Diagnostique d'homogénéité et inférence non-paramétrique pour l'analyse de groupe en imagerie par résonance magnétique fonctionnelle

Sébastien Mériaux

N° D'ORDRE : 8876

THÈSE DE DOCTORAT

SPÉCIALITÉ : PHYSIQUE

École Doctorale « Sciences et Technologies de l'Information des Télécommunications et des Systèmes »

PRÉSENTÉE PAR : Sébastien MÉRIAUX

SUJET : DIAGNOSTIQUE D'HOMOGÉNÉITÉ ET INFÉRENCE NON-PARAMÉTRIQUE POUR L'ANALYSE DE GROUPE EN IMAGERIE PAR RÉSONANCE MAGNÉTIQUE FONCTIONNELLE

Soutenue **le 6 Décembre 2007** devant les membres du jury :

MM.	Michel DOJAT	*Rapporteur*
	Ali MOHAMMAD-DJAFARI	*Président du jury*
	Habib BENALI	*Examinateur*
	Stéphane LEHÉRICY	*Examinateur*
	Alexis ROCHE	*Encadrant de thèse*
	Jean-Baptiste POLINE	*Directeur de thèse*

Remerciements

Avant tout, je voudrais remercier mon entraîneur Jean-Baptiste Poline de m'avoir accordé toute sa confiance en m'intégrant dans son équipe MADIC. Il a su revêtir son maillot de directeur de thèse avec détermination pour me prodiguer conseils et encouragements avisés, afin que je puisse prendre la mesure de mon sujet de thèse, un adversaire coriace pour un match très enrichissant.

Tour à tour pilier droit et pilier gauche pour son soutien permanent dans la mêlée des difficultés à résoudre, talonneur pourvoyeur de bonnes idées, deuxième ligne plus qu'énergique dans son nettoyage de manuscrit poussiéreux, troisième ligne extrêmement présent dans les phases de travail intense, demi de mêlée et demi d'ouverture très inspiré dans ses nombreux choix scientifiques, trois-quart centre et trois-quart aile capable de balayer tout le terrain des compétences requises pour dominer ce sujet, et enfin arrière impeccable dans ses relances incessantes pour maintenir intacte rigueur et motivation, toute ma reconnaissance et ma gratitude à Alexis Roche, un encadrant tout aussi sympathique qu'efficace, avec qui j'ai eu un immense plaisir à travailler. Ce sujet de thèse, sans être ovale, a parfois eu quelques rebonds capricieux mais les coups de pieds tactiques de ce fan de football ont toujours su me remettre dans le sens de la marche. Encore merci à lui.

Cette thèse est sans conteste le fruit d'un travail d'équipe et je veux aussi remercier tous ceux qui n'ont pas hésité à mouiller le maillot et à venir sur le terrain pour me prêter main forte, me conseiller ou tout simplement échanger des idées au détours d'un arrêt de jeu bien mérité. Un grand merci donc à Philippe Ciuciu et à Bertrand Thirion pour leur disponibilité, et foule d'autres mercis pour leur temps et leur bonne humeur à Dimitri Papadopoulos-Orfanos, Édouard Duchesnay, Yann Cointepas, Denis Rivière, Jean-François Mangin, Isabelle Denghien, Pascal Cathier, Salima Makni, Cyril Poupon et Guillaume Flandin. Une bien belle équipe sur le papier comme sur le terrain !

Une pensée émue pour Pierre-Jean LAHAYE, compagnon thésard de la première heure, et plein d'encouragements à Merlin KELLER, qui a su reprendre le flambeau avec brio.

Et comment jouer sans ballon ? Je me dois de remercier tous ceux qui ont eu l'extrême gentillesse de me fournir leurs jeux de données : Philippe PINEL, Claire SERGENT, Claire LANDMANN, Véronique IZARD, Anne-Dominique DEVAUCHELLE, Olivier SIMON, Eric ARTIGES, Gayannée KÉDIA, Isabelle KLEIN, Helmut LAUFS, Sylvain TAKERKART, Russell POLDRACK. Un grand merci tout particulier à Christophe PALLIER, Narly GOLESTANI, Ghislaine DEHAENE-LAMBERTZ, Patricia ROMAIGUÈRE et Jean-Luc ANTON, qui ont eu la curiosité de s'intéresser à mon travail dès ses débuts et la patience nécessaire pour en attendre les premiers résultats.

Et que dire des journalistes sportifs qui acceptent de venir analyser le match ? Keith WORSLEY et Michel DOJAT ont eu la tâche d'examiner scrupuleusement ce document et d'en rapporter la teneur. Qu'ils soient ici vivement remerciés pour ce travail essentiel.

Un grand merci aussi à tous les pensionnaires (Gwënaelle DOUAUD, Albertine DUBOIS, Silke DODEL, Julien DAUGUET, Thierry DELZESCAUX, Sébastien JAN, Frédéric BATAILLE, Luis JANEIRO) de ce vestiaire bien surchargé qu'était le bureau B114 du SHFJ. La bataille du climatiseur y fut parfois intense mais cela n'a jamais empêché l'ambiance de travail d'être toujours chaleureuse. Bonne continuation à tous.

Et bien évidemment je ne peux oublier ce fameux « 16ème homme », ce public fidèle de supportrices et de supporters, toujours enthousiaste, d'un soutien permanent et indéfectible. Ami(e)s, parents et frères, qu'il est doux et agréable de vous savoir toujours à mes côtés.

Et enfin ma tendre et douce Carole, sans aucun doute la plus passionnée des supportrices. Du fond du cœur, je te remercie pour toutes tes attentions si chaleureuses et ta formidable patience dans les moments de doute. Existe-t'il de bien plus précieux ?

TQMA...

Chapitre 1

Introduction

Au cours de cette thèse, nous avons travaillé sur des méthodes statistiques d'analyse de groupe dans le cadre des études de neuro-imagerie par résonance magnétique fonctionnelle (IRMf). Nous présentons succintemment le domaine de la neuro-imagerie, et plus particulièrement l'IRMf, avant de présenter la problématique de cette thèse et nos contributions.

1.1 La neuro-imagerie fonctionnelle

L'imagerie cérébrale est devenue au cours de ces dernières années un nouveau champ d'investigation en plein essor, comme le confirme le prix Nobel de médecine 2003 attribué à Paul Lauterbur et Peter Mansfield pour leurs travaux sur l'imagerie par résonance magnétique (IRM). Les avancées technologiques et les nouvelles méthodes de traitement du signal ont permis le développement de nouvelles techniques d'exploration non invasives de l'activité cérébrale. Le type de signal mesuré (électrique, magnétique, métabolique ou hémodynamique) détermine les résolutions temporelles et spatiales accessibles lors de l'acquisition des images.

Classiquement, il existe deux champs d'investigation principaux en imagerie cérébrale : l'imagerie anatomique qui permet d'étudier les structures corticales et

l'imagerie fonctionnelle qui permet d'étudier des processus cognitifs. En fonction de leurs spécificités (résolutions spatiale et temporelle notamment), les différentes techniques d'imagerie sont plutôt dédiées aux études anatomiques (très bonne résolution spatiale) ou aux études fonctionnelles (très bonne résolution temporelle), voire permettent les deux comme l'IRM.

Suivant le signal mesuré, il est possible de classer les différentes techniques d'imagerie fonctionnelle : certaines d'entre elles mesurent l'activité électromagnétique du cerveau comme l'électroencéphalographie (EEG) et la magnétoencéphalographie (MEG), d'autres mesurent une activité métabolique indirectement liée à l'activité électrique du cerveau comme la tomographie par émission monophotonique (TEMP ou SPECT en anglais pour *Single Photon Emission Computed Tomography*), la tomographie par émission de positons (TEP ou PET en anglais pour *Positron Emission Tomography*) et l'imagerie par résonance magnétique fonctionnelle (IRMf ou fMRI en anglais pour *functional Magnetic Resonance Imaging*). Depuis les années 1990 et les travaux de Ogawa *et al.* [1990] notamment, la TEMP et la TEP sont concurrencées par l'IRMf dont les principaux avantages sont qu'elle ne nécessite pas l'injection de substances radioactives, que le temps d'acquisition d'un volume complet est très court et qu'elle permet d'accéder à des résolutions spatiales plus fines. Les techniques d'EEG et MEG possèdent quant à elles une très bonne résolution temporelle (de l'ordre de la milliseconde) mais la résolution spatiale accessible est de l'ordre du centimètre ce qui rend la localisation des sources d'activité neuronale assez imprécise. De plus amples informations sur la MEG et l'EEG sont disponibles par exemple dans Baillet *et al.* [2001].

Durant cette thèse, nous avons exclusivement travaillé sur des données d'IRMf : cette modalité permet d'obtenir de manière non-invasive et *in vivo* des images tridimensionnelles de l'activité cérébrale chez l'homme, ayant à la fois une bonne résolution spatiale (de l'ordre de quelques millimètres) et une bonne résolution temporelle (de l'ordre de quelques secondes). De plus, elle peut facilement être couplée avec l'IRM anatomique (en utilisant le même scanner), permettant ainsi de détecter les aires activées dans les images fonctionnelles et de les reporter sur l'IRM anatomique pour la visualisation et l'interprétation des résultats. L'IRMf devient donc une des méthodes privilégiées des neuro-sciences pour étudier les bases neurales du fonctionnement cognitif, sensoriel ou moteur, comme le montre

le nombre croissant des publications utilisant cette modalité (voir la figure 1.1).

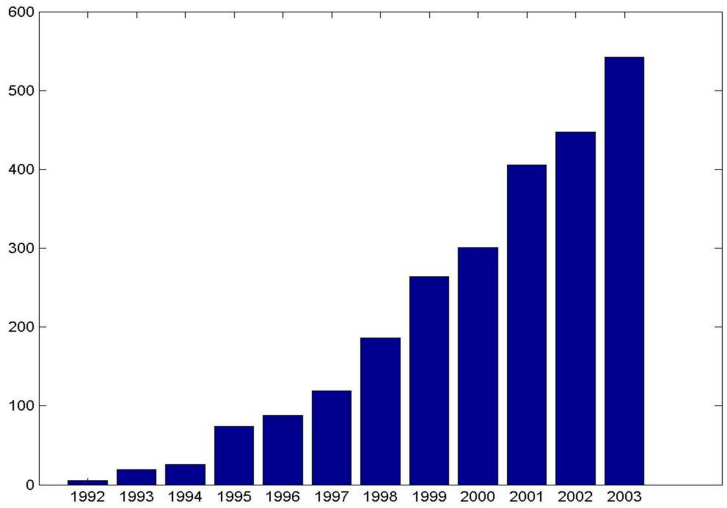

Figure 1.1 – *Nombre d'articles contenant le mot « IRMf » dans leur titre publiés chaque année (Source* PubMed*)*

1.2 Principes généraux d'une étude d'IRM fonctionnelle

L'un des objectifs principaux des neuro-sciences est de localiser les zones associées à certaines fonctions cognitives, ainsi que leurs différents modes d'interaction nécessaires à la réalisation de tâches fondamentales, telles que la mémoire, le langage ou la vision. Le cerveau présentant une très grande variabilité anatomo-fonctionnelle inter-individuelle, les chercheurs tentent de décrire l'anatomie et de localiser précisément les zones fonctionnelles chez chaque sujet, puis une analyse de groupe permet de généraliser les résultats à une population d'intérêt.

1.2.1 L'imagerie par résonance magnétique fonctionnelle

L'imagerie par résonance magnétique (IRM) a été inventée au début des années 1970 par Paul Lauterbur [Lauterbur, 1973] et a connu depuis lors un essor considé-

rable en permettant l'acquisition d'images tridimensionnelles de grande précision anatomique. Sous le terme IRM sont regroupées en fait diverses modalités qui ont en commun d'être toutes acquises avec le même scanner : images anatomiques (pondérées en T_1 ou en T_2), images fonctionnelles (pondérées en T_2^*), images de diffusion (orientation des principaux faisceaux de fibres de matière blanche), angiographie IRM (visualisation des vaisseaux cérébraux généralement en utilisant un produit de contraste), images de perfusion (distribution du sang), spectroscopie RMN (distribution de certains métabolites).

Le signal mesuré en IRM provient du signal de résonance magnétique nucléaire des protons de l'eau (les molécules d'eau représentant environ 80 % du poids du cerveau). Placés dans un champ magnétique principal B_0, ces protons acquièrent une certaine aimantation M_0 proportionnelle à la valeur du champ et animée d'un mouvement de précession autour de l'axe principal du champ B_0. La fréquence de cette précession, appelée fréquence de Larmor, est la fréquence à laquelle une onde radio-fréquence peut exciter les protons et provoquer un basculement de l'aimantation M_0. La dynamique du retour à l'équilibre de M_0 produit une onde électromagnétique (mesurable par la même antenne qui génère l'onde radio-fréquence excitatrice) qui dépend des propriétés locales des tissus via deux constantes de temps T_1 et T_2. Différentes séquences d'imagerie permettent alors d'obtenir des images de contraste différent (T_1, T_2, T_2^* ou densité de protons) en choisissant le temps de répétition T_R (temps séparant deux impulsions radio-fréquence) et le temps d'écho T_E (temps séparant deux mesures de l'aimantation) adéquates. Pour plus de précisions sur les techniques d'imagerie par résonance magnétique nucléaire, se reporter par exemple à Stark et Bradley [1992] ; van de Moortele [1999] ; Houdé *et al.* [2002].

Dans le cas de l'IRM fonctionnelle (IRMf), le temps d'acquisition est essentiel et doit être inférieur à la constante de temps associée au phénomène biologique qui est observé. La séquence EPI (*Echo Planar Imaging*) en écho de gradient, mise au point par Mansfield [1977], permet d'acquérir une coupe entière avec une seule impulsion radio-fréquence, grâce à des inversions rapides de gradient. Il est ainsi possible d'acquérir une coupe en une centaine de millisecondes. Par contre, le temps d'acquisition limité à un seul T_R par volume impose de fait une limite à la résolution spatiale accessible en EPI (classiquement de l'ordre de 3 à 4 mm) ;

la résolution sera moindre qu'avec une séquence en *spin echo* (classiquement de l'ordre de 1 mm), plutôt dédiée à l'acquisition d'une image anatomique pondérée en T_1.

1.2.2 Bases physiologiques du signal en IRM fonctionnelle : l'effet BOLD

À la fin du siècle dernier, Roy et Sherrington [1890] observèrent une modification locale de la couleur du cortex de sujets soumis à des stimulations. Ce changement de couleur s'explique par le fait que le sang artériel (saturé en oxygène) est rouge vif tandis que le sang veineux (appauvri en oxygène) est bleu violacé. Ces observations suggèrent un couplage entre l'activité électrique des neurones et le débit sanguin : c'est le couplage hémodynamique. L'activité cognitive du cerveau engendrerait des variations du volume sanguin cérébral régional mais aussi de la concentration du sang en oxygène.

De plus le glucose, source d'énergie pour le cerveau, n'est quasiment pas stocké par les tissus corticaux, donc un apport permanent par le système sanguin est nécessaire. Ainsi en raison du couplage hémodynamique observé dans des conditions de perfusion normale, une augmentation de la consommation de glucose et donc du débit sanguin cérébral régional reflète une activité synaptique locale.

Or il s'avère que la désoxy-hémoglobine (hémoglobine sans oxygène) présente dans le sang est paramagnétique alors que l'oxy-hémoglobine (hémoglobine avec oxygène) est diamagnétique. La molécule de désoxy-hémoglobine peut être utilisée comme agent de contraste endogène. Une image pondérée en T_2^* sera sensible à la concentration d'oxygène dans le sang : les tissus contenant des vaisseaux riches en désoxy-hémoglobine présenteront un signal IRM moindre que les tissus contenant des vaisseaux riches en oxy-hémoglobine. Le signal ainsi mesuré est le signal BOLD (*Blood Oxygen Level Dependent*) utilisé en IRMf [Ogawa *et al.*, 1990].

L'effet BOLD fait référence à une relation entre débit sanguin cérébral régional et consommation d'oxygène. À la suite d'une activation, une augmentation significative du débit sanguin est observée mais elle n'est pas suivie de la même augmentation significative du métabolisme de l'oxydation du glucose : l'augmentation du débit sanguin semble être totalement disproportionnée avec les besoins

effectifs en oxygène des neurones irrigués [Kim et Ugurbil, 1997]. Une des hypothèses émises est que le métabolisme du glucose serait quasi-anaérobique [Fox *et al.*, 1988], expliquant la faible surconsommation d'oxygène des neurones en activité par rapport aux neurones au repos, mais cela reste à démontrer. Quelle que soit l'explication de ce découplage entre l'augmentation du débit sanguin et la consommation d'oxygène, c'est ce phénomène que mesure le signal BOLD. En effet, l'excès de volume sanguin associé à une faible augmentation de l'extraction de l'oxygène entraîne une diminution de la concentration en désoxy-hémoglobine et donc une modification locale du champ magnétique mesurable par l'IRM T_2^*. En particulier, la diminution de la concentration du sang en désoxy-hémoglobine va engendrer un temps de relaxation T_2^* plus long, puisque le champ magnétique local sera plus homogène, et donc une augmentation du signal IRM.

La fonction de réponse hémodynamique (HRF, *Hemodynamic Response Function*) à un stimulus est observable, en un voxel donné, à partir du décours temporel du signal BOLD mesuré en ce voxel (voir figure 1.2). Cette réponse est faible (de l'ordre de quelques % du signal total), son maximum est retardé d'environ 4 à 7 secondes par rapport au temps de présentation du stimulus et elle peut durer jusqu'à 25 voire 30 secondes.

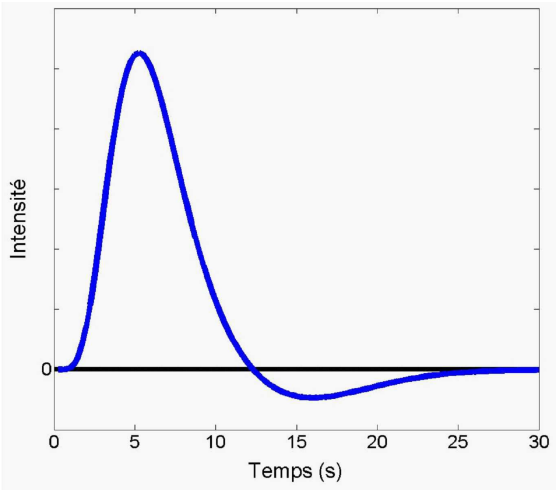

Figure 1.2 – *Fonction de réponse hémodynamique à un stimulus bref mesurée à partir du signal BOLD en un voxel donné*

Les origines physiologiques des phénomènes dits de déplétion du signal (phénomènes d'*undershoot*), observés à partir de 10 à 15 secondes après la présentation du stimulus, ne sont pas encore clairement établis. De manière plus générale, les conditions du couplage hémodynamique restent toujours controversées et les relations liant activité neuronale, débit sanguin, volume sanguin, concentration en oxygène et métabolisme cérébral ne sont toujours pas définitivement comprises.

1.2.3 Mise au point du protocole expérimental

L'IRMf permet de mesurer de manière indirecte l'activité cérébrale d'un sujet par l'intermédiaire du signal BOLD. Le but d'une étude d'IRMf est alors de mettre en évidence les zones cérébrales activées lors de l'exécution d'une tâche spécifique. Pour cela, la mise au point du protocole expérimental est une étape primordiale pendant laquelle l'expérimentateur doit définir précisément la question d'intérêt et concevoir le paradigme expérimental permettant le test de la problématique retenue. De plus, cette étape doit intégrer les contraintes induites par l'utilisation de l'IRMf.

Classiquement les problématiques étudiées en IRMf concernent l'étude de processus cognitifs, sensoriels ou moteurs particuliers, soit sur des sujets sains, soit par comparaison d'un groupe de sujets atteints d'une pathologie particulière avec un groupe de sujets témoins. Ces problématiques cherchent à tester directement l'activité d'une région spécifique du cerveau en étudiant la dynamique de la réponse BOLD mesurée, ou à localiser les aires impliquées dans la réalisation d'un processus cognitif particulier, ou encore à mettre en évidence des relations de connectivité entre régions.

La conception du paradigme expérimental (ensemble des stimuli présentés au sujet et des tâches qu'il a à effectuer) doit prendre en compte la nature du signal BOLD mesuré en IRMf. En tenant compte des caractéristiques fréquentielles différentes entre la réponse BOLD (agissant plutôt comme un filtre passe-bas comme le montrent Friston *et al.* [1995b, 2000]) et le bruit (agissant plutôt comme un filtre passe-haut comme le montrent Aguirre *et al.* [1997] ; Zarahn *et al.* [1997]), D'Esposito *et al.* [1999] distinguent deux catégories de paradigmes : les paradigmes en bloc et les paradigmes événementiels.

Les paradigmes en bloc sont les plus communément utilisés. Dans un paradigme en bloc, les essais relatifs à une tâche donnée de l'expérience sont regroupés ensemble ou sont continus afin de former un bloc dont la durée est étendue [Bandettini *et al.*, 1992]. Généralement le paradigme se compose de plusieurs blocs, chacun d'une durée de 15 à 30 secondes, l'intervalle de temps séparant deux essais étant d'environ 1 à 2 secondes. Ainsi, selon les propriétés de la réponse BOLD, les essais au sein d'un même bloc induisent une réponse soutenue. De plus, les changements de tâches (et en l'occurrence les changements de blocs) se produisent de manière espacée dans le temps. Il est alors possible de démontrer que grâce à ces caractéristiques, ce type de paradigme a un bon pouvoir statistique de décision. Il semble également que du point de vue cognitif, certains types d'expériences donnent de meilleurs résultats dans la version en blocs. Ce sont notamment les paradigmes fondés sur les processus cognitifs à évolution lente (changement d'état attentionnel ou émotionnel, adaptation à une tâche donnée,...). La construction du paradigme et l'analyse des données sont aussi facilitées par la version en blocs, tout comme les instructions et l'exécution des tâches par le sujet. Un de leurs inconvénients majeurs est qu'il est impossible de discriminer certains effets cognitifs importants dus à la répétition de la tâche car ils sont confondus avec l'effet bloc (par exemple, un changement de stratégie dans l'exécution de la tâche). Il est également difficile de mettre en évidence les effets au sein d'un même bloc sachant que plusieurs phénomènes (stratégie, attention,...) peuvent se produire si le bloc dure longtemps [Bandettini *et al.*, 1997].

Les paradigmes événementiels ont été inventés pour pallier les déficiences des paradigmes en bloc et ont notamment pour but d'étudier les changements rapides induits par un seul type d'essai. À l'inverse des paradigmes en bloc, au cours de l'expérience, les différents types de stimuli ou de conditions sont mélangés et présentés généralement aléatoirement. La présentation aléatoire rend les stimuli imprévisibles par le sujet, minimisant les effets d'anticipation et d'habituation [Burock *et al.*, 1998 ; D'Esposito *et al.*, 1999 ; Josephs et Henson, 1999]. D'après les caractéristiques de la réponse hémodynamique et du bruit, ce type de paradigme est moins efficace en terme de détection mais il permet d'estimer la dynamique temporelle du signal. Il est ainsi possible, sous certaines conditions, d'extraire et de comparer les caractéristiques de la réponse BOLD (latence, pic de la réponse,...)

relativement à différents types de stimuli ou de tâches. De plus, grâce à cette focalisation sur la dynamique temporelle, il devient également possible de décomposer le déroulement de l'exécution de la tâche par le sujet en différentes phases correspondant à plusieurs sous-composantes. L'inconvénient majeur de ces paradigmes est leur faible pouvoir statistique. Plus que pour les paradigmes en bloc, ces paradigmes dépendent fortement des hypothèses de linéarité de la réponse BOLD. Plus difficiles à analyser, ils dépendent souvent des *a priori* sur la forme de la réponse.

1.2.4 Choix de la population d'intérêt

Une fois le paradigme expérimental mis au point, l'expérimentateur doit définir la population d'intérêt sur laquelle il souhaite mener son étude. La première contrainte est de définir les critères nécessaires pour choisir les sujets représentatifs de la population d'intérêt : hommes ou femmes, droitiers ou gauchers, niveau d'études, âge,... S'il s'agit de comparer des sujets témoins avec des sujets atteints d'une pathologie particulière, il faut également définir précisément les critères de sélection des malades pouvant participer à l'étude, en fonction de différents paramètres comme l'avancement de la maladie, la nature des symptômes,...

La deuxième contrainte est le choix du nombre de sujets à inclure dans l'étude, sachant que pour des raisons économiques, ce choix ne peut le plus souvent pas excéder 15 à 20 sujets. Mis à part le problème du coût, reste à trouver le compromis entre une puissance statistique suffisante (suggérant de choisir le plus de sujets possibles) et le risque de voir apparaître de nombreux sous-groupes dans la population de sujets scannés à cause de la forte variabilité. Remarquons que cette question du nombre de sujets à inclure dans une étude d'IRMf demeure un sujet de recherche en soi [Woods, 1996 ; Friston *et al.*, 1999 ; Desmond et Glover, 2002 ; Savoy, 2006 ; Thirion *et al.*, 2006b, 2007].

1.2.5 Analyse des données d'IRMf

Pour chaque sujet participant à l'étude, une série temporelle d'images d'IRMf pondérées en T_2^* est acquise pendant qu'il effectue les tâches spécifiées par le paradigme expérimental. Cette série d'images constitue les données brutes du sujet

qu'il faut analyser en tenant compte de la problématique à tester et des paramètres du paradigme expérimental.

L'analyse des données d'IRMf peut se décomposer en deux niveaux : une analyse intra-sujet permet d'obtenir pour chaque sujet une carte d'activation, correspondant à la détection des aires cérébrales impliquées dans l'exécution d'une tâche pour un sujet particulier. Le problème à résoudre pour cette analyse est de déterminer, à partir des décours temporels du signal BOLD dans la série d'images d'IRMf, si certains voxels sont activés lors de l'exécution de la tâche. L'hypothèse de base est que chaque stimulation induit localement une réponse BOLD et à partir de là, il faut construire un modèle de la relation stimuli-réponse BOLD.

Le deuxième niveau, qui constitue le cadre général de travail pour cette thèse, est une analyse de groupe, qui a pour but de résumer les résultats obtenus au niveau intra-sujet en une carte d'activation du groupe, afin de pouvoir détecter les aires cérébrales impliquées dans l'exécution d'une tâche pour la population d'intérêt.

1.3 Problématiques de l'analyse de groupe

L'analyse de groupe à effets aléatoires, par opposition à l'analyse à effets fixes qui sera présentée dans le chapitre 3, vise à généraliser les cartes d'activation obtenues sur une cohorte de sujets pour en tirer des conclusions sur le fonctionnement cérébral au sein de la population d'intérêt. Ce type d'analyse cherche donc à s'affranchir en premier lieu de la variabilité inter-sujets des réponses BOLD.

L'approche standard pour résoudre ce problème, mise en œuvre dans de nombreux logiciels de traitements de données d'IRMf (SPM, FSL, AFNI,...), repose sur un test paramétrique de Student (test t) ou de Fisher (test F), appliqué successivement en chaque voxel d'une grille « de référence » (approche dite massivement univariée). Pour cela, il est nécessaire au préalable de recaler les différents sujets vers un « cerveau moyen », le plus souvent choisi comme le *template* du Montreal Neurological Institute. Nous discuterons au chapitre 2 des hypothèses simplificatrices qui sous-tendent ce recalage, bien que ne les remettant pas en cause dans cette thèse.

En revanche, cette thèse propose un examen critique de l'hypothèse justifiant l'utilisation des tests t (ou F), à savoir que les réponses estimées sont distribuées indépendamment et identiquement selon une loi normale. En effet, cette hypothèse revient à une condition d'*homogénéité* de l'échantillon dont nous montrerons au chapitre 6 qu'elle est fréquemment mise en défaut dans la pratique.

Cette observation corrobore les travaux de plusieurs auteurs tels que Holmes *et al.* [1996] ; Nichols et Holmes [2002], qui ont proposé l'emploi de tests de permutations à la place du test t paramétrique, de façon à garantir un contrôle exact du risque de faux positifs sous des hypothèses non-paramétriques très générales. Ces tests utilisent la statistique de Student comme statistique de décision mais fondent le calcul de sa distribution sur un mécanisme de permutations plutôt que sur l'hypothèse de normalité. Par conséquent, ces méthodes modifient la façon dont est seuillée la carte statistique associée au groupe, mais pas le calcul de la carte elle-même.

Or nous remarquons que, si les effets estimés sont fortement hétérogènes, le choix de la statistique de Student s'avère sous-optimal en terme de pouvoir de détection. Ceci a motivé le développement de nouvelles statistiques de décision plus adaptées à la nature des données d'IRMf.

1.4 Organisation du mémoire et contributions

La suite de ce mémoire est divisée en huit chapitres organisés de la manière suivante.

Le chapitre 2 présente les différentes étapes de l'analyse intra-sujet « standard » des données d'IRMf, telles qu'elles sont mises en œuvre dans le logiciel SPM. Le but de ce chapitre n'est pas de décrire la chaîne des traitements dans ses moindres détails, mais plutôt de mettre en évidence les sources d'erreur que ces traitements peuvent introduire et qui ont des répercussions sur l'analyse de groupe.

Le chapitre 3 est consacré à l'introduction du modèle hiérarchique d'inférence statistique à deux niveaux permettant de rendre compte des deux sources de variabilité intervenant dans l'analyse de groupe. Sont également présentées l'analyse à effets fixes et surtout l'analyse à effets aléatoires « standard » qui repose sur un test paramétrique de Student. L'examen des différentes hypothèses qui sous-tendent ce

type d'analyse permet d'expliquer les développements de nouvelles approches proposées pour l'analyse de groupe lors de cette thèse.

Dans le chapitre 4, nous proposons une méthode multivariée de diagnostique d'influence, fondée sur le calcul d'une matrice des distances inter-sujets à partir de leurs cartes d'effets estimés. Une mesure d'influence, la distance de Cook, est alors appliquée à cette matrice afin de révéler la présence de sous-groupes de sujets ou de sujets atypiques. Cette approche a fait l'objet d'une publication [Kherif *et al.*, 2004], qui est complétée dans ce chapitre par une étude complémentaire sur la valeur critique pour la distance de Cook.

Pour contrôler spécifiquement l'hypothèse d'homogénéité du groupe en chaque voxel, nous consacrons le chapitre 5 à la présentation d'un test de normalité univarié fondé sur la statistique de Grubbs. Afin de mieux contrôler la spécificité de ce test, notamment pour les petits échantillons de données, une méthode de calcul du seuil par simulations de Monte-Carlo est proposée.

Dans le chapitre 6, nous présentons les résultats des études d'homogénéité que nous avons réalisées à l'aide de notre méthode multivariée de diagnostique d'influence et du test de normalité de Grubbs sur une vingtaine de jeux de données d'IRMf, correspondant à un large éventail de paradigmes expérimentaux. La fréquence observée de détection de données atypiques, bien supérieure à celle attendue par simulations, permet de mettre en doute l'hypothèse de normalité sous-jacente à l'analyse de groupe « standard » [Mériaux *et al.*, 2003, 2004]. De plus, nos outils ont permis de donner des informations utiles à la compréhension de la variabilité inter-sujets observée et certains résultats ont ainsi fait l'objet de collaborations à la rédaction d'articles [Golestani *et al.*, 2006 ; Thirion *et al.*, 2007] ou de communications pour des conférences [Kherif *et al.*, 2003 ; Felician *et al.*, 2005].

La remise en cause de l'hypothèse de normalité suggère que le test de Student est non seulement sous-optimal du point de vue de la sensibilité (détection des vrais positifs) mais également biaisé en spécificité (contrôle des faux positifs). Dans le chapitre 7, nous proposons de nouvelles approches pour l'analyse de groupe fondées sur des statistiques de décision robustes pour améliorer la sensibilité, calibrées par un mécanisme de permutations qui permet de contrôler exactement la spécificité. Ces approches ont fait l'objet d'une publication [Mériaux *et al.*, 2006c] mais également d'une contribution à l'article de Dehaene-Lambertz *et al.* [2006b] lorsqu'elles

ont été appliquées à l'étude d'un jeu de données particulièrement hétérogènes impliquant des nourrissons.

Dans le chapitre 8, nous introduisons de nouvelles statistiques de décision dites à « effets mixtes » car elles permettent de prendre en compte les erreurs de mesure sur les contrastes BOLD, ceci afin de pondérer les différentes observations en fonction de leur fiabilité et de réduire ainsi le risque de mauvaise détection lié à l'existence de données atypiques. L'approche à « effets mixtes » a d'abord été envisagée en conservant une hypothèse de distribution normale des effets à travers les sujets [Mériaux *et al.*, 2006b], puis a été généralisée au cas des distributions non-paramétriques [Roche *et al.*, 2007].

Enfin, le dernier chapitre est consacré à une synthèse des différents développements méthodologiques réalisés au cours de cette thèse, ainsi qu'à la discussion de quelques pistes de recherche qui devraient permettre de poursuivre et de compléter ce travail.

Remarquons enfin qu'un important effort d'implémentation informatique a été maintenu tout au long de cette thèse, afin que l'ensemble des développements méthodologiques réalisés soit accessible à la communauté de neuro-imagerie à travers le logiciel *DISTANCE*[1].

[1] *DISTANCE* est un logiciel de diagnostique d'homogénéité et d'analyse de groupe développé sous Matlab et conçu comme une boîte à outil SPM. Il est accessible gratuitement sur le site suivant : `http ://www.madic.org/download/DISTTBx/DISTTBx_main.html`.

Chapitre 2

Analyse intra-sujet « standard »

Dans ce chapitre nous présentons les différentes étapes de l'analyse intra-sujet « standard » des données d'IRM fonctionnelle. Rappelons que pour chaque sujet participant à une étude d'IRMf, un volume cérébral de plusieurs milliers de voxels est acquis tous les T_R (temps de répétition) lors de l'expérience. Cet ensemble d'images 3D d'IRMf (images EPI T_2^*) constitue les données du sujet et leur analyse consiste à relier la mesure des décours temporels de l'activité cérébrale avec le paradigme expérimental.

Cette analyse intra-sujet « standard » peut se décomposer en deux étapes :

- une étape de **pré-traitements** (décrite en 2.1) qui permet de corriger certains artéfacts présents dans les images d'IRMf et de normaliser les données pour l'analyse de groupe ;
- une étape d'**analyse statistique intra-sujet** ou **1er niveau** (décrite en 2.2) qui utilise le modèle linéaire généralisé pour produire une carte d'effet estimé (carte de contraste) et une carte d'erreur d'estimation pour chaque sujet.

Ces étapes sont présentées ici telles qu'elles sont mises en œuvre dans SPM (*Statistical Parametric Mapping*), un des logiciels d'analyse de données d'IRMf le plus utilisé par la communauté de neuro-imagerie [Friston *et al.*, 1995c]. SPM est développé sous Matlab (The MathWorks, États-Unis) par le Wellcome Department

of Imaging Neuroscience (Londres, Royaume Uni) et est distribué gratuitement sur le site : `www.fil.ion.ucl.ac.uk/spm`.

Le but de ce chapitre n'est pas de décrire la chaîne de traitements dans ses moindres détails mais plutôt d'introduire les méthodes d'estimation et d'inférence paramétriques les plus usitées. La mise en évidence de leurs limitations permet également de présenter les différentes approches alternatives qui peuvent être envisagées.

2.1 Pré-traitements des données

Avant de procéder à l'analyse proprement dite des données, il est nécessaire d'appliquer certains pré-traitements aux images acquises. Comme le résume la figure 2.1, ces pré-traitements ont deux objectifs principaux : le premier est de corriger les artéfacts présents dans les images d'IRMf et le deuxième est de normaliser spatialement les données afin d'obtenir un espace commun à tous les sujets, permettant ainsi d'envisager l'analyse de groupe.

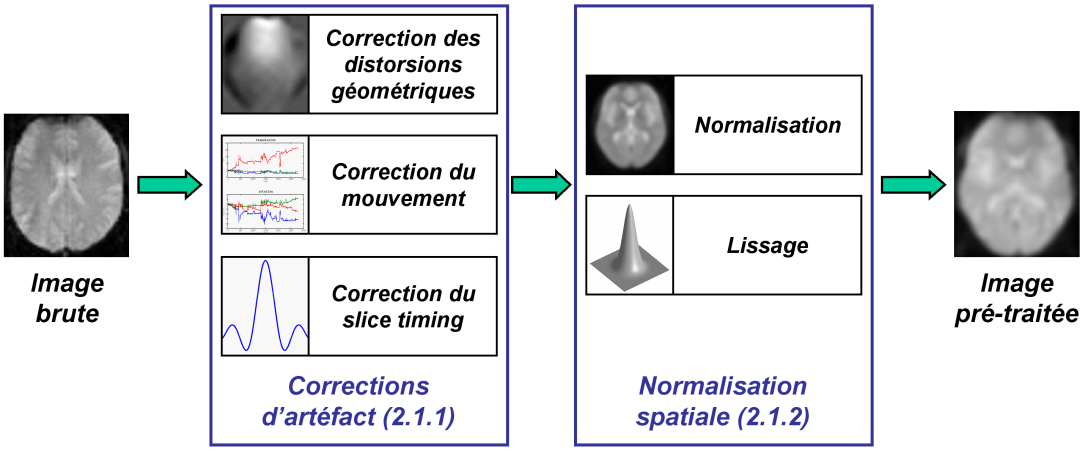

Figure 2.1 – *Schéma récapitulatif des principaux pré-traitements appliqués aux données brutes d'IRM fonctionnelle*

Même si nous discutons dans cette section la validité de l'approche SPM pour réaliser ces pré-traitements et présentons certaines approches alternatives, nous avons considéré tout au long de cette thèse ces pré-traitements comme optimaux. Ainsi, tous les résultats d'analyse de groupe que nous présentons ont été obtenus à partir des images pré-traitées avec SPM, c'est-à-dire corrigées du *slice timing*, réalignées, normalisées et lissées.

2.1.1 Correction d'artéfacts

Le principe de l'analyse « standard » des données d'IRMf est d'ajuster un modèle linéaire généralisé au décours temporel de l'activité cérébrale mesuré en chaque voxel. Une statistique de décision, calculée en formant le rapport du signal modélisé par la variance résiduelle, permet alors de déterminer si le voxel est actif lors de la réalisation du paradigme expérimental. Il est donc nécessaire d'identifier, puis de réduire voire d'éliminer tous les signaux de non-intérêt pour le modèle ajusté aux données. Ces artéfacts peuvent en effet perturber l'estimation du signal modélisé comme de la variance résiduelle et ainsi biaiser le calcul de la statistique de décision.

Tout d'abord, certains artéfacts occasionnels doivent être corrigés préalablement à l'acquisition des images d'IRMf : par exemple la présence d'arcs électriques (*spikes*) à cause des courants induits dans l'antenne radio-fréquence, la présence d'images fantômes (*ghosts*) à cause du phénomène de repli spectral, une perte de signal due à une mauvaise pénétration des ondes radio-fréquence en présence d'un casque EEG, une dérive de signal due aux variations de température des gradients, etc.

Dans cette section, nous présentons les trois types d'artéfacts présents dans les images d'IRMf acquises et que le logiciel SPM permet de corriger :

- les artéfacts dus aux inhomogénéités (susceptibilité magnétique, champ B_0 local) ;
- les artéfacts dus aux mouvements du sujet (volontaire, respiratoire, cardiaque,...) ;
- les artéfacts dus à l'acquisition coupe par coupe (*slice timing*).

