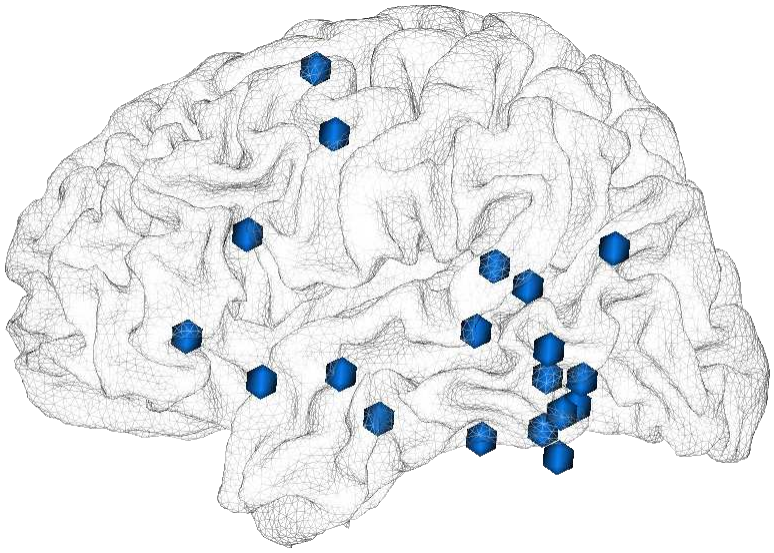


Figure 6.1 – *Locations of the 19 lateralisation indexes extracted from the "reading" contrast map.*

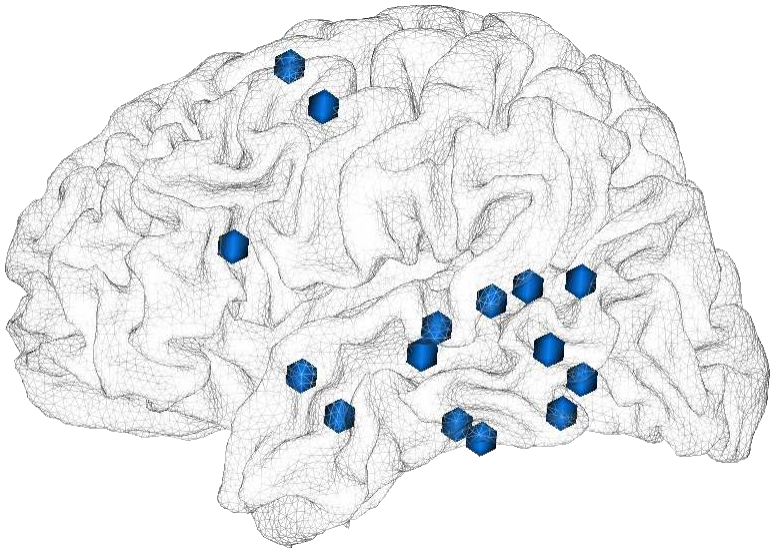


Figure 6.2 – *Locations of the 15 lateralisation indexes extracted from the "speech comprehension" contrast map.*

In order to simulate genotyping data with a genetic structure similar to that of our real data, we considered a simulation method that uses the HapMap CEU panel. We used the gs algorithm proposed by Li and Chen (2008) with the phased (phase III) data for CEU unrelated individuals for chromosome 1 ; we only consider the genotype simulation capability of this software that may also generate linked phenotypes. We generated a dataset consisting in 85,772 SNPs and 500 samples, using the extension method of the algorithm.

We randomly selected 10 SNPs (out of 85,772) having a MAF=0.2 and 8 imaging phenotypes (out of 34). We induced two independent causal patterns: for the first pattern we associated the first 5 SNPs with the first 4 imaging phenotypes; the second pattern was created associating the 5 remaining SNPs with the 4 last phenotypes. For each causal pattern $i \in \{1, 2\}$, we induced a genetic effect using an additive genetic model in-

volving the average of the causative SNPs (x_{ik}): $\bar{x}_i = \sum_{k=1}^{5} \frac{1}{5} x_{ik}$. Then each imaging phenotype y_{ij} ($j \in [1, \ldots, 4]$) of the pattern i was affected using a linear model:

$$y^{\star}_{ij} = y_{ij} + \beta_{ij}\bar{x}_i \tag{6.2}$$

The parameter β_{ij} was set by controlling for the correlation (at a value of 0.5) between the j^{th} affected imaging phenotype ($y^{\star}_{ij}$) and the causal SNPs ($\bar{x}_i$) ie.: $\text{corr}(y^{\star}_{ij}, \bar{x}_i) = 0.5$. Such control of the correlation (or the explained variance) is equivalent to the control of the effect size while controlling for the variances of SNPs ($\text{var}(\bar{x}_i)$) and (unaffected) imaging phenotypes ($\text{var}(y_{ij})$, as well as any spurious covariance between them ($\text{cov}(y_{ij}, \bar{x}_i)$). We favour such control over a simple control for the effect size since the later may result in arbitrary huge or weak associations depending on the genetic/imaging variances ratios.

SNP whose r^2 coefficient with any of the causal SNPs is at least 0.8 is also considered as causal. Such LD threshold, commonly used in the literature (de Bakker et al. 2005), led to 56 causal SNPs: 32 in "pattern 1" and 24 in "pattern 2". We will use those SNPs as "ground truth" of truly causal SNPs to compute sensitivity and specificity of the learning methods. Finally, we striped off 10 blocks of SNPs around the 10 causal SNPs, from the whole genetic dataset, considering that neighbouring SNPs were in LD with the marker if their r^2 were at least 0.2. The 5 first (resp. last) blocks, of pattern 1 (resp. 2), are made of 127 (resp. 71) SNPs and contain all the 32 (resp. 24) SNPs that were declared as causal. The striped blocks were concatenated and moved at the beginning of the dataset leading to 198 (127+71) informative features followed by $85,574$ ($85,772 - 198$) non-informative (noise) features. Such a dataset organisation provides a simple way to study the methods' performances while the dimensionality of the input dataset increases from 200 (mostly made of informative features) to $85,772$ mostly made of noise.

6.2 Comparison study

We compared the performances of the different methods on both simulated and real datasets. Indeed we first compared PLS and CCA, then we investigated how their performance is improved by regularisation with sparse PLS and L2-regularised CCA, and we finally assessed the influence of a first dimension reduction step by PCA or filtering. Note that computations were always limited to the two first pairs of latent variables for computational time purposes. Moreover we were also interested in comparing the different methods with MULM. Table 6.1 summarises the different methods we tested and the acronyms we used.

In this work we investigated in particular the performance of fsPLS on both simulated and real data and we tried to assess how much the performance of fsPLS is influenced by the fact of varying the sparse PLS penalisation parameter $s_{\lambda_{1X}}$ and the number k of SNPs kept by the filter.

Method	Acronym
Mass Univariate Linear Modelling	MULM
Partial Least Squares	PLS
Kernel Canonical Correlation Analysis	KCCA
sparse PLS	sPLS
regularised KCCA	rKCCA
Principal Component Analysis + PLS	PCPLS
Principal Component Analysis + KCCA	PCKCCA
Filtering + (sparse) PLS	f(s)PLS
Filtering + (regularised) KCCA	f(r)KCCA

Table 6.1 – *Summary of the different strategies investigated*

6.3 Performance evaluation

6.3.1 Cross-validation

We decided to evaluate the performances of the different methods by assessing the generalisability of the link they find between the blocks, on both simulated and real data, using a 5-fold and a 10-fold cross-validation (CV) scheme respectively. On the real dataset, we used "training" sets of 84 or 85 subjects and "test" sets of 9 or 10 subjects. In order to have "training" sets of about the same size on the simulated dataset, we used "training" sets of 100 subjects and "test" sets of 400 subjects.

The CV procedure has already been illustrated in Figure 5.2 for filter-based methods and Figure 5.3 for PC-based methods.

We performed a CV for MULM as well, where at each fold the two most significantly associated SNP/phenotype pairs on the training sample were extracted and tested by computing their correlation coefficient on the hold-out sample. We may note at this point that the univariate approach alone did not yield any significant SNP/phenotype associations at the 5% level after Bonferroni or FDR correction on simulated and real datasets.

6.3.2 Positive Predictive Value

Finally, in the case of simulated data, ground truth was known and we could also compare the performances of the different methods by computing the Positive Predictive Value (PPV) when 50 SNPs are selected by each method. This is almost equivalent in our case to the sensitivity of each method when 50 SNPs are selected, since there are 56 causal SNPs in our simulated dataset. In fact, for each method, 5 PPV curves were separately computed on 5 non-overlapping subsamples of 100 observations and averaged over these 5 subsamples. It should be noted that the informative SNPs that are not considered as causal are only slightly correlated to causal SNPs ($0.2<r^2<0.8$). Therefore they were removed to compute the PPV, since they could not really be identified as true or false effects.

6.4 Results

6.4.1 Performance assessment on simulated data

6.4.1.1 Influence of regularisation

We were first interested in comparing the performances of PLS and CCA when the number of SNPs p increases, from 200 (mostly made of the 198 informative features) up to 85,772 SNPs (mostly made of noise), and investigating the influence of L1 regularisation on PLS and of L2 regularisation on CCA.

Figure 6.3, on the left panel, shows the out-of-sample correlation coefficients obtained with the different methods for the two first component pairs, and it shows that in the lower dimensional space ($p = 200$) mostly made of informative features, the pure CCA, rKCCA without regularisation ($\lambda_2 = 0$), has overfitted the "training" data on the first component pair ("training" corr. ≈ 1 and "test" corr. ≈ 0.2). Such a result highlights the limits of pure CCA to deal with situations where the number of training samples (100) is smaller than the dimension ($p = 200$). However, with a suitable regularisation in such a low-dimensional setting, rKCCA($\lambda_2 = 100$) performed better than all other methods, notably all (sparse) PLS. This results was expected since the evaluation criterion (correlation between factorial scores) is exactly the one which is maximised by CCA.

Nevertheless, the increase of space dimensionality (with irrelevant features) clearly highlights the superiority of PLS and more notably sPLS over rKCCA in high-dimensional settings: the performance of rKCCA rapidly decreases while sPLS ($s_{\lambda_{1X}} = 0.1$) tolerates an increase of the dimensionality up to 1000 features before its performance starts to decrease. One may note that as expected theoretically, along with the increase of penalisation (λ_2), rKCCA curves smoothly converge toward PLS.

On the second component pair, the results are less clearly interpretable. However (s)PLS curves are above the rKCCA ones.

The four graphs on the right panel of Figure 6.3 demonstrate the superiority of sPLS methods to identify causal SNPs on the two first genetic components. Indeed, for each method and for different values of p, we computed the PPV for the two first genetic components and for each simulated pattern. PPV curves show a smooth increase of the performance, when moving from unregularised CCA (rKCCA($\lambda_2 = 0$)) to strongly regularised PLS (sPLS($s_{\lambda_{1X}} = 0.1$)). Moreover, while the out-of-sample correlation coefficient was not an appropriate measure to distinguish between the two causal patterns, PPV curves were computed for each pattern separately. One may note that the PPV on the first genetic component appears to be much higher for the first pattern than for the second pattern, especially in low dimensions, while the opposite trend is observed on the second genetic component. This observation tends to show that the first causal pattern is captured by the first component pair, while the second pattern is captured by the second pair. It should be noted that the PPV even reaches one when $p = 200$ for the first pattern on the first component, meaning that only true positives from the first pattern are detected

on this component. Similarly, the PPV reaches one when $p = 200$ for the second pattern on the second component.

6.4.1.2 Influence of the dimension reduction step

Then we investigated the influence of a first step of dimension reduction. Figure 6.4 presents different dimension reduction strategies: Principal Component (PC), filter (f), sparse (s) and combined filter+sparse (fs) methods. The selection parameter setting, 50 selected SNPs, was derived from the known ground truth (56 true causal SNPs). The 50 SNPs were either the 50 best ranked SNPs for fKCCA and fPLS, the 50 non-null weights for sparse PLS, or the 50 SNPs resulting from the combination of filtering and sparsity for fsPLS (10% of the 500 best ranked SNPs or 50% of 100 SNPs).

Figure 6.4, on the left panel, shows that all PC-based methods (green curves) failed to identify generalisable covariations when the number of irrelevant features increases.

Dimension reduction based on filtering slightly improved the performance of CCA and greatly improved the performance of PLS: fPLS($k = 50$) is the second best approach in our comparative study.

Moreover, as previously observed in Figure 6.3, L1 regularisation limits the overfitting phenomenon (see sPLS($s_{\lambda_{1X}} * p = 50$) in Figure 6.4) and delays the decrease of PLS performance when the dimensionality increases. Finally the best performance is obtained by combining filtering and L1 regularisation: fsPLS($k = 100$, $s_{\lambda_{1X}} = 0.5$), which keeps 100 SNPs after filtering and selects 50% of those SNPs by sPLS. Please note that the performance of fsPLS($k = 500$, $s_{\lambda_{1X}} = 0.1$) is lower and similar to that of sPLS(50) in low dimensions, but becomes more robust than sPLS and equivalent to fsPLS($k = 100$, $s_{\lambda_{1X}} = 0.5$) in higher dimensions. However, the purely univariate strategy based on MULM shows poor generalisability, which suggests that even though filtering appears necessary to remove irrelevant features, it is not able to capture the imaging/genetics link by itself and needs to be combined with a multivariate step which will take advantage of the cumulative effects of several SNPs.

Again, on the second component pair, the results are less clearly interpretable. However the curves of the strategies that combine filtering and sparsity are above the other ones.

