Analyse statistique de données en grande dimension : application à l'étude de la variabilité inter-individuelle en neuroimagerie
(résumé français)

La variabilité inter-individuelle est un obstacle majeur à l'analyse d'images médicales, en particulier en neuroimagerie. Il convient de distinguer la variabilité naturelle ou statistique, source de potentiels effets d'intérêt pour du diagnostique, de la variabilité artefactuelle, constituée d'effets de nuisance liés à des problèmes expérimentaux ou techniques, survenant lors de l'acquisition ou le traitement des données. La dernière peut s'avérer bien plus importante que la première : en neuroimagerie, les problèmes d'acquisition peuvent ainsi masquer la variabilité fonctionnelle qui est par ailleurs associée à une maladie, un trouble psychologique, ou à l'expression d'un code génétique spécifique. La qualité des procédures statistiques utilisées pour les études de groupe est alors diminuée car lesdites procédures reposent sur l'hypothèse d'une population homogène, hypothèse difficile à vérifier manuellement sur des données de neuroimagerie dont la dimension est élevée. Des méthodes automatiques ont été mises en oeuvre pour tenter d'éliminer les sujets trop déviants et ainsi rendre les groupes étudiés plus homogènes. Cette pratique n'a pas entièrement fait ses preuves pour autant, attendu qu'aucune étude ne l'a clairement validée, et que le niveau de tolérance à choisir reste arbitraire. Une autre approche consiste alors à utiliser des procédures d'analyse et de traitement des données intrinsèquement insensibles à l'hypothèse d'homogénéité. Elles sont en outre mieux adaptées aux données réelles en ce qu'elles tolèrent dans une certaine mesure d'autres violations d'hypothèse plus subtiles telle que la normalité des données. Un autre problème, partiellement lié, est le manque de stabilité et de sensibilité des méthodes d'analyse au niveau voxel, sources de résultats qui ne sont pas reproductibles.

Nous commençons cette thèse par le développement d'une méthode de détection d'individus atypiques adaptée aux données de neuroimagerie, qui fournit un contrôle statistique sur l'inclusion de sujets : nous proposons une version régularisée d'un estimateur de covariance robuste pour le rendre utilisable en grande dimension. Nous comparons plusieurs types de régularisation et concluons que les projections aléatoires offrent le meilleur compromis. Nous présentons également des procédures non-paramétriques dont nous montrons la qualité de performance, bien qu'elles n'offrent aucun contrôle statistique. La seconde contribution de cette thèse est une nouvelle approche, nommée RPBI (Random-

ized Parcellation Based Inference), répondant au manque de reproductibilité des méthodes classiques. Nous stabilisons l'approche d'analyse à l'échelle de la parcelle en agrégeant plusieurs analyses indépendantes, pour lesquelles le partitionnement du cerveau en parcelles varie d'une analyse à l'autre. La méthode permet d'atteindre un niveau de sensibilité supérieur à celui des méthodes de l'état de l'art, ce que nous démontrons par des expériences sur des données synthétiques et réelles. Notre troisième contribution est une application de la régression robuste aux études de neuroimagerie. Poursuivant un travail déjà existant, nous nous concentrons sur les études à grande échelle effectuées sur plus de cent sujets. Considérant à la fois des données simulées et des données réelles, nous montrons que l'utilisation de la régression robuste améliore la sensibilité des analyses. Nous démontrons qu'il est important d'assurer une résistance face aux violations d'hypothèse, même dans les cas où une inspection minutieuse du jeu de données a été conduite au préalable. Enfin, nous associons la régression robuste à notre méthode d'analyse RPBI afin d'obtenir des tests statistiques encore plus sensibles.

Présentation des études d'IRM fonctionnelle

Imager le cerveau par résonance magnétique

L'*imagerie par résonance magnétique (IRM)* est une techniques d'imagerie médicale non-invasive qui permet d'observer différents types de tissus. Son utilisation est devenue systématique dans les années 1990 et concerne en particulier le diagnostique de la maladie d'Alzheimer, de l'épilepsie, ou la révélation de tumeurs cancéreuses. L'IRM permet d'imager le cerveau complet avec un bon compromis entre résolution spatiale (de l'ordre du millimètre) et résolution temporelle (de l'ordre de la seconde). L'acquisition d'IRM se fait par tranches d'une épaisseur de 1 à 3 millimètres, chaque tranche pouvant être vue comme une grille régulière dont les cases contiennent une valeur de signal. Une image IRM est donc constituée de parallélépipèdes accolés appelés *voxels*. Dans cette thèse, nous distinguons deux principaux types d'images :

les images structurelles possèdent une bonne résolution spatiale, mais nécessite un temps d'acquisition de l'ordre de plusieurs minutes durant lesquelles le sujet doit rester immobile. Sur ces images, on peut discerner la *matière grise* (formée par les synapses –la tête– des neurones implantés en bouquet tout autour du cerveau), la *matières blanche* (formée par les axones –les queues– des neurones recouverts de blanche *myéline*), le crâne et le *liquide cérébro-spinal* qui entoure le cerveau.

les images fonctionnelles acquises à des intervalles de temps très courts (à la manière d'une film, avec un fréquence d'imagerie de l'ordre de la seconde) au prix d'une résolution spatiale plus faible. Le signal représenté dans chaque voxel est une estimation du nombre d'atomes d'hydrogène présent en ce voxel au moment de l'acquisition. Ce nombre est corrélé avec l'activité cérébrale ; Il s'agit de l'*effet BOLD*. Les images fonctionnelles permettent ainsi d'étudier l'activité cérébrale induite par la réalisation de certaines actions contrôlées par l'expérimentateur.