2.1.1.1 Correction des distorsions géométriques

Les données d'IRMf sont affectées par des distorsions géométriques causées par les variations locales du champ magnétique principal B_0 et par les différences de susceptibilité magnétique entre les tissus imagés. Ces différentes sources d'inhomogénéité provoquent des variations dans les fréquences de résonance et donc dans les positions des voxels imagés puisque le codage de la position se fait grâce à la fréquence de résonance. Ces distorsions s'observent surtout au niveau des interfaces où les différences de susceptibilité magnétique sont très marquées, comme les interfaces tissu-air ou tissu-os [Jezzard et Clare, 1999].

De plus, l'acquisition des images d'IRMf se fait à l'aide d'une séquence d'écho de gradient de type EPI (*Echo Planar Imaging*, Mansfield [1977]). Cette séquence consiste à acquérir chaque coupe axiale (suivant l'axe Z) à l'aide d'une seule impulsion radio-fréquence. Le plan de Fourier est parcouru en acquérant une ligne suivant l'axe X en totalité (codage en fréquence à l'aide du gradient G_x) avant de passer à une nouvelle ligne suivant l'axe Y (codage de phase à l'aide du gradient G_y). Ainsi, les inhomogénéités de champ magnétique B_0 créent des erreurs de phase et donc des distorsions géométriques dans la direction du codage de phase puisque la fréquence de lecture suivant l'axe Y est beaucoup plus faible que suivant l'axe X.

Pour corriger ces distorsions géométriques, Jezzard et Balaban [1995] ; Reber *et al.* [1998] ont proposé une méthode fondée sur l'utilisation d'une carte de phase (*phase map*). Cette carte se construit en acquérant deux images EPI pour deux temps d'écho T_E différents, en calculant la différence de phase entre ces deux images puis en dépliant cette image de différence avec une interpolation des zones sans signal afin d'obtenir des valeurs entre $-\pi$ et π. Cette carte de phase est alors utilisée pour déterminer les corrections de position à appliquer aux différents voxels de l'image. Notons que ce type de correction par carte de phase est disponible dans une boîte à outils SPM dédiée (*FieldMap Toolbox*).

2.1.1.2 Correction du mouvement

Le mouvement du sujet pendant l'expérience est inévitable. Tout d'abord, de nombreux mouvements d'origine physiologique sont susceptibles de provoquer des

artéfacts dans les images d'IRMf. Comme l'ont montré Turner *et al.* [1998] ainsi que Nieto-Castanon *et al.* [2003], ces mouvements sont à l'origine d'un bruit basse-fréquence, les mouvements respiratoires provoquant un pic vers $0,25$ Hz, les mouvements cardiaques un pic vers 1 Hz et les autres mouvements physiologiques (péristaltiques,...) provoquant un bruit en $1/f$. Ces artéfacts physiologiques recouvrant des fréquences plus basses que le signal BOLD, ils peuvent être atténués par filtrage passe-haut du signal d'IRMf.

Ensuite, il reste le mouvement volontaire du sujet qui a pour effet de décaler les images acquises pour un même sujet. À cause de la durée relativement importante des séquences d'acquisition en IRMf, les mouvements de la tête sont quasiment inévitables et peuvent provoquer une perte de sensibilité dans la détection de l'activité cérébrale, ainsi que l'apparition de « faux positifs » liés à des fausses corrélations avec le paradigme (du point de vue technique du test d'hypothèse, ce sont en fait de vrais positifs). En effet, des voxels initialement détectés dans la matière blanche peuvent se retrouver dans la matière grise si un mouvement de la tête est mal corrigé. De plus, si ces mouvements se trouvent corrélés au paradigme expérimental, alors une partie de l'activité cérébrale détectée peut finalement se révéler n'être qu'un artéfact de mouvement, comme le montre la figure 2.2.

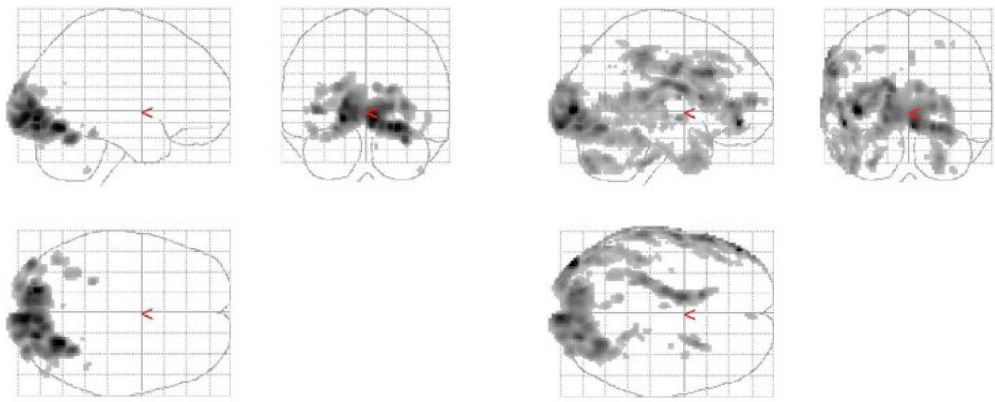

Figure 2.2 – *Cartes statistiques illustrant l'influence du mouvement sur la détection des activations : gauche (pas de mouvement) - droite (mouvement corrélé au paradigme expérimental) (Source* Jesper Andersson*)*

Afin de corriger ces artéfacts dus aux mouvements de la tête du sujet, une opération de « réalignement » consiste à recaler les images fonctionnelles en utilisant une image de référence (souvent une des premières images acquises ou la moyenne de plusieurs images acquises). La méthode la plus utilisée suppose un mouvement rigide de la tête du sujet et calcule alors la transformation rigide (3 paramètres libres de translation et 3 de rotation) qui minimise un critère de similarité (la moyenne des différences d'intensité de niveaux de gris au carré) entre l'image à recaler et l'image de référence. L'image corrigée du mouvement est ensuite calculée par ré-échantillonnage de l'image à recaler à laquelle est appliquée la transformation rigide. Le ré-échantillonnage consiste à estimer l'intensité de chaque voxel de l'image recalée par interpolation des intensités mesurées dans les voxels voisins de l'image à recaler. L'interpolation peut être tri-linéaire (moyenne pondérée des 8 voxels voisins en 3D) ou, plus généralement, réalisée par ajustement à l'image d'une combinaison linéaire de fonctions B-spline tensorielles [Unser, 1999 ; Thévenaz *et al.*, 2000].

Cette méthode de recalage présente plusieurs inconvénients. Tout d'abord, sous l'hypothèse optimiste de mouvements parfaitement rigides, l'étape de ré-échantillonnage introduit des erreurs d'interpolation, même dans le cas d'une interpolation B-spline d'ordre élevé (lorsque l'ordre tend vers l'infini, l'interpolation B-spline tend vers l'interpolation sinus cardinal qui ne permet de reconstruire exactement le signal que sous les hypothèses du théorème de Shannnon).

Ensuite, les coupes ne sont pas acquises au même moment : ainsi les mouvements rapides de la tête (d'une durée inférieure au T_R) ne sont pas pris en compte par un modèle rigide. Le logiciel de traitements de données d'IRMf FSL (*FMRIB's Software Library*) propose une méthode de recalage coupe par coupe développée par Jenkinson *et al.* [2002]. De plus, les artéfacts présents dans les images d'IRMf ne se déplacent pas de manière rigide. Comme l'illustre la figure 2.3 [Andersson *et al.*, 2001], les distorsions géométriques varient suivant l'orientation du sujet dans le scanner IRM et donc les mouvements de la tête vont provoquer des variations apparentes de forme du cerveau. Andersson *et al.* [2001] proposent une méthode qui calcule les paramètres de réalignement ainsi que l'influence de la rotation estimée sur le champ de distorsion géométrique. Cette méthode est disponible dans une boîte à outils SPM dédiée (*Unwarp Toolbox*).

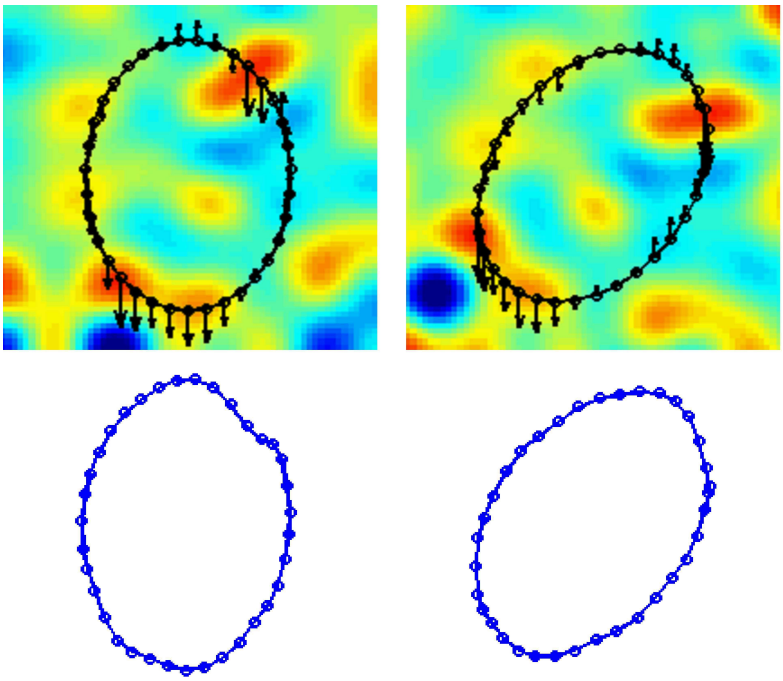

Figure 2.3 – *Illustration de l'interaction entre le mouvement et les distorsions géométriques par l'affichage du contour cérébral sur la carte de phase associée (en haut) et de la déformation résultante du contour cérébral (en bas) : gauche (position initiale) - droite (position recalée) (Source* Jesper Andersson*)*

Enfin, même si la méthode de recalage utilisée par SPM a l'avantage d'être simple à appliquer, elle peut elle-même introduire des artéfacts supplémentaires et notamment conduire à l'apparition de « faux positifs » liés à des corrélations apparentes avec le paradigme expérimental, ceci en l'absence même de mouvement [Freire et Mangin, 2001]. Des méthodes ont été développées pour limiter cet effet, notamment par Roche [2001] qui propose une méthode de réalignement fondée sur une métrique robuste de comparaison d'intensités. Freire *et al.* [2002] ont quant à eux démontré que les mesures de similarité fondées sur l'information mutuelle ou le rapport de corrélation donnent une estimation moins biaisée des paramètres de réalignement.

En pratique, la méthode de recalage de SPM n'étant pas optimale, si les mouvements d'un sujet sont jugés trop importants, la question de le conserver pour les analyses ultérieures peut alors se poser. Une autre stratégie consiste à inclure

les paramètres de mouvement comme régresseurs de non-intérêt dans le modèle linéaire généralisé : ainsi, la variance résiduelle corrélée au mouvement est supprimée ce qui diminue le nombre de faux positifs (meilleure spécificité) mais si le mouvement est corrélé au paradigme expérimental, le fait de supprimer cette variance corrélée au mouvement cause une perte de sensibilité et certaines activations peuvent ne plus être détectées.

2.1.1.3 Correction du délai d'acquisition entre les coupes (*slice timing*)

Lors de l'acquisition d'une image d'IRMf avec une séquence EPI, les différentes coupes d'un même volume ne sont pas acquises simultanément, mais successivement : soit en mode séquentiel (de la première à la dernière), soit en mode entrelacé (les coupes paires puis les coupes impaires ou l'inverse). Ainsi, l'acquisition du volume n'est pas synchrone puisque toutes les coupes n'ont pas été obtenues au même instant. Les décours de l'activité cérébrale mesurés en chaque voxel sont donc décalés temporellement les uns par rapport aux autres en fonction de la coupe à laquelle le voxel appartient et du mode d'acquisition (séquentiel ou entrelacé).

Or la méthode d'analyse statistique « standard » des données d'IRMf consiste à modéliser les décours temporels mesurés à l'aide de fonctions de base ou régresseurs. Le décalage d'acquisition entre les coupes nécessiterait alors de définir un modèle propre à chaque coupe, avec des régresseurs décalés temporellement. En pratique, cette solution est imparfaite à cause des mouvements du sujet et, sans doute parce qu'elle est relativement fastidieuse, elle n'est mise en œuvre à notre connaissance que dans le logiciel *fMRIStat* écrit par Keith Worsley.

La méthode classique de correction du délai inter-coupes consiste en une simple interpolation temporelle des décours mesurés en chaque voxel. Les signaux interpolés sont le plus souvent calés sur la coupe du milieu du volume afin de limiter l'impact des erreurs d'interpolation. Ainsi, lors des analyses ultérieures, toutes les coupes d'une même image seront considérées comme acquises simultanément. Ce recalage temporel est réalisé par interpolation (le plus souvent par une fonction sinus cardinal $\sin(x)/x$) du décours temporel mesuré en chaque voxel.

La figure 2.4 [Henson *et al.*, 1999] illustre l'influence du décalage d'acquisition entre les coupes sur la détection des activations : la carte d'activation obtenue

après correction (au milieu) fait apparaître à la fois les activations motrices (cercle en haut), présentes dans la carte obtenue avec des régresseurs synchronisés sur la coupe la plus élevée (à droite), et les activations visuelles (cercle en bas), présentes dans la carte obtenue avec des régresseurs synchronisés sur la coupe la plus basse (à gauche).

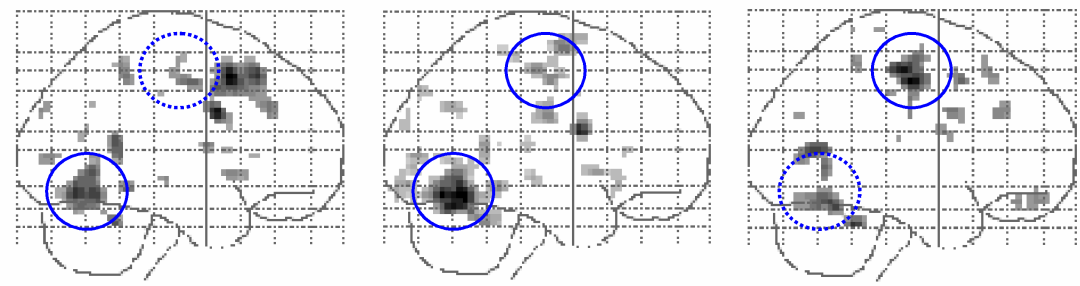

Figure 2.4 – *Cartes statistiques illustrant l'influence du décalage d'acquisition entre les coupes sur la détection des activations : gauche (synchronisation sur la coupe du bas) - milieu (délai inter-coupes corrigé) - droite (synchronisation sur la coupe du haut) (Source* Rik Henson*)*

Plus le temps de répétition T_R est long, plus le décalage temporel dans l'acquisition des coupes est important et la correction du délai inter-coupes devient d'autant plus nécessaire. Mais le problème est que le recalage temporel provoque des erreurs d'interpolation qui sont d'autant plus importantes que le T_R est long. En pratique, il est recommandé ne pas effectuer de correction du délai inter-coupes pour un T_R supérieur à 3 s.

Une autre difficulté majeure à résoudre est l'interaction du mouvement du sujet avec le délai inter-coupes. En pratique, pour une acquisition en mode séquentiel, il est conseillé de commencer par la correction du mouvement avant de corriger le délai inter-coupes. En effet, si l'amplitude du mouvement est non négligeable et que la correction du délai inter-coupes est effectuée en premier, l'interpolation temporelle risque de se faire sur des signaux provenant en réalité de régions différentes du cerveau. Quant à une acquisition en mode entrelacé, il est conseillé

de commencer par la correction du délai inter-coupes avant de corriger le mouvement. En effet, si la correction du mouvement est effectuée en premier, le signal de certains voxels risque d'être déplacé d'une coupe à une coupe adjacente qui a été acquise à un instant très différent.

De plus il faut rappeler que l'acquisition séquentielle des coupes invalide le modèle de transformation rigide pour le mouvement du sujet. Ainsi, l'idéal serait de posséder une méthode capable de corriger les distorsions géométriques, le mouvement du sujet et le délai inter-coupes de manière simultanée afin de prendre en compte les interactions entre ces différents artéfacts et de minimiser les erreurs introduites par les méthodes de correction.

2.1.2 Normalisation spatiale

Cette opération est cruciale pour l'analyse de groupe. En effet, pour pouvoir comparer les activations obtenues pour chaque sujet et être capable de les résumer en une carte d'activation du groupe, il faut pouvoir définir un espace standard commun à tous les sujets. La normalisation spatiale des images d'IRMf peut se décomposer en deux étapes : une première étape de recalage qui permet de positionner les cerveaux dans un même espace de référence, suivie d'une étape de lissage permettant notamment d'atténuer certaines erreurs de normalisation.

2.1.2.1 Recalage

L'objectif principal de la normalisation est de permettre le moyennage inter-sujets et donc de réaliser des études de groupe. Pour ce faire, les images de deux sujets différents doivent être ré-échantillonnées dans un espace commun à tous les sujets, de manière à ce que les régions fonctionnelles homologues soient aussi bien appariées que possible. De plus, si les cerveaux sont normalisés dans un système de coordonnées standard pour la communauté de neuro-imagerie, il devient possible de comparer des études d'IRMf entre elles.

La méthode classiquement utilisée et mise en œuvre dans SPM pour réaliser cette mise en correspondance des différents cerveaux est la « normalisation stéréotaxique » : elle consiste à recaler non-rigidement une image anatomique cérébrale T_1 de chaque sujet vers une image de référence appelée *template* et représentant un

« cerveau moyen ». Le template utilisé par SPM a été réalisé par le MNI (*Montreal Neurological Institute*) : ce template, obtenu à partir des images anatomiques T_1 de 304 sujets sains, est très proche de l'atlas défini par Talairach (10 à 15 % plus large en réalité comme le montre la figure 2.5), ce qui permet de localiser les activations dans un système de coordonnées universel, et de mettre en relation les aires détectées à des aires cérébrales connues comme celles définies par Broadmann.

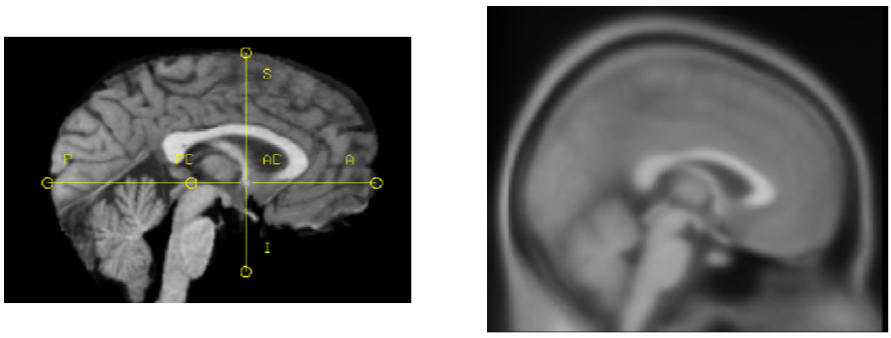

Figure 2.5 – *Templates anatomiques : gauche (Talairach) - droite (MNI)*

Dans SPM, la « normalisation stéréotaxique » est mise en œuvre par un algorithme de recalage iconique similaire à celui permettant de corriger le mouvement, mais suivant un modèle de transformation non-rigide comportant une composante affine et un champ de déformations pures [Friston *et al.*, 1995a ; Ashburner *et al.*, 1997 ; Ashburner et Friston, 1999]. La transformation affine est représentée par 12 paramètres libres (3 de translation, 3 de rotations, 3 de mise à l'échelle et 3 de cisaillement) et permet d'ajuster globalement la forme du cerveau à celle du template. Le champ de déformations pures, modélisé par une combinaison linéaire de fonctions cosinus 3D discrètes, permet en théorie d'améliorer la mise en correspondance des surface corticales ou des noyaux gris. L'estimation de l'ensemble des paramètres de la transformation non-rigide s'effectue par un algorithme de minimisation de la moyenne quadratique des différences d'intensités entre l'image anatomique T_1 à normaliser et le template du MNI, sous la contrainte éventuelle d'une énergie de régularisation. Comme pour la correction de mouvement, l'image normalisée est calculée par ré-échantillonnage de l'image anatomique T_1 du sujet à laquelle est appliquée la transformation non-rigide. Cette même transformation est

utilisée pour normaliser les images fonctionnelles T_2^* du même sujet. La figure 2.6 résume les différentes étapes de la « normalisation stéréotaxique » dans SPM.

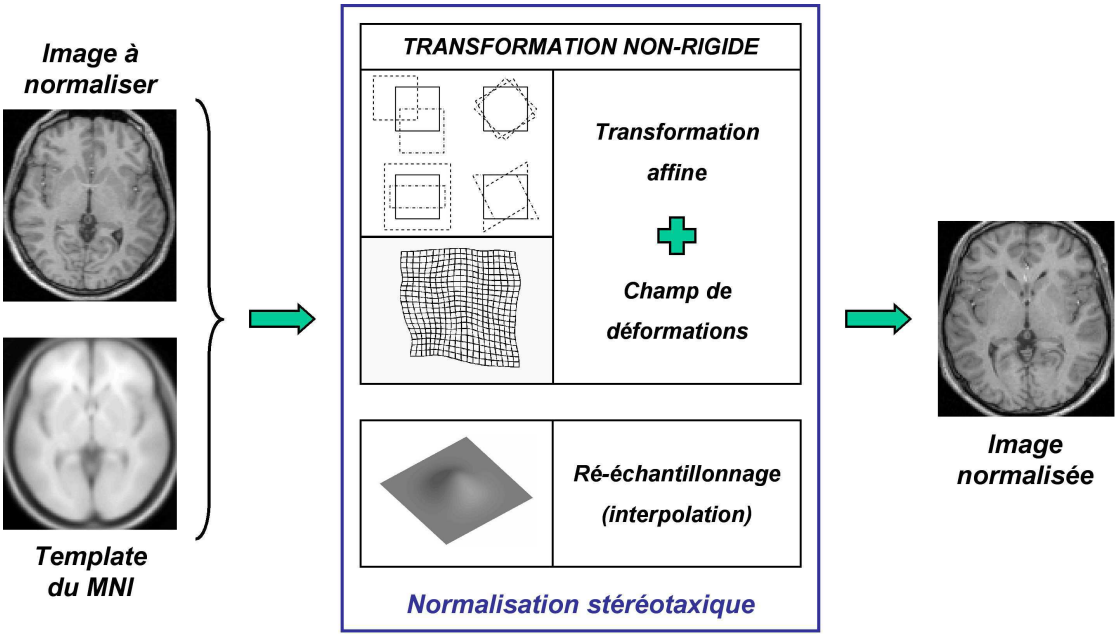

Figure 2.6 – *Principe de la « normalisation stéréotaxique » dans SPM*

La « normalisation stéréotaxique » repose sur une hypothèse très forte d'« homologie » entre les cerveaux, c'est-à-dire sur la possibilité de mettre en correspondance des cerveaux entre eux. Cette hypothèse est notamment critiquée par Brett *et al.* [2002] qui montrent qu'il n'existe pas de correspondance exacte entre la structure et la fonction, ce qui empêche d'utiliser la structure pour faire correspondre des fonctions. Du fait de la grande variabilité anatomique et fonctionnelle observée, il est très difficile de définir des invariants anatomiques et/ou fonctionnels pour permettre de comparer des cerveaux entre eux. De plus, Amunts *et al.* [2000] montrent que l'hypothèse de couplage entre anatomie et fonction ne préserve pas les structures corticales fines et les différences de cytoarchitectonie.

Pour l'analyse de groupe, nous avons besoin de pouvoir définir un espace commun à tous les cerveaux : pour s'en assurer, les images sont donc lissées spatialement (comme décrit dans la section suivante), afin qu'à la résolution finale des

images pré-traitées, l'hypothèse d'« homologie » soit acceptable.

Cependant, le fait d'accepter l'hypothèse d'« homologie », même si elle était vérifiée, ne permet pas de s'affranchir des erreurs de normalisation due au choix de la mesure de similarité (différence d'intensité) et au ré-échantillonnage. Et comme le montre Woods [1996] notamment, ces erreurs de normalisation peuvent introduire des biais significatifs dans les résultats de l'analyse de groupe.

De nombreuses méthodes ont été proposées afin d'améliorer cette procédure de normalisation. Par exemple, la nouvelle version du logiciel SPM (SPM5) utilise la segmentation de l'image anatomique T_1 en matière blanche, matière grise et liquide céphalo-rachidien pour contraindre l'estimation par moindres carrés d'intensité [Ashburner et Friston, 2005]. D'autres méthodes proposent d'utiliser des contraintes anatomiques comme les sillons corticaux [Cachier *et al.*, 2001 ; Corouge *et al.*, 2003], ou des atlas probabilistes [Mazziotta *et al.*, 1995]. Il existe également des méthodes permettant un recalage de la surface du cortex [Fischl *et al.*, 1999b ; Mangin *et al.*, 2004], imposant par la suite des analyses statistiques surfaciques [Fischl *et al.*, 1999a ; Andrade *et al.*, 2001].

2.1.2.2 Lissage

Afin d'augmenter d'une part le rapport signal sur bruit et d'autre part les zones de recouvrement du signal entre sujets (et ainsi atténuer les éventuelles erreurs de normalisation), les images subissent un lissage spatial. Ce lissage consiste en un filtrage spatial passe-bas par convolution de chaque image avec un filtre gaussien isotrope de largeur à mi-hauteur fixé par l'utilisateur.

Une difficulté réside dans la taille du filtre à appliquer car celle-ci détermine directement les activations qui vont pouvoir être détectées. La taille et la forme du filtre optimum doivent correspondre à l'étendue spatiale des activations. Si les activations attendues sont supposées peu étendues, le filtre doit avoir une faible largeur à mi-hauteur alors qu'une plus grande largeur peut être utilisée si les activations sont supposées plus étendues. Malheureusement, il est impossible de connaître à l'avance l'étendue des activations et surtout elle peut varier suivant la position de l'activation dans le cerveau. Ainsi, une heuristique (dérivée d'une contrainte d'approximation des images par des champs aléatoires continus, Frackowiak *et al.*

[1997]) consiste à choisir, comme largeur à mi-hauteur du filtre, une valeur correspondant à 2 ou 3 fois la taille d'un voxel : par exemple une largeur de 5 mm pour un voxel de $2 \times 2 \times 2$ mm^3.

Notons que cette opération augmente bien le rapport signal sur bruit puisqu'elle réduit la décorrélation spatiale du bruit par rapport au signal utile. Mais cette augmentation se fait au détriment de la résolution spatiale des activations détectées puisque le lissage isotrope va mélanger les signaux provenant de tissus de différentes natures. La figure 2.7 illustre l'influence du lissage sur la détection des activations : la sensibilité peut être soit augmentée (exemple A) soit diminuée (exemple B).

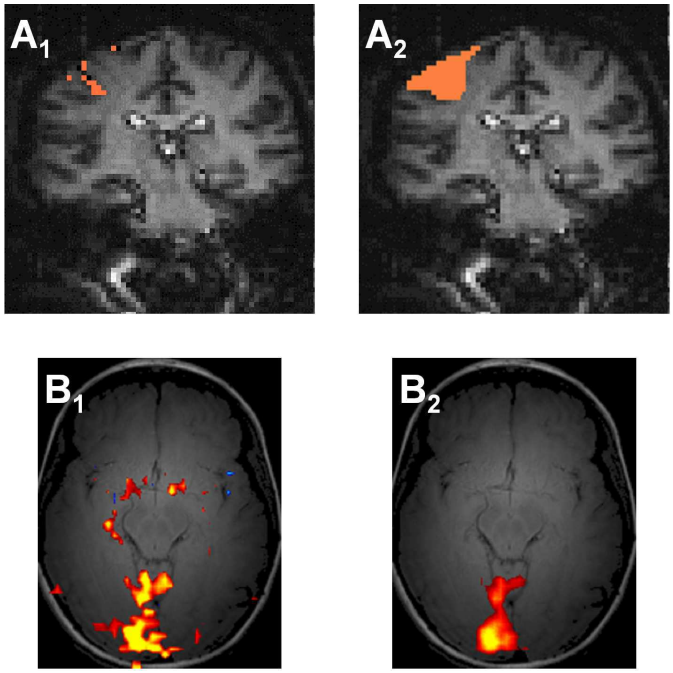

Figure 2.7 – *Exemples de cartes d'activation obtenues sans lissage (A_1 et B_1) et avec un lissage par un noyau gaussien de largeur à mi-hauteur 15 mm (A_2 et B_2) (Source* Katherine Aumer-Ryan*)*

Notons enfin que cette opération permet aussi de justifier la méthode de seuillage des cartes d'activation avec SPM. En effet, cette méthode fait l'hypothèse que le résidu du modèle de régression peut être considéré comme un terme gaussien avec une certaine extension spatiale afin de pouvoir réaliser des inférences sur la significativité des activations détectées. Le filtre spatial diminue l'indépendance entre les

voxels, et de ce fait, les corrections appliquées sur les tests pour prendre en compte le problème des corrections multiples sont moins sévères. Mais la difficulté est que l'hypothèse de résidu gaussien n'est certainement pas vérifiée pour les données brutes.

2.2 Le modèle linéaire généralisé

Lors de l'analyse des données d'une expérience d'IRMf pour un sujet particulier, le problème à résoudre est d'identifier, à partir des décours temporels du signal BOLD enregistrés dans le cerveau d'un sujet, les voxels dont le décours temporel exhibe une forte corrélation avec le paradigme expérimental. En considérant l'hypothèse de base que chaque stimulation induit localement une réponse BOLD, il faut disposer d'un modèle de relation entre stimuli et réponse BOLD. Sur l'exemple de la figure 2.8, cette étape de modélisation doit permettre de relier le décours temporel mesuré dans un voxel aux stimuli auditifs présentés au sujet.

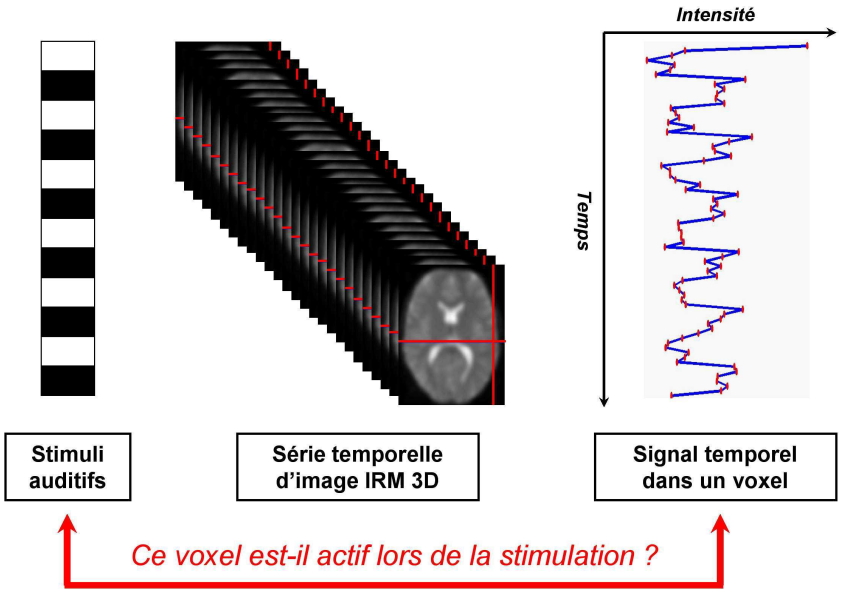

Figure 2.8 – *Principe de la modélisation de la relation entre réponse BOLD et stimuli illustré sur une expérience de stimulation auditive*

2.2.1 Les différentes méthodes de modélisation

Il existe de nombreuses méthodes de modélisation de la relation entre stimuli et réponse BOLD. Elles peuvent se classer suivant plusieurs critères :

- méthodes univariées ou multivariées suivant qu'elles analysent séparément des signaux en un voxel donné ou simultanément l'ensemble des signaux mesurés pour tous les voxels ;
- méthodes volumiques ou surfacique suivant qu'elles travaillent sur un volume 3D ou uniquement sur la surface corticale, ce qui permet un gain en sensibilité car seuls des signaux issus de la matière grise sont pris en compte [Andrade, 2001] ;
- méthodes exploratoires (par nature multivariées) ou guidées par des hypothèses, ce qui les rendent dépendantes du choix du modèle ;
- méthodes paramétriques ou non-paramétriques suivant qu'elles intègrent ou non des hypothèses sur le signal mesuré ou sur la distribution du bruit.

De nombreuses méthodes exploratoires sont utilisées pour analyser les données d'IRMf. Elles correspondent soit à des méthodes de séparation aveugle de sources, comme l'analyse en composantes principales (voir Friston *et al.* [1993] ; Andersson *et al.* [1999] ; Hansen *et al.* [1999] ; Thirion et Faugeras [2003] notamment), l'analyse en composantes indépendantes (voir McKeown *et al.* [1998] ; McKeown [2000] ; Calhoun *et al.* [2001b] ; Beckmann et Smith [2004] notamment) ou l'analyse de corrélation canonique proposée par Friman *et al.* [2001], soit à des méthodes de classification utilisant des mélanges de distributions (voir notamment Almeida et Ledberg [2001] ; Flandin *et al.* [2002a,b] ; Penny *et al.* [2003b, 2005]). Ces méthodes permettent de traiter directement les signaux d'IRMf mesurés sans connaissance *a priori* du paradigme expérimental. Elles considèrent le plus souvent l'ensemble des informations acquises pour tous les voxels du cerveau afin d'extraire différentes structures représentatives de la masse de données. Elles sont alors un moyen d'étude efficace de la connectivité fonctionnelle puisque qu'elles fournissent une information de corrélation des signaux temporels d'IRMf. Cependant l'interprétation des résultats obtenus peut s'avérer parfois difficile à relier au paradigme expérimental utilisé.

L'approche classique pour analyser les données d'IRMf consiste plutôt à utiliser le « modèle linéaire généralisé » (GLM, *Generalized Linear Model*) que nous allons rappeler ici. Ce modèle possède l'avantage d'être facile à mettre en œuvre et de permettre une interprétation directe des résultats en fonction du paradigme expérimental. Cependant, comme toutes les méthodes guidées par des hypothèses, les résultats obtenus restent dépendants du modèle choisi. Dans cette thèse, toutes les données acquises au niveau sujet ont été analysées avec l'implémentation du GLM dans le logiciel SPM [Friston *et al.*, 1995c].

2.2.2 Le modèle linéaire

Le modèle linéaire consiste à expliquer le signal observé par deux composantes : une partie déterministe représentée par une fonction linéaire de variables explicatives, à laquelle s'ajoute une partie probabiliste correspondant à l'erreur de modélisation, comme l'illustre la figure 2.9. Dans le domaine de l'IRMf, ce modèle est principalement utilisé dans sa version univariée, le décours temporel mesuré dans chaque voxel étant analysé indépendamment des autres.

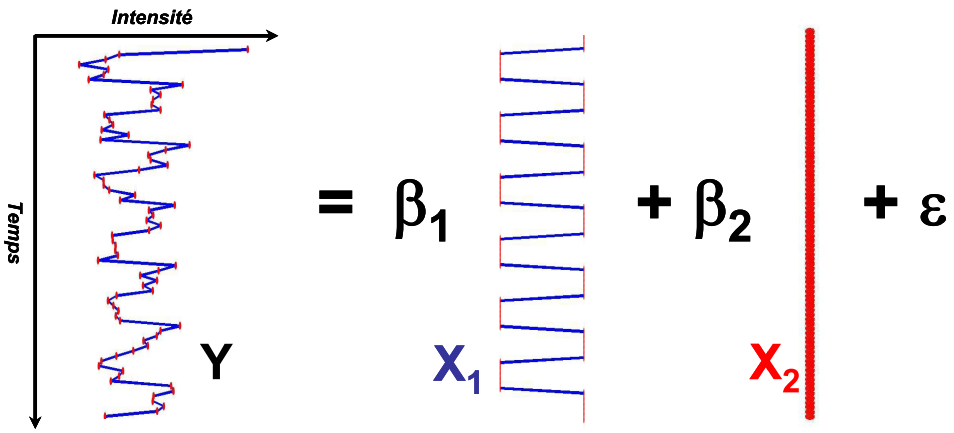

Figure 2.9 – *Exemple de modélisation d'un décours temporel par un modèle linéaire à deux régresseurs pour l'expérience de stimulation auditive présentée figure 2.8*

En notant y_t la valeur du signal BOLD mesuré à l'instant t ($t \in \{1, \ldots, T\}$) dans un voxel donné, le modèle linéaire s'écrit :

$$y_t = \beta_1 \, x_{t1} + \beta_2 \, x_{t2} + \ldots + \beta_p \, x_{tp} + \varepsilon_t$$

où chaque $\mathbf{x_i} = (x_{1i}, x_{2i}, \ldots, x_{Ti})$ pour $i \in \{1, \ldots, p\}$ représente une variable explicative (appelée régresseur) de paramètre inconnu β_i, et ε_t correspond au terme d'erreur de modélisation du signal à l'instant t.

En notation matricielle, soit $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ le décours temporel du signal BOLD mesuré dans un voxel donné, le modèle linéaire s'écrit :

$$\mathbf{y} = \mathbf{X} \, \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

où la matrice $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p})$ est appelée la matrice de dessin expérimental, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ représente le vecteur de paramètres inconnus à estimer et $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T)$ correspond au vecteur d'erreur.

L'intérêt principal du modèle linéaire est de pouvoir choisir les variables explicatives les plus appropriées aux données observées, et notamment de pouvoir construire une matrice de dessin expérimental directement reliée au paradigme. Ainsi, chaque paramètre estimé β_i peut être directement interprétable en terme de contribution d'une condition expérimentale i au signal observé. Cependant, il faut remarquer que les résultats obtenus sont de fait fortement dépendants du modèle choisi.

2.2.3 La matrice de dessin expérimental X

La construction de la matrice de dessin expérimental $\mathbf{X}$ repose sur deux hypothèses implicites : une hypothèse de *linéarité* (les réponses BOLD produites par différentes stimulations s'additionnent) et une hypothèse de *stationnarité* (les réponses BOLD à deux événements de la même stimulation sont identiques au décalage temporel près). Chaque condition expérimentale du paradigme est généralement associée à (au moins) un régresseur dans $\mathbf{X}$: pour le construire, le

signal de stimulation est convolué, comme le montre la figure 2.10, par la fonction de réponse hémodynamique (HRF présentée figure 1.2), qui modélise la réponse BOLD canonique à un événement bref. D'autres régresseurs peuvent être ajoutés : par exemple, des régresseurs de non-intérêt pour modéliser les effets de dérives basse fréquence, ou alors des régresseurs obtenus par convolution du signal de stimulation avec des dérivés de la HRF afin de prendre en compte une éventuelle modulation du délai de réponse.

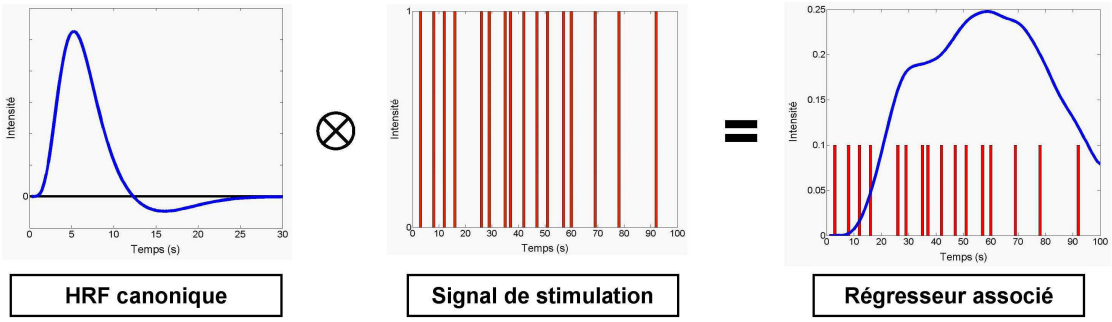

Figure 2.10 – *Construction d'un régresseur par convolution du signal de stimulation avec la fonction de réponse hémodynamique canonique*

2.2.4 La modélisation statistique du bruit

Le terme d'erreur $\boldsymbol{\varepsilon}$ regroupe quant à lui les erreurs dues aux nombreuses sources de variabilité dans les données et les erreurs de modélisation, notamment de la HRF. Il comprend généralement deux sources principales :

- les dérives basse-fréquence (variations lentes du signal BOLD), principalement dues au bruit de mesure du scanner IRM, aux artéfacts physiologiques (activité au repos, respiration, battements cardiaques), aux mouvements du sujet pendant l'expérience ;
- un bruit (variations rapides du signal BOLD) supposé stationnaire dans le temps, centré ($\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$) et non corrélé aux variables explicatives.