The four graphs on the right panel of Figure 6.4 show that the results in terms of PPV performance are similar to cross-validation results. However, it should be noted that the PPV does not take into account the weights/ranks assigned by the different methods to the selected SNPs. Therefore, the PPV curves of fKCCA($k = 50$) and fPLS($k = 50$) are superimposed on the MULM curve in our case, since the 50 SNPs selected by the filter are the 50 best SNPs obtained with MULM.

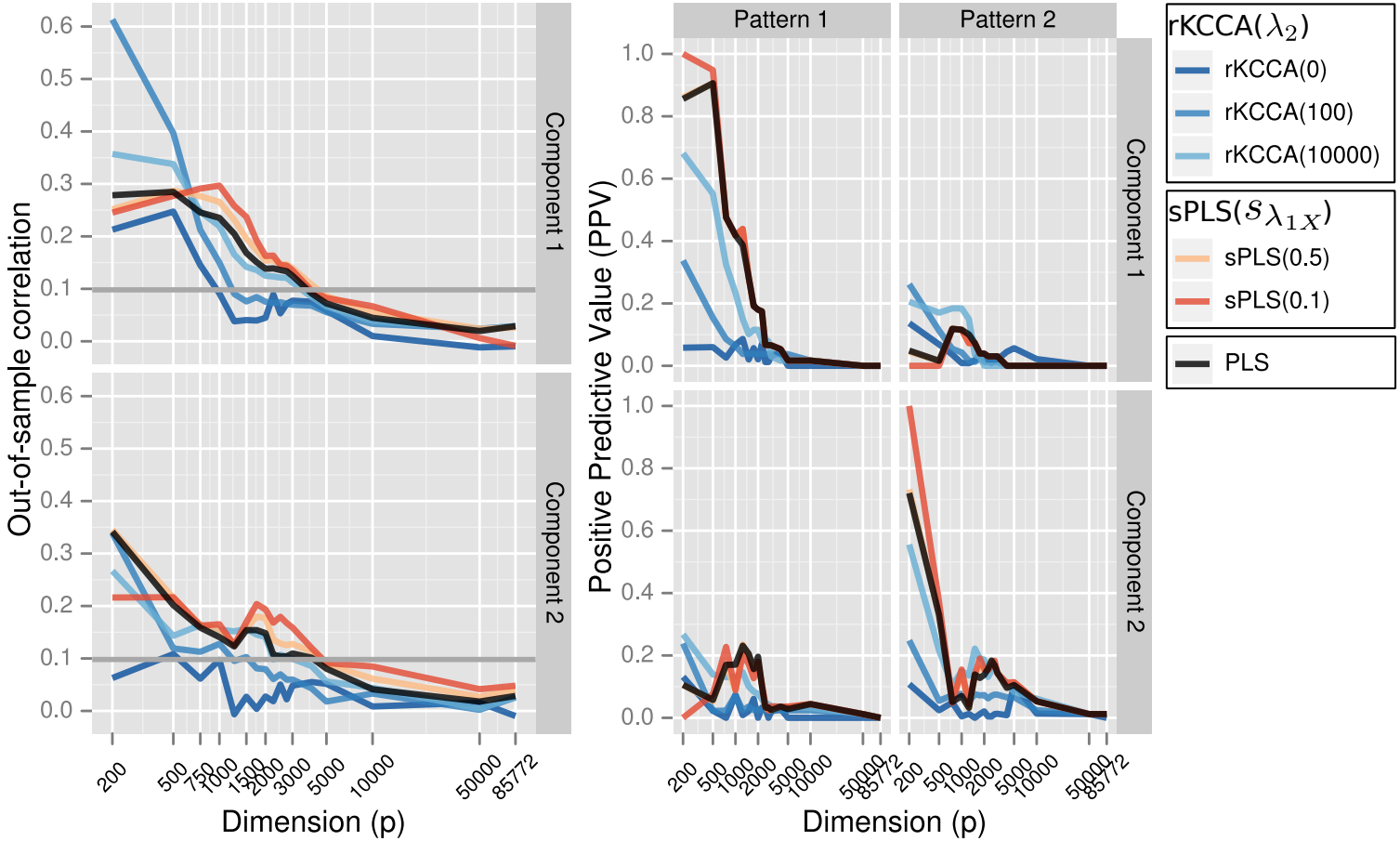


Figure 6.3 – *Comparison of regularisation methods to deal with genetic datasets containing an increasing number of irrelevant features. The total number of features varies along the x-axis between* 200 *and* 85,772 *SNPs. We compared: (i) in blue, regularised kernel CCA (rKCCA) with various L2 regularisation values (λ_2) ranging from* 0 *(pure CCA) to* 10,000*; (ii) in black, PLS; (iii) in red, sparse PLS (sPLS) with various L1 regularisation values ($s_{\lambda_{1X}}$) ranging from* 0.75 *(75% of input features have a non null weight) to* 0.1.
The y-axis of the two left panels shows the (5-fold CV) average out-of-sample correlation coefficients between the two first component pairs.
The four right panels present the power of the methods to identify causal SNPs implied in the two causal patterns. The y-axis depicts the Positive Predictive Values when 50 *SNPs are selected, for each of the two first genetic components: (a_1, a_2).*

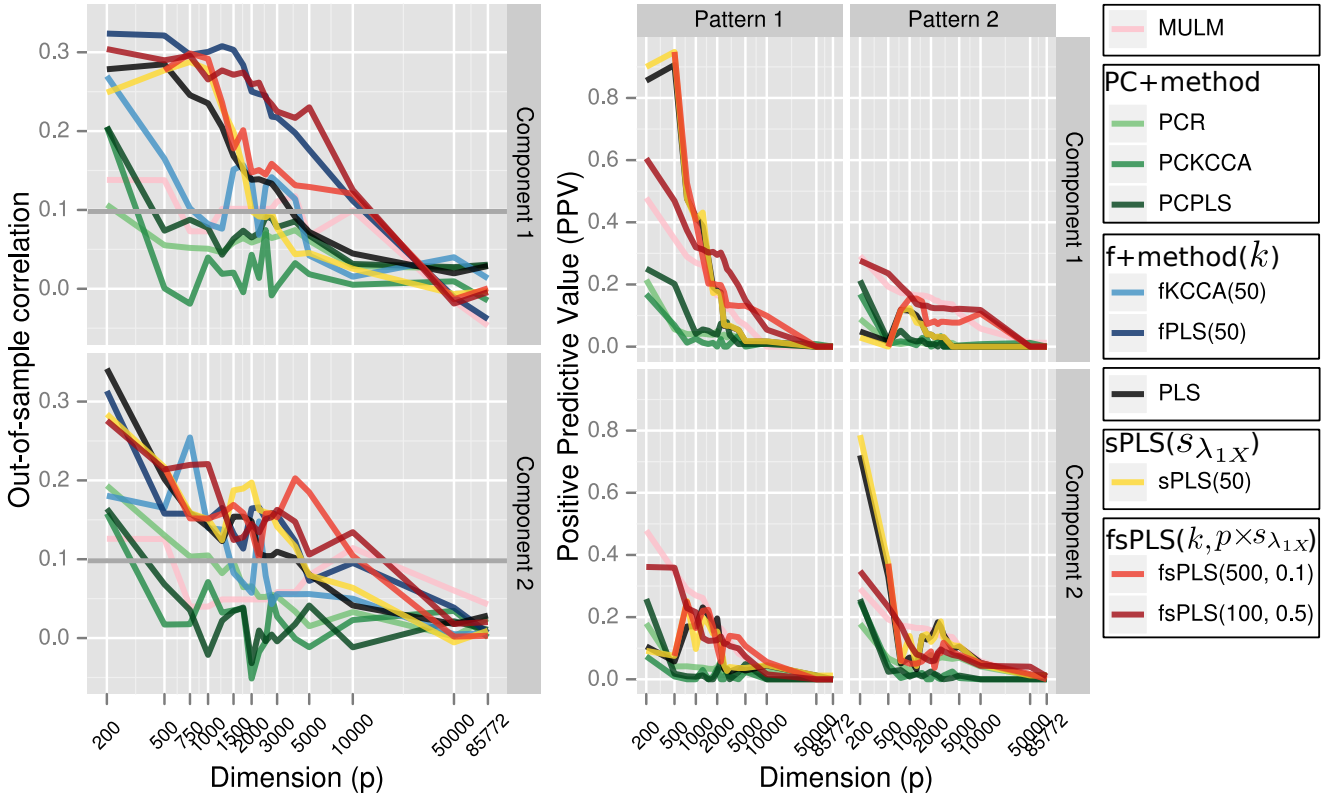


Figure 6.4 – *Comparison of dimension reduction methods to deal with genetic datasets containing an increasing number of irrelevant features. The total number of features varies along the x-axis between* 200 *and* 85,772 *SNPs. We compared:*
(i) in green, Principal Component (PC) based methods: PC regression (PCR), PCA+KCCA (PCKCCA), PCA+PLS (PCPLS).
(ii) in blue, filter (f) based methods: f+KCCA (fKCCA), f+PLS (fPLS). We selected only the 50 *best SNPs, while according to ground truth* 56 *SNPs were identified as causal. They are superimposed with MULM on the PPV curves.*
(iii) in black, PLS.
(iv) in yellow, sparse PLS (sPLS) where $s_{\lambda_{1X}}$ *is such that* 50 *features have a non null weight.*
(v) in red, filter + sparse PLS (fsPLS) with settings both leading to 50 *selected features: fsPLS(*$k = 500$, $s_{\lambda_{1X}} = 0.1$*) (resp. fsPLS(*$k = 100$, $s_{\lambda_{1X}} = 0.5$*)) keeps the* 500 *(resp.* 100*) best ranked features and then* 10% *(*50%*) get a non null weight.*
(vi) finally in pink, we add MULM.
*The y-axis of the two left panels shows the (*5*-fold CV) average out-of-sample correlation coefficients between the two first component pairs.*
The four right panels present the power of the methods to identify causal SNPs implied in the two causal patterns. The y-axis depicts the Positive Predictive Values when 50 *SNPs are selected, for each of the two first genetic components:* (a_1, a_2).

6.4.2 Performance assessment on real data

6.4.2.1 Comparative analysis

Table 6.2 summarises the two first average correlation coefficients obtained on "test" samples (ρ^1_{test} and ρ^2_{test}) for the different methods tested, as well as the two first average correlation coefficients obtained on "training" samples ($\rho^1_{training}$ and $\rho^2_{training}$). The "optimal" parameters for regularisation and filtering chosen here are those that gave the best average cross-validated correlation coefficients, among all parameters tested.

	ρ^1_{test}	ρ^2_{test}	$\rho^1_{training}$	$\rho^2_{training}$
MULM	0.036	-0.104	-0.458	-0.451
PLS	-0.092	0.218	0.990	0.984
sPLS ($s_{\lambda_{1X}} = 0.1\%$)	0.008	0.201	0.938	0.922
PCKCCA	0.010	0.008	1.000	1.000
PCPLS	-0.088	0.217	0.990	0.984
frKCCA ($k = 1000$,$\lambda_2 = 1,000,000$)	0.245	0.324	0.963	0.954
fPLS ($k = 1000$)	0.236	0.268	0.962	0.953
fsPLS ($k = 1000$,$s_{\lambda_{1X}} = 5\%$)	0.432	0.210	0.772	0.788

Table 6.2 – *The two first average correlation coefficients found on left-out "test" samples and on "training" samples.*

Table 6.2 shows that, for the first pair of components, L1 regularisation of PLS cannot solve the overfitting issue by itself. Indeed, like PLS, sparse PLS ($s_{\lambda_{1X}} = 0.1\%$) completely failed to extract a generalisable link in such high dimensions and capture only noise. In such high dimensions, KCCA requires such an extreme L2 regularisation that it is equivalent to PLS in terms of correlation between latent variables (with a proportionality factor of $\frac{1}{1+\lambda_2}$ on weight vectors).

Therefore a first step of dimension reduction appears to be necessary in order to overcome the overfitting issue. Indeed, even though PC-based methods do not succeed either, filtering-based methods perform much better. Among filtering-based methods, fsPLS yields the highest out-of-sample correlation coefficient of 0.43 when 1000 SNPs are left after the univariate filter and respectively 5% of the remaining SNPs are kept by sparse PLS. The second best performance on the first pair of components is obtained with frKCCA with an out-of-sample correlation coefficient of 0.24 ($k = 1000$ and $\lambda_2 = 1,000,000$). However, with such an extreme L2 regularisation, it is almost equivalent to fPLS, as can be seen on Table 6.2.

As for the second component pair, the out-of-sample correlation coefficient obtained by fsPLS is lower than on the first component pair. However for all the other PLS-based methods, the correlation appears to be slightly higher on the second component pair than on the first one. This may be explained by the fact that once the noise leading to overfitting on the first component pair has been removed, some real effects may be observed on further components, while on the opposite, fsPLS prevents from overfitting and can capture some effects on both pairs of components. Finally, MULM and PCA+KCCA do not seem able to capture any generalisable effects on any of the component pairs.

6.4.2.2 Sensitivity analysis of fsPLS and significance assessment

We now detail the sensitivity analysis we performed in order to assess how much the performance of fsPLS is influenced by the sparse PLS penalisation parameter $s_{\lambda_{1X}}$ and by the number k of SNPs kept by the filter, and to select the best pair of parameters. Indeed, we tested different values for the number k of SNPs to be kept by the univariate filter: the 10, 100, 1000 and 10000 "best" ranked SNPs. Seven different sPLS selection rates $s_{\lambda_{1X}}$ were also tested on SNPs ($\mathbf{X}$): 1, 5, 10, 25, 50, 75 and 100%. For instance, when considering 1000 SNPs kept after univariate filtering, the 75% condition means that only 750 SNPs will have non-zero PLS weights. The 10-fold cross-validation procedure presented in 6.3 was repeated for each pair of parameters (k,$s_{\lambda_{1X}}$).