Analyse de données d'IRM fonctionnelle

Les images d'IRM fonctionnelle se présentent sous la forme d'une série temporelle d'images tridimensionnelles. Chacune d'entre elles est composée d'environ 200 000 voxels dont l'étude peut être ramenée à environ 60 000 en considérant un masque du cerveau calculé d'après une image de densité de matière grise de référence (typiquement une image moyenne sur plusieurs sujets). La caractéristique principale d'une image fonctionnelle est son *rapport signal sur bruit*, qui quantifie l'écart relatif entre *(i)* la variabilité liée aux événements cognitifs subits par le sujet et *(ii)* la variabilité du signal qui ne s'y rapporte pas (le *bruit*). Le bruit provient de plusieurs sources incluant les défauts techniques du matériel, des artefacts expérimentaux, une défaillance du sujet (mouvement, forte respiration). Ces problèmes peuvent être raisonnablement corrigés dans une certaine mesure à l'occasion d'un chaîne de *pré-traitements* mais ont une importante influence sur la suite de l'analyse. La normalisation spatiale des images intervient en bout de chaîne et est primordiale pour l'analyse d'un groupe de sujets. Elle consiste en un repositionnement et une mise à l'échelle des images afin d'atteindre la meilleure correspondance spatiale possible entre les images de plusieurs sujets. Cette étape à une conséquence directe sur la puissance de l'analyse statistique effectuée au niveau du groupe.

L'analyse statistique se fait en deux étapes, suivant le concept de *statistique sommaire* (de l'anglais *summary statistic*) [7]. En effet, pour une sensibilité accrue, il convient de prendre en compte dans le modèle statistique d'analyse des données la variabilité intra-individuelle (différences observées entre plusieurs expériences similaires chez un même sujet) d'une part et la variabilité inter-individuelle d'autre part []. Pour simplifier les calculs, une pratique courante est de considérer un premier modèle, un *modèle linéaire généralisé*, appliqué individuellement à chaque sujet :

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{\epsilon}_1, \tag{1}$$

où $\boldsymbol{Y}$ est une matrice $(t \times v)$ qui encode les t images de la série temporelle possédant chacune v voxels. $\boldsymbol{\beta}$ est la matrice des coefficients du modèle, de taille $(m \times v)$, qui doit être estimée. $\boldsymbol{X}_1$ est appelée la *matrice de dessin*, de taille $(t \times m)$. Elle encode m conditions expérimentales. $\boldsymbol{\epsilon}_1$ correspond au bruit du modèle. Le modèle Equation 1 est d'abord appliqué aux données (les coefficients sont estimés, puis un test statistique est appliqué à (un sous-ensemble de) ces coefficients afin d'éventuellement révéler un lien statistiquement fort entre les conditions expérimentales et l'activité cérébrale observée. Chaque test est en fait appliqué à chaque voxel de l'image, en conséquence de quoi l'analyse de premier niveau aboutit à la création d'autant de *cartes statistiques* (images de cerveau dont les valeurs des voxels correspondent à une statistique donnée) que de tests ont été conduits. On parle de *cartes de contrastes fonctionnels* ; chacune reflète l'activité cérébrale relative à une tâche cognitive définie par le test statistique correspondant.

Au deuxième niveau d'analyse, le niveau du groupe, le but est d'exhiber des relations entre des caractéristiques des images (typiquement, les valeurs des voxels mis en correspondance à l'étape de normalisation) et des variables expérimentales, comportementales ou génétiques. Un second modèle linéaire est

considéré :

$$\boldsymbol{B} = \boldsymbol{X}_2\boldsymbol{\gamma} + \boldsymbol{\epsilon}_2, \tag{2}$$

où $\boldsymbol{X}_2$ est la matrice de dessin second niveau qui encode les variables externes, $\boldsymbol{\gamma}$ est la matrice des coefficients à estimer, et $\boldsymbol{B}$ est une matrice de taille $n \times v$ qui contient les cartes issues du premier niveau d'analyse (calculées pour n sujets, et contenant v voxels chacune). En général, les cartes de contrastes sont utilisées, bien qu'elles puissent être remplacées ou complétées par des cartes de variance de l'effet premier niveau.

La base de données Imagen

Les résultats obtenus dans cette thèse ont principalement impliqués les données de la base de données *Imagen*. Image est une étude européenne multicentres effectuée sur environ 2 000 adolescents. La base de données contient 99 contrastes disponibles chacun des sujets à une résolution de 3mm $\times$ 3mm $\times$ 3mm. Les données ont subi les pré-traitements par défaut du logiciel SPM8, utilisé de façon standard par les neuroscientifiques. Des images anatomiques sont également disponibles avec une résolution de 1mm $\times$ 1mm $\times$ 1mm. 600 000 variables génétiques ont été séquencées pour chaque sujet, permettant des études de neuorimagerie génétique, études de groupe qui possèdent la particularité d'impliquer des variables génétiques et qui posent alors des problèmes de grande dimension compte tenu du grand nombre de variables disponibles [18, 15, 1]. Les variables génétiques en questions répondent au nom de *single nucleotide polymorphism (SNP)* et correspondent à des mutations génétiques d'une seule base de la chaîne d'ADN.