Les dérives basse-fréquence sont modélisées de façon déterministe par l'ajout de régresseurs de non-intérêt dans la matrice de dessin expérimental $\mathbf{X}$: une base

de fonctions cosinus discrètes pour filtrer le bruit de mesure et les artéfacts physiologiques, les paramètres de réalignement pour prendre en compte les mouvements du sujet si nécessaire.

La présence du bruit est à l'origine de l'utilisation de statistique dans l'analyse des données d'IRMf. La modélisation de ce bruit nécessite des hypothèses sur la structure de la matrice de variance-covariance $\mathbf{V} = \mathrm{Var}(\boldsymbol{\varepsilon})$, qui vont déterminer le choix des estimateurs du modèle. L'hypothèse la plus simple consiste à considérer un bruit « sphérique », c'est-à-dire indépendamment et identiquement distribué : $\mathbf{V} = \sigma^2 \mathbf{I_T}$, avec $\mathbf{I_T}$ matrice identité de taille $[T \times T]$. Dans ce cas, les erreurs sont supposées normales et le modèle linéaire est dit « général ».

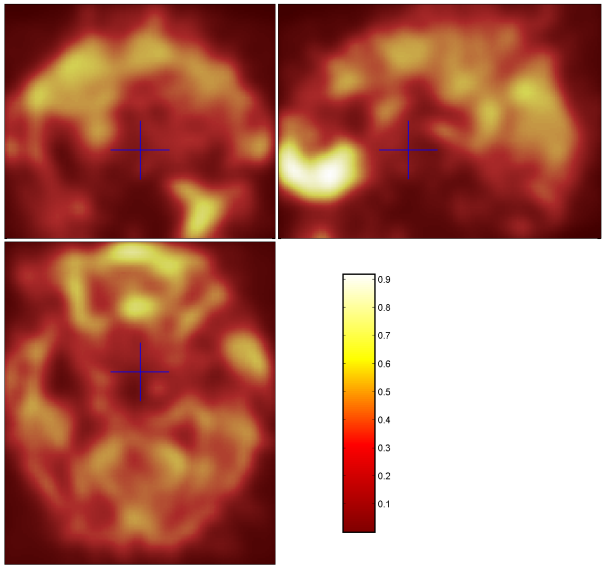

Figure 2.11 – *Exemple de carte d'auto-corrélation temporelle du bruit (Source* Alexis Roche*)*

Cependant, de nombreuses études ont montré que les erreurs en IRMf sont temporellement « corrélées » [Friston *et al.*, 1995b ; Aguirre *et al.*, 1997 ; Zarahn *et al.*, 1997 ; Purdon et Weisskoff, 1998 ; Kruggel et von Cramon, 1999 ; Friston *et al.*, 2000 ; Bullmore *et al.*, 2001 ; Woolrich *et al.*, 2001]. Par exemple, la figure 2.11 montre une carte d'estimation de l'auto-corrélation temporelle du bruit. D'autres exemples de cartes d'auto-correlation temporelle du bruit sont présentées par Bullmore *et al.* [1996] ; Purdon *et al.* [2001] ; Worsley *et al.* [2002].

Cette auto-corrélation temporelle invalide l'hypothèse de bruit identiquement distribué et bien qu'elle ne produise pas de biais sur l'estimation des paramètres $\boldsymbol{\beta}$, elle doit être prise en compte dans le modèle de bruit afin de ne pas biaiser l'estimation de l'erreur sur les paramètres estimés.

Le modèle le plus simple de structure d'auto-corrélation temporelle du bruit est le modèle auto-régressif d'ordre 1 (modèle $AR(1)$), pour lequel l'erreur à un instant donné est corrélée à l'erreur à l'instant précédent. Le modèle $AR(1)$ pour l'erreur à l'instant t ($t \in \{1, \ldots, T\}$) dans un voxel donné s'écrit :

$$\varepsilon_t = \rho\, \varepsilon_{t-1} + \xi_t$$

où ρ est le paramètre d'auto-corrélation ($\mid \rho \mid < 1$) et ξ_t est un terme de bruit blanc (variable aléatoire identiquement et indépendamment distribuée selon une loi normale de moyenne nulle et de variance σ_ε : $\xi_t \sim N(0, \sigma_\varepsilon^2)$). Les hypothèses d'homoscédasticité et d'indépendance des erreurs du modèle ont été transférées aux processus ξ_t, le modèle de bruit devient alors hétéroscédastique vis-à-vis des erreurs et le modèle linéaire est dit « généralisé ». Dans le cas d'un modèle $AR(1)$ stationnaire, l'erreur à l'instant t peut s'écrire en fonction des erreurs précédentes :

$$\varepsilon_t = \xi_t + \rho\, \xi_{t-1} + \ldots + \rho^{t-1}\, \xi_1 + \rho^t\, \varepsilon_0$$

La matrice de variance-covariance $\mathbf{V} = \mathrm{Var}(\boldsymbol{\varepsilon})$ s'écrit alors :

$$\mathbf{V} = \frac{\sigma_\varepsilon^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \vdots & & \ddots & & \vdots \\ \rho^{T-2} & \cdots & \cdots & 1 & \rho \\ \rho^{T-1} & \rho^{T-2} & \cdots & \rho & 1 \end{bmatrix}$$

Cette structure de covariance temporelle est complètement déterminée par les deux paramètres σ^2 et ρ, qui doivent également être estimés à partir des données.

2.2.5 L'estimation des paramètres du modèle

Le modèle linéaire généralisé s'écrit donc :

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \mathbf{V}) \tag{2.1}$$

Il faut estimer les paramètres inconnus du modèle $\hat{\boldsymbol{\beta}}$, ainsi que l'erreur d'estimation $\mathrm{Var}(\hat{\boldsymbol{\beta}})$. D'après le théorème de Gauss-Markov, sous les hypothèses d'absence de corrélation entre l'erreur et les variables explicatives et que la matrice $\mathbf{V}$ est connue et inversible, l'estimateur des Moindres Carrés Généralisés (MCG) est le meilleur estimateur linéaire sans biais :

$$\hat{\boldsymbol{\beta}}_{MCG} = \arg\min \|\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta}\|^2 = (\mathbf{X}^\top\,\mathbf{V}^{-1}\,\mathbf{X})^{-1}\mathbf{X}^\top\,\mathbf{V}^{-1}\,\mathbf{y}$$

L'erreur d'estimation vaut quant à elle :

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{MCG}}) = (\mathbf{X}^\top\,\mathbf{X})^{-1}\mathbf{X}^\top\,\mathbf{V}\,\mathbf{X}(\mathbf{X}^\top\,\mathbf{X})^{-1}$$

Pour obtenir ces estimations, les matrices $\mathbf{X}^\top\,\mathbf{V}^{-1}\,\mathbf{X}$ et $\mathbf{X}^\top\,\mathbf{X}$ doivent être inversibles (pour cela, l'expérimentateur doit définir correctement la matrice de dessin expérimental $\mathbf{X}$) et la connaissance de la matrice $\mathbf{V}$ est nécessaire, ce qui n'est pas toujours possible. Pour résoudre ce problème, une méthode de blanchissement des données est utilisée : elle consiste dans un premier temps à estimer la matrice $\hat{\mathbf{V}}$ par maximum de vraisemblance restreinte (ReML pour *Restricted Maximum Likelihood* en anglais), en considérant le modèle auto-régressif d'ordre 1 et en imposant des contraintes sur sa forme. Puis les données sont « blanchies », c'est-à-dire que chaque membre de l'équation 2.1 est multiplié par $\hat{\mathbf{V}}^{-1/2}$ [Worsley *et al.*, 2002]. L'estimation des paramètres d'effets du modèle $\hat{\boldsymbol{\beta}}$ et de l'erreur d'estimation $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ est enfin réalisée par la méthode des moindres carrés appliquée aux données blanchies.

Cette méthode se justifie dans la mesure où les paramètres de variance (matrice $\hat{\mathbf{V}}$) sont estimés avec un grand nombre de degrés de liberté, une centaine

d'images au moins étant acquises pour un même sujet au cours d'une expérience d'IRMf classique. Le modèle obtenu par blanchissement reste donc à peu près gaussien et les estimations des paramètres qui en découlent sont peu biaisées.

Par la suite, l'expérimentateur est souvent amené à définir un contraste $\mathbf{c}$, qui est un ensemble de coefficients de pondération d'une combinaison linéaire appliqué au vecteur d'effets estimés $\hat{\boldsymbol{\beta}}$. Dans la suite de cette thèse, nous désignerons par effet estimé à la fois un effet estimé particulier $\hat{\beta}_i$ ou un contraste particulier d'effets estimés $\mathbf{c}^\top\hat{\boldsymbol{\beta}}$.

De plus, pour présenter le modèle linéaire généralisé, nous avons considéré la situation en un voxel donné. L'application de ce modèle sur l'ensemble des voxels, puis la spécification d'un contraste d'intérêt par l'expérimentateur, permet d'obtenir, pour chaque sujet étudié, une carte d'effet estimé (carte de contraste) ainsi qu'une carte de variance résiduelle, comme le montre la figure 2.12.

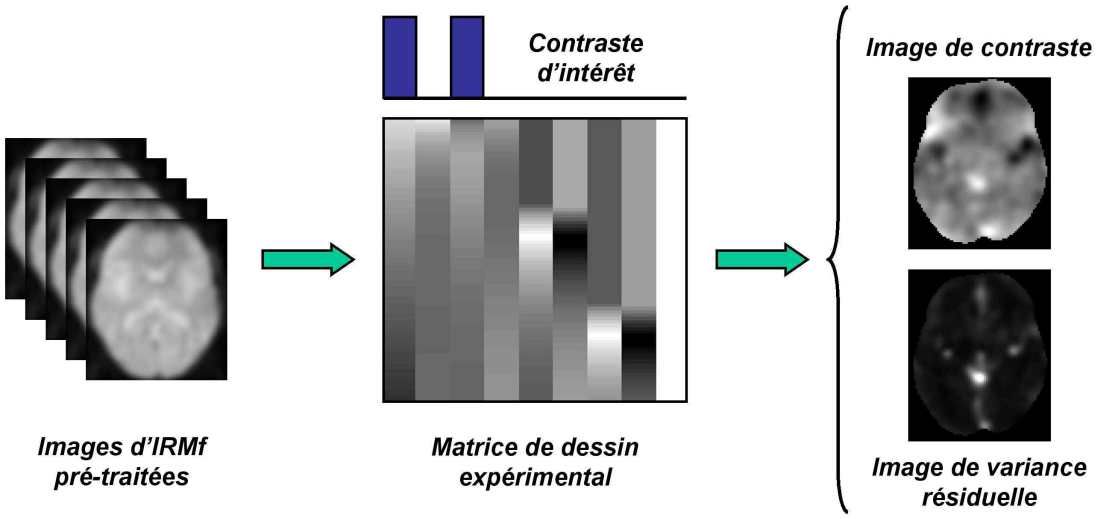

Figure 2.12 – *Obtention d'une carte de contraste et d'une carte de variance résiduelle par application d'un modèle linéaire généralisé et spécification d'un contraste d'intérêt*

2.3 Conclusion

Ce chapitre nous a permis de rappeler rapidement les différentes étapes de prétraitements et de modélisation qui permettent d'obtenir les cartes d'effets estimés et de variance résiduelle pour chaque sujet. Il faut remarquer que chacune de ces étapes de l'analyse intra-sujet constitue une source d'erreurs qui a des répercussions sur l'analyse de groupe. Par exemple, des erreurs de recalage ou de normalisation spatiale peuvent introduire une incertitude sur la localisation de l'effet estimé alors que des erreurs de modélisation de la HRF peuvent biaiser l'estimation de l'amplitude de l'effet.

Dans toute la suite de cette thèse, nous n'avons pas cherché à améliorer l'analyse intra-sujet réalisée avec SPM bien que conscient de son fort impact sur l'analyse de groupe. Cependant, comme nous le verrons dans le chapitre 8, l'utilisation de statistiques à « effets mixtes » permet de prendre en compte l'erreur d'estimation sur l'effet pour corriger les résultats de l'analyse de groupe : ainsi, l'impact d'éventuelles erreurs de modélisation de la HRF se trouve limité.

Chapitre 3

Analyse de groupe

Le principe de l'analyse de groupe est de construire, à partir des séries temporelles d'images IRMf acquises pour plusieurs sujets, une carte d'activation du groupe. Le premier problème à résoudre est de positionner les cerveaux des différents sujets dans un espace commun, ce que l'étape de normalisation spatiale présentée en 2.1.2 permet de réaliser. Ensuite, la deuxième étape consiste à définir une méthode d'inférence statistique pour résumer les effets BOLD mesurés pour chaque sujet en une carte d'activation du groupe.

Il existe de nombreuses méthodes d'analyse de groupe permettant de résumer des informations obtenues au niveau intra-sujet en une information inter-sujets. Lazar *et al.* [2002] distingue les méthodes dites de « combinaison », pour lesquelles les statistiques obtenues au premier niveau sont combinées pour calculer une statistique de groupe, des méthodes dites « d'inférence », pour lesquelles une statistique de groupe est calculée à partir des effets estimés à travers les sujets. La principale difficulté des méthodes de « combinaison » réside dans l'interprétation de la statistique obtenue, et notamment dans la définition de l'hypothèse nulle à tester. Ainsi, les méthodes « d'inférence » sont préférées en IRMf puisqu'elles permettent de tester le signe, l'amplitude ou la reproductibilité des effets à travers les sujets.

Dans ce chapitre, nous introduisons un modèle hiérarchique d'inférence statistique à deux niveaux (3.1) permettant de rendre compte des deux sources de

variabilité intervenant dans l'analyse de groupe : une source intra-sujet et une source inter-sujets. L'analyse à « effets fixes » (3.2) s'intéresse exclusivement à la source de variabilité intra-sujet et cherche à mettre en évidence un effet moyen au sein d'une cohorte *particulière* de sujets.

L'analyse à « effets aléatoires », qui nous intéresse en premier lieu durant cette thèse, tient compte des deux sources de variabilité afin de *généraliser* les résultats obtenus sur la cohorte à l'ensemble de la population dont elle est issue. Les paramètres du modèle hiérarchique ne pouvant pas être identifiés de façon analytique dans le cas général, nous présentons dans ce chapitre la solution approchée couramment utilisée pour l'analyse à effets aléatoires (3.3). La fin de ce chapitre est consacrée à la discussion des hypothèses du modèle d'analyse à effets aléatoires « standard », et notamment des différentes limitations qui en découlent.

3.1 Modèle hiérarchique à deux niveaux

L'objectif général de l'analyse de groupe est de mettre en évidence une dépendance statistique entre un ensemble de variables explicatives $\mathbf{V}$ et le signal d'IRMf $\mathbf{y}$ observé en un point donné de l'espace de référence, ceci au moyen de mesures répétées sur plusieurs sujets. Pour cela, différents auteurs [Worsley *et al.*, 2002 ; Friston *et al.*, 2002 ; Beckmann *et al.*, 2003 ; Neumann *et al.*, 2003 ; Woolrich *et al.*, 2004] ont introduit le modèle hiérarchique à deux niveaux suivant, parfois également appelé « modèle à effets mixtes » :

$$\begin{cases} \mathbf{y}_i &= \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i & 1^{er} \text{ niveau} \\ \boldsymbol{\beta}_i &= \mathbf{V}_i \boldsymbol{\beta}_G + \boldsymbol{\eta}_i & 2^{ème} \text{ niveau} \end{cases} \tag{3.1}$$

Ce modèle, illustré par la figure 3.1, permet de rendre compte de l'aspect composite de la variabilité, avec une source intra-sujet et une source inter-sujets.

Le premier niveau du modèle correspond au modèle linéaire généralisé introduit en 2.2 : il relie statistiquement, pour chaque individu, le vecteur d'effets BOLD $\boldsymbol{\beta}_i$ aux données d'imagerie et modélise donc la variabilité intra-sujet.

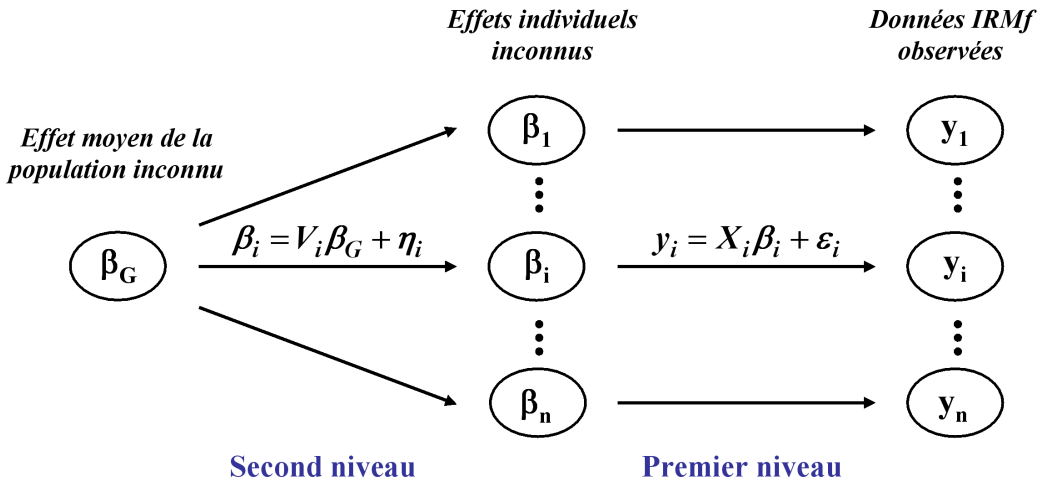

Figure 3.1 – *Modèle hiérarchique à deux niveaux pour l'analyse de groupe*

Le second niveau modélise quant à lui la variabilité inter-sujets en reliant les effets BOLD individuels $\boldsymbol{\beta}_i$ aux prédicteurs représentées par la matrice $\mathbf{V}_i$, via un vecteur de coefficients $\boldsymbol{\beta}_G$ qui constitue le paramètre d'intérêt de l'analyse à effets aléatoires. Le modèle *one-sample* est l'exemple le plus simple de spécification du second niveau pour lequel les prédicteurs sont des matrices $[1 \times 1]$ constantes ($\mathbf{V}_i \equiv 1$) : $\boldsymbol{\beta}_G$ représente alors simplement l'effet moyen de la population. C'est essentiellement ce cas que nous traiterons au cours des chapitres suivants. Parmi les autres types classiques de spécification au deuxième niveau, nous pouvons citer le modèle *two-sample* qui permet de tester des différences de réponses entre deux sous-populations.

3.2 Analyse de groupe à effets fixes (FFX)

Un premier type d'analyse permettant de moyenner les résultats individuels sur un groupe consiste à agréger les données provenant des sujets et à en faire l'analyse comme s'il s'agissait d'un sujet unique. Ce type d'analyse peut être considéré comme une extension de l'analyse intra-sujet à l'analyse de groupe mais elle ne tient pas réellement compte de la variance inter-sujets.

Dans un voxel donné, considérons le modèle linéaire suivant pour le signal y_i recueilli pour le $i^{ème}$ sujet :

$$\mathbf{y_i} = \mathbf{X_i}\,\boldsymbol{\beta_i} + \boldsymbol{\varepsilon_i}$$

Si les données des sujets sont normalisées dans le même espace et que la matrice de dessin expérimental $\mathbf{X_i}$ représente les mêmes conditions expérimentales pour tous les sujets, alors nous pouvons considérer le modèle suivant pour la concaténation de l'ensemble des signaux recueillis pour les n sujets (voir figure 3.2) :

$$\begin{pmatrix} \mathbf{y_1} \\ \mathbf{y_2} \\ \vdots \\ \mathbf{y_n} \end{pmatrix} = \begin{pmatrix} \mathbf{X_1} & 0 & \cdots & 0 \\ 0 & \mathbf{X_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{X_n} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \\ \vdots \\ \boldsymbol{\beta_n} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon_1} \\ \boldsymbol{\varepsilon_2} \\ \vdots \\ \boldsymbol{\varepsilon_n} \end{pmatrix}$$

La résolution de ce modèle fournit $p \times n$ coefficients de régression estimés : $\hat{\boldsymbol{\beta}} = [\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \cdots, \hat{\boldsymbol{\beta}}_n]$. À noter que chaque $\hat{\boldsymbol{\beta}}_i$ représentant les coefficients de régression estimés pour le $i^{ème}$ sujet est identique à celui de l'analyse intra-sujet puisque les régresseurs correspondants sont orthogonaux.

Ce modèle permet, si on fait l'hypothèse que la variance du bruit est identique entre les sujets, de tester individuellement chaque sujet mais surtout de tester si un effet est significatif pour le groupe de sujets considéré, en définissant un contraste englobant l'ensemble des sujets (voir figure 3.2). Cependant, le résultat du test obtenu est spécifique au groupe de sujet étudié et ne peut être généralisé à la population d'intérêt.

3.3 Analyse de groupe à effets aléatoires (RFX)

L'objectif de l'analyse à effets aléatoires est de faire une inférence sur le paramètre $\boldsymbol{\beta}_G$ défini dans l'équation (3.1), modélisant le lien entre les données d'IRMf et les variables explicatives au sein de la population. L'identification de ce paramètre n'admet pas de solution analytique sous les hypothèses générales du modèle

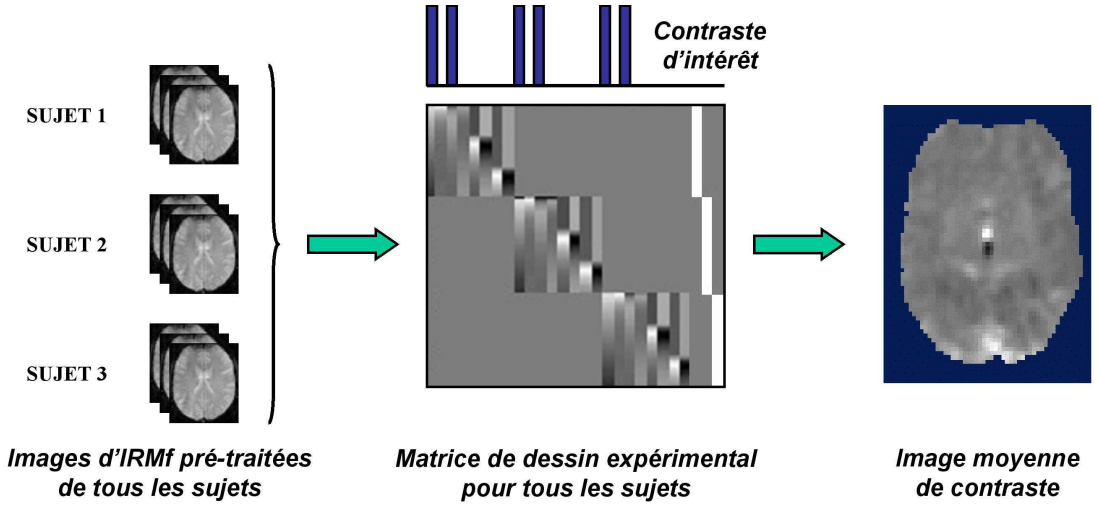

Figure 3.2 – *Principe de l'analyse à effets fixes*

hiérarchique et donne lieu à des techniques coûteuses en temps de calcul [Woolrich *et al.*, 2004 ; Friston *et al.*, 2005].

Afin d'alléger les calculs ainsi que l'encombrement mémoire, les séries temporelles $\mathbf{y}_i$ peuvent être remplacées par des résumés statistiques du premier niveau. Le plus souvent, ces résumés ne sont pas exhaustifs au sens où ils effacent une partie de l'information que $\mathbf{y}_i$ contient sur l'effet individuel $\boldsymbol{\beta}_i$. Nous perdons donc en précision ce que nous gagnons en efficacité calculatoire.

3.3.1 Approche « standard » : la statistique de Student

Ainsi, dans l'approche « standard » mise en œuvre dans le logiciel SPM [Penny *et al.*, 2003a], seuls les effets estimés $\hat{\boldsymbol{\beta}}_i$ sont retenus comme résumés statistiques du premier niveau. L'inférence sur le paramètre de groupe est alors effectuée au moyen d'un test t ou F paramétrique, ce qui revient à supposer que les effets estimés $(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_n)$ sont distribués identiquement et normalement au sein de la population d'intérêt.

Dans le cas d'effets unidimensionnels, l'effet moyen de la population β_G est testé à l'aide de la statistique de Student (ou statistique t). Considérons un échantillon

$(\hat{\beta}_1, \ldots, \hat{\beta}_n)$ d'effets estimés dans un voxel donné pour n sujets. La statistique de Student est égale à :

$$T = \frac{\sqrt{n(n-1)}\bar{\beta}}{\sqrt{\sum_{i=1}^{n}(\hat{\beta}_i - \bar{\beta})^2}} \tag{3.2}$$

où $\bar{\beta}$ représente la moyenne empirique de l'échantillon ($\bar{\beta} = \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_i$).

En faisant l'hypothèse d'une distribution normale des effets à travers les sujets, la statistique de Student suit, sous l'hypothèse nulle H_0 que l'effet moyen de la population est nul, une distribution de Student à $n-1$ degrés de liberté (figure 3.3). Ainsi, il est possible de déterminer, pour un risque de premier espèce α spécifié par l'expérimentateur, le seuil statistique correspondant au-delà duquel l'hypothèse nulle H_0 est rejetée.

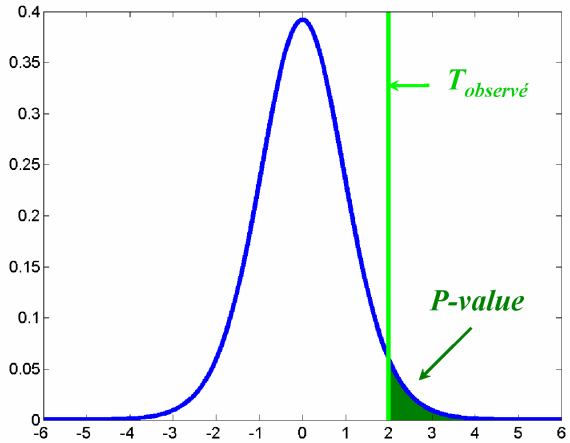

Figure 3.3 – *Distribution de Student à $n-1$ degrés de liberté ($n = 10$ sujets)*

L'intérêt majeur de l'analyse à effets aléatoires (RFX) par rapport à l'analyse à effets fixes (FFX) est de pouvoir généraliser les résultats obtenus au niveau du groupe de sujets à l'ensemble de la population d'intérêt. En effet, dans le cas de l'analyse FFX, l'hypothèse nulle H_0 correspond à « tous les sujets ont un effet nul », ce qui limite les résultats de l'inférence statistique au groupe de sujets étudiés.

De plus, comme le résume le schéma présenté figure 3.4, les deux approches ne considèrent pas les mêmes sources de variabilité : l'analyse FFX ne prend en compte que la variabilité intra-sujet et va alors détecter un effet de groupe significatif pour l'exemple de la figure 3.4 ; l'analyse RFX considère quant à elle les deux sources de variabilité et pour l'exemple de la figure 3.4, il n'est pas certain qu'un effet de groupe significatif puisse être détecté.

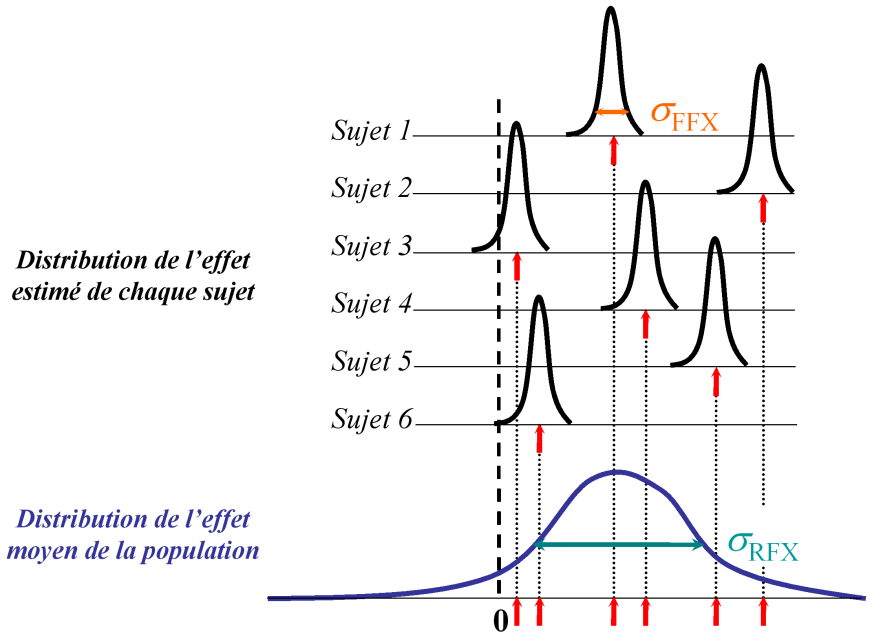

Figure 3.4 – *Comparaison entre l'analyse à effets fixes (FFX) et l'analyse à effets aléatoires « standard » (RFX) (Source* Tom Nichols*)*

Comme nous l'avons déjà dit, l'analyse à effets aléatoires « standard » repose sur une hypothèse forte de normalité, qui fera l'objet d'une analyse critique lors des chapitres 4, 5 et 6. De plus, le fait de ne considérer que les effets estimés $\hat{\boldsymbol{\beta}}_i$ comme résumés statistiques du premier niveau revient à supposer que la variance intra-sujet est constante à travers les sujets. Afin de nous affranchir de cette autre hypothèse et d'utiliser l'information disponible d'estimation de la variance intra-sujet, nous proposons dans le chapitre 8 des statistiques de décision à effets mixtes.

3.3.2 Approche intermédiaire

Une approche intermédiaire [Worsley *et al.*, 2002 ; Beckmann *et al.*, 2003] entre celle de SPM et l'identification exacte du modèle hiérarchique consiste à retenir du premier niveau l'effet estimé $\hat{\boldsymbol{\beta}}_i$ ainsi qu'une estimée de sa matrice de variance $\boldsymbol{\Sigma}_i = \mathrm{Var}(\hat{\boldsymbol{\beta}}_i)$, et de considérer le modèle premier niveau simplifié :

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} = \boldsymbol{\beta}_{\boldsymbol{i}} + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim N(0, \boldsymbol{\Sigma}_i) \tag{3.3}$$

La simplification provient de l'hypothèse d'une densité gaussienne de l'effet estimé conditionnellement à l'effet réel non observé $\boldsymbol{\beta}_i$, ce qui revient à négliger l'incertitude sur $\boldsymbol{\Sigma}_i$ ou, de façon équivalente, à considérer les degrés de liberté intra-sujets comme infinis. Cette hypothèse est justifiée par le fait que les paramètres de variance sont estimés avec un grand nombre de degrés de libertés, une centaine d'images au moins étant acquises au cours d'une expérience d'IRMf classique.

Dans le cas *one-sample*, il est alors possible de résumer les modèles du premier et du deuxième niveau par le modèle unique suivant :

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} = \boldsymbol{\beta}_G + \boldsymbol{\gamma}_{\boldsymbol{i}} \tag{3.4}$$

où $\boldsymbol{\gamma}_{\boldsymbol{i}}$ suit une loi normale $N(0, \sqrt{\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_G})$, si le bruit $\boldsymbol{\eta}_{\boldsymbol{i}}$ du deuxième niveau est considéré comme un bruit blanc gaussien de variance $\boldsymbol{\Sigma}_G$.

Worsley *et al.* [2002] utilisent le modèle 3.4 en considérant les estimations $\boldsymbol{\Sigma}_i$ comme des paramètres fixes, et appliquent la même méthode de blanchissement que pour estimer les paramètres du modèle au premier niveau (voir 2.2.5), ce qui revient en pratique à pondérer chaque observation $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ en la divisant par une estimation de son écart-type $\sqrt{\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_G}$. Il faut alors estimer $\boldsymbol{\Sigma}_G$, ce que Worsley *et al.* [2002] propose de faire par maximum de vraisemblance restreint (ReML), cette méthode permettant d'obtenir une estimation moins biaisée que celle par maximum de vraisemblance. Il n'existe pas de solution analytique pour le ReML, mais des approximations numériques sont calculables par un algorithme d'espérance-maximisation (EM pour *Expectation-Maximisation* en anglais). L'effet moyen de

la population $\boldsymbol{\beta}_G$ est alors estimé par moindres carrés pondérés. Une des principales limitations de cette approche est le petit nombre d'observations disponibles au second niveau qui rend l'estimateur de $\boldsymbol{\Sigma}_G$ extrêmement variable, malgré la méthode de régularisation proposée par Worsley *et al.* [2002].

3.4 Conclusion

L'analyse de groupe « standard » repose sur une hypothèse de normalité des effets estimés à travers les sujets. Cette hypothèse très forte est difficile à vérifier étant donné le petit nombre de sujets participant aux études d'IRMf. De plus, sa mise en défaut a pour conséquence de biaiser la spécificité du test t utilisé pour tester la significativité de l'effet de groupe, mais également de diminuer sa sensibilité.

Dans un premier temps, il convient donc de disposer de procédures permettant de contrôler l'homogénéité des données afin de pouvoir raisonnablement valider cette hypothèse de normalité. Pour cela, nous proposons une méthode multivariée de diagnostique d'influence (chapitre 4), ainsi que l'utilisation d'un test massivement univarié de normalité, le test de Grubbs (chapitre 5).

Ensuite, afin d'éviter l'utilisation d'hypothèses trop fortes sur la distribution des données, il convient de remplacer la statistique de Student par des statistiques « robustes », dans le sens où les performances des tests associés sont stables à travers les distributions. Et le fait que la distribution vraie des données soit inconnue rend naturel l'utilisation de méthodes de calibration statistique non-paramétriques comme les tests de permutations. Ces tests présentent en plus l'avantage de permettre des comparaisons multiples sans l'approximation de la caractéristique d'Euler nécessaire à la théorie des champs aléatoires gaussiens. Ainsi, l'approche proposant des statistiques « robustes » calibrées par permutations (chapitre 7) permet de s'affranchir des hypothèses fortes de l'analyse de groupe « standard ».

Enfin, le modèle d'analyse à effets aléatoires « standard » ne distingue pas la variance intra-sujet (σ_i^2) de la variance inter-sujets (σ_G^2) : il permet d'estimer un seul terme de variance combinant les deux sources de variabilité en supposant la variance σ_i^2 constante à travers les sujets. Or, l'analyse intra-sujet fournit une estimation de la variance résiduelle pour chaque sujet : un modèle à « effets mixtes »

prenant en compte cette mesure d'incertitude, généralement dépendante du sujet, peut être utilisé pour pondérer les sujets dans l'analyse de groupe (chapitre 8).

Chapitre 4

Méthode multivariée de diagnostique d'influence

Comme nous l'avons rappelé dans le chapitre précédent, l'analyse de groupe « standard » repose sur une hypothèse forte d'homogénéité des effets estimés à travers les sujets. Nous avons également montré que si cette hypothèse n'est pas vérifiée, ce que peut révéler la présence de données atypiques (*outliers*) par exemple, alors les résultats de l'analyse à effets aléatoires utilisant le test de Student peuvent présenter un biais de spécificité (mauvais contrôle du taux de faux positifs) ainsi qu'une perte de sensibilité (mauvais contrôle du taux de vrais positifs).

Dans le cas des grands échantillons, la vérification de l'homogénéité (plus précisément, la normalité) n'est pas cruciale car il est possible de montrer que la statistique t suit asymptotiquement une loi de Student même si la distribution des données n'est pas normale. En revanche, dans le cas des études d'IRMf, le nombre de sujets étant le plus souvent compris entre 10 et 20, il devient impératif de vérifier l'homogénéité des données afin d'éviter tout biais statistique. Le principal problème est qu'il est justement très difficile de valider cette hypothèse sur un petit nombre de données.

La première partie de la thèse a donc consisté à développer des procédures de contrôle de l'homogénéité des données, fondées sur l'étude de la structure du groupe, et notamment de la variabilité inter-sujets. Ces méthodes ont été mises en œuvre dans le logiciel *DISTANCE*, développé au cours de cette thèse afin de fournir à la communauté de neuro-imagerie un outil de diagnostique pour la détection des éventuelles données atypiques ou sous-groupes de sujets. L'illustration de l'intérêt de ces méthodes et du logiciel *DISTANCE* a été réalisée sur une vingtaine de jeux de données d'IRMf. Les résultats obtenus sont présentés dans le chapitre 6.

Dans ce chapitre, nous présentons la première approche retenue pour étudier l'homogénéité des données. Il s'agit d'une méthode exploratoire s'appuyant sur une mesure de similarité multivariée, le coefficient *RV*, qui permet le calcul d'une distance inter-sujets (4.1). Les sujets sont ainsi positionnés les uns par rapport aux autres dans un certain espace topologique afin de mettre en évidence des similitudes et des différences entre eux. Pour contrôler l'homogénéité du groupe, et accessoirement détecter d'éventuelles données atypiques, l'influence de chaque sujet sur la moyenne des distances inter-sujets est estimée par le calcul d'une distance de Cook (4.2). Un seuil est alors défini afin de détecter les éventuels sujets présentant une grande influence sur cette moyenne.

Un véritable test de normalité (massivement univarié) sera présenté dans le chapitre suivant. Il s'agit du test de Grubbs qui, plutôt que de détecter des sujets globalement atypiques, évalue *localement* l'hypothèse de normalité et identifie éventuellement des régions cérébrales dans lesquelles la population présente une hétérogénéité « anormale ».

4.1 Étude multivariée de la structure du groupe : le calcul de distances inter-sujets

L'idée de la méthode exploratoire présentée dans ce chapitre repose sur une mesure de similarité multivariée, permettant la comparaison directe des matrices de données obtenues pour chaque sujet.

Il existe dans la littérature de nombreuses méthodes multivariées permettant de comparer les données de neuro-imagerie. Afin de déterminer des régions ana-

logues à travers les sujets et d'étudier la reproductibilité des patterns d'activation, Tegeler *et al.* [1999] ont utilisé une méthode d'analyse discriminante de Fisher, Moeller et Strother [1991] ; Strother *et al.* [1995b,a] ; Hansen *et al.* [1999] ont proposé une méthode d'analyse en composantes principales baptisée SSM (*Scaled Subprofile Model*), McIntosh *et al.* [1996] ont appliqué la méthode PLS (*Partial Least Square*) et Welchew *et al.* [2002] la méthode MDS (*MultiDimensional Scaling*), Strother *et al.* [2002] ont développé une métrique multivariée baptisée NPAIRS (*Nonparametric Prediction, Activation, Influence, and Reproducibility re-Sampling*), McKeown *et al.* [1998] ; Calhoun *et al.* [2001a] ; Nybakken *et al.* [2002] ; Esposito *et al.* [2005] ; Moritz *et al.* [2005] ont adapté des méthodes d'analyse à composantes indépendantes. Fox *et al.* [1999] ont quant à eux mis au point un outil de méta-analyse fondé sur la comparaison des coordonnées de Talairach des activations de chaque sujet, outil appliqué dans Xiong *et al.* [2000] à l'étude d'une tâche de génération sémantique.