Moreover, we calibrated the degree of significance of the out-of-sample correlation coefficients thus obtained using a randomisation procedure where, at each permutation, the rows of $\mathbf{Y}$ were permuted and the cross-validation procedure was repeated on the permuted dataset for each pair of parameters. We performed 1000 permutations in order to get a good estimation of the empirical p-values. We then corrected our empirical p-values for multiple comparisons, because of the different pairs of parameters tested, using a maxT procedure which derives corrected p-values from the empirical distribution of the maximal statistic over tests (Westfall and Young 1993). Table 6.3 summarises the out-of-sample correlation coefficient obtained for the first pair of components using fsPLS, together with its statistical significance, as a function of k and $s_{\lambda_{1X}}$. One can see on Table 6.3 that the best out-of-sample correlation coefficient of 0.43, obtained with $k = 1000$ and $s_{\lambda_{1X}} = 5\%$, happens to be significant after correction ($p = 0.034$). The second best out-of-sample correlation coefficient of 0.41 with $k = 1000$ and $s_{\lambda_{1X}} = 10\%$ is significant as well ($p = 0.043$). Out-of-sample correlation coefficients were not significant for the second component pair.

We also assessed the robustness of selection, across folds, of the SNPs selected in the most significant model with $k = 1000$ and $s_{\lambda_{1X}} = 5\%$. 22 SNPs were selected in more than half of the folds.

		$s_{\lambda_{1X}}$						
		1%	5%	10%	25%	50%	75%	100%
k	10	0.041	0.041	0.041	0.041	0.144	0.112	0.112
	100	0.182	0.074	0.085	0.057	0.069	0.188	0.243
	1000	0.151	0.432 *	0.414 *	0.400	0.317	0.285	0.236
	10000	0.004	0.120	0.130	0.027	−0.006	−0.031	−0.061

Table 6.3 – *Out-of-sample correlation coefficient on the first component pair as a function of k and $s_{\lambda_{1X}}$. Empirical p-values still significant ($p < .05$) after correction are shown here as: *.*

6.4.3 Imaging genetics findings

In order to obtain the SNPs and the brain phenotypes involved in the link between the two blocks, we then applied fsPLS on all the subjects simultaneously for the pair of parameters giving the most significant results on

the first component pair: 1000 SNPs selected with the univariate filter and a sPLS selection rate of 5% ($k = 1000$, $s_{\lambda_{1X}} = 0.05$). It should noted at this point that the significance of the multivariate model has to be considered as a whole and not SNP by SNP, thus we have to be very careful with the interpretation of the results.

After the univariate step, one may observe that each phenotype is associated with at least one of the 1000 best ranked SNPs, if one refers to the univariate p-values (Figure 6.5). This reinforces the idea that the problem is multivariate on both the imaging and genetic sides and that there may exist interactions both between SNPs and between phenotypes, which suggests that the second step of multivariate analysis is useful.

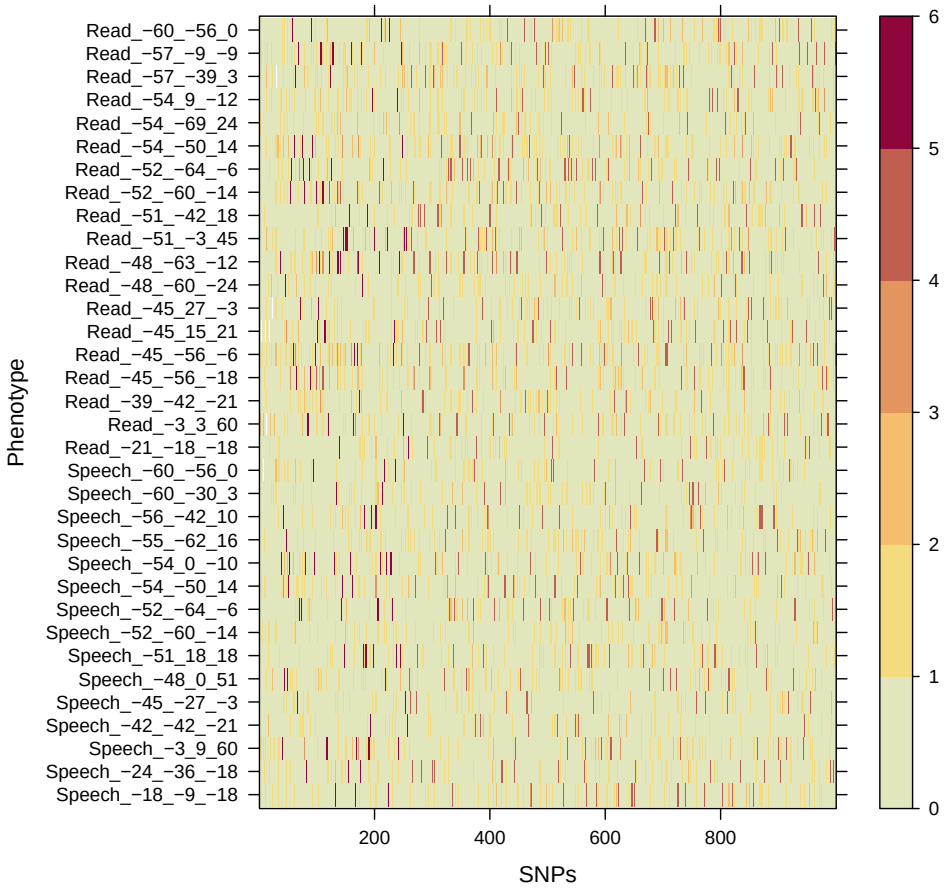


Figure 6.5 – *Distribution of the p-values ($-log(p)$) for the* 1000 *best ranked SNPs after univariate filtering with each of the* 34 *phenotypes (MNI coordinates are reported in brackets for the corresponding task, Reading or Speech comprehension).*

Figure 6.6 and Figure 6.7 provide an illustration of the sPLS weights of SNPs and phenotypes in the genetic and imaging components respectively, after this second step. The two intra-block correlation matrices are shown below the graphs. One may notice that all phenotypes do not contribute to the same extent to the first component and that there seems to be to a stronger involvement of the phenotypes obtained from the "reading" contrast.

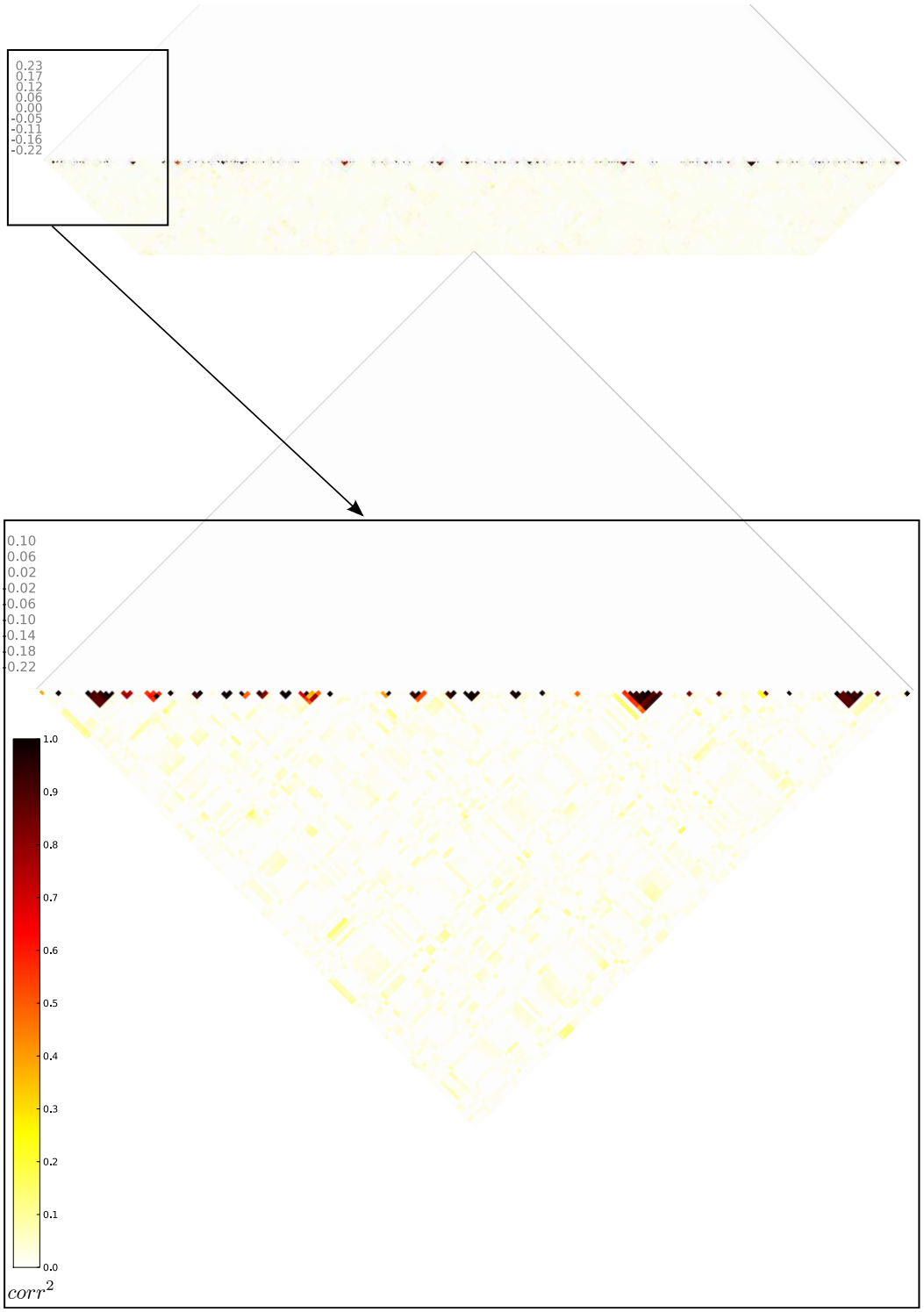


Figure 6.6 – *sPLS weights for SNPs, when considering* $k = 1000$ *SNPs ordered here according to their position along the genome, with* $s_{\lambda_{1X}} = 5\%$ *of selected SNPs. Here we zoom only on the first* 150 *SNPs for visualisation purposes. The matrix of squared pairwise correlations is shown below.*

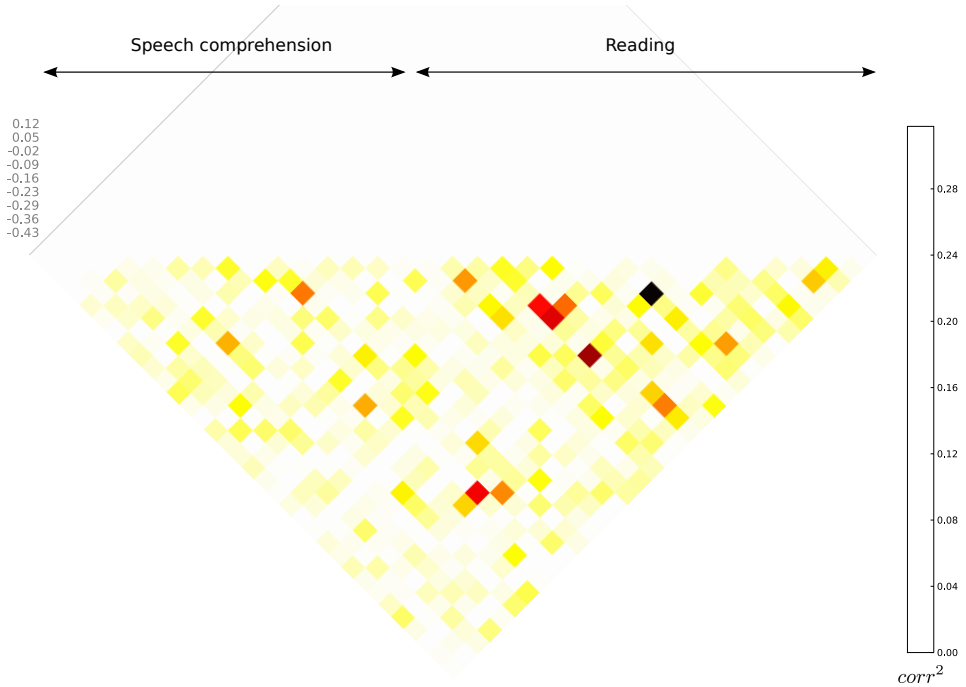


Figure 6.7 – *sPLS weights for phenotypes, when considering* $k = 1000$ *SNPs with* $s_{\lambda_{1X}} = 5\%$ *of selected SNPs. The matrix of squared pairwise correlations is shown below.*

Figure 6.8 and Figure 6.9 show the location of the selected SNPs and of the phenotypes respectively.