Structure statistique des jeux de données de neuroimagerie : écarts à la normalité et détection d'individus aberrants

Observations atypiques en neuroimagerie

L'acquisition d'images médicales est souvent parasitée par des problèmes matériel, des erreurs de mesures, ou des défauts de protocole expérimental. En considérant de surcroît la grande variabilité observée au sein de la population, les jeux de données de neuroimagerie contiennent des observations déviantes, techniquement correctes ou non, dont l'inclusion dans les analyses peut être délicate. Compte tenu de la grande dimension des données et du faible rapport signal-sur-bruit des données de neuroimagerie, il est en effet impossible de procéder à des contrôles qualité manuellement, et il est difficile de placer des seuils permettant de séparer les observations techniquement incorrectes des observations techniquement correctes, mais peu représentatives de la population. Chaque observation déviante peut pourtant par ailleurs avoir une influence potentiellement forte sur les résultats. La détection d'individus aberrants en neuroimagerie est une pratique aujourd'hui très arbitraire, essentiellement massivement univariées (chaque voxel est considéré individuellement et comparé à travers les sujets), la détection multivariée étant un problème mal posé dès lors que le nombre de

variables descriptives est supérieur à cinq fois le nombre de sujets [4]. La vérification manuelle des données reste la technique la plus fiable aux yeux des neuroscientifiques, mais cette dernière est longue, fastidieuse, et impossible sur les grandes cohortes qui émergent à l'heure actuelle (la base de données Imagen en fait partie). Dans ce chapitre, nous cherchons à développer des outils automatiques pour la détection d'individus aberrants, avec la volonté d'obtenir en sus un contrôle statistique sur l'inclusion/l'élimination de sujets. Nous insistons néanmoins sur le fait que l'utilisation de ce type d'outil ne saurait se substituer à l'utilisation d'outil d'analyse robustes. En effet, *(i)* une détection parfaite est utopique et *(ii)* il existe d'autres déviations aux hypothèses d'analyse contre lesquelles il est important de se prémunir (voir chapitre 6).

Détection d'individus aberrants par estimation de covariance

Nous considérons dans cette partie un modèle gaussien multivarié en grande dimension. Une observation $\boldsymbol{x}_i \in \mathbb{R}^p$ du jeu de données $\boldsymbol{X}$ peut être considérée comme atypique lorsqu'elle possède une distance de Mahalanobis $d^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})$ élevée, avec $\boldsymbol{\mu}$ et $\boldsymbol{\Sigma}$ représentant respectivement la moyenne (multidimensionnelle) et la covariance des données. À défaut de connaître leur valeur, il est important que ces deux derniers paramètres soient estimés de manière robuste [2, 12]. L'estimateur *Minimum Covariance Determinant (MCD)* [13] constitue l'état-de-l'art en la matière. Considérant n observations p-dimensionnelles, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, le MCD consiste à trouver les h observations jugées les « plus correctes » en minimisant le déterminant de leur matrice de covariance. Ces observations constituent le *support* du MCD. L'estimateur MCD peut être décrit par le problème d'optimisation alterné suivant :

$$\begin{aligned}(\hat{H}, \hat{\boldsymbol{\mu}}_\mathrm{h}, \hat{\boldsymbol{\Sigma}}_\mathrm{h}) = \underset{\boldsymbol{\mu},\boldsymbol{\Sigma},H}{\operatorname{argmin}}\Bigg(&\log|\boldsymbol{\Sigma}| \\ &+ \frac{1}{h}\sum_{i\in H}(\boldsymbol{x}_i - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\Bigg),\end{aligned} \tag{3}$$

Le problème principal du MCD est que la matrice de covariance des h observations doit être de rang plein, et ce à chaque étape de l'algorithme utilisé pour résoudre le problème 3. De ce fait, h doit être plus grand que $h_\mathrm{min} = \frac{n+p+1}{2}$, et la distribution des données $\boldsymbol{X}$ ne peut être estimée que si plus de h_min observations sont correctes. Lorsque $p = n - 1$ le MCD est équivalent à l'estimateur du maximum de vraisemblance, qui n'est pas robuste. Si $p \geq n$, le MCD n'est pas défini. En pratique, l'utilisation du MCD n'est pas recommandée pour le cas $\frac{p}{n} > 0.2$. Nous proposons donc de fixer $h = \frac{n}{2}$ et de compenser ce choix en régularisant l'estimateur de covariance. Nous baptisons cette méthode *RMCD*. Lorsque RMCD est utilisé, les distances de Mahalanobis des observations ne peuvent plus être comparées à une distribution théorique car celle-ci n'est pas connue. Nous respectons les recommandations qui existent déjà dans le cas du MCD et avons recours à des simulations de Monte-Carlo pour obtenir une distribution empirique de référence [6].

Nous étudions différents types de régularisation pour RMCD, en commençant

par une régularisation ℓ_2 (ou *régularisation ridge*) :

$$(\hat{\boldsymbol{\mu}}_{\mathrm{r}}, \hat{\boldsymbol{\Sigma}}_{\mathrm{r}}|H) = \underset{\boldsymbol{\mu},\boldsymbol{\Sigma}}{\operatorname{argmin}} \Big(\log|\boldsymbol{\Sigma}| + \lambda \operatorname{Tr} \boldsymbol{\Sigma}^{-1} + \frac{1}{h} \sum_{i \in H} (\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \Big), \tag{4}$$

dont la solution en fonction du paramètre de régularisation λ est connue analytiquement et peut donc être calculée instantanément : $\hat{\boldsymbol{\Sigma}}_{\mathrm{r}}|H = \frac{\boldsymbol{X}_H^{\mathsf{T}} \boldsymbol{X}_H}{h} + \lambda \mathrm{I}$ et $\hat{\boldsymbol{\mu}}_r|H = \frac{\boldsymbol{X}_H^{\mathsf{T}} \mathbf{1}}{h}$. $\boldsymbol{X}_H$ correspond au jeu de données restreint au support de l'estimateur RMCD. Le paramètre de régularisation joue un rôle crucial car il incarne la différence entre le MCD classique et le MCD régularisé. λ peut être estimé par validation croisée, bien que nous lui préférions la formule (analytique donc plus rapide, et mieux adapté après étude) de Ledoit-Wolf [9]. L'estimateur RMCD-ℓ_2 est biaisé vers une matrice de covariance sphérique. Cela correspond à une hypothèse sous-jacente d'isotropie des données. Si cette hypothèse est trop fortement violée, des observations atypiques pourraient se trouver dans le support.