Notons aussi que des méthodes univariées dites de *test-retest* ont été proposées par Genovese *et al.* [1997] ; Maitra *et al.* [2002] notamment afin d'étudier la reproductibilité des cartes d'activation à travers les sessions. Cependant, à notre connaissance, aucune de ces méthodes, univariées comme multivariées, n'a encore été utilisée spécifiquement pour étudier l'homogénéité du groupe.

Pour choisir la mesure de similarité la mieux adaptée à la nature des données d'IRMf, nous avons repris les travaux de Kherif *et al.* [2002]. Ils ont suggéré que la mesure de similarité optimale devait répondre aux critères suivants :

- elle doit permettre une comparaison des données dans les deux dimensions (spatiale et temporelle), afin de pouvoir étudier les différentes sources de variabilités ;
- elle doit être insensible aux différences de variance observées entre les voxels ;
- elle doit être sensible à la présence de données atypiques ;
- son calcul doit permettre d'introduire des informations relatives au paradigme expérimental utilisé, comme le choix d'un masque d'analyse (dimension spatiale réduite) ou d'un contraste d'intérêt (dimension temporelle réduite).

Kherif *et al.* [2002] ont étudié plusieurs mesures de similarités (distance de Mahalanobis, test d'égalité des covariances, analyse canonique des corrélations,...)

pour conclure que le coefficient *RV*, proposé par Robert et Escoufier [1976], était la mesure de similarité la plus à même de s'adapter aux contraintes imposées par les données d'IRMf, tout en conservant un temps de calcul raisonnable.

4.1.1 Mesure de similarité : le coefficient *RV*

Considérons deux matrices de données $\mathbf{Y_1}$, de dimension $[p \times N]$, et $\mathbf{Y_2}$, de dimension $[q \times N]$. D'un point de vue géométrique, les deux matrices peuvent être vues comme deux configurations de nuages de N points respectivement dans $\mathbb{R}^p$ et $\mathbb{R}^q$. La problématique à laquelle répond notamment le coefficient *RV* relève des statistiques multi-dimensionnelles introduites par Tucker [1958] : « si N observations sont décrites par p variables d'une part et q variables d'autre part, comment comparer ces observations ? ».

Afin de répondre à cette problématique, plusieurs méthodes ont été proposées : l'Analyse Canonique par Hotelling [1936], l'Analyse Canonique Généralisée par Carroll [1968] et Kettenring [1971], la mesure de liaison de Lingoes et Schönemann [1974], l'indice de redondance de Stewart et Love [1968] et le coefficient de corrélation vectorielle par Escoufier [1973]. L'objectif de ces méthodes consiste à trouver des transformations linéaires des variables de chaque ensemble afin de mesurer leur liaison. Ces différentes mesures de corrélation vectorielle sont présentées et comparées par Lazraq *et al.* [1992].

Pour comparer les deux matrices de données $\mathbf{Y_1}$ et $\mathbf{Y_2}$, Robert et Escoufier [1976] introduisent le coefficient *RV*, qui correspond au coefficient de corrélation vectorielle introduit par Escoufier [1973] dans un contexte de statistique multivariée. La démarche consiste à trouver un « résumé » de chacune des matrices de données afin, dans un deuxième temps, de comparer ces différents « résumés » et de mesurer leur similarité.

Pour « résumer » les matrices de données, Robert et Escoufier [1976] proposent de considérer les distances entre les points à l'intérieur de chaque configuration. Ces distances entre points sont obtenues en calculant les matrices produits $\mathbf{Z_{11}} = \mathbf{Y_1}^\top \mathbf{Y_1}$ et $\mathbf{Z_{22}} = \mathbf{Y_2}^\top \mathbf{Y_2}$. Si les données sont centrées, ces matrices de distances sont les matrices de covariance des données (à un facteur multiplicatif $1/n$ près). La

comparaison entre $\mathbf{Y_1}$ et $\mathbf{Y_2}$ se réduit alors à la comparaison entre leurs matrices de distances respectives $\mathbf{Z_{11}}$ et $\mathbf{Z_{22}}$ de même dimension $[N \times N]$.

Cette comparaison s'effectue en calculant le coefficient RV de la manière suivante :

$$RV_{\mathbf{Y_1},\mathbf{Y_2}} = \frac{\text{trace}(\mathbf{Z_{11}}\mathbf{Z_{22}})}{\sqrt{\text{trace}(\mathbf{Z_{11}}\mathbf{Z_{11}})}\sqrt{\text{trace}(\mathbf{Z_{22}}\mathbf{Z_{22}})}} \tag{4.1}$$

Notons que les matrices $\mathbf{Z_{11}}$ et $\mathbf{Z_{22}}$ sont invariantes par translation ou rotation et qu'en leur associant un opérateur, il est possible de dériver un produit scalaire et une métrique associée, fondés sur la norme de Hilbert-Schmidt. Ainsi, en notant $|\mathbf{A}|_2 = \sqrt{\text{trace}(\mathbf{A}^\top\mathbf{A})}$ la norme de la matrice $\mathbf{A}$ et $\langle \mathbf{A}, \mathbf{B} \rangle = \text{trace}(\mathbf{A}^\top\mathbf{B})$ le produit scalaire entre les matrices $\mathbf{A}$ et $\mathbf{B}$, le coefficient RV peut s'écrire :

$$RV_{\mathbf{Y_1},\mathbf{Y_2}} = \frac{\langle \mathbf{Z_{11}}, \mathbf{Z_{22}} \rangle}{|\mathbf{Z_{11}}|_2 |\mathbf{Z_{22}}|_2} \tag{4.2}$$

Au sens de ce produit scalaire, le coefficient RV est le cosinus carré de l'angle entre $\mathbf{Z_{11}}$ et $\mathbf{Z_{22}}$. Il peut aussi être considéré comme une extension multivariée du coefficient de corrélation de Pearson. Si le coefficient RV a une valeur proche de 1, il indique une grande similarité entre les matrices de données $\mathbf{Y_1}$ et $\mathbf{Y_2}$, alors qu'un coefficient RV nul indique que les deux matrices $\mathbf{Y_1}$ et $\mathbf{Y_2}$ ne sont pas corrélées (et même indépendantes si elles sont normalement distribuées).

Comme le montrent Robert et Escoufier [1976], il est possible de relier le coefficient RV à la distance définie comme suit :

$$\begin{aligned} D(\mathbf{Y_1}, \mathbf{Y_2}) &\triangleq \left\| \frac{\mathbf{Z_{11}}}{\sqrt{\text{trace}(\mathbf{Z_{11}}^2)}} - \frac{\mathbf{Z_{22}}}{\sqrt{\text{trace}(\mathbf{Z_{22}}^2)}} \right\| \\ &= \sqrt{2}\sqrt{1 - \frac{\text{trace}(\mathbf{Z_{11}}\mathbf{Z_{22}})}{\sqrt{\text{trace}(\mathbf{Z_{11}}^2)}\sqrt{\text{trace}(\mathbf{Z_{22}}^2)}}} \\ &= \sqrt{2}\sqrt{1 - RV_{\mathbf{Y_1},\mathbf{Y_2}}} \end{aligned} \tag{4.3}$$

Il est intéressant de noter que la distance calculée par l'équation 4.3 mesure la similarité entre les N observations des deux matrices de données $\mathbf{Y_1}$ et $\mathbf{Y_2}$. Dans le cas des données d'IRMf, les données sujets ont été normalisées et N correspond alors au nombre de voxels dans le masque d'analyse. Ainsi, la distance calculée par l'équation 4.3 est une mesure de la similarité *spatiale* entre les données de deux sujets.

Mais les équations précédentes sont valables tant qu'une dimension au moins est commune aux deux matrices. En considérant le même nombre de variables dans les deux matrices de données $\mathbf{Y_1}$ et $\mathbf{Y_2}$ ($p = q$), il est possible de calculer les matrices produits $\mathbf{Z_{11}} = \mathbf{Y_1}\mathbf{Y_1}^\top$ et $\mathbf{Z_{22}} = \mathbf{Y_2}\mathbf{Y_2}^\top$ de taille $[p \times p]$, puis le coefficient $RV_{\mathbf{Y_1^\top},\mathbf{Y_2^\top}}$. Dans le cas des données d'IRMf, p correspond aux nombres de scans et ainsi, la distance correspondante $D(\mathbf{Y_1^\top}, \mathbf{Y_2^\top})$ est une mesure de similarité *temporelle* entre les données de deux sujets.

4.1.2 Adaptation du coefficient *RV* aux données d'IRM fonctionnelle

Le calcul du coefficient RV tel que définit par l'équation 4.1 utilise les données brutes des sujets pour fournir une mesure de similarité soit spatiale, soit temporelle, selon la dimension commune des matrices comparées.

Cependant, cette comparaison à partir des données brutes peut ne pas être appropriée en raison de l'existence de facteurs de non-intérêt, c'est-à-dire non reliés au paradigme expérimental. Ces facteurs de non-intérêt sont par exemple les dérives basse-fréquence, qui, si elles ne sont pas correctement retirées, peuvent perturber de manière aléatoire toute mesure de similarité.

Plus généralement, la mesure de similarité doit pouvoir être focalisée sur un aspect particulier des données, défini par l'expérimentateur en fonction de l'hypothèse qu'il souhaite tester sur le paradigme. Nous avons donc adapté le calcul du coefficient RV pour prendre en compte le paradigme expérimental, représenté par la matrice de dessin $\mathbf{X}$. Cette adaptation consiste à remplacer chaque matrice de données $\mathbf{Y_i}$ par sa projection dans l'espace défini par le modèle $\mathbf{X}^\top\mathbf{Y_i}$. De plus, l'expérimentateur peut choisir de ne considérer comme espace d'intérêt

qu'un sous-espace G de la matrice de dessin $\mathbf{X}$. Le modèle, comme les données, sont alors projetés dans ce sous-espace d'intérêt G pour devenir $\mathbf{X_G}$ et $\mathbf{Y_{iG}}$.

Il faut aussi que la mesure de similarité intègre le fait que les données d'IRMf ne sont indépendantes ni dans l'espace ni dans le temps. Nous avons donc introduit deux métriques permettant de prendre en compte les corrélations temporelles (métrique $\mathcal{M}$) et spatiales (métrique $\mathcal{N}$).

La métrique $\mathcal{M}$ est définie de la manière suivante :

$$\mathcal{M}^{-\frac{1}{2}} = (\mathbf{X_G}^\top \mathbf{V} \mathbf{X_G})^{-1/2} \tag{4.4}$$

Cette métrique $\mathcal{M}$ permet de corriger les différences d'échelle entre les régresseurs du modèle et de prendre en compte les corrélations temporelles dans les données, représentées par la matrice $\mathbf{V}$.

Même si l'existence de corrélations spatiales dans les données d'IRMf est démontrée, il existe peu de méthodes disponibles pour les corriger. En effet, la principale difficulté à résoudre est l'inversion de la matrice de covariance spatiale de dimension $[N \times N]$, qui ne peut être estimée précisément à partir des données.

Pour la définition de la métrique $\mathcal{N}$, nous avons supposé que les corrélations spatiales peuvent se représenter par une matrice de covariance spatiale diagonale (comme le proposent Worsley *et al.* [1997]) :

$$\mathcal{N}^{-\frac{1}{2}} = \begin{bmatrix} \frac{1}{\hat{\sigma}_1} & 0 & \cdots & \cdots & 0 \\ 0 & \frac{1}{\hat{\sigma}_2} & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & \cdots & \frac{1}{\hat{\sigma}_{N-1}} & 0 \\ 0 & \cdots & \cdots & 0 & \frac{1}{\hat{\sigma}_N} \end{bmatrix} \tag{4.5}$$

Les termes diagonaux $(\hat{\sigma}_1, \ldots, \hat{\sigma}_N)$ sont les estimations de la variance résiduelle en chaque voxel calculées lors de la régression linéaire.

Les deux matrices de données $\mathbf{Y_1}$ et $\mathbf{Y_2}$ sont donc remplacées par les deux matrices suivantes pour le calcul du coefficient RV :

$$\begin{cases} \mathbf{Y_1^*} = \mathcal{M}^{-\frac{1}{2}}\mathbf{X_G}^\top\mathbf{Y_{1G}}\mathcal{N}^{-\frac{1}{2}} \\ \\ \mathbf{Y_2^*} = \mathcal{M}^{-\frac{1}{2}}\mathbf{X_G}^\top\mathbf{Y_{2G}}\mathcal{N}^{-\frac{1}{2}} \end{cases} \tag{4.6}$$

Les équations 4.1 et 4.3 permettent alors de calculer la distance *spatiale* adaptée aux données d'IRMf $D(\mathbf{Y_1^*}, \mathbf{Y_2^*})$, ainsi que la distance *temporelle* $D(\mathbf{Y_1^{*\top}}, \mathbf{Y_2^{*\top}})$.

Enfin, une dernière modification est apportée au calcul du coefficient RV. En raison du grand nombre de voxels dans les images d'IRMf, le calcul des matrices de distances dans le cas de la mesure de similarité *spatiale* ($\mathbf{Z_{ii}} = \mathbf{Y_i}^{*\top}\mathbf{Y_i^*}$ de dimension $[N \times N]$) peut très vite s'avérer très coûteux en temps de calcul et en espace mémoire. En revanche, comme les données sont projetées sur un sous-espace d'intérêt G, le nombre de variables p est réduit. Le calcul des traces de matrices de dimension $[N \times N]$ dans l'équation 4.1 est donc évité en utilisant les produits de Hadamard suivants, qui permettent de n'avoir à calculer que des matrices de dimension $[p \times p]$:

$$\text{trace}(\mathbf{Z_{11}Z_{11}}) = \text{trace}(\mathbf{Y_1}^{*\top}\mathbf{Y_1^*}\mathbf{Y_1}^{*\top}\mathbf{Y_1^*}) = \text{trace}(\mathbf{Y_1^*Y_1^{*\top}Y_1}^{*}\mathbf{Y_1^{*\top}})$$

$$\text{trace}(\mathbf{Z_{22}Z_{22}}) = \text{trace}(\mathbf{Y_2}^{*\top}\mathbf{Y_2^*}\mathbf{Y_2}^{*\top}\mathbf{Y_2^*}) = \text{trace}(\mathbf{Y_2^*Y_2^{*\top}Y_2}^{*}\mathbf{Y_2^{*\top}})$$

$$\text{trace}(\mathbf{Z_{11}Z_{22}}) = \text{trace}(\mathbf{Y_1}^{*\top}\mathbf{Y_1^*}\mathbf{Y_2}^{*\top}\mathbf{Y_2^*}) = \text{trace}(\mathbf{Y_2^*Y_1^{*\top}Y_1}^{*}\mathbf{Y_2^{*\top}})$$

Une fois toutes ces modifications effectuées, il est possible de calculer une distance *spatiale* (ou *temporelle*) pour chaque couple de sujets et d'obtenir ainsi une matrice symétrique des distances inter-sujets. La figure 4.1 montre un exemple de matrice des distances *spatiales* inter-sujets. Chaque ligne (ou colonne) correspond aux distances calculées entre un sujet et chacun des autres sujets du groupe.

Cette simple représentation permet déjà d'étudier la structure du groupe. Sur la figure 4.1, le sujet numéro 8 présente des distances aux autres sujets très élevées alors que des couples de sujets semblent assez proches les uns des autres : le sujet 2 et le sujet 7, le sujet 5 et le sujet 9, le sujet 7 et le sujet 10.

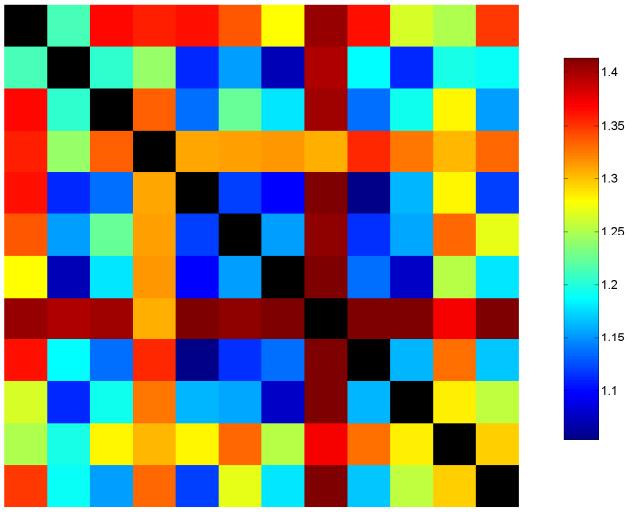

Figure 4.1 – *Matrice des distances spatiales inter-sujets – Contraste « Mémorisation implicite » – Jeu de données* ***K*** *décrit en annexe A*

4.1.3 Visualisation de la matrice des distances : la procédure de MDS

Le petit nombre de sujets impliqués dans les études d'IRMf ne favorisant pas l'utilisation de méthodes de classification, nous avons décidé d'opter plutôt pour un outil de visualisation afin de rendre compte de la structure du groupe résumée par la matrice des distances (*spatiales* ou *temporelles*).

Nous avons choisi d'utiliser une procédure de MDS (*MultiDimensional Scaling*, Gower [1984]) permettant de visualiser la matrice des distances dans un espace euclidien à 2 ou 3 dimensions. Pour cela, la technique de MDS consiste à déterminer la configuration optimale des sujets dans cet espace euclidien, par minimisation de

la différence entre les distances inter-sujets dans la représentation de MDS et les distances inter-sujets dans la matrice d'origine.

Il existe plusieurs algorithmes pour calculer la représentation de MDS. Nous avons décidé d'employer la version « métrique » proposée par Chatfield et Collins [1992]. Cette version repose sur la décomposition en vecteurs propres de la matrice de distances doublement centrée (par rapport à ses lignes et à ses colonnes).

Cet outil de visualisation simple permet d'observer l'homogénéité du groupe ou alors d'éventuelles hétérogénéités, comme la présence de données atypiques ou de sous-groupes. La figure 4.2 représente la matrice des distances inter-sujets (présentée figure 4.1) dans l'espace euclidien engendré par les deux premiers vecteurs propres de la procédure de MDS. Cette représentation confirme que le sujet numéro 8 présente de fortes distances par rapport aux autres sujets du groupe alors que certains sujets semblent assez proches les uns des autres.

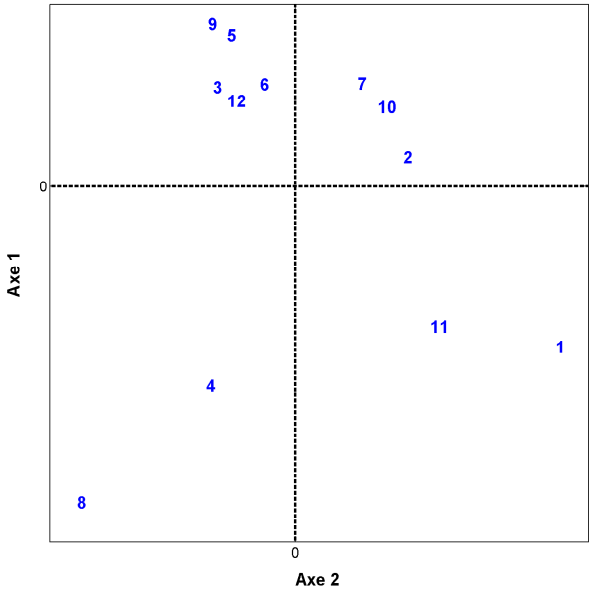

Figure 4.2 – *Procédure de MDS (2 premiers vecteurs propres) sur une matrice des distances spatiales inter-sujets – Contraste « Mémorisation implicite » – Jeu de données* ***K*** *décrit en annexe A*

Cependant, la procédure de MDS appliquée à la matrice des distances ne permet pas toujours de conclure de manière significative quant à la structure du groupe. En effet, les deux premiers vecteurs propres sont souvent insuffisants pour capturer l'essentiel de la variance contenue dans la matrice initiale des distances. Par exemple, la représentation 2D de MDS de la figure 4.2 ne capture que 44 % de la variance totale. Il est également possible d'étudier la structure du groupe en rajoutant une dimension, comme le montre la figure 4.3, mais même cette représentation 3D de MDS ne capture que $54,6$ % de la variance totale.

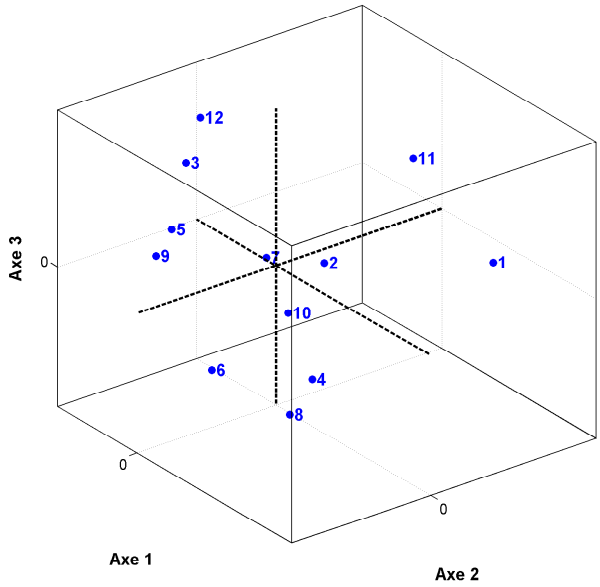

Figure 4.3 – *Procédure de MDS (3 premiers vecteurs propres) sur une matrice des distances spatiales inter-sujets – Contraste « Mémorisation implicite » – Jeu de données* $\boldsymbol{K}$ *décrit en annexe A*

Mis à part les cas où la variance capturée par les deux voire les trois premiers vecteurs propres est élevée, la procédure de MDS ne permet pas de conclure quant à l'homogénéité du groupe. Elle reste malgré tout un outil de visualisation simple et rapide de la matrice des distances calculée.

4.2 Détection multivariée de données atypiques : la distance de Cook

En complément de la procédure de MDS qui permet d'étudier l'homogénéité du groupe en visualisant la matrice des distances inter-sujets, nous avons développé une procédure dédiée à la détection d'éventuelles données atypiques. Cette procédure utilise une mesure d'influence, la distance de Cook, que nous avons adaptée à la matrice des distances inter-sujets.

4.2.1 Mesure d'influence : la distance de Cook

La distance de Cook [Cook, 1977] fait partie des méthodes de diagnostique de régression, fondée sur une approche par « suppression de données ». Afin de mesurer l'influence d'une donnée sur l'estimation d'un modèle de régression, ces méthodes proposent de comparer certains paramètres de régression obtenus avec et sans la donnée considérée.

Ainsi, la mesure DFBETA (*Difference of Fit for BETA*, Belsley *et al.* [1980]) compare les valeurs d'un paramètre β du modèle obtenues avec et sans la $i^{ème}$ observation lors de la régression. La mesure DFITS (*Difference of FITS*, Welsch et Kuh [1977]) quant à elle calcule la différence dans l'estimation de la $i^{ème}$ observation quand la régression du modèle est calculée avec et sans la $i^{ème}$ valeur. Notons qu'il existe encore d'autres mesures disponibles comme la distance de Mahalanobis [Mahalanobis, 1936], la distance de Welsch [Welsch, 1982], la mesure COVRATIO [Belsley *et al.*, 1980], ou encore des mesures utilisant les résidus standardisés ou « studentisés ».

Nous avons choisi d'utiliser la distance de Cook [Cook, 1977] qui permet de mesurer l'influence d'une donnée sur l'ensemble des paramètres du modèle de régression. Ainsi, comme le montrent Neter *et al.* [1985], si les données sont homogènes, elles doivent présenter des distances de Cook de même amplitude car elles sont supposées participer de manière équivalente à l'estimation des paramètres du modèle. En revanche, si une donnée présente une distance de Cook plus élevée que les autres, alors elle peut être déclarée comme atypique, c'est-à-dire comme

ayant une très grande influence susceptible de biaiser l'estimation des paramètres du modèle.

Considérons un modèle linéaire standard $\mathbf{y} = \mathbf{x}\mathbf{b} + \boldsymbol{\varepsilon}$, pour lequel $\mathbf{y}$ est un vecteur de K observations, $\mathbf{x}$ est une matrice de régresseurs de dimension $[K \times p]$, $\mathbf{b}$ est un vecteur de p paramètres inconnus et $\boldsymbol{\varepsilon}$ est un terme d'erreur supposé gaussien. La distance de Cook [Cook, 1977] correspond au carré d'une distance de type Mahalanobis entre le vecteur $\hat{\mathbf{b}}$ des paramètres estimés avec l'ensemble des K observations et le vecteur $\hat{\mathbf{b}_I}$ des paramètres estimés en retirant un sous-ensemble de I observations ($I < K$). Elle se définit ainsi :

$$D_I = \frac{(\hat{\mathbf{b}_I} - \hat{\mathbf{b}})^\top (\mathbf{x}^\top \mathbf{x})(\hat{\mathbf{b}_I} - \hat{\mathbf{b}})}{c\hat{\sigma}^2} = \frac{1}{c\hat{\sigma}^2} \|\mathbf{x}(\hat{\mathbf{b}_I} - \hat{\mathbf{b}})\|^2 \tag{4.7}$$

avec $\hat{\sigma}^2$ représentant une estimation non biaisée de la variance dans les observations et c est une constante laissée au choix de l'utilisateur.

4.2.2 Adaptation de la distance de Cook à la matrice des distances

Considérons la matrice symétrique suivante des distances entre n sujets :

$$\mathbf{M} = \begin{bmatrix} 0 & d_{1,2} & \cdots & \cdots & d_{1,n} \\ d_{1,2} & 0 & d_{2,3} & \cdots & d_{2,n} \\ \vdots & & \ddots & & \vdots \\ d_{1,n-1} & \cdots & \cdots & 0 & d_{n-1,n} \\ d_{1,n} & \cdots & \cdots & d_{n-1,n} & 0 \end{bmatrix}$$

où chaque $d_{j,l}$ correspond à la distance entre le sujet j (avec $1 \leq j \leq n-1$) et le sujet l (avec $j+1 \leq l \leq n$) calculée à l'aide du coefficient RV. La moyenne globale μ de toutes les distances inter-sujets vaut alors :

$$\mu = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{l=j+1}^{n} d_{j,l}$$

Nous avons décidé d'utiliser la distance de Cook pour détecter des sujets atypiques à partir de cette matrice $\mathbf{M}$ des distances inter-sujets. Pour cela, nous avons considéré que si le groupe de sujet est homogène alors toutes les valeurs des distances $d_{j,l}$ doivent être relativement proches. Cela implique que chaque sujet contribue (via ses distances par rapport aux autres sujets) de manière à peu près équivalente à la moyenne globale μ. Ainsi, détecter un sujet atypique revient à détecter un sujet qui a une trop grande influence sur la moyenne globale μ.

Cette mesure d'influence est obtenue en calculant la distance de Cook définie par l'équation 4.7 avec un modèle linéaire simplifié testant la moyenne des observations : $\mathbf{y} = [d_{1,2}, \ldots, d_{1,n}, d_{2,3}, \ldots, d_{2,n}, \ldots, d_{n-1,n}]^\top$ correspond au vecteur des $K = \frac{n(n-1)}{2}$ distances inter-sujets, et $\mathbf{x}$ est le vecteur $[1, 1, \ldots, 1]^\top$ de dimension $[K \times 1]$ afin de tester la moyenne globale des distances. Le vecteur $\hat{\mathbf{b}}$ des paramètres estimés correspond pour ce modèle simplifié à la moyenne globale μ des distances inter-sujets. Afin d'estimer l'influence du sujet i sur cette moyenne globale μ, l'idée est de considérer pour le vecteur $\hat{\mathbf{b}_I}$ la moyenne globale μ_i des distances inter-sujets obtenue en retirant le sujet i (c'est-à-dire en retirant du vecteur $\mathbf{y}$ le sous-ensemble des $I = n - 1$ distances associées au sujet i) :

$$\mu_i = \frac{2}{(n-1)(n-2)} \sum_{j=1, j \neq i}^{n-1} \sum_{l=j+1}^{n} d_{j,l}$$

La distance de Cook D_i mesurant l'influence du sujet i sur la moyenne globale des distances se définit en adaptant l'équation 4.7 de la manière suivante :

$$D_i = \frac{1}{c\hat{\sigma}^2} \|\mathbf{x}(\hat{\mathbf{b}_I} - \hat{\mathbf{b}})\|^2 = \frac{n(n-1)(\mu_i - \mu)^2}{2c\hat{\sigma}^2} \tag{4.8}$$

Pour le choix de la constante c et de l'estimateur $\hat{\sigma}^2$ non biaisée de la variance, il existe plusieurs possibilités détaillées par Cook et Weisberg [1982] : ils suggèrent de favoriser des valeurs « à géométrie fixe », c'est-à-dire ne dépendant pas du sujet i.

Pour le choix de $\hat{\sigma}^2$, nous avons utilisé l'estimateur non biaisé de la variance globale des distances inter-sujets :

$$\hat{\sigma}^2 = \frac{1}{\frac{n(n-1)}{2} - 1} \sum_{j=1}^{n-1} \sum_{l=j+1}^{n} (d_{j,l} - \mu)^2$$

Pour le choix de c, comme proposé par Gray [1993], nous avons choisi une constante indépendante du modèle, $c = n - 1$, correspondant aux nombres de distances non considérées lors du calcul de μ_i.

À partir de l'équation 4.8 et des valeurs choisies pour c et $\hat{\sigma}^2$, nous définissons la distance de Cook associée au sujet i par :

$$D_i = \frac{n(\mu_i - \mu)^2}{2\hat{\sigma}^2} \tag{4.9}$$

En affichant pour chaque sujet i la valeur de la distance de Cook D_i calculée avec l'équation 4.9, ainsi que la moyenne $m_i = \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} d_{i,j}$ de ses distances par rapport aux autres sujets, il est possible de détecter des données atypiques comme le montre la figure 4.4. Cet affichage permet de confirmer les résultats obtenus par la représentation de MDS (figures 4.2 et 4.3) : le sujet 8 présente la plus grande distance moyenne par rapport aux autres sujets et apparaît comme une donnée atypique étant donné sa grande influence sur la moyenne globale des distances (distance de Cook nettement plus élevée que les autres).

4.2.3 Valeur critique de la distance de Cook

La distance de Cook nous permet de mesurer l'influence de chaque sujet sur un paramètre du groupe (ici, la moyenne globale des distances inter-sujets), mais la question qui se pose est : « comment fixer une valeur critique de la distance de Cook au-delà de laquelle le sujet peut être déclaré comme atypique ? ».

Dans la littérature, il existe de nombreuses tentatives d'approximation de la distribution de la distance de Cook, afin de déterminer un seuil statistique au-delà

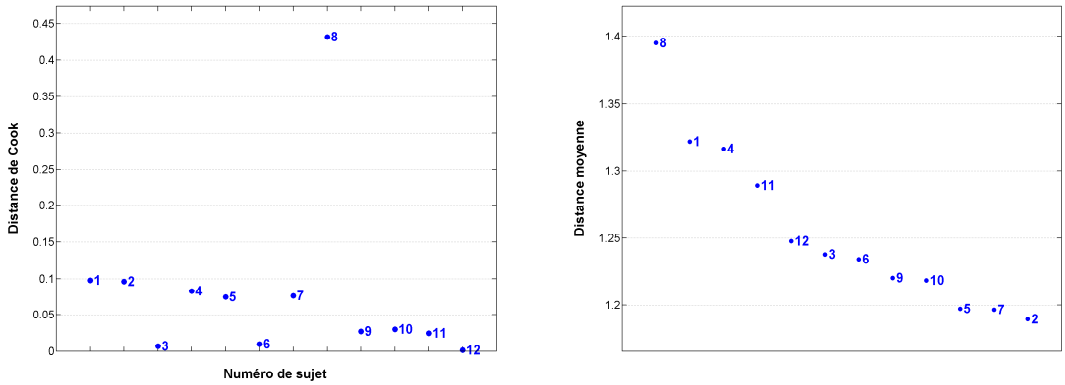

Figure 4.4 – *Distances de Cook (à gauche) et distances moyennes (à droite) calculées à partir de la matrice des distances spatiales inter-sujets – Contraste « Mémorisation implicite » – Jeu de données* $\boldsymbol{K}$ *décrit en annexe A*

duquel une donnée est déclarée atypique. Comme le montrent Jensen et Ramirez [1998], il est possible, sous certaines hypothèses de normalité et pour certains choix de l'estimateur de la variance $\hat{\sigma}^2$, d'exprimer la distance de Cook comme une somme pondérée de distributions de Fisher-Snedecor (distributions F). Cependant, dans le cas général, les propriétés statistiques de la distance de Cook restent inconnues et comme le rappellent Cook et Weisberg [1982], cette distance reste avant tout un outil diagnostique.

La plupart des auteurs proposent alors des valeurs critiques heuristiques. Certains préconisent l'utilisation de $D_i > 1$, considérant qu'une donnée est atypique si son influence dépasse une fois au moins la variance $\hat{\sigma}^2$. D'autres auteurs, notamment Cook et Weisberg [1982], proposent de prendre les 50 % de la distribution $F(p, n-p)$ comme valeur critique, avec p le nombre de paramètres du modèle, soit $F(0.5, 1, n-1)$ dans notre cas. Ce choix de valeur critique fait l'hypothèse qu'une donnée peut être déclarée comme atypique si elle fait sortir l'estimateur d'un intervalle de confiance de 50 %. En étudiant les propriétés de la mesure DFITS et en reliant cette mesure à la distance de Cook [Belsley *et al.*, 1980], Bollen et Jackman [1990] suggèrent d'utiliser $D_i > \frac{4}{n}$. Enfin, Fox [1991] propose de corriger cette dernière valeur critique pour prendre en compte le modèle considéré : $D_i > \frac{4}{n-p-1} = \frac{4}{n-2}$.

Afin de choisir la valeur critique D_c de la distance de Cook, nous avons comparé, par simulations, les différents taux de détection de données atypiques obtenus en utilisant les quatre valeurs présentées précédemment : $D_c = 1$, $D_c = F(0.5, 1, n-1)$, $D_c = \frac{4}{n}$ et $D_c = \frac{4}{n-2}$.

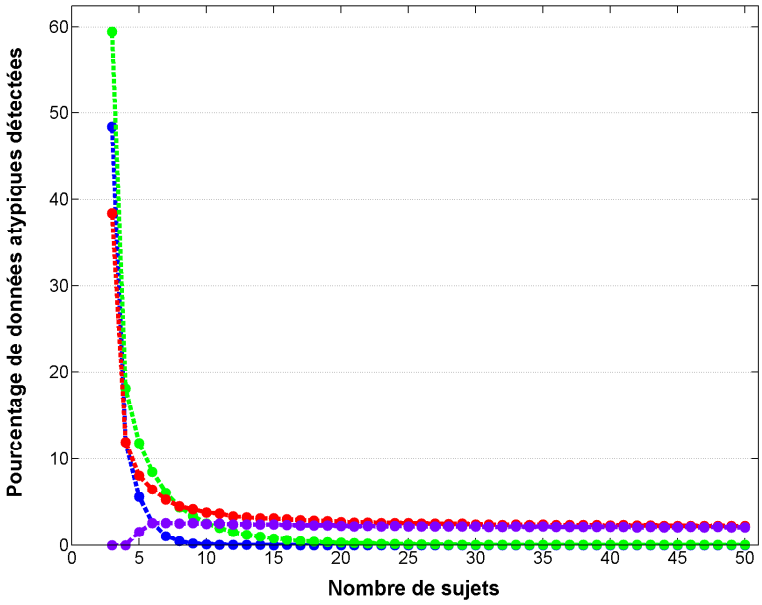

Figure 4.5 – *Taux de détection de données atypiques en utilisant différentes valeurs critiques pour la distance de Cook :* $D_c = 1$ *(en bleu),* $D_c = F(0.5, 1, n-1)$ *(en vert),* $D_c = \frac{4}{n}$ *(en rouge) et* $D_c = \frac{4}{n-2}$ *(en violet)*

Pour la première série de simulations, nous avons généré pour chaque sujet une matrice d'effets de dimension $[1 \times 1000]$, contenant des valeurs normalement distribuées (de moyenne nulle et de variance 1). Nous avons ensuite calculé la matrice des distances euclidiennes entre tous les sujets, et enfin déterminé la distance de Cook associée à chaque sujet. Ces distances ont alors été comparées aux quatre valeurs critiques afin d'estimer leur taux de détection de données atypiques respectif. Ces différents taux sont représentés sur la figure 4.5 pour des populations de 3 à 50 sujets simulés. Les résultats obtenus montrent que les valeurs critiques proposées par Bollen et Jackman [1990] et Fox [1991] permettent de conserver un taux de

détection de données atypiques à peu près constant ($\simeq 2,5$ %) lorsque le nombre de sujets augmente et dépasse 10. De plus, pour des populations de moins de 10 sujets, il faut noter que les valeurs critiques testées, mise à part celle proposée par Fox [1991], semblent anti-conservatives puisque le taux de détection de données atypiques augmente très rapidement à mesure que le nombre de sujets diminue.

Pour la deuxième série de simulations, nous avons généré une matrice des distances pour une population de sujets hétérogènes : les distances pour les sujets homogènes ont été tirées dans une loi uniforme $[0;1]$ et celles simulant des sujets atypiques ont été tirées dans une loi uniforme $[4;5]$. La figure 4.6 représente la comparaison entre le nombre de sujets atypiques simulés dans la matrice des distances et le nombre de sujets atypiques détectés en utilisant les quatre valeurs critiques, pour différents pourcentages de données atypiques dans des populations de 3 à 50 sujets simulés. Les résultats obtenus confirment que les valeurs critiques proposées par Bollen et Jackman [1990] et Fox [1991] possèdent le meilleure pouvoir de détection des données atypiques. De plus, en les comparant, il faut noter que la valeur critique proposée par Bollen et Jackman [1990] semble un peu plus performante pour les pourcentages de données atypiques élevés.

Ces résultats de simulations nous ont conduit à considérer la valeur critique proposée par Bollen et Jackman [1990] pour la distance de Cook :

$$D_c = \frac{4}{n} \tag{4.10}$$

Ainsi, nous privilégions une valeur critique assurant un pouvoir de détection à peu près constant quel que soit le pourcentage de données atypiques et le nombre de sujets dans la population. De plus, lorsque le nombre de sujets est inférieur à 10, nous favorisons une valeur critique anti-conservative, de manière à nous assurer de la détection d'au moins toutes les données atypiques présentes dans la population, étant donné que c'est pour les petits échantillons que la vérification de l'homogénéité est la plus nécessaire.