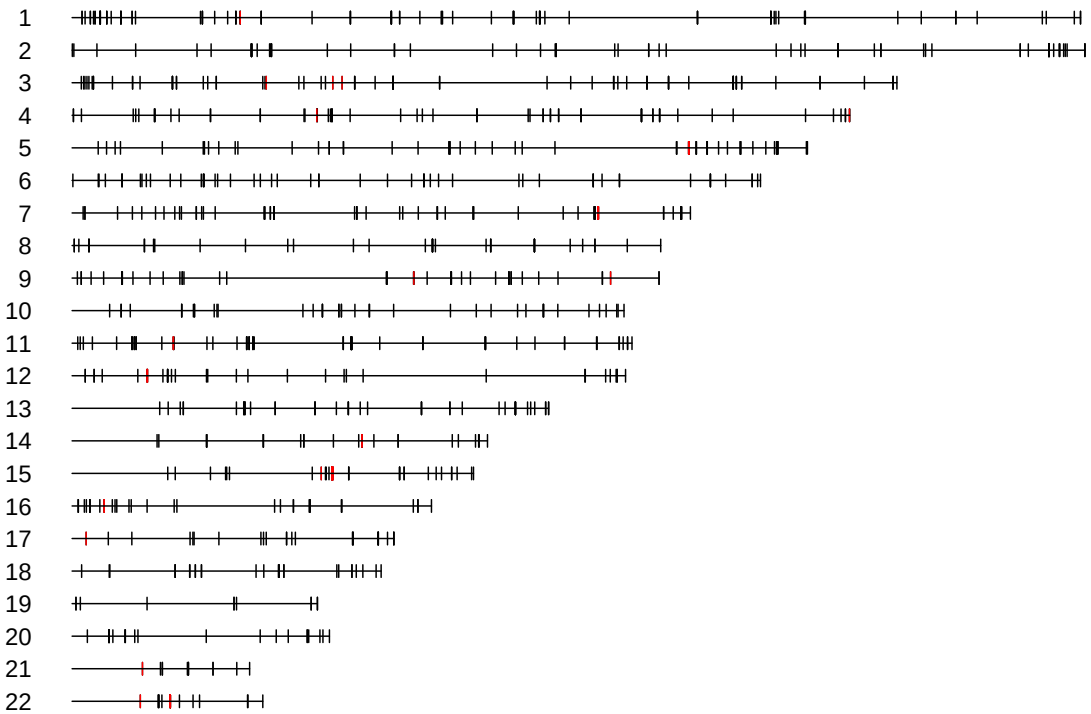


Figure 6.8 – *Distribution of the* 1000 *most significant SNPs (univariate tests) across the genome. The* 50 *SNPs selected by sPLS are highlighted in red.*

The distribution of the 1000 SNPs having the lowest univariate *p*-values along the 22 autosomes is illustrated in Figure 6.8. The 5% of those SNPs that were selected by sPLS are highlighted in red. As can be seen, they spread over all autosomal chromosomes and some of them seem to be in linkage disequilibrium. Among the 50 SNPs selected by fsPLS, some of them were located within a gene (see Table 6.4). Eighteen genes were thus identified (Table 6.5), such as PPP2R2B and RBFOX1, which have been reported to be linked with ataxia and a poor coordination of speech and body movements, or also PDE4B which has been associated with schizophrenia and bipolar disorder.

Reference SNP ID	Ensembl Gene ID	Location within gene	Chrom.	Position
*rs*13047077	C21orf34	Within Non Coding Gene	21	17794297
*rs*2070477	C22orf36	Upstream	22	24990916
*rs*5751901	C22orf36	Upstream	22	24992266
*rs*5760489	C22orf36	Upstream	22	24990646
*rs*6519519	C22orf36	Upstream	22	24991863
*rs*874852	C22orf36	Intronic	22	24987964
*rs*1894702	F5	Intronic	1	169530837
*rs*3934552	FBXL22	Downstream	15	63894400
*rs*12891349	GALNTL1	Intronic	14	69790389
*rs*4902713	GALNTL1	Intronic	14	69770939
*rs*8017671	GALNTL1	Intronic	14	69771213
*rs*2070477	GGT1	Intronic	22	24990916
*rs*5751901	GGT1	Intronic	22	24992266
*rs*5760489	GGT1	Intronic	22	24990646
*rs*5760492	GGT1	Intronic	22	24995202
*rs*6519519	GGT1	Intronic	22	24991863
*rs*874852	GGT1	Intronic	22	24987964
*rs*10519223	HERC1	Intronic or Splice Site	15	63935149
*rs*11630720	HERC1	Intronic or Splice Site	15	63984772
*rs*11635117	HERC1	Intronic	15	64112732
*rs*2228510	HERC1	Non Synonymous Coding	15	63970456
*rs*3764186	HERC1	Intronic	15	64056437
*rs*8034342	HERC1	Intronic	15	64038870
*rs*8034675	HERC1	Intronic	15	64039050
*rs*9972527	HERC1	Upstream	15	64127531
*rs*564249	HPCAL4	Intronic	1	40155623
*rs*2187522	NELL1	Intronic	11	21357112
*rs*4257797	ODZ2	Intronic	5	166869195
*rs*7688580	PAPSS1	Intronic	4	108518005
*rs*12081185	PDE4B	Intronic	1	66321193
*rs*4609402	PDE4B	Intronic	1	66318628
*rs*6684621	PDE4B	Intronic	1	66315450
*rs*1480149	PPP2R2B	Intronic	5	146448551
*rs*1480150	PPP2R2B	Intronic	5	146454825
*rs*6580448	PPP2R2B	Intronic	5	146438035
*rs*6872842	PPP2R2B	Upstream	5	146462839
*rs*1871394	PTPRG	Intronic	3	61931534

Reference SNP ID	Ensembl Gene ID	Location within gene	Chrom.	Position
*rs*12598550	RBFOX1	Intronic	16	7683677
*rs*3785228	RBFOX1	Intronic	16	7679580
*rs*999566	RP11-343J18.2	Within Non Coding Gene	9	128835802
*rs*439339	SLC13A3	Intronic	20	45238334
*rs*7178762	USP3	Intronic	15	63871292
*rs*10834273			11	24091682
*rs*11043662			12	17928813
*rs*13086717			3	46139499
*rs*1480162			5	146471808
*rs*1534101			7	125149625
*rs*17680472			13	71273644
*rs*2120252			15	64136472
*rs*4241767			4	184353138
*rs*4341595			12	18033101
*rs*4477486			12	17939279
*rs*4820001			22	17827684
*rs*4865243			4	58421116
*rs*7044535			9	81969574

Table 6.4 – *SNPs selected by fsPLS.*

Gene name	Function
C21orf34	Non-coding
C22orf36	Unknown
F5	Central regulator of hemostasis. It serves as a critical cofactor for the prothrombinase activity of factor Xa that results in the activation of prothrombin to thrombin
FBXL22	Recognises and binds to some phosphorylated proteins and promotes their ubiquitination and degradation
GALNTL1	May catalyse the initial reaction in O-linked oligosaccharide biosynthesis, the transfer of an N-acetyl-D-galactosamine residue to a serine or threonine residue on the protein receptor (By similarity)
GGT1	Initiates extracellular glutathione (GSH) breakdown, provides cells with a local cysteine supply and contributes to maintain intracelular GSH level. It is part of the cell antioxidant defense mechanism. Catalyses the transfer of the glutamyl moiety of glutathione to amino acids and dipeptide acceptors.
HERC1	This protein is thought to be involved in membrane transport processes.
HPCAL4	May be involved in the calcium-dependent regulation of rhodopsin phosphorylation
NELL1	Involved in the control of cell growth and differentiation
ODZ2	May function as a cellular signal transducer
PAPSS1	Bifunctional enzyme with both ATP sulfurylase and APS kinase activity, which mediates two steps in the sulfate activation pathway
PDE4B	Hydrolyses the second messenger cAMP, which is a key regulator of many important physiological processes. May be involved in mediating central nervous system effects of therapeutic agents ranging from antidepressants

Gene name	Function
	to antiasthmatic and anti-inflammatory agents
PPP2R2B	The B regulatory subunit might modulate substrate selectivity and catalytic activity, and also might direct the localisation of the catalytic enzyme to a particular subcellular compartment. Defects in this gene cause autosoma dominant spinocerebellar ataxia 12 (SCA12), a disease caused by degeneration of the cerebellum, sometimes involving the brainstem and spinal cord, and in resulting in poor coordination of speech and body movements.
PTPRG	Possesses tyrosine phosphatase activity
RBFOX1	RNA-binding protein that regulates alternative splicing events by binding to 5'-UGCAUGU-3' elements. Regulates alternative splicing of tissue-specific exons and of differentially spliced exons during erythropoiesis. This protein binds to the C-terminus of ataxin-2 and may contribute to the restricted pathology of spinocerebellar ataxia type 2 (SCA2). Ataxin-2 is the gene product of the SCA2 gene which causes familial neurodegenerative diseases.
RP11-343J18.2	Non-coding
SLC13A3	High-affinity sodium-dicarboxylate cotransporter that accepts a range of substrates with $4-5$ carbon atoms
USP3	Hydrolase that deubiquitinates monoubiquitinated target proteins such as histone H2A and H2B. Required for proper progression through S phase and subsequent mitotic entry. May regulate the DNA damage response (DDR) checkpoint through deubiquitination of H2A at DNA damage sites. Associates with the chromatin

Table 6.5 – *Genes selected by fsPLS.*

Figure 6.9 shows the location of the phenotypes where lateralisation indexes were computed, for both contrast maps of interest "reading" and "speech comprehension". The weights assigned by sPLS to the imaging phenotypes are illustrated according to the colourbar. The phenotypes that obtained the largest weights (in absolute value) mainly come from the "reading" contrast, especially from the temporal lobe.

Taken altogether, our results show that fsPLS could establish a significant link between a subset of SNPs distributed across the genome and a functional brain network activated during a reading task, some of these SNPs being probably indirectly linked to the neuroimaging phenotypes due to linkage disequilibrium. This suggests that individual variability in the entire genome contains predictors of the observed variability in brain activation during language tasks.

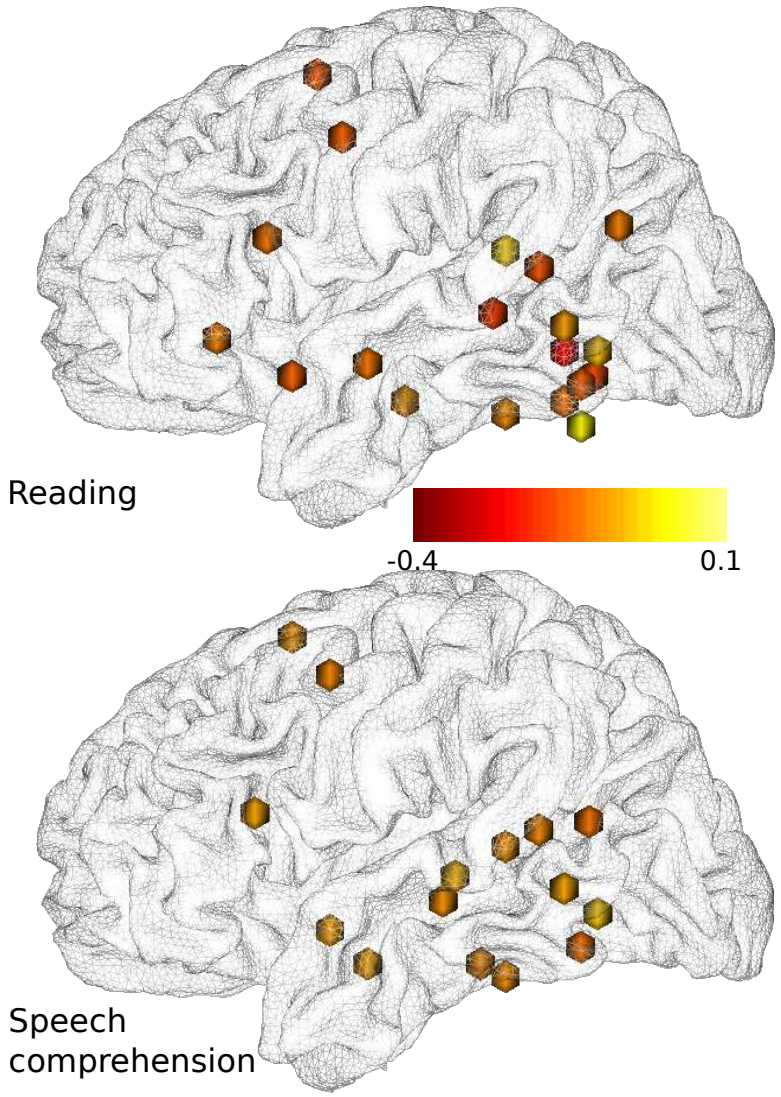


Figure 6.9 – *Location of the 19 phenotypes extracted from the "reading" contrast map and the 15 phenotypes extracted from the "speech comprehension" contrast map. The weights assigned by sPLS to the phenotypes are illustrated according to the colourbar. (The signal that appears outside of the cortical surface belongs to the cerebellum.)*

6.5 Discussion

6.5.1 Performance of the two-step method fsPLS

The originality of this work is to investigate a two-step approach combining univariate filtering with sparse PLS and to show that it performs much better than the other regularisation or dimension reduction strategies combined with PLS or KCCA on both simulated and real high-dimensional imaging genetics data. Indeed even though sparse PLS performs better than PLS and (regularised) KCCA when the dimensionality increases, it does not seem able to overcome the overfitting issue by itself, which suggests that a first step of dimension reduction is also necessary. Univariate filtering appears to be the best solution, especially when combined with

sPLS, while PC-based methods fail in that respect. Moreover, our results on the experimental dataset show that fsPLS was sensitive enough to uncover a generalisable and significant multivariate link between genetic and neuroimaging data.

6.5.2 Influence of the parameters of univariate filtering and L1 regularisation

We performed a sensitivity analysis in order to assess the influence of the parameters of univariate filtering and sPLS selection on the generalisability of the link found by fsPLS between the two blocks of data, which explains why we repeated the cross-validation procedure for all pairs of parameters. We could also have tried to add a nested CV loop in order to select, at each fold of our external 10-fold CV, the best pair of model parameters (filtering and sPLS selection rate) corresponding to that fold. The role of the external 10-fold CV would then become the assessment of the generalisability of the whole procedure: fsPLS and parameter selection. But because of the computational load of such a procedure, we did not assess its significance by permutations.