De manière similaire, nous étudions l'utilisation de régularisation ℓ_1 pour RMCD-ℓ_2. La matrice de covariance solution est connue pour posséder une inverse parcimonieuse [17]. Il n'existe malheureusement aucun algorithme rapide pour l'estimation ℓ_1 et aucune formule analytique n'a été proposée dans la littérature. Nous utilisons l'algorithme GLasso [3] implémenté dans la librairie Python scikit-learn [11].

Une approche par projections aléatoires vient compléter notre étude de différents types de régularisation. Le principe est de réduire la dimension du problème par des projections aléatoires des données sur des espaces de dimension suffisamment réduite pour permettre l'utilisation du MCD non régularisé. Un consensus est formé entre les différentes décisions effectuées dans les sous-espaces pour aboutir à une décision dans l'espace d'origine. Après une étude empirique, nous déterminons qu'un choix raisonnable est de considérer p projections dans des sous-espaces de dimension $p/5$.

Détection d'individus aberrants par méthodes non-paramétriques

Les données de neuroimagerie ne respectent pas nécessairement une distribution gaussienne. Il est donc judicieux d'étudier des méthodes de détection d'observations aberrantes qui ne sont pas basées sur la distance de Mahalanobis. L'utilisation des méthodes par estimation de densité en neuroimagerie a été restreinte à la détection d'individus présentant une pathologie, par opposition aux individus sains [10]. Ces applications sont donc (semi-)supervisées. Nous considérons quant à nous une tâche non supervisée car nous n'avons pas d'hypothèse a priori sur la distribution ni sur le nombre d'observations aberrantes. En supposant que nous ayons un modèle de densité correspondant à la distribution des données, les observations atypiques sont celles présentent dans des régions de faible densité. De tels modèles de densité peuvent être estimés sur la base des distances entre couples d'observations, entre plus proches voisins, ou résulter d'estimation de densité par noyau. Les *machines à vecteur de support (SVMs)* peuvent aussi être utilisées, bien qu'elles soient conçues à l'origine

pour de la classification supervisée. Nous étudions cette dernière approche ainsi qu'une extension de l'estimation de densité par noyau où le paramètre du noyau est calculé en fonction des données.

L'algorithme *One-Class SVM* permet de définir une frontière autour d'une classe d'observation bien identifiée. Si une nouvelle observation est examinée, il est alors possible de dire si oui ou non il elle appartient à la même classe que les données étudiées avec une certaine probabilité. Par extension, en supposant que nous arrivons à isoler une proportion de données certifiées correctes, nous pouvons estimer une frontière au-delà de laquelle toute nouvelle observation serait douteuse. Nous utilisons ce mécanisme pour identifier les observations atypiques. Nous ne parvenons malheureusement pas à réaliser un test statistique sur la qualité de chaque observation et sommes limités à pouvoir ordonner le niveau de confiance en chaque observation. Le même problème survient avec l'utilisation de l'algorithme *Local Component Analysis (LCA)* [14], qui estime par validation croisée le paramètre de noyau optimal (moyennant éventuellement une régularisation pour le cas de la grande dimension) à utiliser dans une estimation de densité par noyau. En pratique, ce paramètre peut être vu comme une matrice de covariance locale des données $\mathbf{\Sigma}$:

$$\mathbf{\Sigma}^* = \underset{\mathbf{\Sigma}}{\operatorname{argmin}} \left[-\sum_{i=1}^{n} \log \left(\frac{|\mathbf{\Sigma}|^{-\frac{1}{2}} (2\pi)^{-\frac{p}{2}}}{n-1} \sum_{j \neq i} \exp \left(-\frac{1}{2} d^2_{\boldsymbol{x}_j, \mathbf{\Sigma}}(\boldsymbol{x}_i) \right) \right) \right]. \quad (5)$$

Dans le cas de LCA, puisque nous somme en présence d'une estimation de densité à proprement parler (et pas d'une distance à la frontière comme dans le cas du One-Class SVM), nous appliquons une transformation de type *soft-thresholding* à notre modèle de densité afin d'accentuer le contraste entre les zones à faible densité et les zones à haute densité. Nous définissons alors un score pour chaque observation qui permet d'identifier la structure des données par l'observation du spectre de celles-ci. Cette approche permet au praticien de mieux connaître son jeu de données et d'éventuellement choisir un nombre fixé d'observations à retirer du jeu de données considéré.