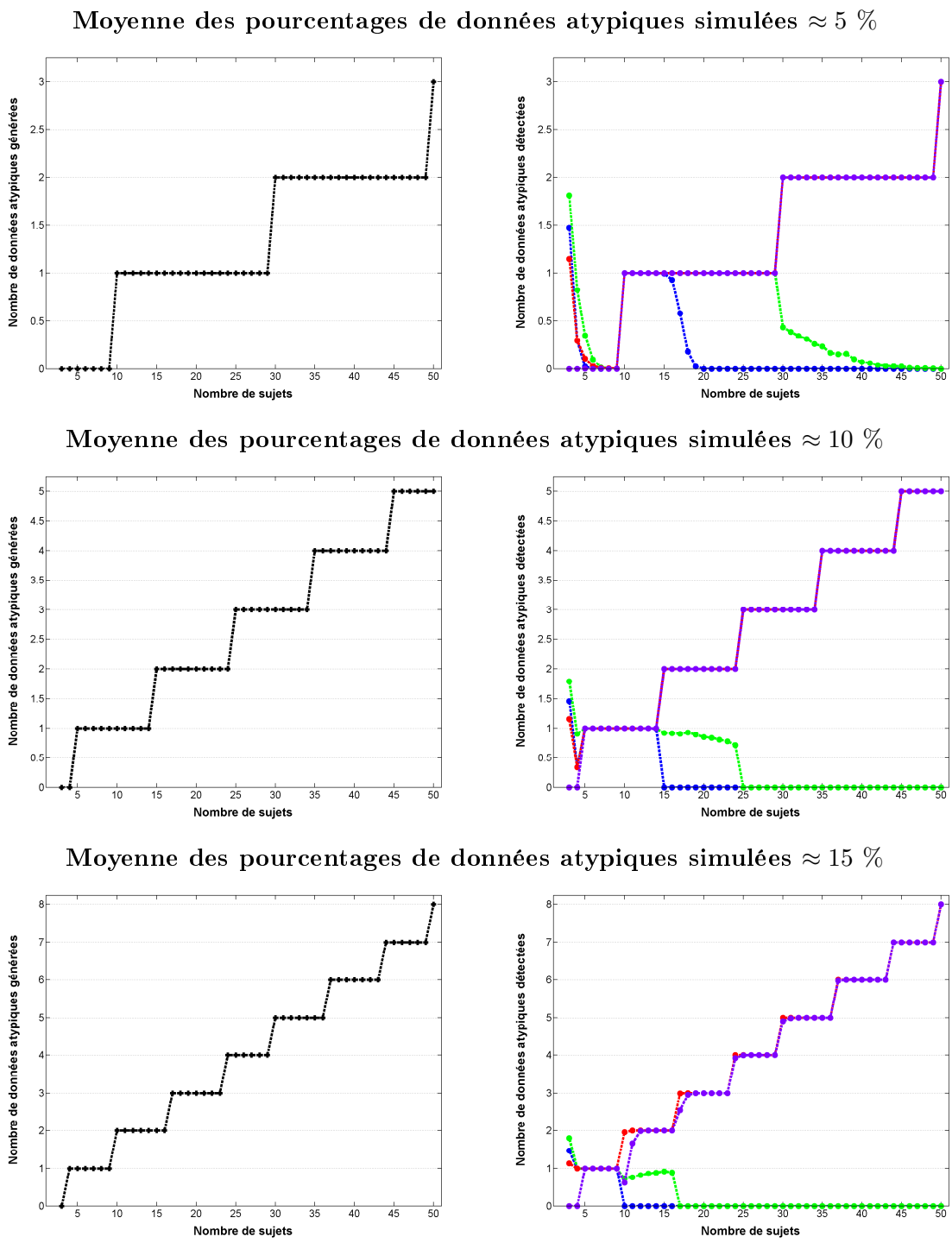

Figure 4.6 – *Nombre de données atypiques simulées (en noir) et nombre de données atypiques détectées en utilisant différentes valeurs critiques pour la distance de Cook :* $D_c = 1$ *(en bleu),* $D_c = F(0.5, 1, n-1)$ *(en vert),* $D_c = \frac{4}{n}$ *(en rouge) et* $D_c = \frac{4}{n-2}$ *(en violet)*

Si nous reprenons la figure 4.4 en y ajoutant la valeur critique calculée par l'équation 4.10, nous confirmons sur la figure 4.7 suivante que le sujet 8 peut être considéré comme un sujet atypique et qu'il est donc susceptible de biaiser les résultats de l'analyse de groupe du fait de sa trop grande influence.

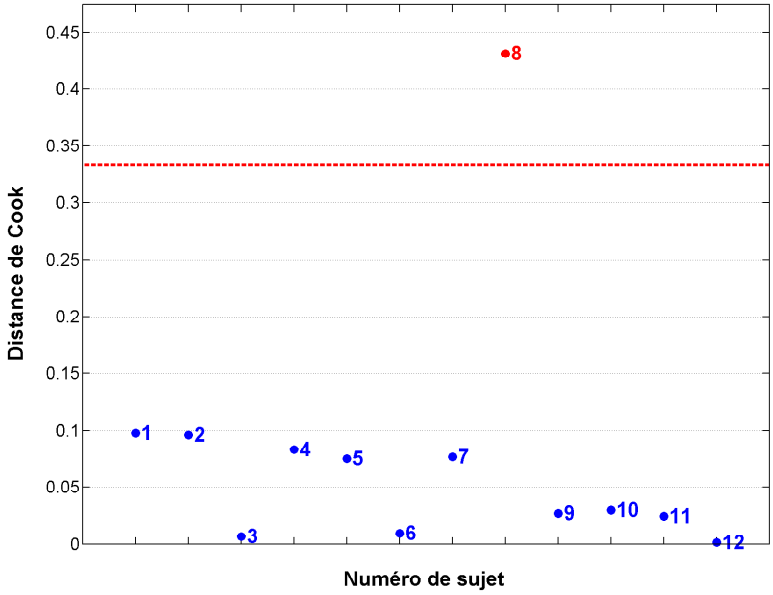

Figure 4.7 – *Distances de Cook et valeur critique associée (ligne pointillée rouge) pour la matrice des distances spatiales inter-sujets – Contraste « Mémorisation implicite » – Jeu de données* $\boldsymbol{K}$ *décrit en annexe A*

4.3 Conclusion

Dans ce chapitre, nous avons défini une distance entre sujets fondée sur la ressemblance entre les cartes d'effets estimés. Ainsi, il est possible d'étudier l'homogénéité du groupe de manière globale en calculant une matrice de distances inter-sujets. Cette matrice peut révéler la présence de sous-groupes de sujets ou de sujets atypiques lorsqu'elle est visualisée par une procédure de MDS ou lorsque les distances moyennes de chaque sujet par rapport aux autres sont comparées.

La difficulté est que nous voudrions pouvoir utiliser cette matrice pour tester l'hypothèse de normalité des effets à travers les sujets. Notre idée est que si la population de sujets est relativement « homogène » alors les distances inter-sujets calculées doivent être relativement proches. Nous proposons d'utiliser la distance de Cook pour estimer l'influence de chaque sujet (via sa distance aux autres) sur la moyenne globale des distances. Une valeur critique de distance de Cook est alors choisie pour évaluer l'homogénéité du groupe de sujets.

Cependant, notre approche ne constitue pas pour l'instant une méthode de test d'homogénéité : elle permet un diagnostique d'influence en détectant les éventuels sujets atypiques qui présentent une trop grande influence sur la moyenne globale. Nous proposons alors d'utiliser un vrai test de normalité au niveau du voxel afin de valider l'hypothèse d'homogénéité des effets à travers les sujets (chapitre suivant).

Chapitre 5

Test d'homogénéité de Grubbs

Dans le chapitre précédent, nous avons présenté une procédure multivariée de contrôle de l'homogénéité des données, permettant de comparer directement les matrices de données de chaque sujet entre elles. Cette procédure nous renseigne sur la présence de sujets atypiques ou de sous-groupes dans la population de sujets étudiés, mais elle ne constitue pas un test de normalité à proprement parler.

L'analyse de groupe « standard » utilise un test statistique univarié. Il est donc nécessaire de contrôler l'hypothèse d'homogénéité du groupe en chaque voxel : en effet, même si aucun sujet n'apparaît comme atypique en calculant la matrice des distances inter-sujets et les distances de Cook associées, les effets estimés à travers les sujets peuvent être hétérogènes seulement dans une zone spécifique du cerveau, les résultats de l'analyse de groupe « standard » risquant par là-même d'être biaisés dans cette zone.

Nous proposons pour cela d'utiliser un test de normalité univarié (5.1), permettant de valider (ou d'infirmer) l'hypothèse de normalité des effets à travers les sujets au niveau du voxel. Ce test utilise la statistique de Grubbs (5.2), dont la distribution est tabulée par simulations de Monte-Carlo afin de définir un seuil statistique (5.3).

5.1 Tests de normalité

Il existe dans la littérature un très large éventail de tests de normalité ; par exemple, Thode [2002] recense au moins 40 tests différents sans compter les extensions et variantes de ceux-ci. Parmi les plus usités se trouvent le test de Kolmogorov-Smirnov, le test d'Anderson-Darling, le test Cramér-von Mises et le test de Shapiro-Wilk.

Le test de conformité de Kolmogorov-Smirnov consiste à mesurer l'écart maximum entre la fonction de répartition empirique de la variable testée et celle d'une variable normale. L'hypothèse de normalité est alors rejetée si au moins un des écarts calculés est supérieur à une valeur limite fournie par tabulation. Lilliefors [1967] a adapté le test de Kolmogorov-Smirnov au cas où les paramètres (moyenne et variance) de la distribution normale sont inconnus, ce qui permet d'augmenter la puissance du test.

Le test d'Anderson-Darling [Anderson et Darling, 1952] modifie quant à lui le test de Kolmogorov-Smirnov en donnant plus de poids aux valeurs extrêmes de la distribution. Les valeurs limites utilisées pour rejeter l'hypothèse de normalité dans le cas du test de Kolmogorov-Smirnov sont indépendantes de la distribution testée. Le test d'Anderson-Darling propose de calculer ces valeurs en tenant compte de la distribution testée ce qui permet d'augmenter la sensibilité du test.

Le test de Cramér-von Mises est lui aussi une adaptation du test de Kolmogorov-Smirnov. La différence entre ces deux tests réside dans le fait que pour le test de Kolmogorov-Smirnov, seul l'écart maximum entre la distribution empirique et la distribution normale d'ajustement entre en considération, alors que l'indicateur d'écart du test de Cramér-von Mises prend mieux en compte l'ensemble des données en ce sens que la somme des écarts intervient. Le test de Kolmogorov-Smirnov est donc beaucoup plus sensible à l'existence de points aberrants dans un échantillon que le test de Cramér-von Mises. Ce dernier test est généralement plus puissant, mais cela n'a pas été démontré théoriquement.

À la différence des tests de Kolmogorov-Smirnov, Anderson-Darling et Cramér-von Mises qui utilisent la fonction de distribution empirique des données, la statistique W du test de Shapiro-Wilk [Shapiro *et al.*, 1968] est construite par la régression des valeurs observées rangées par ordre croissant sur les valeurs cor-

respondantes attendues dans le cas d'une distribution normale. Lorsque la statistique W est proche de 1, cela indique une régression presque linéaire et donc une distribution des données proche d'une distribution normale. La statistique W peut aussi s'écrire comme le rapport de deux statistiques, estimant chacune la variance des données sous l'hypothèse de la normalité de la distribution. À l'origine, Shapiro *et al.* [1968] ont montré la validité de cette statistique W pour des échantillons de 3 à 50 données, et elle a été étendue aux plus grands échantillons par Royston [1982].

De manière générale, les tests d'Anderson-Darling et de Shapiro-Wilk apparaissent comme les plus puissants pour tester la normalité, principalement dans le cas des petits échantillons. Nous n'avons pas encore évalué ces méthodes expérimentalement. Notre objectif premier étant, non pas de tester la normalité, mais plutôt de détecter les régions pour lesquelles le test t paramétrique risque d'être fortement biaisé, il nous est appraru plus naturel d'utiliser comme statistique de test une mesure de « discordance », comme la définissent Barnett et Lewis [1994]. Un test de « discordance » correspond à une procédure de détection permettant de décider si une valeur donnée peut être considérée comme faisant partie de la population principale. Barnett et Lewis [1994] classent les différents tests de « discordance » en sept catégories, sachant qu'ils peuvent être distingués en fonction du type de distribution de la population-parent dont provient l'échantillon analysé (normale, exponentielle, gamma, uniforme,...) mais aussi des critères retenus pour former la statistique du test (écart entre les valeurs, écart entre les extrêmes,...). Parmi l'ensemble de ces tests, nous avons choisi le test de Grubbs qui utilise une statistique liée au rapport *écart / étalement*.

5.2 Définition de la statistique de Grubbs

Le test de Grubbs, aussi appelé test ESD (*Extreme Studentized Deviation*), a été introduit par Grubbs [1950, 1969] pour tester la présence de données aberrantes dans un échantillon, c'est-à-dire en fait tester l'hypothèse nulle que la distribution dont proviennent les données est normale. La statistique de Grubbs est construite pour détecter un écart à la normalité en terme d'influence d'une observation sur la

moyenne, à la différence de la statistique de Cochran [Cochran, 1950], qui détecte les données atypiques en terme de dispersion.

Considérons un échantillon $(\hat{\beta}_1, \ldots, \hat{\beta}_n)$ d'effets estimés dans un voxel donné pour n sujets. La statistique de Grubbs [Grubbs, 1950, 1969] se définit par :

$$G = \frac{\max_i \mid \hat{\beta}_i - \bar{\beta} \mid}{s} \tag{5.1}$$

Dans cette équation, $\bar{\beta}$ représente la moyenne empirique de l'échantillon d'effets estimés et s l'écart-type :

$$\bar{\beta} = \frac{\sum_{i=1}^{n} \hat{\beta}_i}{n}, \qquad s = \sqrt{\frac{\sum_{i=1}^{n} (\hat{\beta}_i - \bar{\beta})^2}{n-1}}$$

Ainsi, la statistique de Grubbs correspond au plus grand écart absolu à la moyenne, en multiples de l'écart-type. Appliquée en chaque voxel, l'équation 5.1 permet d'obtenir une carte statistique de Grubbs pour le groupe de sujets, qu'il reste à seuiller pour détecter les régions où l'hypothèse de normalité est mise en défaut.

Tietjen et Moore [1972] ont montré que la statistique de Grubbs est optimale en un certain sens pour la détection d'une seule donnée atypique. Pour la détection de plusieurs données atypiques, la procédure classique consiste à itérer le test de Grubbs, en retirant de l'échantillon la donnée atypique détectée après chaque itération. Cependant, cette approche n'est pas optimale puisque la probabilité de détecter une donnée atypique varie à chaque itération. D'autres tests sont dédiés à la détection simultanée de plusieurs données atypiques, comme le test de Dixon [Dixon, 1950], le test de Rosner [Rosner, 1975] ou encore le test de Grubbs étendu proposé par Tietjen et Moore [1972]. Notons enfin que le test de Grubbs (avec celui de Cochran) est exploité par la norme internationale ISO 5725 lors d'applications de méthodes statistiques pour la maîtrise de la qualité, et plus spécifiquement pour la détermination de la répétabilité et la reproductibilité d'une méthode de mesure.

5.3 Distribution de la statistique de Grubbs

Sous l'hypothèse de normalité des données, il est possible, pour un risque de première espèce α choisi par l'expérimentateur, de définir une valeur critique G_c pour la statistique de Grubbs au-delà de laquelle l'hypothèse que l'échantillon ne contient pas au moins une donnée atypique est rejetée.

Grubbs [1950, 1969] a proposé la valeur critique G_c analytique suivante, en fonction du risque α :

$$G_c = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{(\frac{\alpha}{2n},n-2)}}{n-2+t^2_{(\frac{\alpha}{2n},n-2)}}} \tag{5.2}$$

dans laquelle $t_{(\frac{\alpha}{2n},n-2)}$ est la valeur critique d'une distribution de Student à $n-2$ degrés de liberté pour un risque de première espèce $\frac{\alpha}{2n}$.

Afin de vérifier la valeur critique G_c proposée par Grubbs [1950, 1969], nous avons estimé la distribution de la statistique de Grubbs par simulations de Monte-Carlo. 10^6 échantillons de n données normales ont été générées pour différentes valeurs de n, et la fonction de distribution de la statistique de Grubbs obtenue a été comparée à la fonction de distribution calculée avec l'équation 5.2 (voir la figure 5.1).

Les résultats de simulation obtenus montrent que la valeur critique G_c proposée par Grubbs [1950, 1969] est plutôt anti-conservative pour les valeurs élevées de la statistique de Grubbs et plutôt conservative pour les valeurs faibles. Il apparaît de plus que cette valeur critique G_c approche la distribution estimée par simulations de Monte-Carlo uniquement pour les grands échantillons ($n \geq 50$) et pour les valeurs élevées de la statistique de Grubbs ($G \geq 2,75$).

À la vue de ces résultats, la mise en œuvre informatique dans le logiciel *DISTANCE* du calcul de la valeur critique pour le test de Grubbs a été réalisée par simulations de Monte-Carlo, le choix du nombre de simulations à effectuer étant laissé à l'expérimentateur.

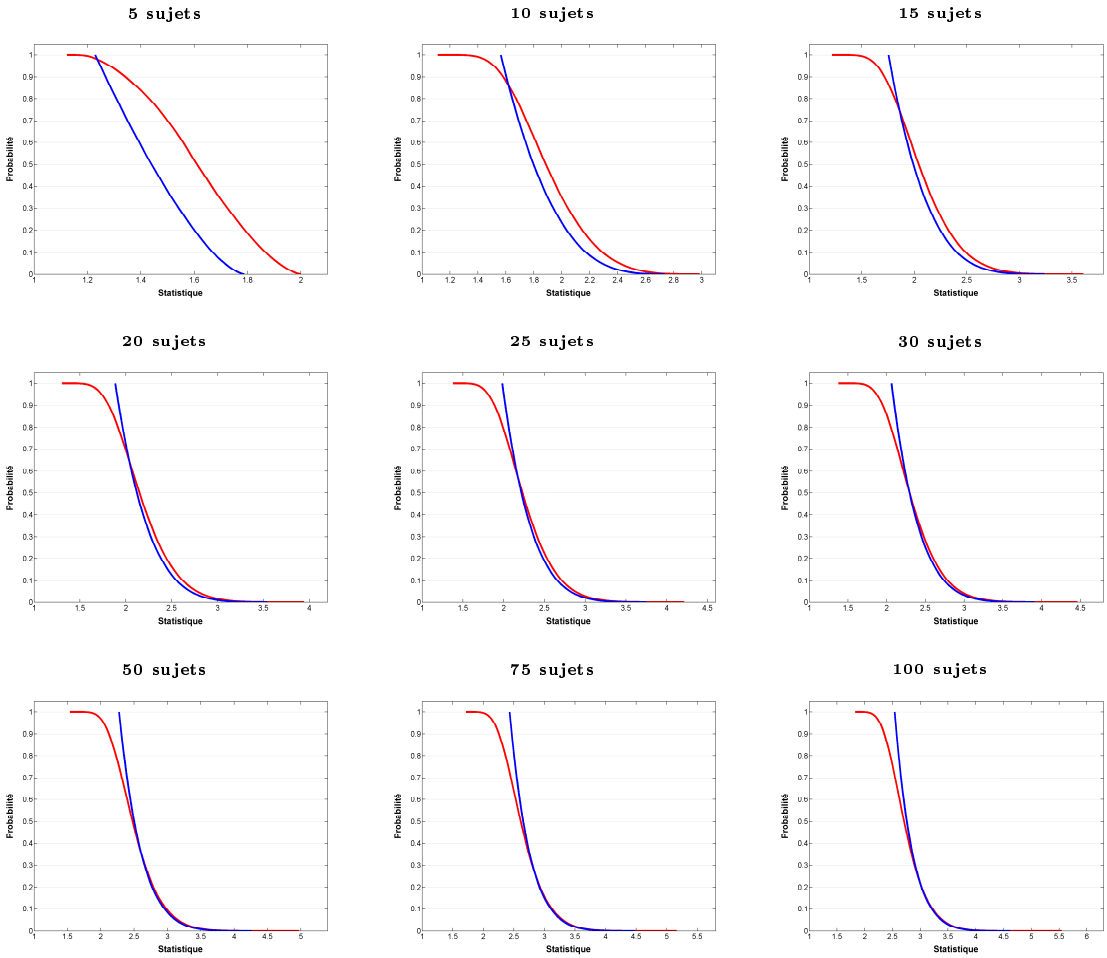

Figure 5.1 – *Fonction de distribution de la statistique de Grubbs : estimée par simulations de Monte-Carlo (rouge) ou calculée avec l'équation 5.2 (bleu)*

La figure 5.2 présente la carte statistique de Grubbs obtenue pour le jeu de données **K** décrit en annexe A, seuillée pour la valeur critique $G_c = 2,75$ estimée par simulations de Monte-Carlo en considérant un risque de première espèce de 1 % non corrigé des comparaisons multiples ($P_{non\ corrigée} < 0,01$). Seuls les voxels dépassant ce seuil et appartenant à des clusters d'extension spatiale supérieure à 50 voxels sont affichés ($E \geq 50$). Cette carte seuillée est présentée pour quatre positions différentes dans l'espace de Talairach, correspondant au maximum local de la statistique de Grubbs dans les quatre clusters les plus étendus. Sont également présentés les effets estimés à travers les sujets pour les quatre positions considérées.

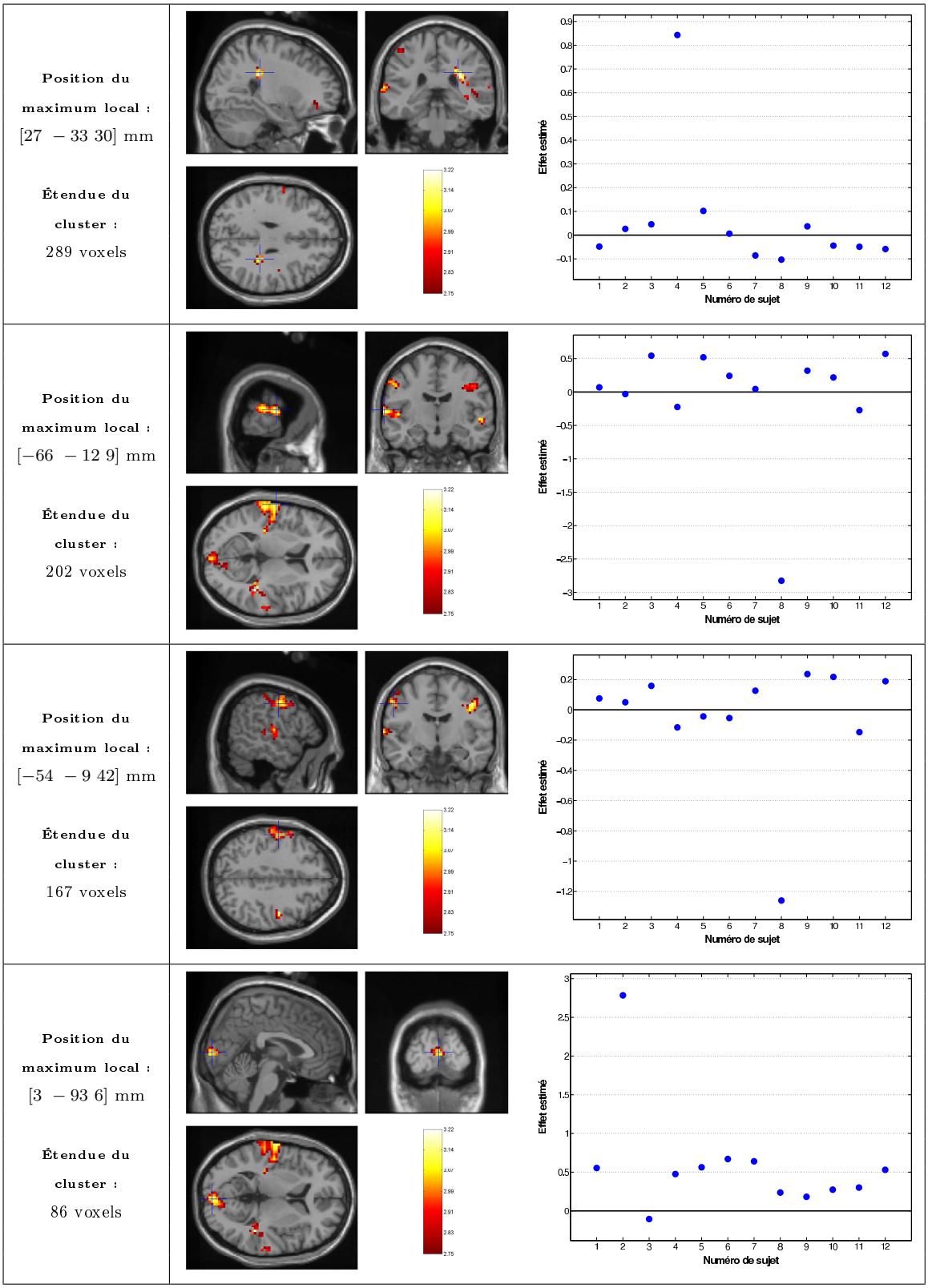

Figure 5.2 – *Carte statistique de Grubbs et liste des effets estimés à travers les sujets pour quatre positions dans l'espace de Talairach – Contraste « Mémorisation implicite » – Jeu de données* ***K*** *décrit en annexe A*

La figure 5.2 permet tout d'abord de confirmer que l'hypothèse d'homogénéité des données n'est pas valide dans plusieurs régions cérébrales. De plus, il est intéressant de noter que le sujet numéro 8, déjà identifié comme atypique par les résultats de la procédure de MDS (figure 4.2) et de calcul des distances de Cook (figure 4.7), est aussi détecté comme atypique par le test de Grubbs dans deux régions distinctes (maxima locaux en $[-66\ \ -12\ 9]$ et $[-54\ \ -9\ 42]$ mm dans l'espace de Talairach). Enfin, le test de Grubbs fait apparaître deux autres données atypiques (les sujets 2 et 4) dans deux autres régions cérébrales (maxima locaux en $[27\ \ -33\ 30]$ et $[3\ \ -93\ 6]$ mm dans l'espace de Talairach). Ces résultats illustrent l'intérêt de compléter le diagnostique d'homogénéité multivarié par un test univarié.

5.4 Conclusion

À la différence de la méthode de diagnostique d'influence présentée dans le chapitre précédent, le test statistique de Grubbs permet de tester l'hypothèse de distribution normale massivement univariée des effets à travers les sujets, hypothèse de base de l'analyse de groupe « standard ». Notre méthode de calcul du seuil par simulations de Monte-Carlo a permis de montrer que la formule analytique proposée dans la littérature n'est vérifiée qu'asymptotiquement ce qui rendait en pratique le test de Grubbs anti-conservatif pour les petits échantillons de données.

L'exemple proposé confirme l'intérêt d'une approche massivement univariée, qui a l'avantage de donner une information locale quant à la présence éventuelle de données atypiques. En effet, l'hétérogénéité à travers les sujets peut ne se révéler que localement, par exemple à cause d'erreurs de recalage ou de normalisation dans certaines régions. Ainsi la conception selon laquelle certains sujets sont globalement atypiques nous apparaît simpliste, ce qui nous amène à favoriser des méthodes capables de pondérer les sujets localement, et non globalement, en fonction de leurs effets estimés.

Chapitre 6

Résultats des études d'homogénéité

La neuro-imagerie fonctionnelle par IRM est relativement récente puisqu'elle débute dans les années 1990 avec les travaux d'Ogawa sur l'effet BOLD [Ogawa *et al.*, 1990]. Cette relative jeunesse explique sans doute que les travaux de méta-analyse soient encore relativement peu fréquents. Comme nous l'avons expliqué dans l'introduction, l'utilisation la plus courante de l'IRMf est la découverte ou l'étude des réseaux cérébraux impliqués dans un processus cognitif ou sensori-moteur pour une certaine population d'intérêt. Nous avons vu que la méthodologie actuelle propose d'échantillonner environ 15 sujets dans la population à étudier, mais de nombreuses études ne sont réalisées qu'avec 10-12 sujets. Dans une population parente homogène, suivant l'échantillonnage réalisé, les résultats vont différer quelque peu. Mais si la population parente est hétérogène, dans le sens où elle contient plusieurs sous-populations, les hypothèses classiques de l'analyse de groupe ne seront pas respectées, et les résultats obtenus seront difficilement interprétables. Enfin, si l'échantillonnage conduit à sélectionner un ou plusieurs sujets placés dans les extrêmes de la distribution de la population, il se peut aussi que l'analyse de groupe ne reflète pas de manière fidèle l'effet moyen de la population.

Il nous est apparu essentiel de savoir si les études d'IRMf étaient plus ou moins

fréquemment sujettes à ce type de problème. En effet, l'interprétation du fonctionnement cérébral par les neuro-scientifiques se fonde en grande partie sur ces résultats, et il n'est pas rare d'observer dans les discussions des articles de la littérature de longs débats sur l'implication de telle ou telle aire cérébrale dans tel ou tel processus cognitif, sur la base du résultat d'un groupe de 10 sujets soumis à 2 stimuli. Nous nous posons donc la question de savoir si ces études sont majoritairement fidèles et proposent une bonne représentation de l'effet sur la population parente (homogène), ou si au contraire, les effets que nous évoquions précédemment sont importants. Il pourrait paraître étrange qu'aucune étude n'ait tenté de quantifier la fréquence avec laquelle on peut raisonnablement conclure que l'analyse classique n'est pas représentative de l'effet sur la population.

Pourtant, cet état de fait n'est pas si surprenant. L'observation de ce type de phénomène nécessite de travailler avec un assez grand nombre d'études d'IRMf. L'accès et la gestion des données de ce grand nombre d'études constituent en soi une difficulté. L'accès nécessite la mise en contact et la confiance des expérimentateurs. La gestion nécessite généralement le stockage et l'organisation d'un large volume de données d'imagerie.

Dans ce chapitre, nous résumons les différents résultats obtenus en appliquant la méthode de diagnostic d'influence (présentée chapitre 4) et le test d'homogénéité de Grubbs (présenté chapitre 5) à un ensemble de 20 jeux de données. Bien que d'autre tests ou techniques auraient pu être appliqués, ces méthodes devraient suffire à quantifier le problème de la représentativité de l'analyse de groupe. Les données utilisées ont été collectées tout au long de la thèse auprès d'équipes travaillant dans des laboratoires et sur des thématiques différents, afin de couvrir le plus large éventail possible de paradigmes expérimentaux.

Après avoir décrit les principales caractéristiques des jeux de données étudiés, nous présentons les différentes études d'homogénéité que nous avons réalisées : recherche de données atypiques, détection de sous-groupes, comparaisons inter-sessions, inter-contrastes,... Le test de Grubbs est utilisé comme outil inférentiel pour valider (ou infirmer) localement l'hypothèse de normalité pour l'analyse de groupe « standard » avec SPM. La matrice des distances et les distances de Cook associées à chaque sujet servent quant à elles d'outils exploratoires pour essayer d'expliquer les sources d'hétérogénéité observées.

Les résultats obtenus montrent la présence de manière fréquente de données atypiques remettant en cause l'hypothèse d'homogénéité utilisée pour l'analyse de groupe « standard ». Ceci justifie l'utilisation de statistiques de décision robustes dans un cadre non-paramétrique. De plus, l'étude des variances résiduelles estimées au premier niveau révèle la présence d'une forte variabilité inter-sujets, confirmant l'intérêt de développer les méthodes d'analyse de groupe à effets mixtes. À notre connaissance, il s'agit de la première étude aussi complète sur le sujet.

6.1 Présentation des jeux de données utilisés

Nous avons pu constituer au cours de la thèse une banque de 20 jeux de données, collectés auprès de différentes équipes de recherche à travers le monde. Les principales caractéristiques de ces jeux de données sont résumées dans la table 6.1 : la provenance du jeu de données, le nombre de sujets impliqués dans l'étude et l'éventuel article présentant les résultats obtenus.

6.1.1 Présentation générale

Ce travail de collecte a débuté en utilisant la base de données du *fMRI Data Center*, accessible à l'adresse `www.fmridc.org/f/fmridc`. Le *fMRI Data Center* se propose de réunir les données produites par les auteurs du *Journal of Cognitive Neuroscience* et de les rendre accessible à l'ensemble de la communauté de neuro-imagerie. Un prix scientifique est décerné chaque année à la meilleure « ré-analyse » de ces données.

Les données proposées correspondent le plus souvent aux données brutes d'IRMf, alors que les procédures de contrôle d'homogénéité que nous avons développées travaillent à partir des données pré-traitées et analysées avec un modèle linéaire généralisé, comme décrit dans le chapitre 2. Nous avons donc été contraint, pour chaque jeu de données collecté auprès du *fMRI Data Center*, de relancer les analyses du premier niveau.

Nous avons récupéré trois jeux de données correspondant à différentes publications dans le *Journal of Cognitive Neuroscience* : le jeu de données **L** (article de Klein *et al.* [2000]), le jeu de données **R** (article de Poldrack *et al.* [2001]) et un

dernier jeu de donnée (article de Fiebach *et al.* [2002]), qui n'a malheureusement pas pu être exploité par la suite car la modélisation utilisée au premier niveau n'était pas compatible avec notre méthode de calcul des distances inter-sujets.

Ensuite, nous avons réussi à collecter de nombreux jeux de données en établissant des collaborations avec des équipes de recherche de notre laboratoire (Service Hospitalier Frédéric Joliot, Orsay, France) : l'équipe INSERM U562 de Stanislas Dehaene pour les jeux de données **A**, **B**, **C**, **D**, **F**, **I**, **J**, **O**, **Q** et **T**, et l'équipe INSERM ERM 02-05 de Jean-Luc Martinot pour les jeux de données **E** et **H**.

D'autres jeux de données proviennent de collaborations avec des équipes de recherche extérieures : l'équipe d'Andreas Kleinschmidt et Helmut Laufs (université Goethe, Allemagne) pour le jeu de données **M**, l'équipe de Patricia Romaiguère (CNRS UMR6149, France) pour le jeu de données **P** et l'équipe de Sylvain Takerkart (université de Princeton, USA) pour le jeu de données **S**.

Nous avons également travaillé sur deux jeux de données pour lesquels notre équipe de recherche a participé à la mise au point du paradigme expérimental et à l'acquisition des données. Ainsi, le jeu de données **G** a été acquis en collaboration avec le centre d'IRMf de Marseille pour le FIAC (*Functional Imaging Analysis Contest*), organisé à l'occasion de la conférence de l'OHBM (*Organization for Human Brain Mapping*) de 2005 à Toronto. Le traitement de ces données a permis de comparer directement les résultats obtenus par différentes méthodes d'analyse et a fait l'objet d'un numéro spécial de la revue *Human Brain Mapping* (voir notamment Poline *et al.* [2006] et Dehaene-Lambertz *et al.* [2006a]). Quant au jeu de données **N**, il correspond à une étude de localisation fonctionnelle sur une grande population de sujets, menée en collaboration avec l'équipe INSERM U562 de Stanislas Dehaene, et permet d'aborder d'une autre manière la problématique de l'analyse de groupe (voir notamment Thirion *et al.* [2005, 2007]).

Enfin, nous avons décidé d'étudier le jeu de données proposé par Henson *et al.* [2002] sur le site `www.fil.ion.ucl.ac.uk/spm` pour illustrer l'analyse de groupe « standard » utilisant le logiciel SPM (jeu de données **K**). Il est intéressant de noter que ce jeu de données nous a servi d'exemple pratique pour présenter dans les chapitres 4 et 5 les procédures de contrôle d'homogénéité que nous avons développées, car il contient une donnée atypique.

	Unité(s) de recherche	**Nombre de sujets**	**Publication**
A	INSERM U562, Orsay, France	*21*	-
B	INSERM U562, Orsay, France	*16*	Pallier *et al.* [2003]
C	INSERM U562, Orsay, France	*16*	Landmann *et al.* [2007]
D	INSERM U562, Orsay, France	*22*	-
E	INSERM ERM 02-05, Orsay, France	*31*	Artiges *et al.* [2006]
F	INSERM U562, Orsay, France	*9*	Simon *et al.* [2002]
G	UNAF, INSERM U562, IFR 131, Orsay et Marseille, France	*15*	Dehaene-Lambertz *et al.* [2006a] (Expérience 2)
H	INSERM ERM 02-05, Orsay, France	*31*	-
I	INSERM U562, Orsay, France	*10*	Dehaene-Lambertz *et al.* [2006a] (Expérience 1)
J	INSERM U562, Orsay, France	*10*	Dehaene-Lambertz *et al.* [2006b]
K	Wellcome Department of Cognitive Neurology, Londres, Royaume-Uni	*12*	Henson *et al.* [2002]
L	UNAF, Orsay, France	*8*	Klein *et al.* [2000]
M	Cognitive Neurology Unit, Goethe University, Frankfort, Allemagne	*15*	Laufs *et al.* [2003]
N	UNAF, INSERM U562, Orsay, France	*88*	-
O	INSERM U562, Orsay, France	*12*	Golestani *et al.* [2006]
P	CNRS UMR6149, Marseille, France	*15*	Felician *et al.* [2005]
Q	INSERM U562, Orsay, France	*10*	-
R	MGH-NMR Center and Harvard Medical School, Los Angeles, USA	*8*	Poldrack *et al.* [2001]
S	Princeton University, Princeton, USA	*24*	Wager *et al.* [2004]
T	INSERM U562, Orsay, France	*11*	-

Table 6.1 – *Principales caractéristiques des jeux de données étudiés*

6.1.2 Description des études réalisées

Nous allons présenter rapidement les études de neuro-imagerie correspondant aux jeux de données collectés, afin de montrer le large éventail de paradigmes expérimentaux auxquels nous avons appliqué nos procédures de contrôle d'homogénéité. Il était important en effet de ne pas se limiter à un aspect particulier des neuro-sciences cognitives.

En résumé, les principales fonctions cognitives étudiées sont le langage (jeux de données **A**, **B**, **G**, **I**, **J**, **O**, **Q** et **R**), la déduction logique (jeu de données **C**), l'accès à la conscience (jeu de données **D**), la perception des émotions (jeu de données **H**), la mémorisation des visages (jeu de données **K**), l'activité spontanée au repos (jeu de données **M**), la désignation des parties du corps (jeu de données **P**), la perception de la douleur (jeu de données **S**), la représentation des nombres (jeu de données **T**). D'autres études sont quant à elles dédiées à une région cérébrale spécifique, comme le cortex cingulaire antérieur (jeu de données **E**), le cortex pariétal (jeu de données **F**) ou le cortex visuel (jeu de données **L**). Enfin, le jeu de données **N** correspond à une étude sur la variabilité inter-sujets de la localisation fonctionnelle de certains processus cognitifs et sensori-moteurs.

Pour plus de détails, un court descriptif présentant la problématique cognitive, le paradigme expérimental et les éventuels résultats obtenus est disponible en annexe A pour chacune des études d'IRMf correspondant aux jeux de données collectés au cours de cette thèse.

Notons enfin que si la détection de sujets (ou de contrastes) atypiques est essentiel pour une meilleure interprétation des résultats de l'analyse de groupe, il peut être tout aussi crucial de déterminer la cause de la « déviance » observée, ou tout au moins de la caractériser, et en particulier de différencier une cause « instrumentale » (liée aux processus d'acquisition et de traitements des données) d'une cause « comportementale » (liée au sujet lui-même). Cette information est importante soit pour corriger d'éventuels artéfacts d'acquisition, soit pour extraire de l'information pertinente du lien entre le pattern d'activation atypique avec d'autres données comportementales. Nous avons, quand cela était possible, mis en œuvre certains outils pour expliquer la cause de la « déviance » observée.

6.2 Études d'homogénéité

Pour chacun des jeux de données collectés, les expérimentateurs ont défini plusieurs contrastes d'intérêt spécifiques d'une problématique cognitive différente. Pour chacun de ces contrastes, une recherche de donnée atypique a été effectuée en calculant la matrice des distances inter-sujets (comme décrit en 4.1), puis en comparant les distances de Cook associées à chaque sujet (comme décrit en 4.2) à la valeur critique définie par l'équation 4.10. Les résultats globaux de cette recherche de données atypiques sont présentés et discutés à la fin de cette section (6.2.4). Bien que le statut des études différentes (autre groupe de sujet) et des contrastes différents (même groupe de sujet) ne soient pas équivalents, nous présentons l'ensemble des résultats dans un tableau récapitulatif. Il est en effet possible qu'un sujet soit atypique vis à vis d'un réseau cérébral mais pas vis à vis d'un autre. Nous mentionnerons par la suite si un sujet est détecté atypique quelque soit le contraste utilisé.