Our main results on the experimental dataset show that fsPLS extracted the most generalisable and significant neuroimaging/genetics link when considering 1000 SNPs after filtering and 5% of these SNPs selected by sPLS. The intersection between the 50 best SNPs after the univariate ranking step and of the 50 SNPs finally selected by fsPLS is of 6 SNPs. Those results as well as those obtained on simulated data raise the question of the relative contribution of the univariate filtering and the sparsity constraint to select relevant features. A relatively large number of SNPs kept after filtering seems to be required, up to a trade-off between the numbers of true and false positives, to allow sPLS to extract a robust association between a multivariate pattern of SNPs and a multivariate neuroimaging pattern. However, univariate filtering appears to be a mandatory step to filter out the vast majority of irrelevant features. Indeed, the results on experimental datasets demonstrated that a looser threshold on filtering (more than 1000 SNPs) always leads to an overfitting behaviour of PLS regardless of sparsity (see Table 6.3).

Another reason to perform univariate filtering is that PLS and even sparse PLS are too sensitive to a large number of irrelevant features, as they try to explain the variance of each block while they try to find some link between the blocks. Indeed, let us remind the criterion that is maximised by PLS regression: $\max_{\|\mathbf{a}\|_2=\|\mathbf{b}\|_2=1} \underbrace{\mathrm{corr}(\mathbf{Xa},\mathbf{Yb})}_{\text{Inter-block corr}} \underbrace{\sqrt{\mathrm{var}(\mathbf{Xa})}}_{\text{Intra-block stdev}} \underbrace{\sqrt{\mathrm{var}(\mathbf{Yb})}}_{\text{Intra-block stdev}}$, where the first term is the inter-block correlation between the two latent variables of each block and the two last terms the intra-block standard deviations of the latent variable of each block. In the case of very large blocks, the two terms of intra-block standard deviations weigh too much compared to the term of inter-block correlation, as discussed by Tenenhaus and Tenenhaus (2011). For the same reason, the non-stationarity of LD may also have an influence on the selection performed by sPLS. Indeed, sparse PLS tends to select large blocks of correlated features, such as the SNPs from a large LD

block, even though they are not strongly associated with the phenotypes, to favour the genetic intra-block standard deviation term at the expense of the correlation term. Univariate filtering helps to solve these problems by reducing the number of SNPs and partially breaking the LD structure, since it selects only the SNPs that are the most correlated to the imaging phenotypes.

6.5.3 Potential limitations of fsPLS

However, although common practice in genome wide association studies, univariate tests may not be the best filter and it could be interesting to consider multivariate filters that account for specific interactions between potential predictors (e.g., for a review see Díaz-Uriarte and Alvarez de Andrés 2006). For instance a limitation of univariate filtering may be that it filters out suppressor variables. Indeed such variables are useful to remove the non-specific variability of the relevant SNPs, improving their predictive power, while being themselves not correlated (and thus not detectable) with imaging phenotypes.

As for penalisation, even though it is well-known that it plays an important role when trying to uncover specific relationships among high-dimensional data, the choice of the penalisation is also important. For instance in the classical regression framework, an L1, L2 or L1-L2 (elastic net) penalisation scheme does not give rise to the same results when data are correlated. Indeed, in the case of correlated variables grouping into a few clusters, L1 penalisation tends to select one "representative" variable of each cluster, which facilitates the interpretation of the results but may lead to an unstable solution, whereas L2 penalisation and the elastic net criterion tend to give similar weights to the whole set of correlated predictors. In our case, we observed that the introduction of an L1 penalisation in PLS had the same effects as an Elastic Net regularisation, due to the built-in extreme L2-regularisation of PLS. In Figure 6.6, one may observe the correlation matrix of the 1000 best ranked SNPs ordered according to their position along the genome and the weights of the 5% of these SNPs that are selected by sPLS are shown in blue. It shows indeed that L1 penalisation cannot offset the PLS tendency to explain the variance of the SNPs and thus often leads to the selection, with similar weights, of several SNPs from the same block (dark red blocks) that are spatially correlated due to linkage disequilibrium. Similarly, in Figure 6.7, we plotted the correlation matrix of the 34 phenotypes. We may notice that there exists a structure of correlation between the variables obtained from the "reading" contrast (the last 19 variables of the matrix) which happen to be the variables that got the largest weights. One could investigate more sophisticated penalisations that take into account the correlation structure of the data, like in group-sparse multitask regression (Wang et al. 2012a).

Another limitation of our method may be that on the experimental dataset it could not distinguish between different pairs of covarying sub-networks on the first pair of PLS components. Even on further dimensions, subtle sub-networks were not visible in such high-dimensional settings.

Moreover, it should be noted that some non-linear effects of the number of minor alleles may also be missed by fsPLS, with the additive ge-

netic coding that we used. A different genetic coding, such as dominant/recessive or genotypic coding, could be investigated in further work.

6.6 Conclusion

To conclude, in this study, we investigated a two-step method combining univariate filtering and sparse PLS, called fsPLS, and we showed that it performed much better than other regularisation or dimension reduction strategies combined with PLS or KCCA, on both simulated and real high-dimensional imaging genetics data. Moreover, on the experimental dataset, it allowed us to detect a significant link between a set of SNPs and a functional brain network activated during a reading task, in a whole genome analysis framework. This suggests that individual variability in the genome contains predictors of the observed variability in brain activation during language tasks. We showed that we could generalise our model on left out subjects, and that this two-step multivariate technique is useful to select associated SNPs that may not be detected by a univariate screening only. However the interpretation of the results is still a very difficult issue and the neuroscientific relevance of these findings should be investigated in further research. As for the fsPLS method itself, more elaborated filtering rules and more sophisticated types of penalisation should also be investigated, which could hopefully help for the interpretation of the results.

This work has been published in Le Floch et al. (2012c;b;a).

Conclusion and perspectives

The purpose of this work was to find methods that account for the potential multivariate nature of imaging genetics data, either locally or at a longer range, while facing the very high dimensionality of the problem and increasing the detection power compared with the classical massive univariate approach.

1D SNP clusters and 4D clusters

Our first contribution was to improve sensitivity of the univariate approach by taking advantage of the multivariate nature of SNP data, in a local way, looking for 1D clusters of adjacent SNPs associated with the same imaging phenotype.

The originality of the 1D cluster approach was to adapt cluster-level inference techniques from neuroimaging to SNP data, using SNP-wise statistics combined with a minP correction to deal with LD non-stationarity along the genome. We applied this approach on both simulated and real data. The results are very preliminary but the proposed method seems promising in order to improve the sensitivity of genetic association studies while controlling for type I error.

Then, we pushed further the concept of clusters and we combined voxel clusters and SNP clusters, by using a simple 4D cluster test that detects conjointly brain and genome regions with high associations. We calibrated the test by permutations. This test showed greater sensitivity than the SNP-wise voxel-wise association test or the 3D cluster test on a real dataset. We hope that it will improve the detection power of imaging genetics studies in even larger datasets.

Limitations and extensions

However, these two approaches encounter a few limitations, which are similar to those occurring with cluster-level inference in neuroimaging.

Choice of the threshold and localisation of the activation First, the choice of the cluster-forming threshold remains critical, leading to different results.

Second, the 4D cluster-size test provides a statistic at the cluster level, such that even if a cluster is significant, one cannot reject the null hypothesis for any specific voxel and SNP. The interpretation should be that in this brain area and this set of contiguous SNPs, there is an abnormally strong imaging genetics association.

A possible extension of this work would be to investigate the strategy recently proposed by Smith and Nichols (2009) which deals with these two issues, by using a threshold free procedure that leads to enhanced voxel-wise statistics, but as mentioned earlier there may be pivotality issues associated with this technique that still have to be investigated to ensure an accurate specificity control.

Size of the morphology element Moreover, the choice of the size of the morphology element for 1D clusters is critical as well, similarly to the degree of smoothing in neuroimaging. It should be also noted that the size of the morphology element is expressed as a number of SNPs and not as a real biologic distance such as the number of base pairs or the genetic distance (in cM). It would be worth investigating the use of a morphology element expressed in cM or as a number of base pairs.

Non-stationarity and subset pivotality Another limitation of 4D cluster size inference is related to the non-stationarity of both the image and the LD along the genome. Indeed, if it is well-known that smooth brain regions will get a higher sensitivity than rough regions, we noticed as well in our results that more clusters are detected in high LD areas.

Conversely, the 1D SNP cluster approach deals with LD non-stationarity, using a SNP-wise cluster-size statistic and minP correction. Unfortunately dealing with non-stationarity often induce subset pivotality issues, since the statistic at one SNP depends on the statistics of the neighbouring SNPs. Therefore, as mentioned in section 4.2.4, minP correction may not be valid anymore in that case.

Computational load Finally, these two approaches are very computationally intensive, especially the 1D SNP cluster approach because of the minP correction, which requires more permutations than the maxT strategy in order to precisely assess uncorrected p-values before correction.

Other possible extensions Other extensions of these tests would be interesting to pursue. For instance, one could try to combine the intensity of the association with the cluster size. Strategies equivalent to those already proposed in the neuroimaging literature to test for both the spatial extent and the average intensity or maximum value (Poline et al. 1997) within the cluster may be also developed in the context of imaging genetics studies.

Because strict family wise correction may lead to very high thresholds and very low sensitivity in datasets with large voxel and SNP regions, it would be worth extending this work to other less stringent procedures such as False Discovery Rate (FDR) which may also be easily applied on the 4D imaging genetics data (Nichols 2006). Note that even with a less stringent FDR procedure, the detection across one million SNPs is likely to have a poor sensitivity.

Finally, these methods could be applied to any kind of quantitative phenotype such as neuroimaging data and even easily extended to qualitative phenotypes.

Interpretation of the results In this work, we did not study the implications of our findings in the cognitive neuroscience field. However, it is known that language asymmetry is a heritable phenotype, and that the DYX2 and DYX5 genetic regions are likely to be involved in the phenotype we observe, such that the findings obtained here are likely to be interpretable.

However, we need to be cautious with the interpretation of the results, since these approaches may detect direct associations (possibly due to one or several causal SNPs) but also indirect ones due to correlations between SNPs within a LD block.

Sparse Partial Least Squares regression combined with preliminary dimension reduction

Our second contribution was to investigate the use of two-block multivariate methods, namely Partial Least Squares regression and Canonical Correlation Analysis, to increase the detection power of imaging genetics studies, by accounting for the potential multivariate nature of the association, at a longer range, on both the imaging and the genetics sides. In order to face the very high dimensionality of the problem, the originality of this work was to compare different strategies combining both regularisation and preliminary dimension reduction, with PLS or CCA, on simulated and real imaging genetics datasets.

Performance of the two-step method fsPLS

We showed that the two-step approach called fsPLS, combining univariate filtering with sparse PLS, performed much better than the other multivariate strategies on both simulated and real data.

Indeed on the simulated dataset, even though sparse PLS performed better than PLS and (regularised) CCA when the dimension increased, it could not overcome the overfitting issue by itself, which suggested that a first step of dimension reduction was also necessary. Univariate filtering appeared to be the best solution, especially when combined with sPLS, while methods based on PCA failed in that respect.

Moreover, our results on the experimental dataset showed that fsPLS was sensitive enough to uncover a generalisable and significant multivariate link between genetic and neuroimaging data, while univariate screening only could not detect any significant SNP/imaging phenotype association.

Influence of the parameters of univariate filtering and L1 regularisation

We also investigated the influence of the parameters of univariate filtering and L1 regularisation on the generalisability of the link found by fsPLS between the two blocks of data.

We observed on the experimental dataset that fsPLS extracted the most generalisable and significant neuroimaging/genetics link when considering 1000 SNPs after filtering and 50 of these SNPs selected by sPLS. Those

results as well as those obtained on simulated data raise the question of the relative contribution of the univariate filtering and the sparsity constraint to select relevant features. A relatively large number of SNPs kept after filtering seems to be required, up to a trade-off between the numbers of true and false positives, to allow sPLS to extract a robust association between a multivariate pattern of SNPs and a multivariate neuroimaging pattern. However, the results on both simulated and experimental datasets demonstrated that a too loose threshold on filtering leads to a important increase of the overfitting behaviour of PLS regardless of sparsity.

Another reason to perform univariate filtering is related to the fact that PLS and even sparse PLS try to explain the variance of each block of data, while looking for a correlation between the two blocks. Consequently, in very high-dimensional settings such as imaging genetics, the two intra-block standard deviation terms weigh too much compared to the inter-block correlation term. Moreover, PLS and sPLS may be influenced by the non-stationarity of LD, and tend to give higher weights (and to select in the case of sPLS) large locks of LD. Univariate filtering helps to solve these problems by reducing the number of SNPs and partially breaking the LD structure, since it selects only the SNPs that are the most correlated to the imaging phenotypes. The link between the optimal threshold and the sampling density of LD should be further studied.

Thus, even though it may seem to contradict with the multivariate nature of methods such as PLS, univariate filtering seems necessary to allow PLS to overcome the overfitting issue and to extract a generalisable and significant imaging genetics link. In fact, it would be interesting to investigate the influence of univariate filtering on other multivariate approaches. For instance, one could try to combine univariate filtering and multitask regression or parallel ICA, and to compare them with our fsPLS approach.

Potential limitations of fsPLS

However, univariate filtering may not be the best dimension reduction technique and it could be compared in further work with multivariate filters that account for potential interactions between predictors (e.g., for a review see Díaz-Uriarte and Alvarez de Andrés 2006).