Applications à la neuroimagerie

Nous mesurons la capacité des méthodes présentées ci-dessus à identifier les observations aberrantes. Dans un premier temps, il s'agit de vérifier que la structure principale des données est bien capturée, à un facteur d'échelle près. Nous nous intéressons à la spécificité de la détection et à son intérêt pratique dans un second temps. Nous considérons trois modèles de données déviantes :

- données à forte variance
- données multimodales
- données issues d'une autre distribution de même moyenne

Nous confrontons les méthodes à des données générées avec un mélange de gaussiennes, ce qui permet d'obtenir des données non gaussiennes. À la différence des modèles ci-dessus, les différents modes constituent dans ce cas des observations valides. Nous ajoutons des données déviantes distribuées uniformément dans l'espace (elles sont déviantes vis-à-vis de chacune des composantes du modèle génératif). La qualité des méthodes est déterminée par des courbes ROC [21] et l'aire sous ces courbes [5].

Les données réelles sont des images de contraste issues de la base de données Imagen. Nous construisons une vérité terrain en utilisant un estimateur MCD sur l'ensemble des 2 000 sujets à contraste fixé. Nous restreignons ensuite le nombre de sujets pour nous placer dans un cas où p avoisine, voire dépasse n. Il est important de noter que nous travaillons avec des atlas anatomiques qui définissent une centaine de régions d'intérêt dans le cerveau. Au lieu du signal par voxel, nous considérons ainsi $p = 100$ signaux moyens par région dans nos expérience. L'approche par parcelle a en outre l'intérêt d'améliorer la correspondance des descripteurs à travers les sujets.

L'étude des résultats révèle que RMCD-ℓ_2 se comporte rarement mal et peut être utilisé en pratique. On lui préférera néanmoins l'approche par projections aléatoires qui offre plus de robustesse vis-à-vis du cas où les données correctes ne sont pas distribuées selon une loi gaussienne. L'approche peut en outre être utilisée dans des cas extrêmes où $p \gg n$. La structure des données est pourtant mieux capturée par l'algorithme LCA, qui ne permet en revanche pas de prendre une décision contrôlée statistiquement sur l'inclusion de données. L'analyse de cette structure (présence de coudes dans le spectre des observations) vient confirmer les résultats sur données réelles qui suggèrent entre 30 et 40 % de données déviantes.

Rendre les analyses de groupe reproductibles grâce à des partitions aléatoires du cerveau

En neuroimagerie, les études de groupe sont utilisées pour associer des variables issues des images à des variables comportementales ou génétiques et ainsi permettre de prévenir certaines maladies et facteurs à risque. Les méthodes d'analyse niveau voxel utilisées actuellement manquent de stabilité et de sensibilité, si bien que les résultats qu'elles fournissent sont difficilement reproductibles. Nous présentons une nouvelle méthode d'analyse qui repose sur l'utilisation de plusieurs partitions du cerveau en parcelles. Un test de permutation sur une statistique d'agrégation contrôle le taux de faux positifs.

Partitionner le cerveau

Les modèles d'analyse spatiaux sont une réponse au manque de correspondance entre les sujets au niveau voxel. La technique la plus simple et la plus utilisée consiste à lisser les données pour augmenter le recouvrement entre les zones activées à travers les sujets. Une approche alternative consiste à considérer des parcelles et à moyenner le signal à l'intérieur de chaque pour construire de nouveaux descripteur. La limite principale de l'approche reste le manque de consensus sur la parcellisation à utiliser, sachant que les résultats varient en fonction de la parcellisation utilisée. Thirion et al. [16] ont proposé d'utiliser des parcellisations calculées de manière déterministe à partir des données : cette méthode améliore la sensibilité des analyses de groupe mais implique néanmoins de fixer la parcellisation. Travailler par régions d'intérêt possède néanmoins plusieurs avantages : *(i)* il s'agit d'un modèle simple et facilement interprétable, *(ii)* le problème de comparaison multiples est diminué avec le nombre de descripteurs ; *(iii)* il est possible de choisir un algorithme de parcellisation qui s'adapte à la structure locale des données. Concernant ce dernier point, nous

utilisons l'*algorithme de classification de Ward* [20] qui minimise la variance à l'intérieur de chaque classe, et possède la particularité de pouvoir prendre en compte plusieurs sujets simultanément afin d'ajouter une contrainte spatiale au problème de partitionnement. Le nombre de sujets considérés a une influence sur la compacité des parcelles. Il ne s'agit donc pas d'obtenir une parcellisation de référence –puisque celle-ci dépend donc des sujets considérés– mais d'obtenir une parcellisation fonctionnelle adaptée aux données.

Inférence statistiques à partir de multiples parcellisations aléatoires

Considérant que plusieurs parcellisations légèrement variables peuvent donc être obtenues pour un même jeu de données selon le nombre de sujets utilisé dans l'algorithme de Ward, il devient possible d'utiliser ces parcellisations comme autant de représentations d'une même variable aléatoire et d'intégrer les résultats obtenues sur celle-ci, marginalisant ainsi ce paramètre. Cette approche revient à stabiliser l'analyse de groupe niveau parcelle et à ainsi rendre les résultats associés plus reproductibles. La méthode est nommée *Randomized Parcellation Based Inference (RPBI)*.

Soit $\mathcal{P}$ un ensemble fini de parcellisations, et V l'ensemble des voxels considérés dans l'analyse. Étant donné un voxel v et une parcellisation P, la fonction de seuillage au niveau parcelle θ_t est définie par :

$$\theta_t(v, P) = \begin{cases} 1 \text{ if } F(\Phi_P(v)) > t \\ 0 \text{ otherwise} \end{cases} \tag{6}$$

où $\Phi_P : V \to P$ est une fonction de correspondance qui associe chaque voxel à une parcelle de la parcellisation P ($\forall v \in P^{(i)}$, $\Phi_P(v) = P^{(i)}$). Pour un test prédéfini, F renvoie la statistique F associée à la valeur du signal moyen pour une parcelle donnée. Enfin, la statistique d'agrégation au voxel v est données par la fonction de comptage :

$$C_t(v, \mathcal{P}) = \sum_{P \in \mathcal{P}} \theta_t(v, P). \tag{7}$$

$C_t(v, \mathcal{P})$ représente le nombre de fois (à travers les analyses individuelles) où le voxel v fait partie d'une parcelle dont la statistique associée est plus grande que t. t est choisi de façon à assurer un contrôle de type Bonferroni à $p < 0.1$. La valeur de la fonction de comptage est convertie en p-valeur par un test de permutation.