Dans un premier temps, nous présentons certains résultats significatifs obtenus en appliquant nos procédures de contrôle d'homogénéité sous forme d'études de cas. Ces résultats illustrent différentes sources d'hétérogénéité qui peuvent être détectées dans les jeux de données d'IRMf.

6.2.1 Étude de cas : jeu de données O

Il s'agit d'une étude sur les processus de production syntaxique chez des sujets bilingues, avec des conditions « Phrases » et « Mots ». Pour ce jeu de données, le calcul des distances moyennes et des distances de Cook, à partir de la matrice des distances spatiales inter-sujets, a révélé la présence d'un sujet atypique pour le contraste « Phrases moins Mots » (voir la figure 6.1).

La représentation des deux premiers vecteurs propres obtenus par la procédure de MDS appliquée à la matrice des distances spatiales inter-sujets (voir la figure 6.2) semble confirmer ce résultat bien que seulement $34,0$ % de la variance totale soit capturée par cette représentation.

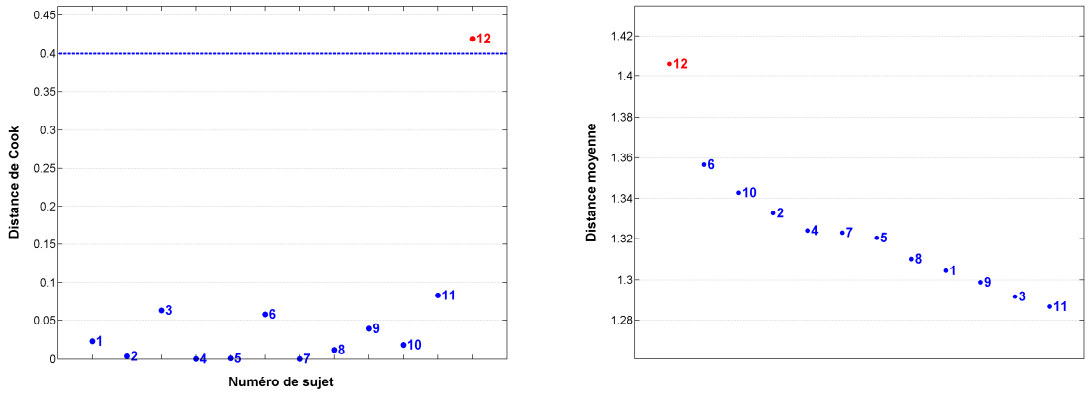

Figure 6.1 – *Distances de Cook (à gauche) et distances moyennes (à droite) calculées à partir de la matrice des distances spatiales inter-sujets – Contraste « Phrases moins Mots » – Jeu de données* ***O*** *décrit en annexe A*

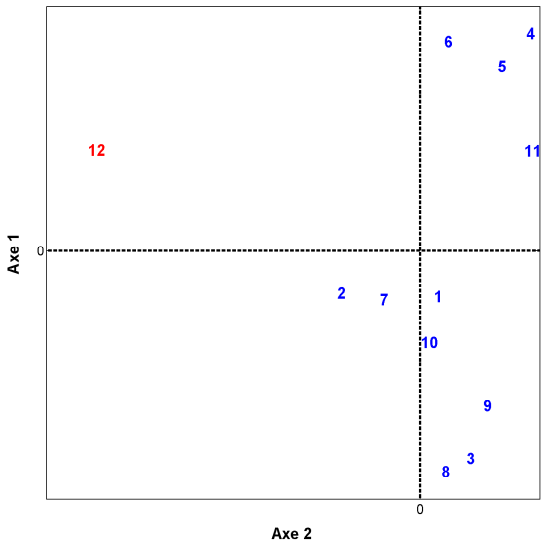

Figure 6.2 – *Procédure de MDS (2 premiers vecteurs propres) sur la matrice des distances spatiales inter-sujets – Contraste « Phrases moins Mots » – Jeu de données* ***O*** *décrit en annexe A*

Afin d'expliquer le fait que le sujet 12 est détecté comme atypique pour le contraste « Phrases moins Mots », nous représentons la projection du maximum d'intensité (MIP, *Maximum Intensity Projection*) des cartes statistiques associées à ce contraste pour quatre sujets différents. Ces cartes statistiques sont seuillées à $P_{non\ corrigée} < 0,001$ au niveau voxel, et uniquement les clusters dont l'extension spatiale dépasse 10 voxels sont affichés sur la figure 6.3 ($E \geq 10$) : nous remarquons que le sujet 12 ne présente que très peu d'activation pour le contraste considéré.

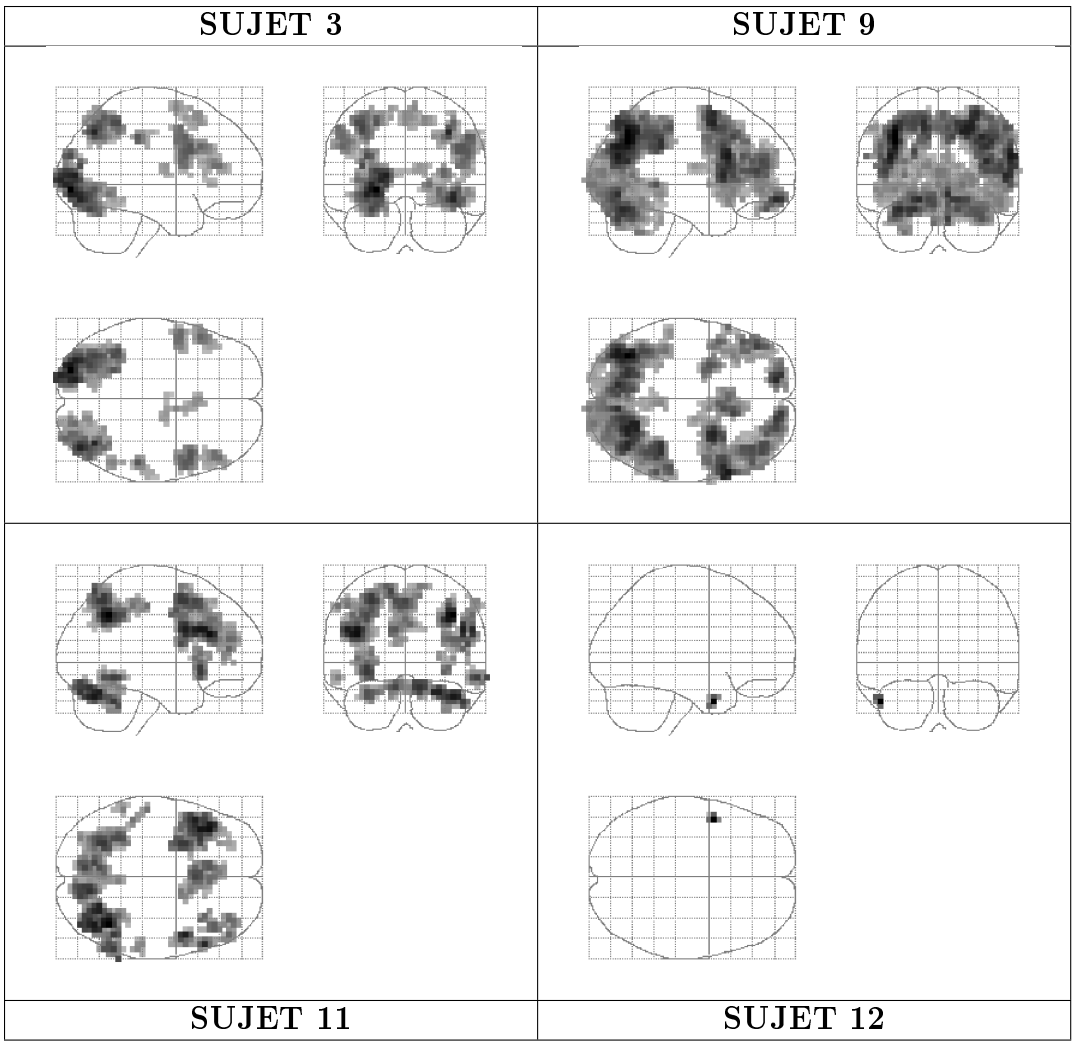

Figure 6.3 – *Cartes statistiques obtenues pour quatre sujets – Contraste « Phrases moins Mots » – Jeu de données* ***O*** *décrit en annexe A*

Pour tenter d'identifier la raison de cette hypo-activation du sujet 12, nous avons calculé des distances inter-contrastes : la matrice des distances spatiales a été calculée entre les contrastes « Phrases Anglais » et « Phrases Français » à travers les sujets. Les distances de Cook correspondantes et la représentation des deux premiers vecteurs propres obtenus par la procédure de MDS (voir la figure 6.4) montrent que seul le sujet 12 présente une grande différence entre ces deux conditions : ainsi, il est possible que le sujet 12 ait eu des difficultés à réaliser une des deux tâches de production de phrases, sans doute celle avec des mots anglais puisque sa langue maternelle est le français. Ce sujet pourrait alors être déclaré comme « outlier comportemental » puisque qu'il ne semble pas avoir effectué la tâche de la même manière que les autres sujets.

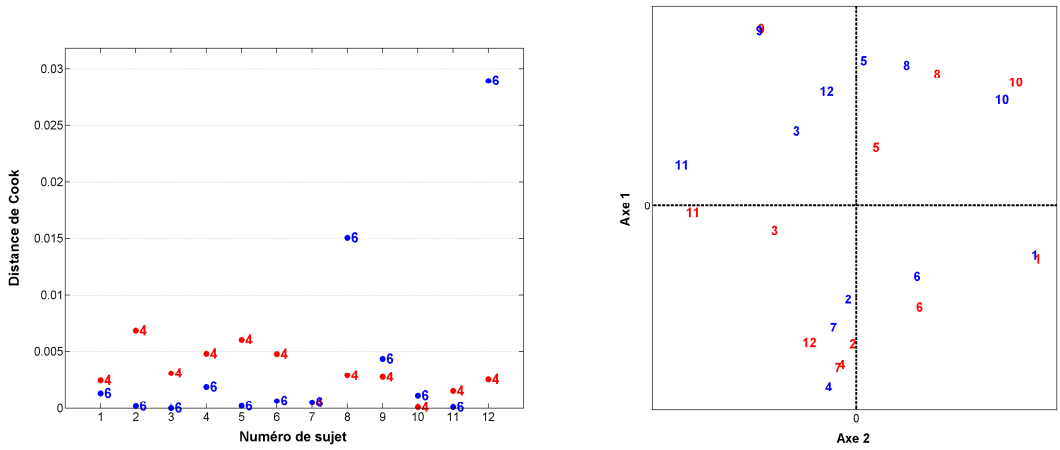

Figure 6.4 – *Distances de Cook et procédure de MDS (2 premiers vecteurs propres capturant* 41,0 % *de la variance totale) calculées à partir de la matrice des distances spatiales entre les contrastes « Phrases Anglais » (bleu) et « Phrases Français » (rouge) – Jeu de données* ***O*** *décrit en annexe A*

Notons également que de forts mouvements ont été détectés pendant l'acquisition des données d'IRMf pour ce sujet, mouvements qui n'ont pu être complètement corrigés par la procédure de réalignement mise en œuvre dans SPM. Ainsi, ce sujet pourrait aussi être un « outlier artéfactuel » puisque que ces différences avec les autres sujets peuvent provenir de données mal estimées à cause du mouvement.

Nous présentons enfin la projection du maximum d'intensité des cartes statistiques de groupe obtenues par un test t paramétrique (voir la figure 6.5) pour le contraste d'intérêt de l'étude (« Phrases Moins Mots/Anglais Moins Français »), en considérant l'ensemble des sujets (à gauche) et en retirant le sujet atypique 12 du groupe. Ces cartes ont été seuillées à $P_{non\ corrigée} < 0,01$ au niveau voxel, et uniquement les clusters dont l'extension spatiale dépasse 5 voxels sont affichés ($E \geq 5$). Cette figure permet de confirmer l'influence qu'une donnée peut avoir sur les résultats de l'analyse de groupe utilisant un test t paramétrique. Dans cet exemple, le fait de retirer la donnée atypique fait apparaître plus de clusters significatifs pour les seuils considérés alors qu'*a priori*, diminuer le nombre de données devrait faire baisser la puissance statistique de détection.

Il faut également noter que cette approche consistant à retirer un sujet jugé atypique après une analyse *conjointe* des effets n'est pas valide puisque la mesure d'incertitude est alors dépendante des autres sujets et la condition d'indépendance statistique n'est plus vérifiée.

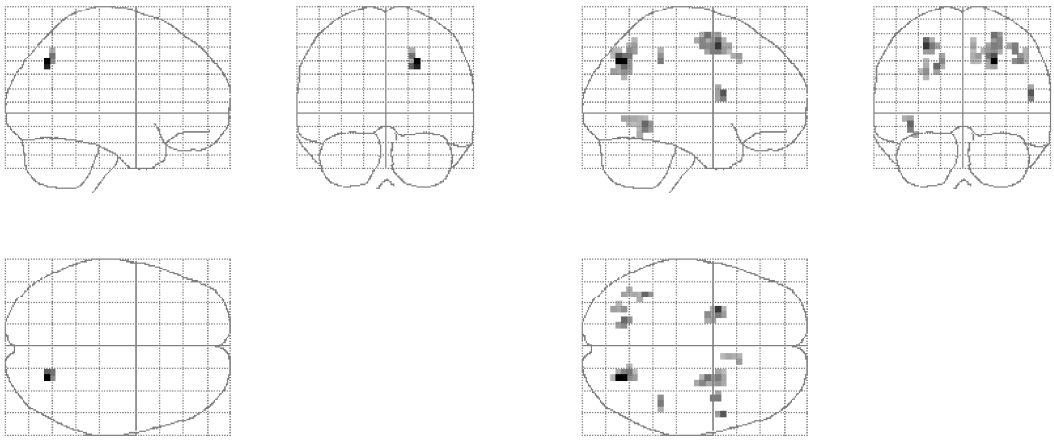

Figure 6.5 – *Cartes d'activation du groupe obtenues par un test de Student paramétrique : à gauche sur l'ensemble des 12 sujets, à droite en retirant du groupe le sujet atypique numéro* 12 *– Contraste « Phrases Moins Mots / Anglais Moins Français » – Jeu de données* ***O*** *décrit en annexe A*

6.2.2 Étude de cas : jeu de données C

Nous présentons les résultats obtenus pour ce jeu de données afin d'illustrer la notion d'effet de masquage auquel est sujet le calcul de la distance de Cook. En effet, la détection d'éventuels sujets atypiques en utilisant la distance de Cook se fait par analyse conjointe des données de tous les sujets. Ainsi, si deux sujets sont très différents du reste du groupe mais que l'un des deux reste néanmoins plus proche des autres sujets que l'autre, alors il est fort probable que l'utilisation de la distance de Cook ne permette de détecter qu'un seul sujet atypique, celui le plus éloigné du groupe.

Pour ce jeu de données, le sujet 5 est détecté comme atypique en calculant les distances moyennes et les distances de Cook associées à la matrice des distances spatiales inter-sujets pour le contraste « Mots » (voir la figure 6.6).

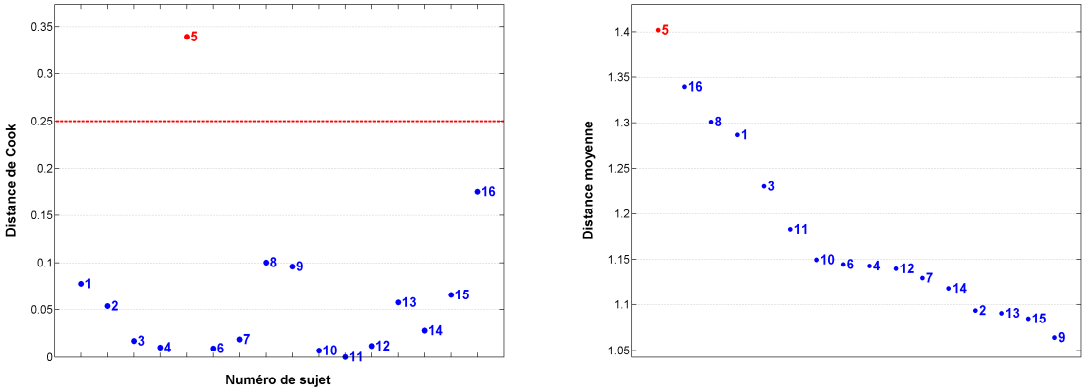

Figure 6.6 – *Distances de Cook (à gauche) et distances moyennes (à droite) calculées à partir de la matrice des distances spatiales inter-sujets – Contraste « Mots » – Jeu de données* ***C*** *décrit en annexe A*

Cependant, la figure 6.6 révèle également qu'un petit groupe de sujets semble assez homogène alors que les sujets 1, 3, 8 et 16 présentent une distance moyenne et une distance de Cook élevées. La représentation des deux premiers vecteurs propres obtenus par la procédure de MDS appliquée à la matrice des distances spatiales

inter-sujets (voir la figure 6.7) semble confirmer ce résultat bien que seulement $42,9$ % de la variance totale soit capturée par cette représentation.

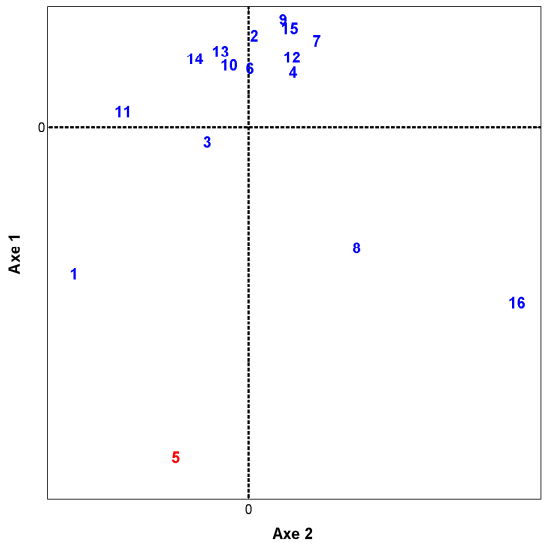

Figure 6.7 – *Procédure de MDS (2 premiers vecteurs propres) sur la matrice des distances spatiales inter-sujets – Contraste « Mots » – Jeu de données* $\boldsymbol{C}$ *décrit en annexe A*

Le calcul des distances inter-sujets a donc été relancé pour ce contraste en retirant le sujet 5, puis de manière itérative, en retirant tous les éventuels sujets détectés comme atypiques, jusqu'à ce qu'aucun sujet ne soit plus détecté comme atypique (voir la figure 6.8).

Sont détectés comme atypiques successivement le sujet 16, les sujets 1 et 8 et enfin le sujet 3, cet ordre correspondant à l'ordre décroissant des distances moyennes calculées sur le groupe entier (voir le graphique de droite sur la figure 6.6). Nous mettons ainsi en évidence l'effet de masquage dans une population de sujets hétérogènes : seul le sujet le plus différent des autres est détecté comme atypique durant la première passe, alors que d'autres sujets présentent également des différences significatives par rapport au sous-groupe relativement homogène de sujets.

Nous pouvons remarquer que la distance de Cook n'est pas une méthode adaptée à ce type de situation et que nous aurions pu tout aussi bien ne détecter aucun

sujet atypique. D'autres modèles capables d'estimer la vraisemblance de plusieurs sous-groupes au sein d'une population seront nécessaires pour analyser ce type de données. Notons enfin que nous ne pouvons pas facilement différencier entre un échantillonnage atypique (par exemple, ces 5 sujets sont échantillonnés parmi les valeurs extrêmes d'une distribution uni-modale qui modélise correctement la population) et un échantillonnage typique dans une distribution bi-modale.

Figure 6.8 – *Distances de Cook calculées à partir de la matrice des distances spatiales inter-sujets obtenues pour différents sous-groupes de sujets – Contraste « Mots » – Jeu de données* ***C*** *décrit en annexe A*

6.2.3 Étude de cas : jeu de données F

L'étude correspondant au jeu de données **F**, acquis par Olivier Simon et Stanislas Dehaene au SHFJ d'Orsay, s'intéresse à un assez grand nombre de contrastes. Comme expliqué en annexe A, six contrastes d'intérêt mettant chacun en jeu le cortex pariétal sont testés : la saisie d'objet (« Saisie » ou *grasping*), le pointage d'objet (« Pointage » ou *pointing*), des saccades visuelles (« Saccades »), l'attention visuelle (« Attention »), la soustraction mentale (« Calcul » ou *calculation*) et la détection de phonèmes (« Langage » ou *language*).

Pour le contraste « Calcul », le sujet 4 est détecté comme atypique lors du calcul de la matrice des distances temporelles inter-sujets, puis du calcul des distances moyennes et des distances de Cook associées à chaque sujet (voir la figure 6.9). Rappelons que la distance dite « temporelle » est ici une distance entre les profils fonctionnels à travers les voxels, tandis que la distance « spatiale » est une distance entre les cartes d'activation. Pour la distance temporelle, la dimension des données est donc au plus égale au nombre des contrastes testés, soit six dans ce cas.

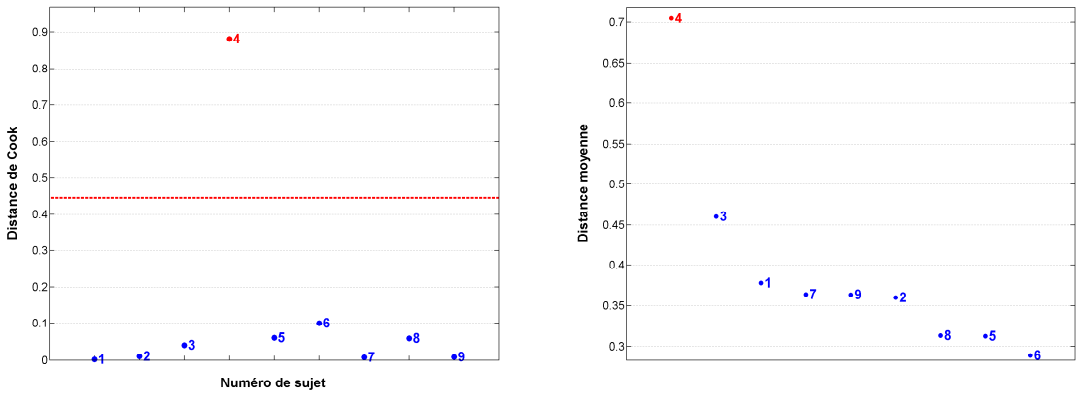

Figure 6.9 – *Distances de Cook (à gauche) et distances moyennes (à droite) calculées à partir de la matrice des distances temporelles inter-sujets – Contraste « Calcul » – Jeu de données **F** décrit en annexe A*

La figure 6.10 représente les deux premiers vecteurs propres de la procédure de MDS appliquée à la matrice des distances temporelles inter-sujets, qui confirme

la position particulière du sujet 4. Notons que sur cette figure, 99,3 % de la variance totale est capturée : la représentation reproduit donc presque parfaitement la variance des données initiales. Ceci n'est guère étonnant vue la faible dimension initiale de ces données.

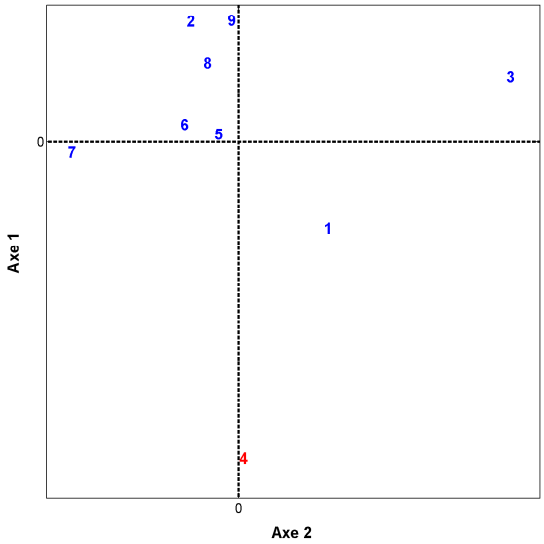

Figure 6.10 – *Procédure de MDS (2 premiers vecteurs propres) sur la matrice des distances temporelles inter-sujets – Contraste « Calcul » – Jeu de données* ***F*** *décrit en annexe A*

Sur ce jeu de données, les sujets ont réalisé la même tâche lors de deux sessions identiques et un calcul de distances temporelles inter-sessions à travers les sujets a été effectué. Les distances moyennes et les distances de Cook associées à la matrice des distances inter-sessions montrent c'est la deuxième session du sujet 4 qui est atypique pour la dimension temporelle (voir la figure 6.11). Comme précédemment, ce résultat est confirmé par la représentation des deux premiers vecteurs propres de la procédure de MDS appliquée à la matrice des distances temporelles inter-sessions (voir la figure 6.12, pour laquelle 99,3 % de la variance totale est capturée). La cause de cette variation sur la deuxième session n'a pas pu être identifiée.

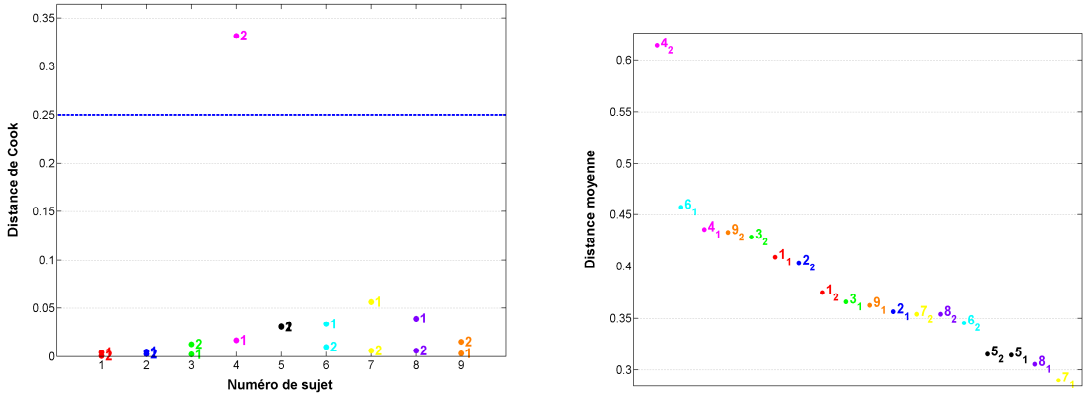

Figure 6.11 – *Distances de Cook (à gauche) et distances moyennes (à droite) calculées à partir de la matrice des distances temporelles inter-sessions – Contraste « Calcul » – Jeu de données* ***F*** *décrit en annexe A*

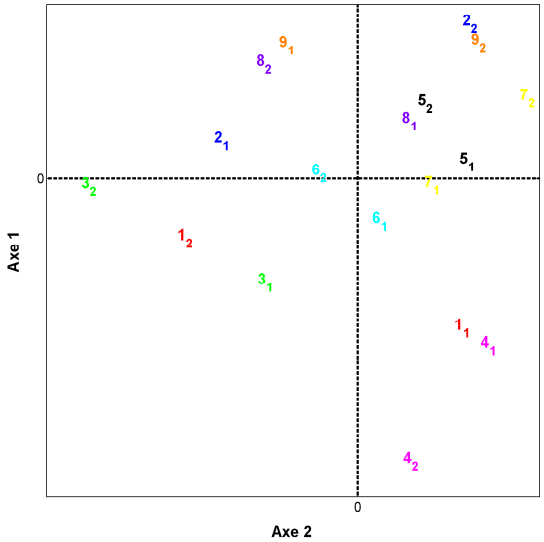

Figure 6.12 – *Procédure de MDS (2 premiers vecteurs propres) sur la matrice des distances temporelles inter-sessions – Contraste « Calcul » – Jeu de données* ***F*** *décrit en annexe A*

Par comparaison, les distances inter-sessions à travers les sujets ont aussi été calculées pour la dimension spatiale. Les distances moyennes et les distances de Cook associées à la matrice des distances spatiales inter-sessions montrent que la variabilité inter-sujets semble plus importante que la variabilité inter-sessions (voir la figure 6.13). De plus, aucun sujet (et aucune session d'un sujet) n'est atypique pour la dimension spatiale. Les mêmes résultats sont observés pour la représentation des deux premiers vecteurs propres de la procédure de MDS appliquée à la matrice des distances spatiales inter-sessions (voir la figure 6.14), mais seulement $23,0$ % de la variance totale est capturée.

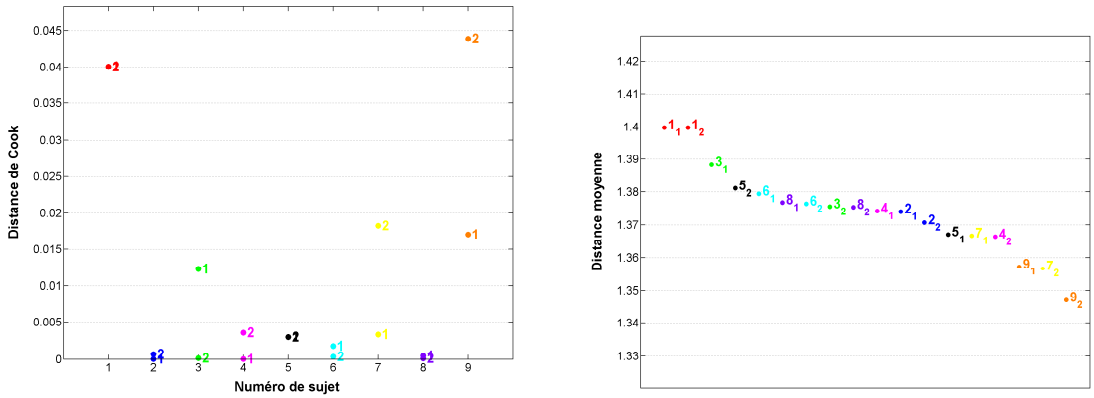

Figure 6.13 – *Distances de Cook (à gauche) et distances moyennes (à droite) calculées à partir de la matrice des distances spatiales inter-sessions – Contraste « Calcul » – Jeu de données* ***F*** *décrit en annexe A*

Pour illustrer la notion de distance *temporelle*, nous avons également calculé la matrice des distances spatiales entre les six tâches du protocole (« Saisie », « Pointage », « Saccades », « Attention », « Calcul » et « Langage ») à travers les sujets. La figure 6.15 représente les deux premiers vecteurs propres de la procédure de MDS appliquée aux matrices des distances spatiales inter-contrastes de trois sujets différents : nous pouvons remarquer que les deux premiers sujets (en haut) donnent la même topographie qui est très différente de celle observée pour le troisième sujet.

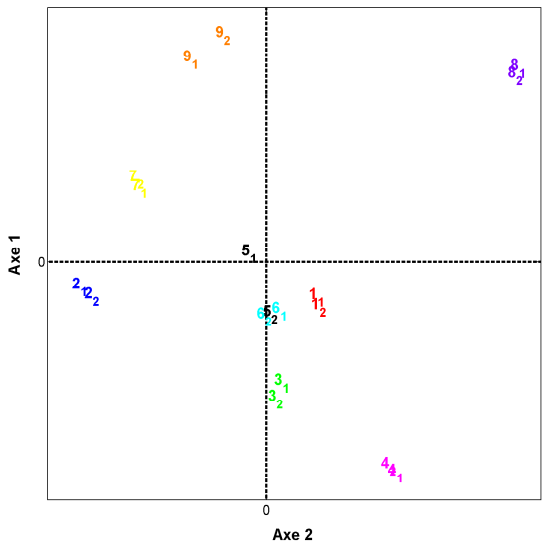

Figure 6.14 – *Procédure de MDS (2 premiers vecteurs propres) sur la matrice des distances spatiales inter-sessions – Contraste « Calcul » – Jeu de données* ***F*** *décrit en annexe A*

La distance temporelle entre deux sujets peut s'interpréter à partir de la figure 6.14 comme étant une mesure de la similarité entre les topographies des différentes tâches obtenues pour chaque sujet. Pour les résultats présentés précédemment, les topographies ont été comparées en regardant la contribution de la tâche de « Calcul ». Ainsi, la comparaison des patterns spatiaux entre les tâches peut permettre d'étudier la variabilité du réseau d'aires cérébrales impliquées par chacune des tâches.

Le calcul de la matrice des distances spatiales entre les six tâches a été effectué pour neuf sujets, puis la moyenne des neuf matrices ainsi obtenues a été soumise à une procédure de MDS. La représentation en 2D du résultat de la procédure de MDS ne capturant pas plus de 60 % de la variance totale, nous avons soumis la matrice des distances moyenne à une méthode de classification hiérarchique (distance de Ward) afin d'obtenir l'arbre représenté sur la figure 6.16.

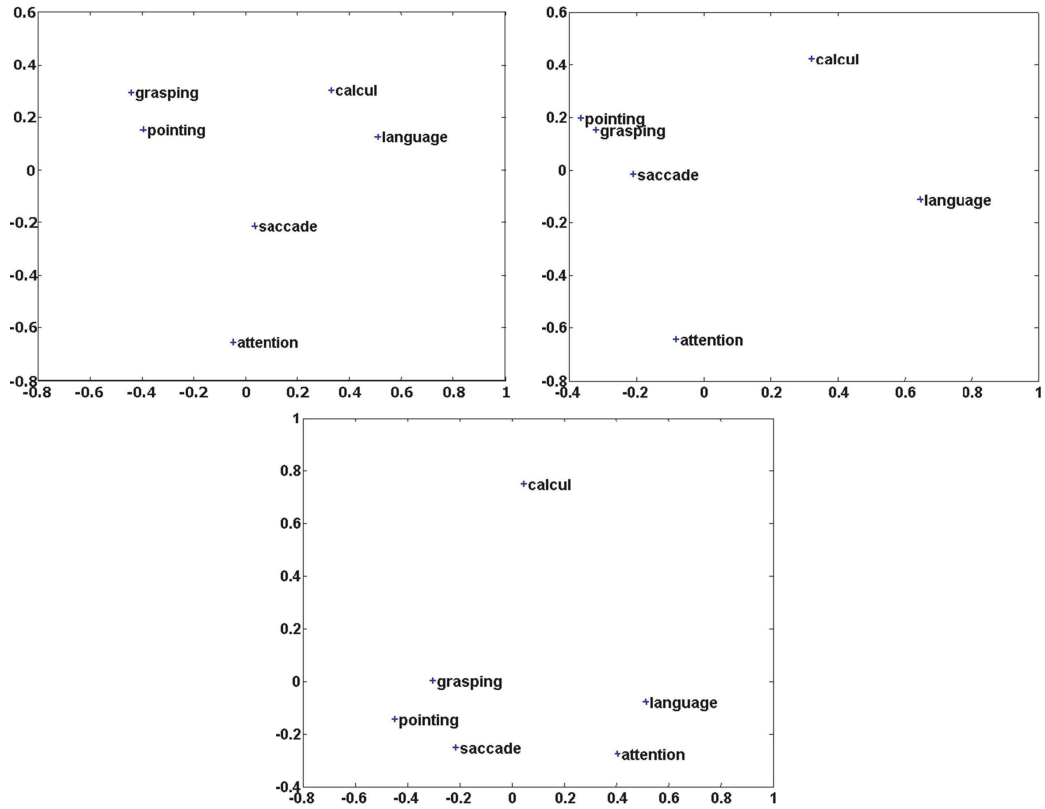

Figure 6.15 – *Procédure de MDS (2 premiers vecteurs propres) sur les matrices des distances spatiales entre les six tâches du protocole obtenues pour trois sujets différents – Jeu de données* ***F*** *décrit en annexe A*

Nous observons sur cet arbre une séparation des tâches motrices (« Saisie » et « Pointage ») d'une part et des tâches visio-spatiales (« Saccades » et « Attention ») d'autre part. L'arbre hiérarchique met également en évidence une composante motrice du langage (« Langage » proche de « Saisie » et « Pointage »). La tâche de calcul implique des activations principalement dans des régions pariétales, également impliquées dans les tâches à composante visio-spatiale : cette observation se retrouve aussi de manière synthétique sur l'arbre hiérarchique puisque la tâche « Calcul » est proche des tâches « Saccades » et « Attention ». Bien que cette étude soit très préliminaire [Kherif *et al.*, 2003], elle ouvre de nouvelles perspectives quant à la classification des taches cognitives à partir des données de neuro-imagerie.

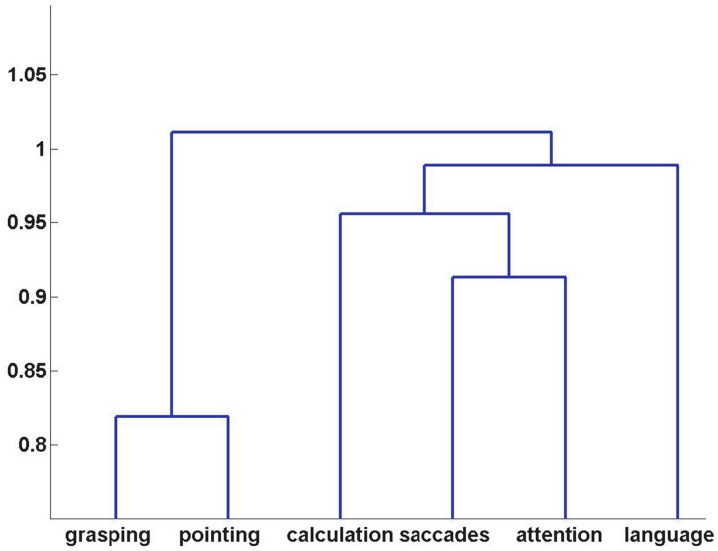

Figure 6.16 – *Arbre de classification hiérarchique correspondant à la moyenne des matrices de distances inter-contrastes obtenue pour neuf sujets – Jeu de données* ***F*** *décrit en annexe A*

6.2.4 Détection de données atypiques

En résumé, si nous considérons les groupes de sujets, nous obtenons que 6 groupes de sujet sur les 20 jeux de données testés ont présenté au moins une donnée atypique pour au moins un des contrastes d'intérêt testés, soit 30 % comme le montre la table 6.2. Ce résultat est important : il est maintenant démontré que le phénomène d'hétérogénéité n'est pas rare dans les études d'IRMf, et donc que les méthodes courantes d'analyse de groupe doivent s'adapter pour le prendre en compte.

Les résultats de l'analyse de groupe sur la totalité des sujets comparés à ceux obtenus sur un sous-groupe « homogène » de sujets (en retirant le sujet atypique) montrent des différences significatives (intensité des activations et extension spatiale des clusters). Certains résultats publiés actuellement risquent d'être significativement biaisés par la présence de sujets atypiques.

Jeu de données	**Nombre de contrastes testés**	**Nombre de données atypiques détectées**
A	*28*	***0***
B	*11*	***0***
C	*14*	***1*** (pour un contraste)
D	*11*	***0***
E	*13*	***0***
F	*8*	***1*** (pour un contraste)
G	*15*	***0***
H	*20*	***0***
I	*11*	***0***
J	*5*	***0***
K	*1*	***1*** (pour un contraste)
L	*14*	***1*** (pour sept contrastes)
M	*8*	***0***
N	*27*	***3*** (1 pour deux contrastes et 2 pour un contraste)
O	*20*	***1*** (pour un contraste)
P	*10*	***0***
Q	*6*	***0***
R	*8*	***0***
S	*5*	***0***
T	*6*	***0***

Table 6.2 – *Données atypiques détectées*

Dans certains cas, il a été possible d'expliquer pourquoi un sujet apparaissait comme atypique : artéfacts d'acquisition des données dus au scanner, fort mouvement du sujet pendant l'expérience, données comportementales anormales (temps de réaction très longs), incapacité à réaliser la tâche,... Dans d'autre cas, nous n'avons pu expliquer facilement l'aspect atypique du pattern d'activité. Il serait alors intéressant de pouvoir scanner le sujet de nouveau, ou d'obtenir plus d'information comportementale sur l'ensemble du groupe.