The choice of the type of regularisation is critical as well. We have seen that sparse PLS is similar to an Elastic Net penalty. One could investigate more sophisticated penalisations that would incorporate prior knowledge on the correlation structure of the data (such as linkage disequilibrium or functional connectivity) or on biologically meaningful groups of variables, such as sets of genes belonging to the same pathway or anatomically connected brains regions. For instance it would be interesting to compare sparse PLS with group-sparse multitask regression (Wang et al. 2012a).

As mentioned earlier in section 5.4.2, another limitation of such multivariate methods is that they do not provide any variable-wise degree of significance or any explicit control for false positives. In further work, the robustness of the selection could be interestingly investigated instead, using bootstrap techniques for instance.

Moreover, the additive genetic coding that we used may not be the

most appropriate to capture some non-linear effects of the number of minor alleles. A different genetic coding, such as dominant/recessive or genotypic coding, should be investigated in further work.

Interpretation of the results

On the experimental dataset, fsPLS allowed us to detect a significant link between a set of SNPs and a functional brain network activated during a reading task, in a whole genome analysis framework. This suggests that individual variability in the genome contains predictors of the observed variability in brain activation during language tasks. We showed that we could generalise our model on left-out subjects. However, it would be very interesting to replicate these results on an independent and larger dataset.

Moreover, the interpretation of the results is still a very difficult issue and the neuroscientific relevance of these findings should be investigated in further research. As for the fsPLS method itself, more elaborated filtering rules and more sophisticated types of penalisation should also be investigated, which could hopefully help for the interpretation of the results.

Finally, the ultimate goal of this work would be to relate imaging genetics results to final phenotypes such as cognitive scores or clinical phenotypes. For instance, methods designed for multi-block analysis, such as Regularised Generalized Canonical Correlation Analysis (Tenenhaus and Tenenhaus 2011), could be interestingly used to investigate the relationships between these three types of data .

Publications

Le Floch E., Guillemot V., Lalanne C., Frouin V., Pinel P., Trinchera L., Tenenhaus A., Moreno A., Zilbovicius M., Bourgeron T., Dehaene S., Thirion B., Poline J.B. and Duchesnay E., Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. *NeuroImage*, 63: 11-24.

Le Floch E., Trinchera L., Tenenhaus A., Poline J.B., Frouin V. and Duchesnay E., Dimension reduction and regularisation combined with Partial Least Squares in high dimensional Imaging Genetics studies. *International Conference on Partial Least Squares and Related Methods 2012, Houston*.

Le Floch E., Pinel P., Tenenhaus A., Trinchera L., Poline J.B., Frouin and Duchesnay E., Discovering associations in high dimensional imaging-genetics data: a comparison study of dimension reduction and regularisation strategies combined with Partial Least Squares. *ISBI 2012 (International Symposium on Biomedical Imaging), Barcelona*.

Le Floch E., Pinel P., Thirion B., Poline J.B., Frouin V. and Duchesnay E., A multivariate investigation in high-dimensional Imaging Genetics studies using sparse Partial Least Squares and dimension reduction. *International Imaging Genetics Conference 2012, Irvine*.

Fouque A.L., Fillard P., Bargiacchi A., Cachia A., Zilbovicius M., Thyreau B., Le Floch E., Ciuciu P. and Duchesnay E., Voxelwise Multivariate Statistics and Brain-Wide Machine Learning Using the Full Diffusion Tensor. *MICCAI 2011 (Medical Image Computing and Computer Assisted Intervention), Toronto*.

Trinchera L., Le Floch E. and Tenenhaus A., Variable Selection via Correlated Component PLS-type Regression. *CLADAG 2011 (CLAssification and Data Analysis Group of the Italian Statistical Society), Pavia (Italy)*.

Le Floch E., Lalanne C., Pinel P., Moreno A., Trinchera L., Tenenhaus A., Thirion B., Poline J.B., Zilbovicius M., Frouin V. and Duchesnay E., Bridging the gap between imaging and genetics : a multivariate statistical investigation. *Oral presentation at Human Brain Mapping 2011, Québec city*.

Trinchera L., Le Floch E. and Tenenhaus A., Variable Selection in Partial Least Squares Methods: overview and recent developments. *ISBIS 2010 (International Symposium on Business and Industrial Statistics), Portoroz (Slovenia)*.

Le Floch E., Lalanne C., Pinel P., Moreno A., Trinchera L., Tenenhaus A., Thirion B., Poline J.B., Frouin V. and Duchesnay E., Cluster-level Infer-

ence and Resampling-Based Multiple Testing applied to Imaging Genetics Studies. *Human Brain Mapping 2010, Barcelona*.

Pinel P., Bourgeron T., Moreno A., Le Floch E., Fauchereau F., Barbot A., Le Bihan D., Poline J.B. and Dehaene S., From functional brain mapping to genetic mapping: genetic determinants of reading and sentence comprehension. *Wiring the brain : from genetic to neuronal networks 2009, Adare, County Limerick (Ireland)*.

Appendices

Connection of RRR with PLS-SVD and PLS regression

A

Multivariate Reduced-Rank Regression (RRR) (Reinsel and Velu 1998) consists in transforming the classical multivariate multiple linear regression model of a $n \times q$ response matrix $\mathbf{Y}$ on a $n \times p$ design matrix $\mathbf{X}$, by imposing a rank $R \leq min(p, q)$ on regression coefficients and taking into account the multivariate nature of the response matrix. The criterion optimised by multivariate RRR is:

$$\widehat{\mathbf{A}}, \widehat{\mathbf{B}} = \underset{\mathbf{A},\mathbf{B}}{\arg\min} \operatorname{Tr} \left\{ (\mathbf{Y} - \mathbf{XBA})\, \mathbf{\Gamma}\, (\mathbf{Y} - \mathbf{XBA})' \right\} \tag{A.1}$$

where regression coefficients are decomposed into a matrix $\mathbf{B}$ with R linearly independent columns and a matrix $\mathbf{A}$ with R linearly independent rows. $\mathbf{\Gamma}$ is a weight matrix, commonly set to be the identity matrix. The solutions for $\mathbf{B}$ and $\mathbf{A}$ are derived from the Singular Value Decomposition (SVD) of the matrix $(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\mathbf{X}'\mathbf{Y}\mathbf{\Gamma}^{\frac{1}{2}}$.

In the implementation of sparse (multivariate) RRR by Vounou et al. (2010), $\mathbf{\Gamma}$ and $\mathbf{X}'\mathbf{X}$ are approximated by identity matrices because of the very high dimensionality of the data, which makes RRR equivalent to PLS-SVD (described in section 2.2.3). Please note that the notations $\mathbf{A}$ and $\mathbf{B}$ are inverted in PLS-SVD.

However, instead of performing an SVD, they recast the PLS-SVD problem into an iterative procedure using the NIPALS algorithm, in order to apply L1-penalisation on weight vectors $\mathbf{a}$ and $\mathbf{b}$ for each rank, by using soft-thresholding within the NIPALS inner-loop. Indeed, for the rank-one model, the criterion optimised becomes:

$$\widehat{\mathbf{a}}, \widehat{\mathbf{b}} = \underset{\mathbf{a},\mathbf{b}}{\arg\min} -2\mathbf{a}\mathbf{Y}'\mathbf{X}\mathbf{b} + \mathbf{a}\mathbf{a}'\mathbf{b}'\mathbf{b} + \lambda_1 \parallel \mathbf{b} \parallel_1 + \lambda_2 \parallel \mathbf{a}' \parallel_1 \tag{A.2}$$

They obtain further ranks by optimising the same criterion on the residuals of the matrices $\mathbf{X}$ and $\mathbf{Y}$ after regression on their own latent variables. However, this replaces the PLS-SVD orthogonality constraints on coefficients $\hat{\mathbf{A}}_r$ and $\hat{\mathbf{B}}_r$ (see section 2.2.3) by the orthogonality between the latent variables within each block $\mathbf{X}\hat{\mathbf{B}}_r$ and $\mathbf{Y}\hat{\mathbf{A}}_r$, since both sparsity and orthogonality of the coefficients cannot be easily obtained at the same time (Zou et al. 2006). Thus, the sparse RRR algorithm departs also from the PLS-SVD framework and becomes equivalent to the sparse version of PLS regression in its canonical mode, described in section 5.2.2.

B

Proof of the equivalence between L1-regularised PLS and soft-thresholding

If one considers the optimisation of the Lagrangian function corresponding to equation 5.10, computes its partial derivative with respect to $\mathbf{u}$ and searches when it equals zero, it leads to the soft-thresholding solution. Indeed the Lagrangian function corresponding to equation 5.10 is:

$$\mathcal{L} = \mathbf{a}'\mathbf{X}'\mathbf{Y}\mathbf{b} - \lambda_{1X}\|\mathbf{a}\|_1 - \lambda_{2X}(\|\mathbf{a}\|_2^2 - 1) \tag{B.1}$$

$$= \mathbf{a}'\mathbf{X}'\mathbf{Y}\mathbf{b} - \lambda_{1X}\mathbf{a}'\text{sign}(\mathbf{a}) - \lambda_{2X}(\mathbf{a}'\mathbf{a} - 1) \tag{B.2}$$

Then we compute the partial derivative of the Lagrangian function with respect to $\mathbf{a}$ and search when it equals zero:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}} = \mathbf{X}'\mathbf{Y}\mathbf{b} - \lambda_{1X}\text{sign}(\mathbf{a}) - 2\lambda_{2X}\mathbf{a} = 0 \tag{B.3}$$

This leads to the following solution for each coefficient a_i of vector $\mathbf{a}$:

$$a_i = \begin{cases} \dfrac{\mathbf{X}_i'\mathbf{Y}\mathbf{b} - \lambda_{1X}}{2\lambda_{2X}} \text{ if } \mathbf{X}_i'\mathbf{Y}\mathbf{b} - \lambda_{1X} > 0 \\ \dfrac{\mathbf{X}_i'\mathbf{Y}\mathbf{b} + \lambda_{1X}}{2\lambda_{2X}} \text{ if } \mathbf{X}_i'\mathbf{Y}\mathbf{b} + \lambda_{1X} < 0 \\ 0 \text{ otherwise} \end{cases} \tag{B.4}$$

where $\mathbf{X}_i$ is the i-th column of $\mathbf{X}$.
Finally for each coefficient a_i of vector $\mathbf{a}$:

$$a_i = \frac{\text{sign}(\mathbf{X}_i'\mathbf{Y}\mathbf{b})\ \max(0, |\mathbf{X}_i'\mathbf{Y}\mathbf{b}| - \lambda_{1X})}{2\lambda_{2X}} = \frac{g_{\lambda_{1X}}(\mathbf{X}_i'\mathbf{Y}\mathbf{b})}{2\lambda_{2X}} \tag{B.5}$$

where λ_{2X} will be chosen such that $\|\mathbf{a}\|_2 = 1$.
The same demonstration holds for the calculation of $\mathbf{b}$.

References

Bibliography

T.W. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*, 22: 327–351, 1951. (Cited on pages 14 and 73.)

A. Argyriou, T. Evgeniou, , and M Pontil. Multi-task feature learning. *Advances in Neural Information Processing Systems*, page 41–48, 2007. (Cited on pages 15 and 76.)

F. Bach. Bolasso : Model consistent lasso estimation through the bootstrap. *In Proceedings of the 25th International Conference on Machine Learning, number 2004*, 2008. (Cited on page 117.)

J.C. Barrett, B. Fry, J. Maller, and M.J. Daly. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2):263–265, 2005. (Cited on pages 99 and 101.)

A.J. Bell and T.J. Sejnowski. An information maximisation approach to blind separation and blind deconvolution. *Neural Comput.*, 7(6): 1129–1159, 1995. (Cited on page 66.)

V. Calhoun, P. Maciejewski, G. Pearlson, and K. Kiehl. Temporal lobe and default hemodynamic brain modes discriminate between schizophrenia and bipolar disorder. *Human Brain Mapping*, 29(11):1265–1275, 2008. (Cited on page 66.)

V.D. Calhoun, T. Adali, G.D. Pearlson, and J.J. Pekar. A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping*, 14(3):140–151, 2001. (Cited on page 66.)

V.D. Calhoun, J. Liu, and T. Adali. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage*, 45(1(Suppl 1)):S163–S172, 2009. (Cited on page 67.)

M.K. Carroll, G.A. Cecchi, I. Rish, R. Garg, and A. Ravishankar Rao. Prediction and interpretation of distributed neural activity with sparse models. *NeuroImage*, 44(1):112–122, 2009. (Cited on pages 11 and 63.)

H. Chun and S. Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society - Serie B*, 72 (1):3–25, 2010. (Cited on pages 107 and 109.)

D. Clayton and H.-T. Cheung. An R package for analysis of whole-genome association studies. *Human Heredity*, 64:45–51, 2007. (Cited on page 91.)

N. Cope, D. Harold, G. Hill, V. Moskvina, J. Stevenson, P. Holmans, M.J. Owen, M.C. OĎonovan, and J. Williams. Strong evidence that KIAA0319 on chromosome 6p is a susceptibility gene for developmental dyslexia. *American Journal of Human Genetics*, 76(4):581–591, 2005. (Cited on page 92.)

D.D. Cox and R.L. Savoy. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, 19(2):261–270, 2003. (Cited on pages 64 and 68.)

P.I.W. de Bakker, R. Yelensky, I. Pe'er, S.B. Gabriel, M.J. Daly, and D. Altshuler. Efficiency and power in genetic association studies. *Nature Genetics*, 37(11):1217–1223, 2005. (Cited on page 123.)

F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage*, 43(1):44–58, 2008. (Cited on page 68.)