Nous étudions également l'intérêt de la méthode par parcellisations aléatoires pour d'autres tâches que l'analyse de groupe. Nous revenons en particulier sur la détection d'individus aberrants évoquées précédemment.

Résultats / Discussion

Nous comparons RPBI à trois types d'analyse de l'état de l'art : *(i)* l'analyse niveau voxel classique, *(ii)* l'analyse niveau cluster (où la statistique de test ne repose plus sur l'intensité d'un voxel, mais sur la taille du cluster d'activation auquel il appartient, les clusters étant définis grâce à un seuil donné), *(iii)* la méthode TFCE, développement le plus abouti à l'heure actuelle de l'analyse

niveau cluster, dans laquelle aucun choix de seuil n'est nécessaire. Nous générons des groupes de 20 sujets simulés possédant une activation en forme de cube de $4 \times 4 \times 4$ voxels dans une grille de taille $40 \times 40 \times 40$ voxels. Nous ajoutons une variabilité aléatoire (variance = 2 voxels) sur la position exacte de cette activation. Nous mesurons la capacité de chacune des quatre méthodes d'analyse à retrouver l'activation au niveau des groupes. Dans un second temps, nous remplaçons l'activation cubique par une forme plus réaliste qui ne présente aucun isotropie. Dans les deux expériences sur données simulées, nous ajoutons un bruit gaussien que dont nous contrôlons la corrélation en appliquant un lissage. Nous contrôlons également le niveau de signal sur bruit des images.

Nous utilisons les données réelles de la base de données Imagen grâce à laquelle il est possible de construire une pseudo-vérité terrain pour l'analyse de groupe en utilisant les 2000 sujets disponibles dans une analyse niveau voxel. La sensibilité de ce type d'analyse pour un si grand nombre de sujets est en effet élevé. En portant le nombre de sujets considérés à 20, en revanche, nous pouvons comparer les performances des différentes méthodes d'analyse évoquées ci-dessus. Tout comme dans le cas des données simulées, nous construisons des courbes précision-rappel qui montrent les performances des méthodes dans différents régimes. Nous réalisons également une étude de reproductibilité des résultats ainsi qu'un contrôle de la spécificité des méthodes (afin d'être sûrs de pouvoir comparer leur précision/rappel de manière juste). Nous utilisons également les 2000 sujets disponibles pour réaliser une détection d'individus atypiques, puis nous réduisons artificiellement le nombre de sujets disponibles pour des nouvelles détections en suivant le principe des parcellisations aléatoires. Nous construisons des diagrammes en boîte afin d'évaluer la qualité de la détection obtenue par cette dernière méthode. Enfin, nous appliquons RPBI à une étude de neuroimagerie génétique.

RPBI offre une meilleure reproductibilité des résultats, bien que sa précision ne dépasse pas significativement celle du TFCE. Les approches niveau voxel ou cluster sont en tout cas largement dépassées. La détection d'individus aberrants ne bénéficie que très peu de notre approche et semble donc ne pas valoir l'effort, excepté dans le cas où très peu de sujets sont disponibles ($p \gg n$). La sensibilité offerte par RPBI est accrue par rapport à celle des autres méthodes, ce qui se traduit dans l'étude de neuroimagerie génétique par l'observation d'un effet significatif nouveau dans le thalamus gauche. Nous montrons que RPBI fonctionne mieux sur des données non lissées. Nous discutons également l'implémentation que nous avons réussi à optimiser pour obtenir un temps total d'exécution inférieur à trois minutes pour une expérience impliquant 20 sujets, 100 parcellisations aléatoires et 10 000 permutations.

La régression robuste s'accommode des différentes sources de variabilité rencontrées en neuroimagerie

Comme nous l'avons mentionné dans les chapitres précédents, la présence d'individus déviants ne constitue pas la seule violation d'hypothèse potentiellement néfaste. D'autres déviations aux hypothèses du modèle, plus subtiles, méritent d'être anticipées. De ce fait, malgré toutes les précautions prises pour épurer les jeux de données, il convient d'utiliser des méthodes d'analyse robustes, sus-

ceptibles de s'adapter aux nuisances insoupçonnées rencontrées dans les données.