Dans ce travail, nous avons également recherché des cartes de variance résiduelle atypiques en calculant une matrice des distances inter-images de variance. Bien que la moyenne de l'activité soit le paramètre le plus direct sur lequel chercher des données atypiques, la variance nous informe sur un paramètre *a priori* indépendant de la moyenne. Sur les 19 jeux de données pour lesquels les cartes de variance résiduelle sont disponibles, 12 ont présenté au moins une carte de variance résiduelle atypique, soit 63 %. Cette très forte proportion (plus de 50 %) pourrait être due à plusieurs facteurs. D'une part, la variance pourrait être un paramètre plus sensible que la moyenne pour ce type de détection. D'autre part, la distance de Cook a été calibrée sur des données normales. Bien que le degré de liberté des variances impliquées soit très haut et donc les variables assimilables à des variables normales, il est possible que le test se révèle trop sensible. D'autres travaux sont à envisager pour répondre à ces questions.

6.3 Discussion et conclusion

Sauf erreur de notre part, ce travail constitue la première étude, menée sur un nombre suffisamment important de jeux de données d'IRMf, qui cherche à quantifier la présence de sujets ou de contrastes atypiques vis-à-vis de la moyenne du groupe, cette recherche étant fortement motivée par le fait que ces données atypiques sont susceptibles d'influencer de manière importante les résultats de l'analyse de groupe « standard ». Nous avons observé que 30 % des jeux de données présentent au moins un sujet atypique. Cette observation nous semble *absolument cruciale*, car si ce phénomène n'est pas marginal, il doit nécessairement être pris en compte.

Nous avons réalisé des simulations de la distance de Cook dans le cas de distances normales (voir 4.2.3). Ces simulations prédisent en moyenne $2,5$ % de détection de sujets atypiques pour la valeur critique choisie. Les résultats obtenus (30 % des jeux de données présentent au moins un sujet atypique) constituent un indicateur que les hypothèses de distribution normale des effets à travers les sujets sont sans doute fausses, même s'il reste à étudier la conséquence de la normalité/non normalité des données sur la distribution des distances. L'hypothèse retenue est que pour une population homogène, les observations influencent également la moyenne du groupe, et notre étude montre que ce n'est pas le cas pour 30 % des jeux de données. De plus, la même étude sur les cartes de variance résiduelle montre là aussi une hétérogénéité entre sujets : ainsi l'hypothèse d'une variance intra-sujet à peu près constante à travers les sujets ne semble pas non plus valable.

Une des causes de l'observation de données atypiques peut être une mauvaise normalisation spatiale dans le repère de Talairach (espace du template du MNI). Il serait donc intéressant de poursuivre ces travaux en utilisant des techniques d'analyse permettant de prendre en compte d'éventuelles approximations. Les techniques de parcellisation développées par Flandin *et al.* [2002a,b] ; Thirion *et al.* [2006a] par exemple, sont capables de définir des petites régions homogènes du point de vue spatial et fonctionnel, qui s'adaptent ainsi à l'anatomie fonctionnelle de chaque sujet. Les distances entre sujets étudiées dans ce chapitre pourraient donc être remplacées par des distances fondées sur cette parcellisation. Il se pourrait alors que des sujets à l'anatomie fonctionnelle légèrement différente ne soient plus détectés comme sujets atypiques, mais cela reste à vérifier.

Pour conclure, nous avons proposé au cours de ce chapitre des méthodes pour aborder la variabilité inter-individuelle et les causes de cette variabilité. Nous avons montré que celle-ci était présente assez fréquemment en neuro-imagerie par IRMf. Cet axe de recherche reste peu exploré, nos travaux restent préliminaires. Dans les chapitres suivants, nous présenterons des techniques capables de mieux appréhender cette variabilité et donc d'accroître la confiance que le neuro-scientifique peut avoir dans les résultats de l'analyse de groupe des données d'IRMf.

Chapitre 7

Statistiques de décision robustes et tests de permutations

Comme nous l'avons expliqué dans le chapitre 3, l'analyse de groupe « standard » utilise un test de Student paramétrique appliqué à la moyenne de la population pour détecter un effet de groupe. Cette procédure repose sur l'hypothèse que les effets estimés sont distribués normalement au sein de la population d'intérêt.

L'hypothèse de normalité est très difficile à vérifier du fait du petit nombre de sujets impliqués dans les études d'IRMf. Néanmoins, l'application de procédures de contrôle d'homogénéité à un ensemble de 20 jeux de données (chapitre 6) a montré la présence fréquente de données atypiques ou de sous-groupes disparates dans les cohortes étudiées, mettant ainsi en doute la généralité de cette hypothèse. Ceci suggère que le test de Student est non seulement sous-optimal du point de vue de la sensibilité (détection des vrais positifs) mais également biaisé en spécificité (contrôle des faux positifs).

Il est tentant mais statistiquement incorrect d'appliquer le test de Student à un groupe réduit dont certains sujets ont été exclus sur la foi d'une analyse *a posteriori*, car la condition essentielle d'indépendance statistique des observations n'est

alors plus vérifiée. Les tests que nous proposons ici combinent des statistiques robustes pour améliorer la sensibilité (7.1), avec un mécanisme de permutations (7.2) impliquant *toutes les observations* et qui permet de contrôler exactement la spécificité sous des hypothèses non-paramétriques concernant la distribution des effets. L'intérêt pratique de ces tests non-paramétriques est illustré à la fin de ce chapitre (7.3) sur le jeux de données **J** décrit en annexe A.

7.1 Statistiques de décision robustes

Le terme « *robuste* » a été introduit pour la première fois en statistique par George Edward Pelham Box en 1953. Un estimateur est dit « *robuste* » s'il est insensible à des petits écarts sur les hypothèses pour lesquelles il a été optimisé. Il y a deux sens au terme « petit » : de petites variations sur toutes les données, ou des écarts importants sur un petit nombre de données. C'est le deuxième aspect qui est le plus mal pris en compte par les estimateurs classiques. Ainsi, la robustesse traduit le plus souvent la résistance de l'estimation aux données atypiques. Elle peut se définir mathématiquement par le plus petit nombre de données extrêmes qui modifie la valeur de l'estimation ramené à la taille de l'échantillon.

L'exemple théorique présenté par la figure 7.1 montre que la modification d'une seule donnée dans l'échantillon suffit à faire passer le test de Student de infiniment significatif à non significatif. Ainsi, la présence éventuelle de données atypiques dans les études d'IRMf justifie l'utilisation de statistiques plus robustes que la statistique de Student afin améliorer la sensibilité des analyses de groupe.

Cependant, dans le cadre de l'analyse de groupe en IRMf, la principale difficulté est qu'aucune information n'est disponible *a priori* concernant la distribution des effets à travers les sujets. Afin de tester un effet de groupe, il est donc raisonnable de choisir une statistique qui conserve un bon pouvoir de détection à travers un large éventail de distributions, incluant notamment des distributions à décroissance lente ou multi-modales. C'est en ce sens général que nous entendons ici le terme « *robuste* », et non au sens classique plus restrictif de la résistance aux données atypiques, comme le mettent par exemple en œuvre les M-estimateurs [Rousseeuw et Leroy, 1987] qui ont inspiré de précédentes méthodes d'analyse de groupe en neuro-imagerie [Brammer *et al.*, 1997 ; Wager *et al.*, 2005].

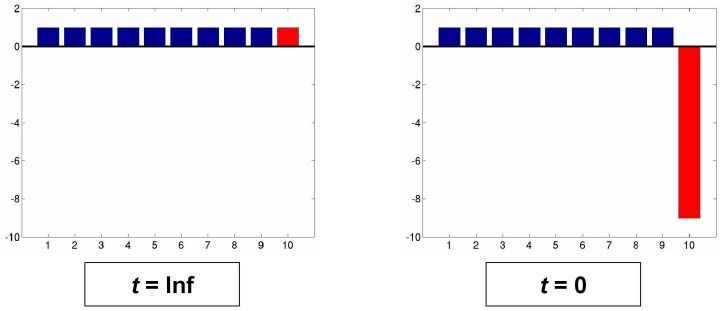

Figure 7.1 – *Exemple théorique pour lequel la présence d'une donnée atypique (à droite) suffit à rendre un test de Student non significatif*

L'exemple théorique présenté par la figure 7.2 illustre cette notion de robustesse. Les courbes ROC (*Receiver Operating Characteristic)* obtenues par simulations de Monte-Carlo permettent de comparer le pouvoir de détection des tests de permutations (présentés en 7.2), fondés respectivement sur la statistique de Student et sur celle de Wilcoxon (présentée en 7.1.2), pour différentes distributions d'effets échantillonnées sur 10 sujets. Les courbes ROC de gauche confirment que la statistique de Student est bien la plus sensible dans le cas d'une distribution normale et ce quel que soit le risque de première espèce (taux de faux positifs accepté). Pour une distribution de Laplace (courbes ROC en haut à droite), la statistique de Wilcoxon devient légèrement plus sensible que la statistique de Student pour les risques de première espèce élevés, tendance qui s'accentue nettement pour une distribution symétrique bi-modale (courbes ROC en bas). Ces simulations illustrent le caractère robuste de la statistique de Wilcoxon au sens où sa sensibilité est relativement plus stable à travers les distributions que celle de la statistique de Student.

Nous étudions dans ce chapitre plusieurs statistiques robustes pour l'analyse d'un effet moyen (analyse *one-sample*) : la statistique du signe [Fisher, 1925] (7.1.1), la statistique du rang signé [Wilcoxon, 1945] (7.1.2), et le rapport de vraisemblances empiriques proposé plus récemment par Owen [2001] (7.1.3). À noter que Rorden *et al.* [2007] ont récemment recensé différentes statistiques de rang pour l'analyse de comparaison de populations (analyse *two-samples*).

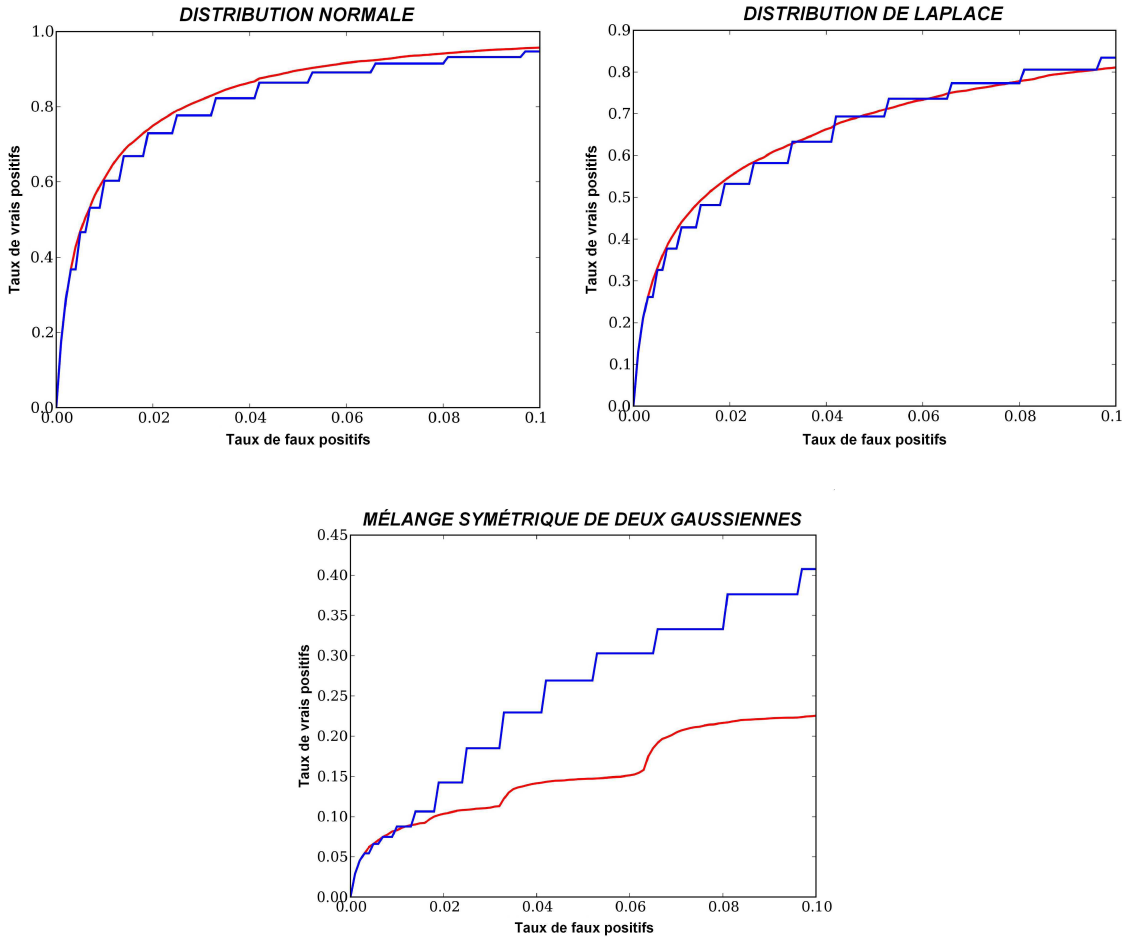

Figure 7.2 – *Courbes ROC des tests de permutations fondés sur la statistique de Student (rouge) et sur la statistique de Wilcoxon (bleu) pour des échantillons de 10 observations tirées dans différentes distributions*

7.1.1 La statistique du signe (Fisher)

La statistique du signe semble être la plus ancienne statistique jamais proposée pour tester la significativité d'un effet puisque le docteur anglais John Arbuthnott l'utilisait déjà en 1710 dans un de ces articles [Arbuthnott, 1710]. Son nom provient du géodésiste allemand Friedrich Robert Helmert qui l'utilise à son tour en 1905, mais elle est aujourd'hui associée au scientifique anglais Sir Ronald Aylmer Fisher, qui l'a popularisée grâce à son ouvrage *Statistical Methods for Research Workers*

[Fisher, 1925], dans lequel il compare notamment ses performances avec celles du test t. Nous présentons ici la statistique du signe telle que la formulent Hollander et Wolfe [1999].

Considérons un échantillon $(\hat{\beta}_1, \ldots, \hat{\beta}_n)$ d'effets estimés dans un voxel donné pour n sujets. La statistique du signe se définit comme la proportion de valeurs positives n^+ dans cet échantillon de données, avec la convention qu'une donnée nulle compte pour 1/2 :

$$T_s = \frac{n^+}{n} = \frac{\mathrm{Card}\{i, \hat{\beta}_i > 0\} + \frac{1}{2} \times \mathrm{Card}\{i, \hat{\beta}_i = 0\}}{n} \tag{7.1}$$

La statistique du signe permet de tester la moyenne d'une population symétrique ou plus généralement la médiane d'une population arbitraire. Elle peut être utilisée à la place d'un test t lorsque l'hypothèse de normalité des données est remise en question.

Sous l'hypothèse nulle H_0 que la médiane est nulle, les valeurs positives et négatives sont équiprobables, ce qui implique que T_s suit une loi binomiale $\mathcal{B}_{n,\frac{1}{2}}$ quelle que soit la distribution effective des observations. Ainsi il est possible de tabuler la distribution de la statistique T_s par permutations de signes indépendamment des données, comme le montre la figure 7.3 pour un échantillon de 10 sujets.

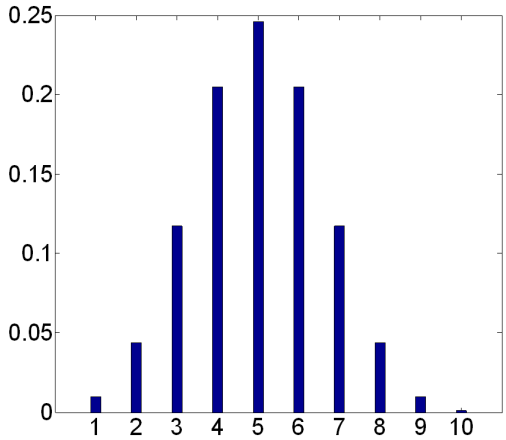

Figure 7.3 – *Distribution de la statistique du signe obtenue par permutations (échantillon de 10 sujets)*

7.1.2 La statistique du rang signé (Wilcoxon)

La statistique du rang signé a été proposée pour la première fois par Wilcoxon [1945]. Dans cet article, elle est utilisée, tout comme le test de la somme des rangs, pour tester la significativité de la différence entre deux échantillons indépendants. Comme pour la statistique du signe, nous présentons ici la statistique du rang signé telle que la formulent Hollander et Wolfe [1999] dans le cas d'un seul échantillon.

Considérons un échantillon $(\hat{\beta}_1, \ldots, \hat{\beta}_n)$ d'effets estimés dans un voxel donné pour n sujets. La statistique du rang signé consiste à ranger les valeurs absolues des effets estimés par ordre croissant, puis à additionner les rangs pondérés par le signe de l'effet correspondant :

$$T_w = \sum_{i=1}^{n} \text{rank}(|\hat{\beta}_i|) \times \text{sign}(\hat{\beta}_i) \tag{7.2}$$

La statistique de Wilcoxon est réputée plus sensible que la statistique du signe pour des distributions modérément hétérogènes. Sa distribution sous l'hypothèse de moyenne nulle peut être tabulée par permutations de signes, à condition de supposer la population symétrique. Nous vérifions aisément que la distribution de référence qui en découle (voir figure 7.4) est indépendante des données, tout comme dans le cas de la statistique du signe, ce qui confère à ces deux statistiques un attrait calculatoire important.

7.1.3 La statistique du rapport de vraisemblances empiriques (Owen)

Le concept de vraisemblance empirique introduit par Owen [2001] correspond à une extension non-paramétrique des méthodes classiques de vraisemblance paramétrique.

7.1.3.1 Construction de la statistique

Considérons un échantillon $(\hat{\beta}_1, \ldots, \hat{\beta}_n)$ d'effets estimés dans un voxel donné pour n sujets, et notons $f(\beta)$ la distribution inconnue de ces effets. Sous l'hypothèse

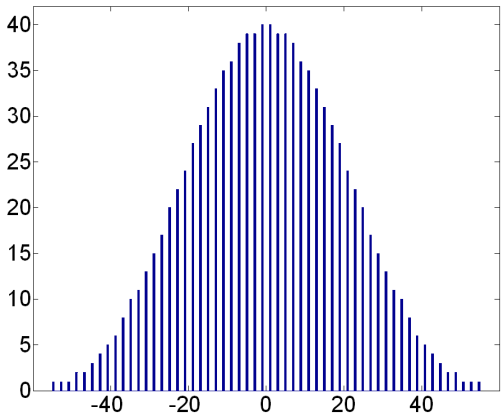

Figure 7.4 – *Distribution de la statistique du rang signé obtenue par permutations (échantillon de 10 sujets)*

usuelle d'indépendance des observations, la fonction de vraisemblance associée à f s'écrit :

$$L(f) = \prod_{i=1}^{n} f(\beta_i)$$

Classiquement, f est supposée appartenir à une famille paramétrique $\mathcal{F}$ et nous cherchons à tester une hypothèse nulle du type $f \in \mathcal{F}_0$, où $\mathcal{F}_0$ est par exemple la sous-famille de $\mathcal{F}$ constituée des distributions à moyenne nulle. Il est alors courant d'utiliser comme statistique de décision le rapport de vraisemblances maximales :

$$R = \frac{\max_{f \in \mathcal{F}_0} L(f)}{\max_{f \in \mathcal{F}} L(f)}$$

L'idée de la vraisemblance empirique est d'appliquer exactement la même démarche dans le cas où aucune contrainte de forme n'est disponible sur la distribution, de sorte que $\mathcal{F}$ est l'espace de toutes les distributions réelles. Bien que cela implique de passer alors dans un espace de dimension infinie, Owen [2001] montre que le rapport de vraisemblances maximales reste mathématiquement bien défini. En effet, le problème de maximisation associé au dénominateur admet pour

solution la « distribution empirique » :

$$\hat{f}(\beta) = \frac{1}{n}\sum_{i=1}^{n} \delta(\beta - \hat{\beta}_i)$$

définie *au sens des distributions* et qui est simplement la moyenne des distributions de Dirac centrées sur les observations. De manière analogue, la solution du problème de maximisation associé au numérateur est un mélange à déterminer des mêmes distributions de Dirac :

$$\hat{f}_0(\beta) = \sum_{i=1}^{n} w_i \delta(\beta - \hat{\beta}_i) \qquad \text{avec} \quad \begin{cases} \forall i,\ w_i \geq 0 \\ \sum_i w_i = 1 \end{cases}$$

Owen [2001] justifie de définir le rapport de vraisemblances empiriques associées respectivement à $\hat{f}$ et $\hat{f}_0$ comme :

$$R_e = \prod_{i=1}^{n} \frac{w_i}{1/n} = \prod_{i=1}^{n} n w_i \tag{7.3}$$

Nous vérifions aisément que R_e est toujours compris dans l'intervalle $[0, 1]$, comme c'est le cas pour un rapport de vraisemblances maximales paramétrique. Les coefficients w_i du mélange sont obtenus en écrivant la contrainte de moyenne nulle ($\sum_i w_i \hat{\beta}_i = 0$), d'où nous tirons l'équation $w_i = \frac{1}{n(1+\lambda\hat{\beta}_i)}$ où λ est racine de l'équation implicite :

$$\sum_{i=1}^{n} \frac{\hat{\beta}_i}{1 + \lambda\hat{\beta}_i} = 0$$

Cette équation admet une racine unique sauf dans le cas où toutes les observations sont de même signe, pour lequel nous posons conventionnellement $R_e = 0$ pour assurer la continuité. Dans les autres cas, la solution unique peut être approchée numériquement par la méthode de Newton.

7.1.3.2 Test unilatéral

Par construction, la statistique du rapport de vraisemblances empiriques permet de réaliser un test bilatéral (*two-tailed*) car elle est insensible au signe de l'effet moyen. Afin d'en définir une version signée, nous exploitons le phénomène de Wilks, une propriété asymptotique générale des rapports de vraisemblances maximaux dont Owen [2001] montre qu'elle reste vraie dans le cas du rapport non-paramétrique défini par l'équation 7.3 : sous l'hypothèse de moyenne nulle, la quantité $-2\log R_e$ converge en distribution vers un χ^2 lorsque la taille de l'échantillon tend vers l'infini, ce χ^2 possédant le même nombre de degrés de liberté que la dimension du paramètre d'intérêt, soit 1 dans le cas d'un test univarié de la moyenne.

Notons $\bar{\beta}$ la moyenne empirique de l'échantillon. La démonstration du théorème exploite le fait que la statistique signée que nous appellerons « t-empirique » :

$$T_e = \mathrm{sign}(\bar{\beta})\sqrt{-2\log R_e} \tag{7.4}$$

dont le carré vaut $-2\log R_e$, suit asymptotiquement une loi normale standard $N(0,1)$ et peut donc s'interpréter comme un score z dans le cas de « grands » échantillons. Dans la situation malheureusement plus habituelle en IRMf de « petits » échantillons, la distribution de T_e peut s'écarter sensiblement d'une gaussienne selon la vraie distribution des observations $f \in \mathcal{F}_0$; il est donc préférable de la coupler à un test de permutations comme pour les statistiques du signe et du rang signé. Notons que la loi de référence de T_e est alors dépendante des données contrairement au cas des deux autres statistiques citées.

7.2 Méthode de calibration statistique : les tests de permutations

Nous avons déjà évoqué plus haut le principe des permutations de signes dont nous allons voir qu'elles constituent une méthode de calibration statistique très

générale, pouvant être appliquée à n'importe quelle statistique de décision et permettant de se libérer de l'hypothèse de normalité sous-jacente au test t classique, tout ceci au prix d'un accroissement du temps de calcul rendu possible par la puissance des ordinateurs actuels.

7.2.1 Généralités

Apparus dans les années 1930 sous l'initiative des travaux de Sir Ronald Aylmer Fisher [Fisher, 1935] et de Edwin James George Pitman [Pitman, 1937a,b, 1938], les tests de permutations ont été largement étudiés depuis. Longtemps peu utilisées en raison de la quantité importante de calculs qu'elles requièrent, les techniques de ré-échantillonage se sont répandues au cours des dernières décennies et sont aujourd'hui d'utilisation courante dans de nombreux domaines d'application des statistiques. Les ouvrages de Lehmann [1986] ; van der Vaart [1998] entre autres contiennent des parties consacrées aux propriétés théoriques des tests de permutations pour des observations univariées. Certaines de ces propriétés ont été étendues par Strasser et Weber [1999] à des observations multivariées.

Les tests de permutations ont été introduits en neuro-imagerie par Holmes *et al.* [1996] qui ont étudié différents mécanismes de permutations correspondant à différents types d'analyse : permutations de signes pour l'analyse *one-sample* qui nous occupe ici, permutations de labels de classe pour l'analyse *two-sample*, permutations de conditions expérimentales pour l'analyse à effets fixes,...

Les tests de permutations offrent l'avantage d'être exacts sous des hypothèses minimales (non-paramétriques), ce qui signifie qu'ils garantissent un contrôle au pire conservatif des faux positifs. Leur puissance est en outre comparable à celle des tests paramétriques les plus puissants lorsque les hypothèses qui sous-tendent ces derniers sont vérifiées. Enfin, et c'est un aspect essentiel en neuro-imagerie, les tests de permutations résolvent de façon exacte le problème des comparaisons multiples contrairement aux techniques paramétriques fondées sur la théorie des champs aléatoires gaussiens [Worsley, 1994], qui peuvent biaiser considérablement les niveaux de significativité corrigés [Hayasaka et Nichols, 2003].

La figure 7.5 illustre, dans le cas d'une distribution uniforme des effets estimés de 10 sujets, le biais de spécificité observé pour le test de Student paramétrique

(courbe rouge). En effet, dès que le taux de faux positifs théorique (TFP) est inférieur à 10^{-1}, le TFP estimé par le test de Student paramétrique est anti-conservatif (ratio $\text{TFP}_{\textit{éstimé}}/\text{TFP}_{\textit{théorique}}$ supérieur à 1), et ce d'autant plus que le $\text{TFP}_{\textit{théorique}}$ diminue. En revanche, pour la même distribution uniforme, la figure 7.5 confirme que le test de Student par permutations (courbe bleue) n'introduit pas de biais de spécificité et qu'au pire, le TFP estimé par permutations est toujours conservatif (ratio $\text{TFP}_{\textit{éstimé}}/\text{TFP}_{\textit{théorique}}$ toujours inférieur à 1). Notons aussi que le TFP estimé par permutations devient nul pour un $\text{TFP}_{\textit{théorique}}$ inférieur à 10^{-3}, car pour une distribution uniforme de 10 sujets, il est possible de générer au maximum $2^{10} = 1024$ permutations et donc la précision maximale accessible sur le TFP estimé vaut $1/2^{10} \approx 10^{-3}$.

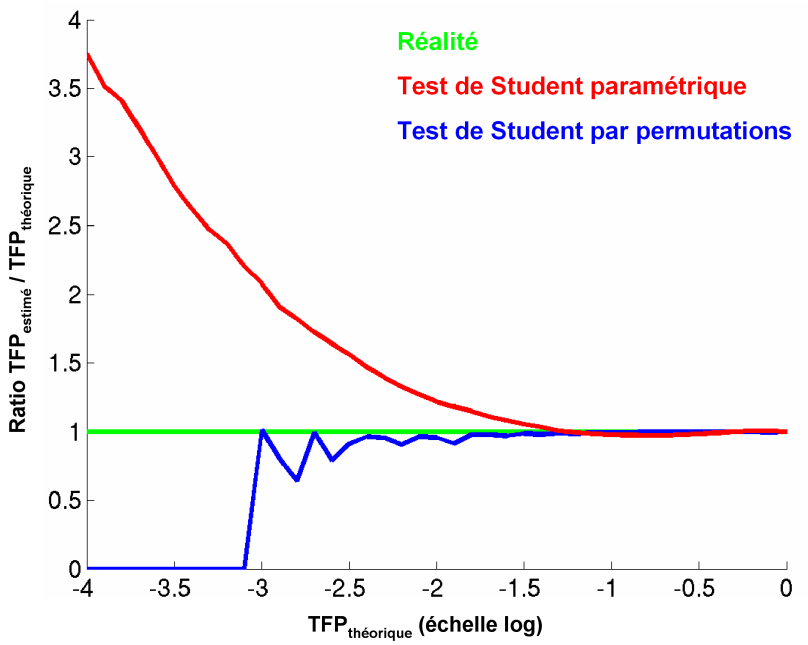

Figure 7.5 – *Comparaison des taux de faux positifs (TFP) théoriques et estimés par le test de Student paramétrique (rouge) ou le test de Student par permutations (bleu) pour une distribution uniforme de 10 sujets*

7.2.2 Permutations de signes

Considérons un échantillon $(\hat{\beta}_1, \ldots, \hat{\beta}_n)$ d'effets estimés dans un voxel donné pour n sujets. Faisons l'hypothèse H que les effets estimés sont distribués indépendamment et identiquement de manière symétrique autour de l'effet moyen de la population β_G. En toute généralité, comme nous le verrons au chapitre 8, nous devons distinguer l'effet estimé de l'effet réel en considérant que le premier est une mesure éventuellement bruitée du second. Dès lors, nous pouvons formuler deux conditions suffisantes pour que H soit vérifiée : (i) le bruit d'estimation sur l'effet est additif et de distribution symétrique ; (ii) les *vrais* effets (non observés) ont une distribution symétrique dans la population d'intérêt [Holmes *et al.*, 1996 ; Nichols et Holmes, 2002].

L'hypothèse nulle H_0 à tester est que l'effet moyen de la population est nul : $\beta_G = 0$. Si H et H_0 sont vraies, alors il est possible de modifier arbitrairement les signes des effets estimés $(\hat{\beta}_i \to -\hat{\beta}_i)$ sans modifier la distribution conjointe de tous les effets estimés. Autrement dit, les signes observés sont échangeables.

Considérons maintenant une statistique de décision :

$$T \equiv T\Big(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_n\Big)$$

La condition d'échangeabilité des signes suggère de considérer l'ensemble des valeurs de T obtenues en effectuant tous les changements de signes possibles dans l'échantillon, au nombre de 2^n. L'histogramme qui en résulte s'interprète comme la distribution de T conditionnelle à la partie non échangeable des observations, soit les valeurs absolues des effets estimés $(|\hat{\beta}_1|, \ldots, |\hat{\beta}_n|)$. Dans ce sens conditionnel, le test de permutations est exact, sa sensibilité étant intrinsèquement limitée par le nombre fini de permutations possibles.

La figure 7.6 illustre la calibration par permutations sous l'hypothèse nulle H_0 de la statistique de Student pour un échantillon de 10 effets estimés dans un voxel donné (échantillon d'origine représenté en haut à gauche). La distribution représentée est obtenue en calculant la statistique de Student pour chaque nouvel échantillon d'effets estimés obtenu en échangeant le signe d'une ou plusieurs données.

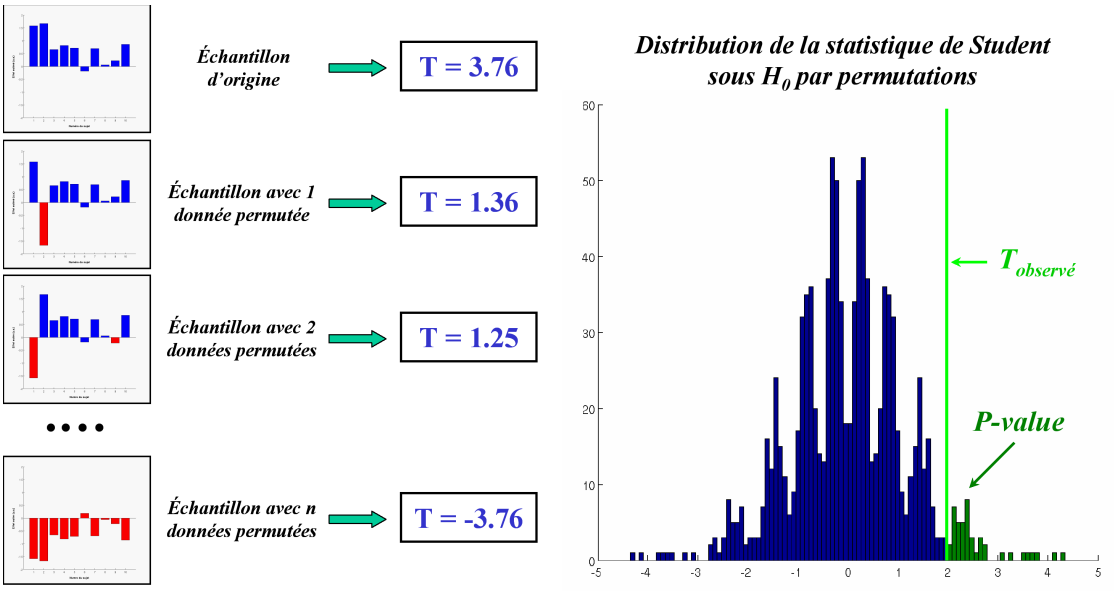

Figure 7.6 – *Distribution de la statistique de Student obtenue par permutations (échantillon de 10 sujets)*

7.2.3 Seuil statistique

Pour obtenir la carte d'activation du groupe, un test statistique est appliqué en chaque voxel d'un certain volume de recherche (contenant K voxels). Ce test consiste à choisir un seuil statistique u et à détecter tous les voxels du volume de recherche pour lesquels la statistique de décision T, calculée à partir des effets estimés à travers les sujets, dépasse ce seuil.

Même s'il est tout à fait envisageable de choisir un seuil statistique variant spatialement, il est préférable, pour diminuer le temps de calcul, de fixer un seuil uniforme sur tout le volume de recherche. Ce choix permet de garder un bon pouvoir de détection à la condition que la distribution de la statistique de décision T retenue pour le test soit raisonnablement stationnaire. Cette condition justifie en pratique l'utilisation de statistiques de décision invariantes par changement d'échelle.

Le seuil statistique uniforme u est fixé en contrôlant le taux de faux positifs (TFP) à un certain niveau α fixé par l'expérimentateur. Pour cela, il faut résoudre

l'équation suivante :

$$\alpha = \mathbb{E}(\text{TFP}|H_0) = \frac{1}{K}\sum_{k=1}^{K} P(T_k \geq u|H_0) \tag{7.5}$$

où T_k est la valeur de la statistique de décision T calculée pour le voxel k et $P(T_k \geq u|H_0)$ est la probabilité marginale de référence que T_k dépasse le seuil statistique uniforme u, probabilité *a priori* dépendante du voxel k considéré. En considérant que la distribution de la statistique de décision sous l'hypothèse nulle H_0 est stationnaire, chaque distribution $p_k(T|H_0)$ est calibrée en permutant les effets estimés comme décrit en 7.2.2. En reprenant l'équation 7.5, le seuil statistique uniforme u est alors égal au $100 \times (1 - \alpha)$ percentile de la distribution moyenne des statistiques de décision à travers les voxels :

$$\bar{p}(T) = \frac{1}{K}\sum_{k=1}^{n} p_k(T|H_0)$$

En pratique, afin d'accélérer les calculs, il est possible d'obtenir une bonne approximation de $\bar{p}(T)$ en se limitant à un certain nombre $\tilde{K}$ de voxels, choisis aléatoirement dans le volume de recherche.

7.2.4 Comparaisons multiples

Comme nous venons de l'expliquer en 7.2.3, la carte d'activation du groupe est obtenue en appliquant un test statistique en chaque voxel d'un certain volume de recherche, afin de déterminer s'il est activé ou non. En pratique, le volume de recherche peut contenir plusieurs milliers de voxels, ce qui correspond à autant d'hypothèses testées simultanément. Le test statistique est alors sujet au problème bien connu des comparaisons multiples. En déclarant comme « activé » chaque voxel k pour lequel $T_k \geq u$ (où u est le seuil statistique correspondant au risque α d'après l'équation 7.5), il est possible de voir apparaître en moyenne αK voxels déclarés comme « activés » à tort (faux positifs).

Afin d'éviter l'apparition de ces faux positifs, une méthode de correction pour le problème des comparaisons multiples est nécessaire. Dans le logiciel SPM, la méthode de correction retenue utilise la théorie des champs aléatoires gaussiens (RFT pour *Random Field Theory*, Worsley [1994]). Cette théorie se fonde sur l'hypothèse que la carte statistique peut être représentée par un champ aléatoire gaussien continu et stationnaire, dont il est possible de calculer la caractéristique d'Euler. Comme le rappellent Hayasaka et Nichols [2003], de nombreuses conditions doivent être vérifiées pour valider l'utilisation de cette méthode ou sinon de sévères biais peuvent être observés dans les niveaux de significativité corrigés : la carte statistique doit notamment être dérivée d'un champ gaussien (données multivariées normalement distribuées) et être régulière spatialement (données suffisamment lissées). De plus, la correction RFT doit s'effectuer pour des seuils statistiques u élevés, ce qui limite intrinsèquement la puissance de la méthode.

Les tests de permutations permettent aisément de calculer des « p-values » corrigées en fonction de la probabilité d'un faux positif dans le volume de recherche (*familywise error rate*). Il suffit pour cela de calibrer par permutations la statistique $T_{\max}$ définie comme le maximum des statistiques sur le volume de recherche [Holmes *et al.*, 1996 ; Nichols et Holmes, 2002], comme cela est illustré sur la figure 7.7. Comparativement à la correction RFT, le calcul des « p-values » corrigées au niveau voxel par permutations est exact et ne nécessite pas d'hypothèses supplémentaires.

De façon analogue, nous pouvons associer aux régions connexes obtenues après seuillage (les *clusters*, définis au sens de la connectivité des 18 plus proches voisins) un niveau de significativité fonction de leur taille, en comparant celle-ci à la distribution par permutations de la statistique $S_{\max}$ définie comme la taille de la plus grande région détectée [Bullmore *et al.*, 1999 ; Hayasaka et Nichols, 2003]. Nous obtenons ainsi des « p-values » corrigées au niveau cluster.

7.2.5 Randomisation

Le nombre de permutations de signes augmentant de manière exponentielle avec le nombre de sujets, il peut être à la fois très coûteux et superflu d'énumérer toutes les permutations. En pratique, dès que le nombre de sujets dépasse 13, nous

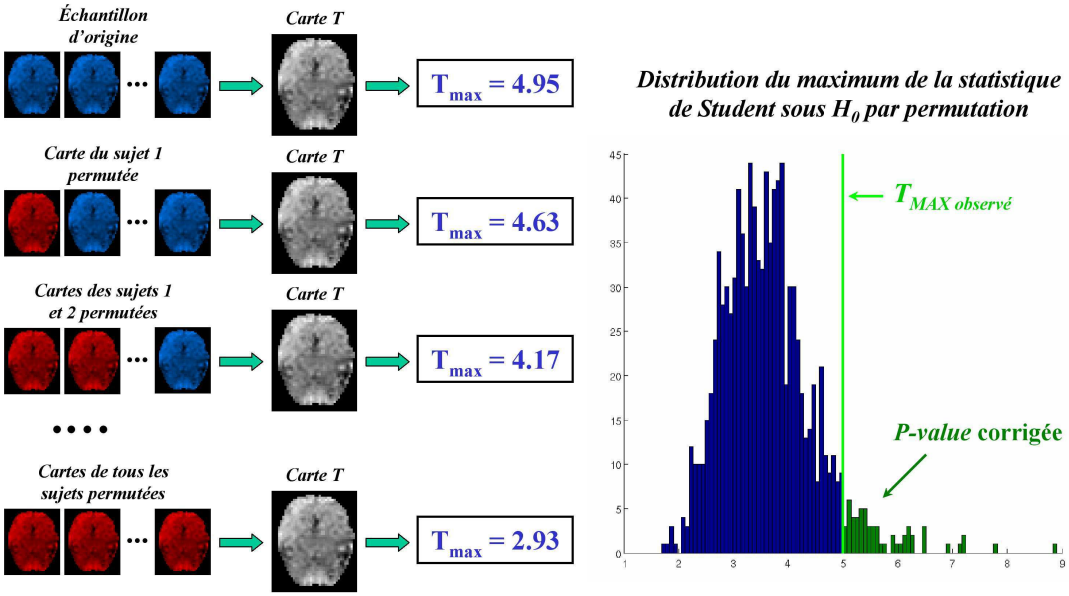

Figure 7.7 – *Distribution du maximum de la statistique de Student obtenue par permutations (échantillon de 10 sujets)*

remplaçons le calcul exhaustif par un tirage aléatoire de $N = 10000$ permutations indépendantes. Cette randomisation introduit une erreur statistique d'écart-type $\sqrt{(P - P^2)/N}$ sur toute « p-value » P, corrigée ou non. Ainsi, la précision sur une « p-value » de 5 % est de $0,2$ %. Sur une « p-value » de $0,1$ %, elle est de $0,03$ %.