R. Díaz-Uriarte and S. Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(3), 2006. (Cited on pages 35, 139, and 144.)

S. Dudoit, M. J. van der Laan, and K. S. Pollard. Multiple testing. part I. single-step procedures for control of general type I error rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 13, 2004. (Cited on pages 17 and 90.)

B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1):1–26, 1979. (Cited on page 116.)

B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Applied Statistics*, 32(2):407–451, 2004. (Cited on pages 61 and 62.)

K.A. Frazer, S.S. Murray, N.J. Schork, and E.J. Topol. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10: 241–251, 2009. (Cited on pages 2 and 37.)

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007. (Cited on pages 61 and 62.)

K. Friston, A. Holmes, K. Worsley, J.-B. Poline, C. Frith, and R. Frackowiak. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4):189–210, 1994. (Cited on page 46.)

K.J. Friston, K.J. Worsley, R.S.J. Frackowiak, and A.C. Evans J.C. Mazziotta. Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1(3):210–220, 1993. (Cited on pages 16 and 87.)

C. Furlanello, M. Serafini, S. Merler, and G. Jurman. An accelerated procedure for recursive feature ranking on microarray data. *Neural Networks*, 16:641–648, 2003. (Cited on pages 1 and 37.)

Y. Ge, S. Dudoit, and T. Speed. Resampling-based multiple testing for microarray data analysis. *Sociedad de Estadistica e Investigacion Operativa Test*, 12(1):1–77, 2003. (Cited on pages 32 and 88.)

C. Giessing, G.R. Fink, F. Rösler, and C.M. Thiel. fMRI data predict individual differences of behavioral effects of nicotine: A partial least square analysis. *Journal of Cognitive Neuroscience*, 19(4):658–670, 2007. (Cited on page 63.)

D. C. Glahn, P. M. Thompson, and J. Blangero. Neuroimaging endophenotypes : Strategies for finding genes influencing brain structure and function. *Human Brain Mapping*, 28:488–501, 2007. (Cited on pages 1 and 37.)

I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, editors. *Feature Extraction: Foundations And Applications*. Springer-Verlag, 2006. (Cited on page 67.)

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3): 389–422, 2002. (Cited on page 67.)

Y. Hakak, J.R. Walker, C. Li, W. Hung Wong, K.L. Davis, J.D. Buxbaum, V. Haroutunian, and A.A. Fienberg. Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8):4746–4751, 2001. (Cited on page 64.)

A.R. Hariri, V.S. Mattay, A. Tessitore, B. Kolachana, F. Fera, D. Goldman, M.F. Egan, and D.R. Weinberger. Serotonin transporter genetic variation and the response of the human amygdala. *Science*, 297:400–403, 2002. (Cited on page 45.)

T. Hastie, R. Tibshirani, M.B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W.C. Wing C Chan, D. Botstein, and P. Brown. 'gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, 1(2):research0003.1–0003.21, 2000. (Cited on page 66.)

A.C. Haury, P. Gestraud, and J.P. Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE*, 6(12):e28210, 2011. (Cited on page 68.)

S. Hayasaka, K. Luan Phan, I. Liberzon, K.J. Worsley, and T.E. Nichols. Nonstationary cluster-size inference with random field and permutation methods. *NeuroImage*, 22(2):676–687, 2004. (Cited on pages 88 and 101.)

D.P. Hibar, J.L. Stein, O. Kohannim, N. Jahanshad, A.J. Saykin, L. Shen, S. Kim, N. Pankratz, T. Foroud, M.J. Huentelman, S.G. Potkin, C.R. Jack Jr., M.W. Weiner, A.W. Toga, P.M. Thompson, and the Alzheimer's Disease Neuroimaging Initiative. Voxelwise gene-wide association

study (vGeneWAS): Multivariate gene-based association testing in 731 elderly subjects. *NeuroImage*, 56:1875–1891, 2011. (Cited on pages 2, 14, 38, and 72.)

A.E Hoerl and R.W. Kennard. Ridge regression : biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. (Cited on pages 11 and 60.)

J. Hoh and J. Ott. Scan statistics to scan markers for susceptibility genes. *Proceedings of the National Academy of Sciences*, 97:9615–9617, 2000. (Cited on pages 17 and 89.)

A.P. Holmes, R.C. Blair, J.D.G. Watson, and I. Ford. Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow and Metabolism*, 16(1):7–22, 1996. (Cited on pages 16 and 87.)

B.D. Horne and N.J. Camp. Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. *Genetic Epidemiology*, 26(1):11–21, 2004. (Cited on page 66.)

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936. (Cited on page 107.)

J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, and P.S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7(6):673–679, 2001. (Cited on page 66.)

O. Kohannim, D.P. Hibar, J.L. Stein, N. Jahanshad, C.R. Jr. Jack, M.W. Weiner, A.W. Toga, P.M. Thompson, and the Alzheimer's Disease Neuroimaging Initiative. Boosting power to detect genetic associations in imaging using multi-locus, genome-wide scans and ridge regression. *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pages 1855–1859, 2011. (Cited on pages 2, 14, 38, and 72.)

J.W. Koten Jr., G. Wood, P. Hagoort, R. Goebel, P. Propping, K. Willmes, and D. I. Boomsma. Genetic contribution to variation in cognitive function: An fMRI study in twins. *Science*, 323:1737–1740, 2009. (Cited on page 45.)

K.-A. Lê Cao, P. G. Martin, C. Robert-Granié, and P. Besse. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, 10(34), 2009. (Cited on page 107.)

K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse. A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008. (Cited on pages 23, 107, and 111.)

E. Le Floch, V. Guillemot, C. Lalanne, V. Frouin, P. Pinel, L. Trinchera, A. Tenenhaus, A. Moreno, M. Zilbovicius, T. Bourgeron, S. Dehaene, B. Thirion, J.B. Poline, and E. Duchesnay. Significant correlation between

a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse partial least squares. *NeuroImage*, 63: 11–24, 2012a. (Cited on page 140.)

E. Le Floch, C. Lalanne, P. Pinel, A. Moreno, L. Trinchera, A. Tenenhaus, J.B. Poline, V. Frouin, and E. Duchesnay. Cluster-level inference and resampling-based multiple testing applied to imaging genetics studies. *Proceedings of Human Brain Mapping 2010, Barcelona*, 2010. (Cited on page 97.)

E. Le Floch, P. Pinel, A. Tenenhaus, L. Trinchera, J.B. Poline, V. Frouin, and E. Duchesnay. Discovering associations in high dimensional imaging-genetics data: a comparison study of dimension reduction and regularisation strategies combined with partial least squares. *Proceedings of ISBI 2012 (International Symposium on Biomedical Imaging), Barcelona*, 2012b. (Cited on page 140.)

E. Le Floch, L. Trinchera, A. Tenenhaus, J.B. Poline, V. Frouin, and E. Duchesnay. Dimension reduction and regularisation combined with partial least squares in high dimensional imaging genetics studies. *Proceedings of International Conference on Partial Least Squares and Related Methods 2012, Houston*, 2012c. (Cited on page 140.)

S.E. Leurgans, R.A. Moyeed, and B.W. Silverman. Canonical correlation analysis when the data are curves. *Journal of The Royal Statistical Society, Series B*, 55:725–740, 1993. (Cited on pages 23 and 110.)

J. Li and Y. Chen. Generating samples for association studies based on HapMap data. *BMC Bioinformatics*, 9(44), 2008. (Cited on page 122.)

Y. Li, T. Adali, and V.D. Calhoun. Estimating the number of independent components for functional magnetic resonance imaging data. *Human Brain Mapping*, 28(11):1251–1266, 2007b. (Cited on page 78.)

Y. Liang and A. Kelemen. Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex diseases. *Statistics Surveys*, 2:43–60, 2008. (Cited on pages 17 and 89.)

H. Liu, R. Kustra, and J. Zhang. A novel dimensionality reduction technique based on independent component analysis for modeling microarray gene expression data. *International Conference on Artificial Intelligence*, 2004. (Cited on page 66.)

J. Liu, G. Pearlson, A. Windemuth, G. Ruano, N.I. Perrone-Bizzozero, and V. Calhoun. Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. *Human Brain Mapping*, 30(1):241–255, 2009. (Cited on pages 2, 15, 38, 77, and 78.)

T. W. McAllister, L. A. Flashman, B. C. McDonald, and A. J. Saykin. Mechanisms of cognitive dysfunction after mild and moderate TBI (MTBI): Evidence from functional MRI and neurogenetics. *Journal of Neurotrauma*, 23(10):1450–1467, 2006. (Cited on pages 1 and 37.)

A.R. McIntosh, F.L. Bookstein, J.V. Haxby, and C.L. Grady. Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage*, 3:143–157, 1996. (Cited on pages 12 and 64.)

S.A. Meda, B. Narayanan, J. Liu, N.I. Perrone-Bizzozero, M.C. Stevens, V.D. Calhoun, D.C. Glahn, L. Shen, S.L. Risacher, A.J. Saykin, and G.D. Pearlson. A large scale multivariate parallel ICA method reveals novel imaging–genetic relationships for alzheimer's disease in the ADNI cohort. *NeuroImage*, 60(3):1608–1621, 2012. (Cited on page 78.)

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010. (Cited on page 117.)

T.M. Mitchell, R. Hutchinson, R.S. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman. Learning to decode cognitive states from brain images. *Machine Learning*, 57(1-2):145–175, 2004. (Cited on page 64.)

J.H. Moore, J.C. Gilbert, C.T. Tsai, F.T. Chiang, T. Holdena, N. Barneya, and B.C. White. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology*, 241(2): 252–261, 2006. (Cited on page 64.)

B. Neale and P. Sham. The future of association studies : Gene-based analysis and replication. *American Journal of Human Genetics*, 75:353–362, 2004. (Cited on pages 17 and 89.)

T. Nichols. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, chapter False Discovery Rate procedures, pages 246–252. Academic Press, New York, 2006. (Cited on pages 33, 102, and 142.)

J. Nopola-Hemmi, B. Myllyluoma, T. Haltia, M. Taipale, V. Ollikainen, T. Ahonen, A. Voutilainen, J. Kere, and E. Widén. A dominant gene for developmental dyslexia on chromosome 3. *Journal of Medical Genetics*, 38(10):658–664, 2001. (Cited on pages 98 and 103.)

J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A.R. Boyko, A. Auton, A. Indap, K.S. King, S. Bergmann, M.R. Nelson, M. Stephens, and C.D. Bustamante. Genes mirror geography within europe. *Nature*, 456(7218): 98–101, 2008. (Cited on page 66.)

G. Obozinski, B. Taskar, and M. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252, 2010. (Cited on pages 15 and 76.)

E. Parkhomenko, D. Tritchler, and J. Beyene. Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proceedings*, 1(Suppl 1):S119, 2007. (Cited on pages 75, 107, and 112.)

E. Parkhomenko, D. Tritchler, and J. Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1):Article 1, 2009. (Cited on pages 75, 107, and 112.)

E. Paulesu, J.-F. Demonet, F. Fazio, E. McCrory, V. Chanoine, N. Brunswick, F. Cappa, G. Cossu, M. Habib, C.D. Frith, and U. Frith. Dyslexia: Cultural diversity and biological unity. *Science*, 291:2165–2167, 2001. (Cited on pages 90 and 121.)

L. Pezawas, B.A. Verchinski, V.S. Mattay, J.H. Callicott, B.S. Kolachana, R.E. Straub, M.F. Egan, A. Meyer-Lindenberg, and D.R. Weinberger. The brain-derived neurotrophic factor val66met polymorphism and variation in human cortical morphology. *The Journal of Neuroscience*, 24(45): 10099–10102, 2004. (Cited on page 45.)

P. Pinel and S. Dehaene. Beyond hemispheric dominance: Brain regions underlying the joint lateralization of language and arithmetic to the left hemisphere. *Journal of Cognitive Neuroscience.* , 2009. Posted Online January 13, 2009. (doi:10.1162/jocn.2009.21184). (Cited on pages 90 and 121.)

P. Pinel, B. Thirion, S. Meriaux, A. Jobert, J. Serres, D. Le Bihan, J.-B. Poline, and S. Dehaene. Fast reproducible identification and large-scale databasing of individual functional cognitive networks. *BMC Neuroscience*, 8(91), 2007. (Cited on pages 90, 98, and 121.)

J.-B. Poline and B.M. Mazoyer. Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise-ratio pixel clusters. *Journal of Cerebral Blood Flow and Metabolism*, 13:425–437, 1993. (Cited on pages 16 and 87.)

J.-B. Poline, K.J. Worsley, A.C. Evans, and K.J. Friston. Combining spatial extent and peak intensity to test for activations in functional imaging. *NeuroImage*, 5:83–96, 1997. (Cited on pages 33, 102, and 142.)

J.B. Poline and B.M. Mazoyer. Analysis of individual brain activation maps using hierarchical description and multiscale detection. *IEEE Trans Med Imaging*, 13(4):702–710, 1994. (Cited on page 102.)

R Development Core Team. R: A language and environment for statistical computing. http://www.R-project.org, 2009. (Cited on page 91.)

G. Reinsel and R. Velu. *Multivariate Reduced-Rank Regression, Theory and Applications*. Springer, New York, 1998. (Cited on pages 74 and 151.)

J. L. Roffman, A. P. Weiss, D. C. Goff, S. L. Rauch, and D. R. Weinberger. Neuroimaging-genetic paradigms: a new approach to investigate the pathophysiology and treatment of cognitive deficits in schizophrenia. *Harvard Review of Psychiatry*, 14(2):78–91, 2006. (Cited on pages 1 and 37.)