Régression robuste en neuroimagerie

Nous revenons au modèle d'analyse de second niveau présenté en (2) :

$$\boldsymbol{B} = \boldsymbol{X}_2\boldsymbol{\gamma} + \boldsymbol{\epsilon}_2. \tag{8}$$

Une solution classique est la solution donnée par l'estimateur des moindres carrés (OLS) :

$$\hat{\boldsymbol{\gamma}}_{\text{OLS}} = \underset{\boldsymbol{\gamma}}{\text{argmin}} \|\boldsymbol{B} - \boldsymbol{X}_2\boldsymbol{\gamma}\|^2. \tag{9}$$

Cet estimateur n'est pas robuste à la présence d'individus aberrants, c'est pourquoi les jeux de données sont souvent nettoyés au préalable. Mais d'autres déviations aux hypothèses du modèle sont potentiellement dangereuse, en particulier lorsque l'on considère des études de neuroimagerie génétique et le problème associé du déséquilibre entre classes. Plusieurs méthodes de régression robustes sont alors utilisables pour repousser les limites de la régression de type OLS. Comme [19], nous nous concentrons sur la régression robuste préconisée par Huber [8], en nous concentrant davantage sur le cas des grandes cohortes de sujets, problème émergent dans la communauté. L'intérêt du type de régression que nous considérons et qu'une procédure de test analytique existe. Nous avons en effet pour but d'utiliser la régression robuste dans des procédures d'analyse plus complexes et déjà coûteux. Nous ne pouvons donc nous permettre d'utiliser une méthode de régression qui ne fournisse pas des p-valeurs rapidement (typiquement, un test de permutation est inenvisageable et met hors-jeu des méthodes telles que la *régression sur vecteurs de support*, ou les *least trimmed squares*.

La régression de Huber consiste à remplacer la fonction carré de l'OLS par une fonction ρ qui diminue l'influence des observations atypiques :

$$\hat{\boldsymbol{\gamma}}_{\text{RLM}} = \underset{\boldsymbol{\gamma}}{\text{argmin}} \sum_{i=1}^{n} \rho\left(\frac{b_i - \sum_{j=1}^{p} x_{2_i j}\gamma_j}{\sigma}\right). \tag{10}$$

σ est la déviation standard des résidus, qui doit en réalité est estimée en même temps que γ. Le choix classique pour ρ est :

$$\rho(x) = \begin{cases} \frac{1}{2}x^2, & \text{if } |x| < c, \\ c|x| - \frac{1}{2}c^2, & \text{if } |x| \geq c, \end{cases} \tag{11}$$

avec $c = 1.345$ pour 95% de fidélité au modèle gaussien en l'absence de contamination. Nous utilisons l'algorithme IRLS pour résoudre le problème et réalisons les tests statistiques associés à notre modèle en utilisant la méthode préconisée une fois encore par Huber. Il s'agit de remplacer la matrice de covariance du modèle OLS par une covariance robuste fonction de $\hat{\boldsymbol{\gamma}}_{\text{RLM}}$ et σ.

Validation de la procédure statistique

Dans un premier temps, nous validons la procédure de test proposée par Huber étant donné qu'il s'agit en réalité d'une approximation qui peut ne pas correspondre sous certaines conditions. Nous la validation dans le contexte de

la neuroimagerie. Nous sommes particulièrement attachés au contrôle de la spécificité du test sous l'hypothèse nulle, et ce quel que soit le niveau de contamination par ailleurs. Nous générons des observations distribuées selon une loi gaussienne, des covariables sans effet corrélé, ajoutons une proportion contrôlée d'observations déviantes et appliquons un modèle de régression robuste. Dans ces conditions, le contrôle nominal est bien respecté à tous les niveaux de précision considérés (jusqu'à 10^{-7}). Nous réalisons la même expérience avec cette fois un effet d'une covariable, qui est testée ensuite. Nous sommes intéressés par la capacité de la régression robuste à retrouver cet effet, que nous choisissons subtile pour une meilleure illustration des performances vis-à-vis de l'OLS. Les courbes ROC obtenues prouvent la supériorité de la régression robuste, notamment quand beaucoup d'observations sont aberrantes. Les résultats sont du même ordre si nous considérons les données synthétiques plus réalistes présentées au chapitre précédent.

Applications

Nous appliquons la régression robuste à une expérience de neuroimagerie génétique contenant 400 sujets et 10 covariables. L'une d'entre elle, un SNP, est testée. Nous effectuons plusieurs analyses par régions d'intérêt en faisant varier le nombre de parcelles (nous réalisons 10 expériences pour un nombre de parcelles fixé). Nous obtenons des parcellisations aléatoires en suivant la méthode exposée au chapitre précédent : nous choisissons aléatoirement des sujets que nous utilisons avec un algorithme de classification de Ward. La régression robuste trouve toujours plus de parcelles possédant un effet significatif que OLS. Comme nous contrôlons la spécificité du test, nous interprétons ce résultats comme la preuve d'une sensibilité accrue. Nous observons que les résultats relatifs ne varient plus à partir du moment où 1000 parcelles sont considérées. Nous poussons l'analyse plus loin en utilisant RPBI sur le même problème, ainsi qu'une version robuste de RPBI, *RPBI (RLM)*. Nous remplaçons simplement la régression OLS par une régression robuste de type Huber dans la procédure, ce qui est possible parce que nous avons une procédure de test analytique (et donc moins coûteuse en calcul) à notre disposition. Les résultats obtenus avec la version robuste montrent plus d'activité et offrent encore une fois plus de sensibilité.

Nous considérons ensuite une analyse de groupe avec RPBI sur un problème mettant en oeuvre une variable comportementale à confronter aux données d'imagerie. 100 parcellisations aléatoires de 1000 parcelles chacunes sont envisagées. Nous utilisons RPBI avec de la régression OLS, comme au chapitre précédent, mais aussi avec de la régression robuste, afin de comparer les performances des deux procédures complexes. Il est à noter que RPBI classique offre déjà plus de sensibilité que n'importe quel autre type d'analyse. Il apparaît une activation dans le thalamus droit avec RPBI (OLS) et pas avec RPBI (RLM). En regardant de plus près les résultats, nous voyons que le signal dans les parcelles correspondant au thalamus souffre de la présence d'une observations très largement déviantes qui n'est pourtant pas détectée par n'importe quelle procédure de nettoyage des données. Ses caractéristiques son trop complexe pour qu'elles puissent être considérées comme anormales. L'intégration de la régression robuste à la méthode RPBI permet de corriger d'éventuelle erreurs et de rattraper des violations d'hypothèses fortes mais néanmoins indétectables autrement que par chance ou exploration exhaustive des données. Nous prouvons donc l'intérêt

de combiner régression robuste avec des méthodes d'analyse plus complexes.