P.E. Roland, B. Levin, R. Kawashima, and S. Åkerman. Three-dimensional analysis of clustered voxels in 15-o-butanol brain activation images. *Human Brain Mapping*, 1(1):3–19, 1993. (Cited on pages 16 and 87.)

S Ryali, K. Supekar, D.A. Abrams, and V. Menon. Sparse logistic regression for whole-brain classification of fMRI data. *NeuroImage*, 51(2):752–764, 2010. (Cited on pages 64 and 68.)

S. Sanna, A.U. Jackson, R. Nagaraja, C.J. Willer, W.M. Chen, L.L. Bonnycastle, H. Shen, N. Timpson, G. Lettre, G. Usala, P.S. Chines, H.M. Stringham, L.J. Scott, M. Dei, S. Lai, G. Albai, L. Crisponi, S. Naitza, K.F. Doheny, E.W. Pugh, Y. Ben-Shlomo, S. Ebrahim, D.A. Lawlor, R.N. Bergman, R.M. Watanabe, M. Uda, J. Tuomilehto, J. Coresh, J.N. Hirschhorn, A.R. Shuldiner, D. Schlessinger, F.S. Collins, G. Davey Smith, E. Boerwinkle, A. Cao, M. Boehnke, G.R. Abecasis, and K.L. Mohlke. Common variants in the GDF5-UQCC region are associated with variation in human height. *Nature Genetics*, 40:198–203, 2008. (Cited on page 91.)

M. Sarkis, K. Diepold, and F. Westad. A new algorithm for gene mapping: Application of partial least squares regression with cross model validation. *Genomic Signal Processing and Statistics, 2006. GENSIPS '06. IEEE International Workshop on*, pages 89–90, 2006. (Cited on page 63.)

G. Shi, E. Boerwinkle, A.C. Morrison, C.C. Gu, A. Chakravarti, and D.C. Rao. Mining gold dust under the genome wide significance level: a two-stage approach to analysis of GWAS. *Genetic Epidemiology*, 35(2): 111–118, 2011. (Cited on pages 11 and 63.)

S. M. Smith and T. E. Nichols. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, 44(1):83–98, 2009. (Cited on pages 32, 88, 100, and 142.)

C. Soneson, H. Lilljebjörn, T. Fioretos, and M. Fontes. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. *BMC Bioinformatics*, 11(191), 2010. (Cited on pages 66 and 107.)

Terry Speed. *Statistical Analysis of Gene Expression Microarray Data*. CRC Press Inc, 2003. (Cited on page 64.)

J. Stein, X. Hua, S. Lee, A. Ho, A. Leow, A. Toga, A. Saykin, L. Shen, T. Foroud, N. Pankratz, M. Huentelman, D. Craig, J. Gerber, A. Allen, J. Corneveaux, B. DeChairo, S. Potkin, M. Weiner, and P. Thompson. Voxelwise genome-wide association study (vGWAS). *NeuroImage*, 53: 1160–1174, 2010. (Cited on pages 1, 13, 37, 71, and 73.)

S.C Strother, J. Anderson, L. Hansen, U. Kjems, R. Kustra, J. Sidtis, S. Frutiger, S. Muley, S. LaConte, and D. Rottenberg. The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. *NeuroImage*, 15(4):747–771, 2002. (Cited on page 66.)

A. Tenenhaus and M. Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257–284, 2011. (Cited on pages 36, 138, and 145.)

P.M. Thompson, T.D. Cannon, K.L. Narr, T. van Erp, V.P. Poutanen, M. Huttunen, J. Lönnqvist, C.G. Standertskjöld-Nordenstam, J. Kaprio, M. Khaledy, R. Dail, C.I. Zoumalan, and A.W. Toga. Genetic influences

on brain structure. *Nature Neuroscience*, 4(12):1253–1258, 2001. (Cited on page 45.)

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, B*, 58(1):267–288, 1996. (Cited on pages 11, 23, 61, and 111.)

L.R. Tucker. An inter-battery method of factor analysis. *Psychometrika*, 23 (2):111–136, 1958. (Cited on pages 12 and 64.)

H.D. Vinod. Canonical ridge and econometrics of joint production. *Journal of Econometrics*, 4:147–166, 1976. (Cited on pages 23 and 110.)

M. Vounou, E. Janousova, R. Wolz, J.L Stein, P.M. Thompson, D. Rueckert, G. Montana, and the Alzheimer's Disease Neuroimaging Initiative. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *NeuroImage*, 60 (1):700–716, 2012. (Cited on pages 2, 38, and 76.)

M. Vounou, T.E. Nichols, and G. Montana. Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank approach. *NeuroImage*, 53:1147–1159, 2010. (Cited on pages 2, 14, 15, 38, 73, 75, 76, 107, and 151.)

S. Waaijenborg, P. Verselewel de Witt Hamer, and A. Zwinderman. Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, 7(1):Article 3, 2008. (Cited on pages 75, 107, and 112.)

H. Wang, F. Nie, H. Huang, S. Kim, K. Nho, S.L. Risacher, A.J. Saykin, L. Shen, and the Alzheimer's Disease Neuroimaging Initiative. Identifying quantitative trait loci via group-sparse multi-task regression and feature selection: An imaging genetics study of the ADNI cohort. *Bioinformatics*, 28:229–237, 2012a. (Cited on pages 2, 15, 35, 38, 76, 77, 139, and 144.)

H. Wang, F. Nie, H. Huang, S.L. Risacher, A.J. Saykin, L. Shen, and the Alzheimer's Disease Neuroimaging Initiative. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics*, 28(12):i127–i136, 2012b. (Cited on pages 2, 38, and 77.)

P.H. Westfall and S.S. Young, editors. *Resampling-based Multiple Testing*. Wiley, New York, 1993. (Cited on pages 17, 88, 90, and 130.)

C.J. Willer, S. Sanna, A.U. Jackson, A. Scuteri, L.L. Bonnycastle, R. Clarke, S.C. Heath, N.J. Timpson, S.S. Najjar, H.M. Stringham, J. Strait, W.L. Duren, A. Maschio, F. Busonero, A. Mulas, G. Albai, A.J. Swift, M.A. Morken, N. Narisu, D. Bennett, S. Parish, H. Shen, P. Galan, P. Meneton, S. Hercberg, D. Zelenika, W.M. Chen, Y. Li, L.J. Scott, P.A. Scheet, J. Sundvall, R.M. Watanabe, R. Nagaraja, S. Ebrahim, D.A. Lawlor, Y. Ben-Shlomo, Davey-Smith G., A.R. Shuldiner, R. Collins, R.N.

Bergman, M. Uda, J. Tuomilehto, A. Cao, F.S. Collins, E. Lakatta, G.M. Lathrop, M. Boehnke, D. Schlessinger, K.L. Mohlke, and Abecasis G.R. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genetics*, 40:161–169, 2008. (Cited on page 91.)

D.M. Witten and R. Tibshirani. Extensions of sparse canonical correlation analysis, with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1):Article 28, 2009. (Cited on pages 75, 107, and 112.)

H. Wold. *Multivariate Analysis*, chapter Estimation of principal components and related models by iterative least squares, pages 391–420. Academic Press, New York, 1966. (Cited on pages 22 and 108.)

H. Wold. Partial least squares. In *Encyclopedia of Statistical Sciences*, pages 581–591. Wiley, 1985. (Cited on page 110.)

S. Wold, H. Martens, and H. Wold. The multivariate calibration problem in chemistry solved by the PLS method. In A. Ruhe and B. Kastrøm, editors, *Proceedings Conference Matrix Pencils*, volume Lecture Notes in Mathematics, pages 286–293. Springer-Verlag, 1983. (Cited on page 107.)

K.J. Worsley. Non-stationary FWHM and its effect on statistical inference of fMRI data. *NeuroImage, 8th International Conference on Functional Mapping of the Human Brain*, 16(2):779–780, 2002. (Cited on page 88.)

K.J. Worsley, A.C. Evans, S. Marrett, and P. Neelin. A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow and Metabolism*, 12:900–918, 1992. (Cited on pages 16, 87, and 88.)

K.J. Worsley, S. Marrett, P. Neelin, and A.C. Evans. Searching scale space for activation in PET images. *Human Brain Mapping*, 4:74–90, 1996. (Cited on page 102.)

O. Yamashita, M.A. Sato, T. Yoshioka, F. Tong, and Y. Kamitani. Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *NeuroImage*, 42(4):1414–1429, 2008. (Cited on page 64.)

J. Yang, B. Benyamin, B.P. McEvoy, S. Gordon, A.K. Henders, D.R. Nyholt, P.A. Madden, A.C. Heath, N.G. Martin, G.W. Montgomery, M.E. Goddard, and P.M. Visscher. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42:565–569, 2010. (Cited on pages 2 and 37.)

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1): 49C–67, 2006. (Cited on page 77.)

J. Zhu and T. Hastie. Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3):427–443, 2004. (Cited on page 64.)

H. Zou and T. Hastie. Regression and variable selection via the elastic net. *Journal of the Royal Statistical Society, B*, 67:301–320, 2005. (Cited on pages 11 and 61.)

H. Zou, T. Hastie, and Tibshirani R. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006. (Cited on pages 111 and 151.)

List of Figures

List of Tables

Titre Méthodes multivariées pour l'analyse jointe de données de neuroimagerie et de génétique

Résumé L'imagerie cérébrale connaît un intérêt grandissant, en tant que phénotype intermédiaire, dans la compréhension du chemin complexe qui relie les gènes à un phénotype comportemental ou clinique. Dans ce contexte, un premier objectif est de proposer des méthodes capables d'identifier la part de variabilité génétique qui explique une certaine part de la variabilité observée en neuroimagerie. Les approches univariées classiques ignorent les effets conjoints qui peuvent exister entre plusieurs gènes ou les covariations potentielles entre régions cérébrales. Notre première contribution est de chercher à améliorer la sensibilité de l'approche univariée en tirant avantage de la nature multivariée des données génétiques, au niveau local. En effet, nous adaptons l'inférence au niveau du cluster en neuroimagerie à des données de polymorphismes d'un seul nucléotide (SNP), en cherchant des clusters 1D de SNPs adjacents associés à un même phénotype d'imagerie. Ensuite, nous prolongeons cette idée et combinons les clusters de voxels avec les clusters de SNPs, en utilisant un test simple au niveau du "cluster 4D", qui détecte conjointement des régions cérébrale et génomique fortement associées. Nous obtenons des résultats préliminaires prometteurs, tant sur données simulées que sur données réelles. Notre deuxième contribution est d'utiliser des méthodes multivariées exploratoires pour améliorer la puissance de détection des études d'imagerie génétique, en modélisant la nature multivariée potentielle des associations, à plus longue échelle, tant du point de vue de l'imagerie que de la génétique. La régression Partial Least Squares et l'analyse canonique ont été récemment proposées pour l'analyse de données génétiques et transcriptomiques. Nous proposons ici de transposer cette idée à l'analyse de données de génétique et d'imagerie. De plus, nous étudions différentes stratégies de régularisation et de réduction de dimension, combinées avec la PLS ou l'analyse canonique, afin de faire face au phénomène de sur-apprentissage dû aux très grandes dimensions des données. Nous proposons une étude comparative de ces différentes stratégies, sur des données simulées et des données réelles d'IRM fonctionnelle et de SNPs. Le filtrage univarié semble nécessaire. Cependant, c'est la combinaison du filtrage univarié et de la PLS régularisée L1 qui permet de détecter une association généralisable et significative sur les données réelles, ce qui suggère que la découverte d'associations en imagerie génétique nécessite une approche multivariée.

Mots-clés Imagerie Génétique, Analyse multivariée, Inférence au niveau du cluster, Régression Partial Least Squares, Analyse Canonique, Sélection d'attributs, Régularisation

Title Multivariate methods for the joint analysis of neuroimaging and genetic data

Abstract Brain imaging is increasingly recognised as an interesting intermediate phenotype to understand the complex path between genetics and behavioural or clinical phenotypes. In this context, a first goal is to propose methods to identify the part of genetic variability that explains some neuroimaging variability. Classical univariate approaches often ignore the potential joint effects that may exist between genes or the potential covariations between brain regions. Our first contribution is to improve the sensitivity of the univariate approach by taking advantage of the multivariate nature of the genetic data in a local way. Indeed, we adapt cluster-inference techniques from neuroimaging to Single Nucleotide Polymorphism (SNP) data, by looking for 1D clusters of adjacent SNPs associated with the same imaging phenotype. Then, we push further the concept of clusters and we combined voxel clusters and SNP clusters, by using a simple 4D cluster test that detects conjointly brain and genome regions with high associations. We obtain promising preliminary results on both simulated and real datasets . Our second contribution is to investigate exploratory multivariate methods to increase the detection power of imaging genetics studies, by accounting for the potential multivariate nature of the associations, at a longer range, on both the imaging and the genetics sides. Recently, Partial Least Squares (PLS) regression or Canonical Correlation Analysis (CCA) have been proposed to analyse genetic and transcriptomic data. Here, we propose to transpose this idea to the genetics vs. imaging context. Moreover, we investigate the use of different strategies of regularisation and dimension reduction techniques combined with PLS or CCA, to face the overfitting issues due to the very high dimensionality of the data. We propose a comparison study of the different strategies on both a simulated dataset and a real fMRI and SNP dataset. Univariate selection appears to be necessary to reduce the dimensionality. However, the generalisable and significant association uncovered on the real dataset by the two-step approach combining univariate filtering and L1-regularised PLS suggests that discovering meaningful imaging genetics associations calls for a multivariate approach.

Keywords Imaging Genetics, Multivariate analysis, Cluster-level inference, Partial Least Squares, Canonical Correlation Analysis, Feature selection, Regularisation