Conclusion

Dans cette thèse, nous montrons qu'il est possible de détecter les observations aberrantes automatiquement en neuroimagerie, bien que les données mise en oeuvre soient de grande dimension. Nous proposons des outils adaptés pour ce faire, ainsi que d'autres outils plus fins permettant non pas de prendre une décision, mais d'explorer la structure des données pour guider la décision. Nous développons par ailleurs une méthode d'analyse qui rende les résultats de neuroimagerie plus reproductibles, ce qui constitue une autre forme de robustesse pour les outils. Cette méthode peut en outre bénéficier d'évolutions comme l'intégration en son sein de méthodes de régression robustes. Nous avons montré que ce type de régression a tout intérêt à être utilisé de manière générale en neuroimagerie, et à plus forte raison couplé avec les méthodes complexes d'analyse, dont celle que nous avons proposée. Nos applications démontrent qu'en combinant détection d'observations déviante, analyse reproductible et méthodes statistiques robustes, les études de neuroimagerie sont prémunies contre un ensemble beaucoup plus large de déviations aux hypothèses classiques et deviennent ainsi plus fiables. Aucun des points ci-dessus ne saurait être laissé pour compte pour des résultats optimaux, bien que chacun d'entre eux peut bénéficier de développements futurs visant à les améliorer encore. Nous pensons ainsi ouvrir la voie à une recherche plus approfondie dans le domaine des analyses robustes en neuroimagerie et petit à petit diminuer l'influence de la variabilité inter-individuelle en neuroimagerie.

Bibliographie

[1] Da Mota, B., Frouin, V., Duchesnay, E., Laguitton, S., Varoquaux, G., Poline, J.B., Thirion, B., et al. : A fast computational framework for genome-wide association studies with neuroimaging data. In : 20th International Conference on Computational Statistics (COMPSTAT 2012) (2012)

[2] Daszykowski, M., Kaczmarek, K., Heyden, Y.V., Walczak, B. : Robust statistics in data analysis – A review : Basic concepts. Chemometrics and Intelligent Laboratory Systems 85(2), 203–219 (2007)

[3] Friedman, J., Hastie, T., Tibshirani, R. : Sparse inverse covariance estimation with the lasso. ArXiv e-prints (Aug 2007)

[4] Hamilton, W.C. : The revolution in crystallography. Science 169, 133–141 (jul 1970)

[5] Hanley, J.A., McNeil, B.J. : The meaning and use of the area under a receiver operating (ROC) curve characteristic. Radiology 143(1), 29–36 (1982)

[6] Hardin, J., Rocke, D.M. : The distribution of robust distances. Journal of Computational and Graphical Statistics 14(4), 928–946 (2005)

[7] Holmes, A., Friston, K. : Generalisability, random effects & population inference. Neuroimage 7, S754 (1998)

[8] Huber, P.J. : Robust Statistics, chap. 7, p. 149. John Wiley & Sons, Inc. (2005)

[9] Ledoit, O., Wolf, M. : A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis 88(2), 365–411 (2004)

[10] Mourão-Miranda, J., Hardoon, D.R., Hahn, T., Marquand, A.F., Williams, S.C., Shawe-Taylor, J., Brammer, M. : Patient classification as an outlier detection problem : An application of the one-class support vector machine. NeuroImage 58(3), 793–804 (2011)

[11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É. : Scikit-learn : Machine Learning in Python. Journal of Machine Learning Research (Oct 2011)

[12] Peña, D., Prieto, F.J. : Multivariate outlier detection and robust covariance matrix estimation. Technometrics 43, 286–310 (2001)

[13] Rousseeuw, P.J. : Least median of squares regression. J. Am Stat Ass 79, 871–880 (1984)

[14] Roux, N.L., Bach, F. : Local component analysis. CoRR abs/1109.0093 (2011)

[15] Stein, J.L., Hua, X., Lee, S., Ho, A.J., Leow, A.D., Toga, A.W., Saykin, A.J., Shen, L., Foroud, T., Pankratz, N., et al. : Voxelwise genome-wide association study (vGWAS). Neuroimage 53(3), 1160–1174 (2010)

[16] Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., Poline, J.B. : Dealing with the shortcomings of spatial normalization : multi-subject parcellation of fMRI datasets. Hum Brain Mapp 27(8), 678–693 (Aug 2006)

[17] Tibshirani, R. : Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288 (1994)

[18] Vounou, M., Nichols, T.E., Montana, G. : Discovering genetic associations with high-dimensional neuroimaging phenotypes : a sparse reduced-rank regression approach. Neuroimage 53(3), 1147–1159 (2010)

[19] Wager, T.D., Keller, M.C., Lacey, S.C., Jonides, J. : Increased sensitivity in neuroimaging analyses using robust regression. NeuroImage 26, 99 (2005)

[20] Ward, J. : Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301), 236–244 (1963)

[21] Zweig, M., Campbell, G. : Receiver-operating characteristic (ROC) plots : A fundamental evaluation tool in clinical medicine. Clin Chem 39(4), 561–577 (1993)