7.3 Résultats illustrant l'intérêt des statistiques robustes calibrées par tests de permutation

L'intérêt des tests de permutations utilisant des statistiques de décision robustes est illustré sur le jeu de données **J** décrit en annexe A. Ce jeu de données a été choisi car il présente une très forte variabilité inter-sujets. Cette variabilité peut s'expliquer par le fait que chez le nourrisson de 11 à 17 semaines, le développement des aires corticales est encore à un stade très précoce. De plus, il est difficile d'exiger d'un nourrisson une attention continue à la tâche expérimentale à effectuer.

7.3.1 Traitements des données

L'analyse intra-sujet des données d'IRMf acquises pour chaque nourrisson a été réalisée avec le logiciel SPM99. Les pré-traitements suivants ont été appliqués aux données brutes d'IRMf, comme décrit en 2.1 : correction du mouvement, correction du délai d'acquisition entre les coupes, normalisation spatiale à l'aide d'un template dédié [Dehaene-Lambertz *et al.*, 2002] et lissage spatial par un filtre gaussien de largeur à mi-hauteur $5 \times 5 \times 5$ mm^3. Ensuite, comme décrit en 2.2, un modèle linéaire généralisé a été appliqué aux données pré-traitées, en construisant une matrice de dessin expérimental constituée de deux conditions convoluées par la fonction de réponse hémodynamique canonique d'un adulte : la première condition correspondant à l'écoute de la phrase (condition *Première phrase*) et la seconde condition correspondant à sa répétition (condition *Seconde phrase*). Pour prendre en compte les mouvements importants des nourrissons lors de l'expérience, les six paramètres utilisés pour calculer la correction du mouvement ont été ajoutés comme régresseurs dans la matrice de dessin expérimental. L'estimation des paramètres du modèle linéaire généralisé a permis d'obtenir pour chaque nourrisson deux cartes d'effets d'estimés, correspondant à deux contrastes d'intérêt pour l'expérimentateur : le contraste « Activation à la parole » (*Première phrase* + *Seconde phrase*) et le contraste « Effet de la répétition » (*Seconde phrase* − *Première phrase*).

Les résultats d'analyse de groupe rapportés ici ont été obtenus en utilisant cinq méthodes différentes : le test t paramétrique standard tel qu'il est mis en œuvre dans le logiciel SPM, un test t calibré par permutations à l'aide du logiciel *DISTANCE* et les trois tests robustes présentés en 7.1 (test du signe, test du rang signé et test t-empirique), également calibrés par permutations à l'aide de *DISTANCE*. Notons que le test t calibré par permutations que nous utilisons est équivalent au test *one-sample* proposé par la boîte à outils SnPM de Andrew Holmes et Tom Nichols, dans le cas où aucun lissage de la variance n'est effectué. Toutes les analyses de groupe ont été restreintes à un masque de 3050 voxels englobant les aires périsylviennes des deux hémisphères, afin de se concentrer sur les zones corticales connues pour être activées lors d'une tâche d'écoute passive chez l'adulte.

Pour les deux contrastes d'intérêt, les cartes statistiques de groupe ont été seuillées pour un risque de première espèce de 1 % non corrigé des comparaisons multiples ($P_{non\ corrigée} \leq 0,01$), puis pour les voxels dépassant le seuil, les « p-values » corrigées au niveau voxel et au niveau cluster ont été calculées. Pour le test t paramétrique, le seuil correspondant au risque spécifié a été obtenu à partir de la distribution de Student à $n-1$ degrés de liberté (voir figure 3.3), avec $n = 10$ nourrissons participant à l'étude, et les « p-values » corrigées au niveau voxel et au niveau cluster ont été calculées en utilisant la théorie des champs aléatoires gaussiens mise en œuvre dans SPM. En ce qui concerne les tests calibrés par permutations, les calculs du seuil correspondant au risque spécifié et des « p-values » corrigées au niveau voxel et au niveau cluster ont été effectués en générant l'ensemble des permutations possibles ($2^{10} = 1024$ permutations) et en conservant tous les voxels du masque d'analyse. Quelle que soit la statistique de décision calibrée par permutations, le temps de calcul global n'a pas excédé 30 s (calculs effectués par un PC standard avec un processeur de 2,80 GHz tournant sous Linux).

Pour comparer les différents méthodes d'analyse de groupe entre elles, nous nous restreignons dans la présentation des résultats aux clusters dont la « p-value » corrigée est inférieure à 5 % pour au moins une des méthodes utilisées. Quant aux cartes d'activation du groupe présentées, elles correspondent à la projection du maximum d'intensité (MIP, *Maximum Intensity Projection*) des cartes statistiques seuillées à $P_{non\ corrigée} \leq 0,01$ (pas de seuil sur l'extension spatiale des clusters).

7.3.2 Contraste « Activation à la parole »

La table 7.1 et la figure 7.8 résument les résultats obtenus pour le contraste d'intérêt « Activation à la parole ». Deux clusters sont détectés pour les seuils considérés, le premier dans le gyrus frontal inférieur gauche qui correspond à l'aire de Broca (en rouge), et le second dans le lobe temporal gauche (en bleu).

Le premier cluster est significatif pour le seuil de 5 % en probabilité corrigée en utilisant les trois statistiques de décision robustes présentées en 7.1. Nous pouvons également remarquer que ce premier cluster est presque détecté comme significatif par le test t calibré par permutations. Pour le second cluster, seule la statistique du signe permet de le déclarer comme significatif pour le seuil de 5 % en probabilité

Position anatomique du cluster	Procédure de test statistique	$P_{corrigée}$ au niveau cluster	Étendue du cluster (en voxels)	$P_{corrigée}$ au niveau voxel (maximum)	Position du maximum (en mm)		
					x	y	z
Gyrus frontal inférieur gauche (Broca)	Test t paramétrique	0.34	27	0.91	-48	24	-8
	Test t permutations	0.08	54	0.27	-48	24	-8
	Test du signe	**0.04**	9	0.55	-36	12	-4
	Test du rang signé	**0.02**	89	0.55	-44	24	-8
	Test t-empirique	**0.02**	88	0.55	-44	24	-8
Lobe temporal gauche	Test t paramétrique	0.27	31	0.64	-56	-8	-12
	Test t permutations	0.13	39	0.06	-56	-8	-12
	Test du signe	**0.04**	9	0.55	-64	-12	-12
	Test du rang signé	0.12	36	0.55	-64	-12	-12
	Test t-empirique	0.12	36	0.55	-64	-12	-12

Table 7.1 – *Résultats de l'analyse de groupe pour différentes procédures de test statistique – Contraste « Activation à la parole » – Jeu de données* ***J*** *décrit en annexe A*

corrigée, alors que paradoxalement cette même statistique le détecte comme le cluster de plus faible étendue. Nous avons remarqué que la distribution des effets estimés à travers les sujets est particulièrement hétérogène dans la région de ce cluster, ce qui met en évidence la grande robustesse de la statistique du signe comparativement aux autres statistiques de décision utilisées.

Pour le test t paramétrique, aucune probabilité corrigée (au niveau voxel comme au niveau cluster) n'est significative pour les seuils considérés. Ce résultat suggère que la théorie des champs aléatoires gaussiens n'est pas valide pour le risque de première espèce relativement faible (1 %) que nous avons choisi. De plus, le fait que la calibration du test t par permutations permette de détecter des clusters de plus grande étendue confirme le résultat déjà établi [Holmes *et al.*, 1996 ; Nichols et Holmes, 2002] que le test t paramétrique peut présenter un biais de spécificité important, qui conduit à estimer dans le cas présent des « p-values » conservatives.

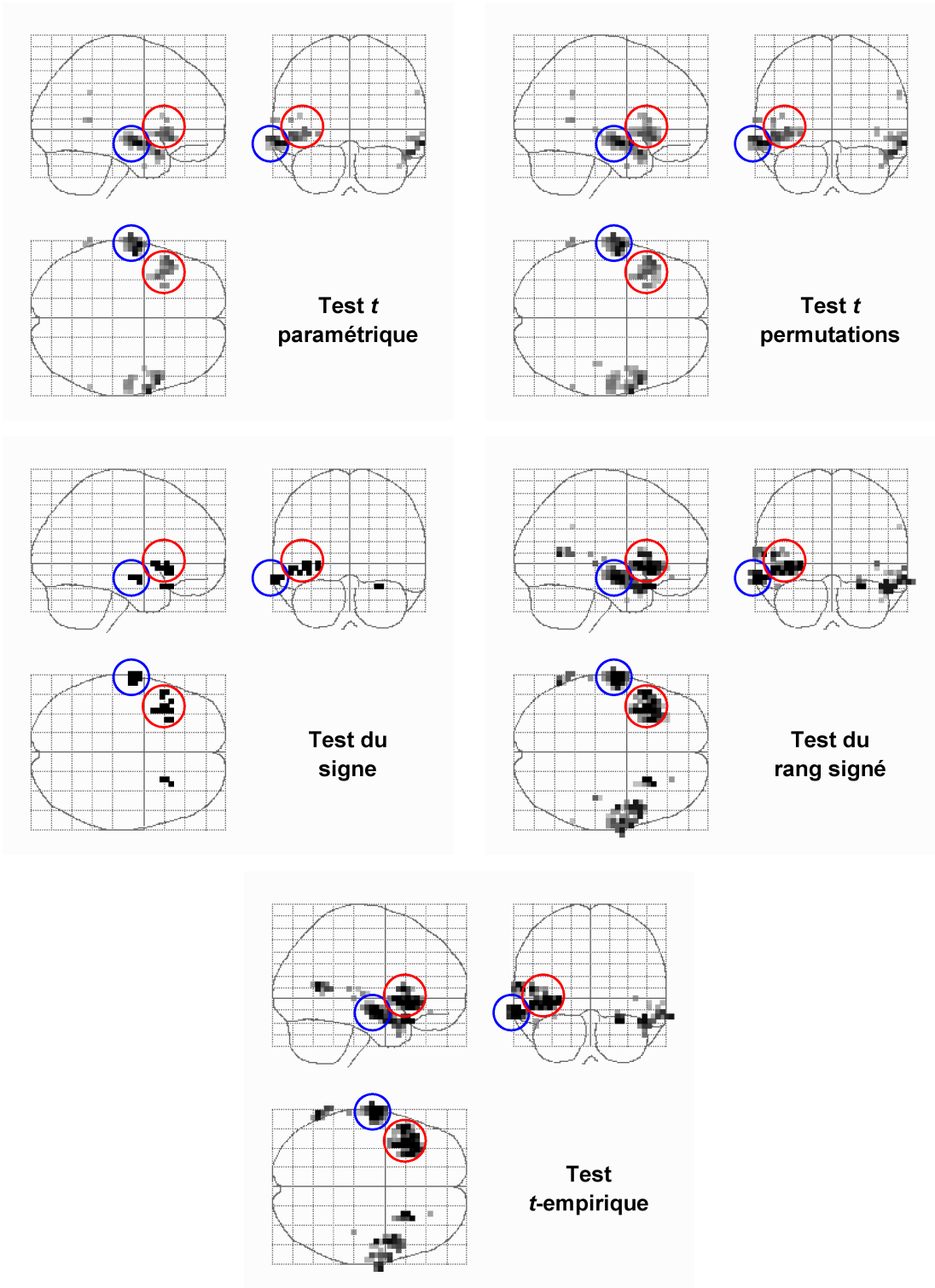

Figure 7.8 – *Cartes d'activation du groupe obtenues par différentes procédures de test statistique – Contraste « Activation à la parole » – Jeu de données* ***J*** *décrit en annexe A*

Les résultats obtenus pour ce contraste d'intérêt confirment qu'un gain de sensibilité substantiel peut être obtenu en utilisant des statistiques de décision plus robustes que la statistique de Student : les « p-values » corrigées au niveau cluster sont alors plus significatives. En revanche, il est intéressant de remarquer un comportement inverse pour les « p-values » corrigées au niveau voxel, qui sont plus significatives en utilisant la statistique de Student. Ce comportement s'explique par le fait que les statistiques de décision robustes que nous utilisons atteignent leur borne supérieure dans les voxels pour lesquels tous les effets estimés à travers les sujets sont positifs. Or il est très probable de trouver par hasard au moins un voxel pour lequel une permutation arbitrairement choisie fournit un échantillon d'effets estimés tous positifs.

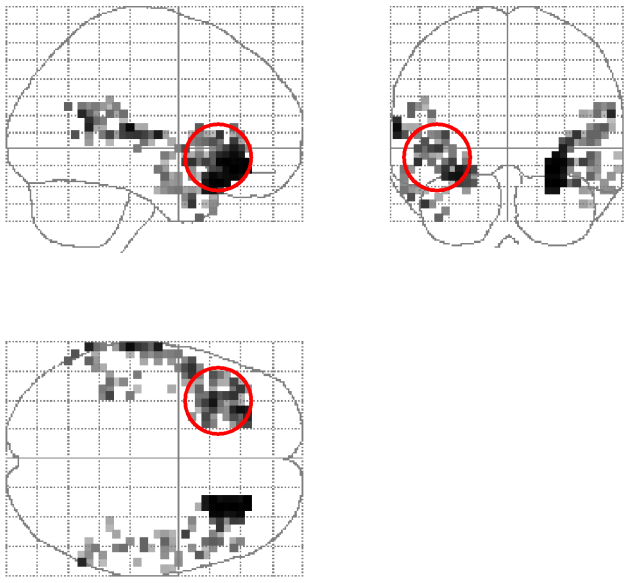

Figure 7.9 – *Carte statistique de Grubbs – Contraste « Activation à la parole » – Jeu de données* $\boldsymbol{J}$ *décrit en annexe A*

La figure 7.9 présente la carte statistique de Grubbs obtenue pour le contraste « Activation à la parole », seuillée pour la valeur critique $G_c = 2,62$ estimée par simulations de Monte-Carlo en considérant un risque de première espèce de 1 % non corrigé des comparaisons multiples ($P_{non\ corrigée} < 0,01$). Cette carte seuillée

montre que l'hypothèse de normalité peut être rejetée dans de nombreuses régions critiques, indiquant une très forte variabilité inter-sujets pour ce jeu de données. De plus, il est intéressant de noter la présence de l'aire de Broca (en rouge) parmi ces régions critiques, confirmant le gain en sensibilité des tests fondés sur des statistiques robustes dans les régions où les effets estimés ne sont pas homogènes.

7.3.3 Contraste « Effet de la répétition »

La table 7.2 et la figure 7.10 résument les résultats obtenus pour le contraste d'intérêt « Effet de la répétition ». Seul un cluster situé dans le gyrus frontal inférieur gauche correspondant à l'aire de Broca est détecté (en rouge), pour les seuils considérés, en utilisant la statistique du rang signé et la statistique t-empirique.

Position anatomique du cluster	Procédure de test statistique	$P_{corrigée}$ au niveau cluster	Étendue du cluster (en voxels)	$P_{corrigée}$ au niveau voxel (maximum)	Position du maximum (en mm)		
					x	y	z
Gyrus frontal inférieur gauche (Broca)	Test t paramétrique	0.29	26	0.94	-44	24	-4
	Test t permutations	0.08	44	0.31	-44	24	-4
	Test du signe	0.66	1	0.66	-44	28	-4
	Test du rang signé	**0.04**	62	0.66	-44	28	-8
	Test t-empirique	**0.03**	66	0.66	-44	28	-8

Table 7.2 – *Résultats de l'analyse de groupe pour différentes procédures de test statistique – Contraste « Effet de la répétition » – Jeu de données* ***J*** *décrit en annexe A*

Les résultats obtenus pour ce contraste confirment les résultats établis pour le contraste précédent, et notamment le même gain de sensibilité est observé au niveau cluster en utilisant des statistiques de décision plus robustes que la statistique de Student. La seule différence provient de la statistique du signe qui présente ici un faible pouvoir statistique.

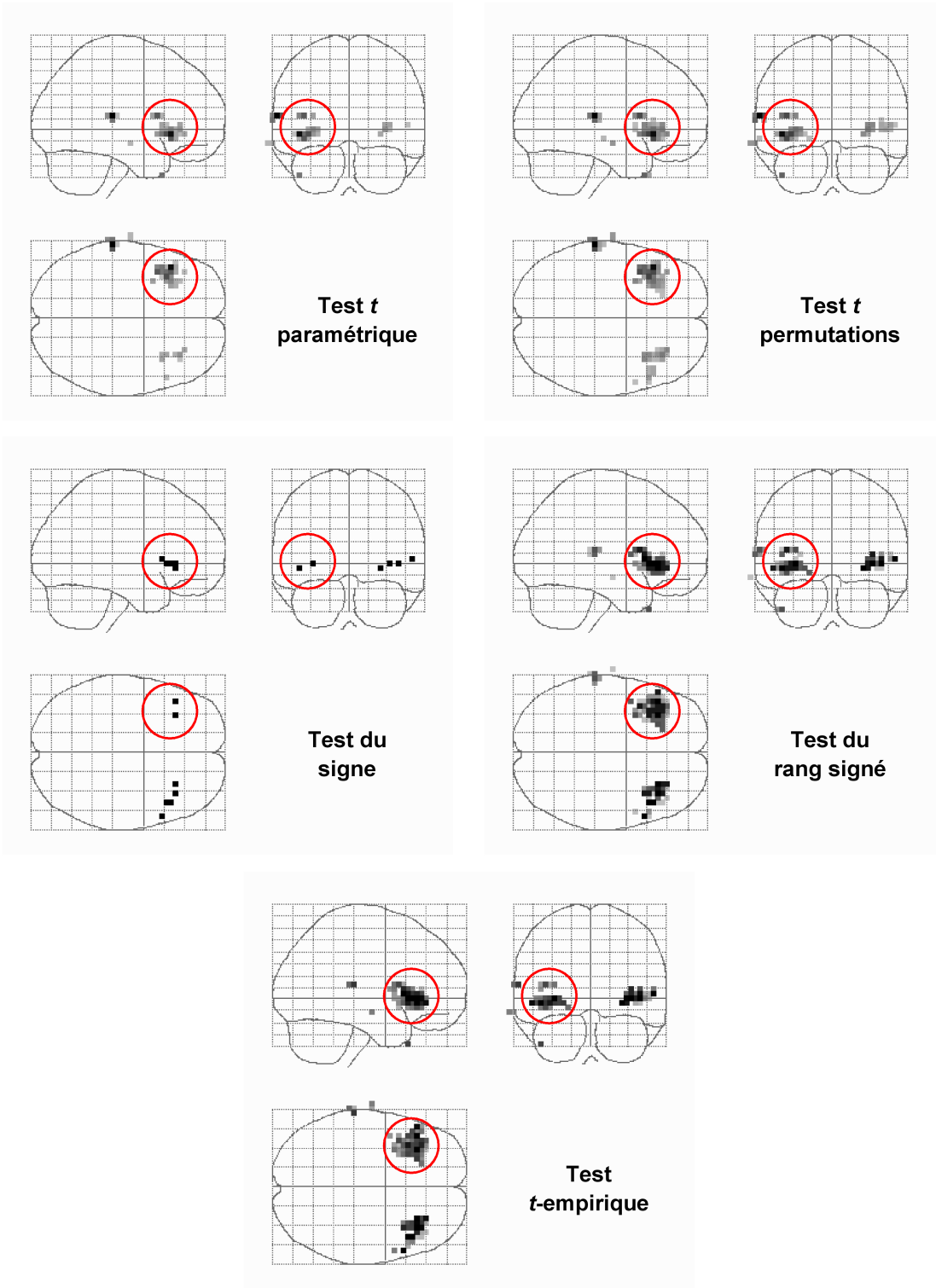

Figure 7.10 – *Cartes d'activation du groupe obtenues par différentes procédures de test statistique – Contraste « Effet de la répétition » – Jeu de données* ***J*** *décrit en annexe A*

D'un point de vue cognitif, les résultats obtenus pour les deux contrastes d'intérêt à l'aide de statistiques robustes permettent de mettre en évidence le rôle de l'aire de Broca dans la perception du langage et notamment sa possible implication dans les processus de mémorisation linguistique [Dehaene-Lambertz *et al.*, 2006b]. Il est très intéressant de noter que l'organisation fonctionnelle des régions temporales supérieures observée ici chez le nourrisson est la même que celle déjà observée chez l'adulte (voir notamment les résultats de l'expérience 1 dans l'article de Dehaene-Lambertz *et al.* [2006a]).

7.4 Conclusion

Nous proposons l'utilisation de statistiques robustes, la robustesse se définissant dans le contexte du test d'hypothèses comme la stabilité de la sensibilité de la statistique à travers les distributions possibles des effets. En effet, le petit nombre de sujets rend difficile l'estimation de la distribution vraie des effets à travers les sujets. De plus, nous avons observé que cette distribution peut varier suivant les voxels. L'utilisation de tests « robustes » présentant des performances stables à travers les distributions permet d'augmenter la sensibilité des résultats de l'analyse de groupe.

Les deux premières statistiques utilisées (statistique du signe et statistique du rang signé) sont plus connues pour être robustes au sens des estimateurs mais permettent également un gain de sensibilité comme le montre l'exemple proposé. La statistique ELR a en revanche été construite spécifiquement pour être robuste au sens des distributions.

Enfin, la méthode de calibration statistique par permutations permet de limiter les hypothèses sur la distribution des données à une hypothèse de symétrie. Même si elle n'est pas forcément vérifiée, elle reste beaucoup moins forte que l'hypothèse de normalité et permet un gain sensible en spécificité du test. Elle présente également l'avantage de permettre des comparaisons multiples dans un cadre exact, sans ajouter d'hypothèses supplémentaires. Notons enfin que Romano [1990] montre que sous certaines conditions le test de permutations reste asymptotiquement exact pour une distribution non-symétrique.

Chapitre 8

Analyse à effets mixtes

Les statistiques de test étudiées au cours des chapitres précédents sont justifiées par un modèle d'échantillonnage simple dans lequel les mesures de contraste BOLD associées aux différents sujets sont considérées comme exactes. Une telle hypothèse est simpliste du fait des différentes sources d'imprécision affectant le procédé d'imagerie (bruit thermique, biais magnétiques, artéfacts de mouvement,...), ainsi que du manque de réalisme des modèles utilisés lors des analyses intra-sujets en termes de forme de la réponse hémodynamique, d'activité au repos ou encore d'artéfacts physiologiques. Ces sources d'erreur propres à chaque sujet, et donc assimilables à une *variabilité intra-sujet*, masquent en partie la *variabilité inter-sujets* que les études de groupe cherchent à caractériser. On parle ainsi d'« effets mixtes » pour souligner l'existence de deux sources distinctes de variabilité.

L'objectif de ce chapitre est d'introduire des statistiques de test tenant compte des erreurs de mesure sur les contrastes BOLD, ceci afin de pondérer les différentes observations en fonction de leur fiabilité et de réduire ainsi le risque de mauvaise détection lié à l'existence de données atypiques. Les statistiques robustes étudiées au chapitre 7 répondent en partie à cette problématique mais sans exploiter les niveaux d'incertitude fournis par les analyses intra-sujets, dont nous observons

qu'ils sont souvent assez hétérogènes au sein d'un groupe de sujets, comme illustré par la figure 8.1.

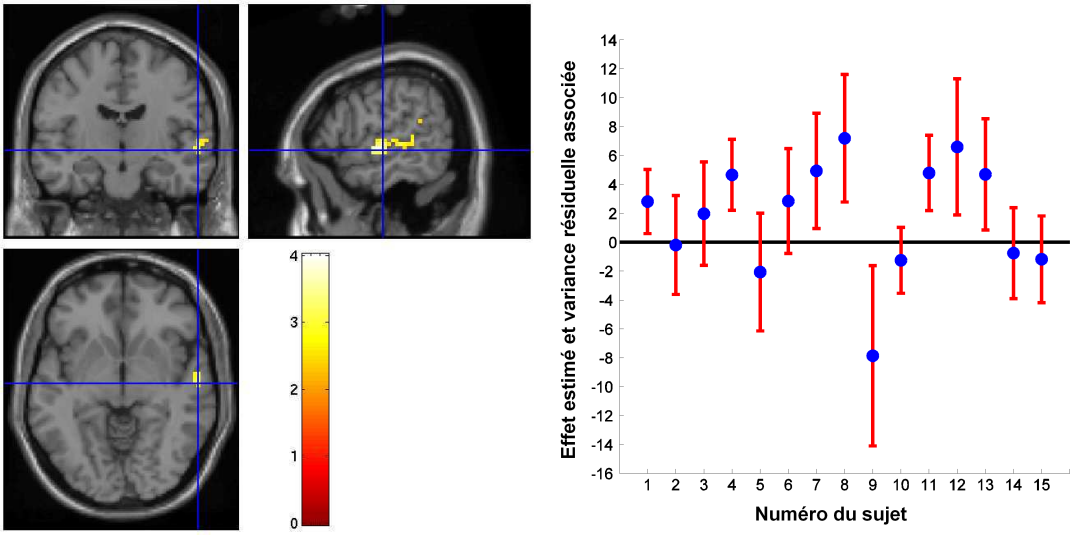

Figure 8.1 – *Exemple de variance résiduelle non constante à travers les sujets pour le jeu de données* ***G*** *décrit en annexe A – Contraste « Interaction Phrase × Locuteur » – Position de la croix* $[60 - 15 - 6]$ *mm en coordonnées de Talairach*

Ce travail s'inscrit dans la suite de [Worsley *et al.*, 2002 ; Friston *et al.*, 2002 ; Beckmann *et al.*, 2003 ; Neumann *et al.*, 2003 ; Woolrich *et al.*, 2004] qui ont élaboré des statistiques « à effets mixtes » à partir de modèles hiérarchiques (8.1) du type de ceux évoqués au chapitre 3. Nous construisons ici deux variantes nouvelles de ces statistiques fondées sur le concept générique de rapport de vraisemblances maximales (8.2). L'une, qui peut être vue comme une généralisation « effets mixtes » de la statistique de Student, repose sur le même modèle hiérarchique que dans les articles pré-cités, dans lequel la distribution inter-sujets des effets est supposée gaussienne. L'autre statistique découle d'une version non-paramétrique du modèle hiérarchique, et généralise ainsi le rapport de vraisemblances empiriques introduit par Owen [2001] et discuté au chapitre 7. Nous montrons notamment que la deuxième statistique est plus robuste que la première au prix d'un temps de calcul fortement accru.

Afin de garantir la validité sous des hypothèses très générales des tests statistiques qui découlent de ces statistiques de décision, nous reprenons le mécanisme des permutations introduit dans le cadre des statistiques robustes et l'adaptons à la calibration des statistiques à effets mixtes (8.3). L'intérêt pratique de ces statistiques de décision à effets mixtes est illustré à la fin de ce chapitre (8.4) sur le jeux de données **G** décrit en annexe A.

8.1 Modèle hiérarchique à deux niveaux

En chaque voxel de l'espace de Talairach, l'analyse intra-sujet fournit pour chaque sujet i de l'étude une estimation $\hat{\beta}_i$ de son effet β_i (contraste) ainsi qu'une estimation $\hat{\sigma}_i$ de l'erreur standard d'estimation sur β_i, cette dernière statistique étant classiquement ignorée dans les analyses de groupe. Il faut noter que le couple de statistiques $(\hat{\beta}_i, \hat{\sigma}_i)$ n'est pas exhaustif, puisqu'il ne prend pas en compte par exemple le nombre de degrés de liberté avec lequel a été estimé $\hat{\sigma}_i$, ni l'information de covariance spatiale provenant de l'analyse intra-sujet. Cependant, l'utilisation de $\hat{\sigma}_i$ permet d'exploiter plus d'informations dans les données d'IRMf et d'obtenir ainsi des inférences statistiques plus sensibles.

Notre objectif est de faire une inférence sur l'effet moyen de la population à partir de l'échantillon d'observations $(\hat{\beta}_1, \ldots, \hat{\beta}_n)$ et des incertitudes correspondantes $(\hat{\sigma}_1, \ldots, \hat{\sigma}_n)$ qui seront traitées ici comme des paramètres fixes et non des variables aléatoires. Nous considérons un modèle hiérarchique à deux niveaux permettant de relier les observations à la distribution inconnue de l'effet dans la population, notée $f(\beta)$:

$$\left\{ \begin{array}{lcl} \hat{\beta}_i & = & \beta_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \hat{\sigma}_i) \\ \beta_i & \sim & f(\beta) \end{array} \right. \tag{8.1}$$

Le premier niveau de ce modèle décrit le processus d'estimation du vrai effet inconnu β_i et modélise la variabilité intra-sujet. Pour des raisons calculatoires, nous faisons l'hypothèse simplificatrice d'une densité gaussienne de l'effet estimé

conditionnellement à l'effet réel non observé β_i, ce qui revient à négliger l'incertitude sur $\hat{\sigma}_i$ ou, de façon équivalente, à considérer les degrés de liberté intra-sujets comme infinis. Nous verrons néanmoins au paragraphe 8.3 que cette simplification n'impacte en rien la validité des tests statistiques (mais éventuellement leur sensibilité). La densité du premier niveau dépend du sujet via l'erreur standard $\hat{\sigma}_i$ ce qui permet de prendre en compte l'hétéroscédasticité, c'est-à-dire la possibilité d'avoir des niveaux de bruit différents pour chaque sujet.

Au second niveau, la densité f représente la distribution de l'effet β_i à travers les sujets et permet de modéliser la variabilité inter-sujet. Nous introduisons l'espace $\mathcal{F}$ dans lequel f est recherchée. Le premier cas particulier que nous considérerons est le cas classique où $\mathcal{F}$ est la famille gaussienne décrite par deux paramètres inconnus (moyenne et variance de groupe). Nous traiterons ensuite le cas non-paramétrique où aucune hypothèse n'est faite sur f de sorte que $\mathcal{F}$ est un espace de dimension infinie.

8.1.1 Fonction de vraisemblance

En marginalisant les effets cachés, nous voyons que chaque observation $\hat{\beta}_i$ a pour densité non-conditionnelle le produit de convolution $g_i \otimes f$, avec $g_i \triangleq N(0, \hat{\sigma}_i)$, mettant en œuvre les variabilités intra- et inter-sujets. Ceci montre que les observations ne sont généralement pas distribuées de façon identique sauf dans le cas improbable d'un bruit homoscédastique.

Sous l'hypothèse classique d'un échantillonnage indépendant des sujets, nous pouvons écrire la fonction de vraisemblance du modèle :

$$L(f) = p(\hat{\beta}_1, \ldots, \hat{\beta}_n | f) = \prod_{i=1}^{n} g_i \otimes f\,(\hat{\beta}_i)$$

Cette fonction joue un rôle central dans le calcul des rapports de vraisemblance maximaux que nous détaillons maintenant.

8.2 Rapports de vraisemblances maximales

Nous recherchons une statistique de décision permettant de tester efficacement l'hypothèse nulle d'un effet moyen nul, autrement dit :

$$H_0 : \quad \int \beta f(\beta)\, d\beta = 0$$

Lorsque aucune statistique de décision ne s'impose de façon évidente, il est classique de considérer le rapport de vraisemblances maximales :

$$R = \frac{\max_{f \in \mathcal{F}_0} L(f)}{\max_{f \in \mathcal{F}} L(f)} \tag{8.2}$$

où $\mathcal{F}_0 \subset \mathcal{F}$ est le sous-espace des densités vérifiant l'hypothèse nulle. Nous vérifions aisément que R est toujours compris entre 0 et 1, une valeur faible indiquant que l'hypothèse nulle est peu vraisemblable alors qu'une valeur proche de 1 suggère que les informations expérimentales ne permettent pas de la rejeter. Nous choisirons donc une région critique de la forme $R \leq r$, où r est un seuil à déterminer en fonction d'un certain taux d'erreur.

8.2.1 Cas gaussien

Le calcul du rapport de vraisemblance (équation 8.2) nécessite de résoudre deux problèmes d'optimisation, l'un sur l'espace de recherche entier (dénominateur), l'autre sous contrainte (numérateur). Dans le cas où $\mathcal{F}$ choisi est la famille gaussienne, ces problèmes n'admettent de solution analytique que dans le cas particulier du bruit homoscédastique où la variance intra-sujet est constante. En règle générale, la maximisation doit donc être effectuée au moyen d'une méthode numérique comme, par exemple, l'algorithme EM (*Expectation-Maximization*) dû à Dempster *et al.* [1977], dont une instantiation spécifique à notre problème est détaillée en annexe B.

Comme toute méthode itérative, l'algorithme EM doit être initialisé par une première estimée $\hat{f}_0$ de la solution. Nous choisissons la densité gaussienne de

moyenne et variance respectivement égales à la moyenne empirique et à la variance empirique de l'échantillon. Il s'agit en fait de la densité qui réaliserait le maximum de vraisemblance *si les variances intra-sujets étaient identiquement nulles*, ce qui suggère que l'initialisation est assez précise lorsque la variabilité inter-sujets domine la variabilité intra-sujet.

Bien que l'algorithme EM n'assure pas l'obtention du maximum global de la vraisemblance, il garantit la convergence vers un maximum local avec une valeur de vraisemblance nécessairement plus élevée que celle de l'estimée initiale. En pratique, de 3 à 5 itérations sont nécessaires pour atteindre une précision de 1 % sur les variations de la vraisemblance.

8.2.2 Cas non-paramétrique

Le cas où $\mathcal{F}$ est l'espace de toutes les densités de probabilité sur $\mathbb{R}$ s'avère relativement aisé à traiter grâce à un résultat d'analyse convexe donné par Lindsay [1983] qui permet d'affirmer que la densité maximisant la vraisemblance existe et se réduit à un mélange *fini* de distributions de Dirac, le nombre de composantes distinctes étant au plus égal au nombre d'observations :

$$f(\beta) = \sum_{i=1}^{n} w_i \delta(\beta - z_i) \tag{8.3}$$

où les w_i sont des poids positifs et de somme unitaire et les z_i sont des points de support. Ce théorème généralise la propriété déjà mentionnée au chapitre 7 dans le contexte d'observations *exactes* (variances intra-sujets nulles), selon laquelle la vraisemblance non-paramétrique est maximisée par la « distribution empirique » : $\hat{f}_0(\beta) = \frac{1}{n}\sum_{i=1}^{n} \delta(\beta - \hat{\beta}_i)$. Nous pouvons noter que la « distribution empirique » correspond à des poids uniformes et à des points de support coïncidant avec les observations.

Pas plus que dans le cas gaussien, le calcul explicite des paramètres inconnus (ici, les poids et les points de support) n'est possible. Nous proposons donc un deuxième algorithme EM, décrit dans l'annexe C, pour calculer aussi bien le numérateur que le dénominateur de l'équation 8.2. Bien que relativement simple, cet

algorithme EM est environ 100 fois plus coûteux en temps de calcul que l'algorithme EM utilisé dans le cas gaussien. Les simulations révèlent néanmoins que le rapport de vraisemblance non-paramétrique peut s'avérer significativement plus sensible que sa version gaussienne lorsque la distribution du groupe est fortement non-gaussienne (voir figure 8.2). Des exemples comparatifs d'ajustements gaussiens et non-paramétriques sont également donnés à la figure 8.3.

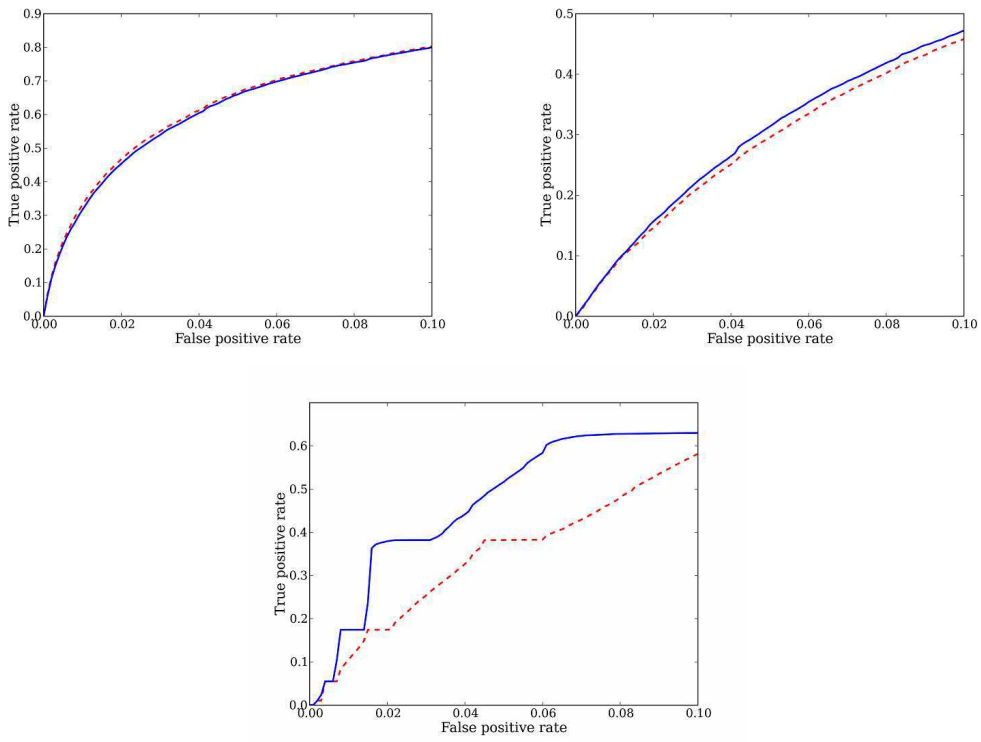

Figure 8.2 – *Courbes ROC simulées comparant la sensibilité des tests de permutations unilatéraux basés respectivement sur le rapport de vraisemblance non-paramétrique (en bleu) et sur le rapport de vraisemblance gaussien (en rouge pointillé). Pour chaque cadre,* 10000 *échantillons de taille* 10 *ont été tirés dans une loi* $f(\beta)$ *de moyenne* 1*, et dégradés par un bruit additif gaussien homoscédastique de variance unitaire. En haut à gauche,* $f(\beta) = N(1,1)$ *est gaussienne. En haut à droite,* $f(\beta) = 1/2\,[N(-1,1) + N(3,1)]$ *est un mélange symétrique de deux gaussiennes. En bas,* $f(\beta) = 1/2\,[N(-1,0.1) + N(3,0.1)]$ *est un mélange du même type avec des pics plus fins. À noter que le rapport de vraisemblance gaussien est ici équivalent à la statistique de Student car le bruit est homoscédastique (voir 8.2.3).*

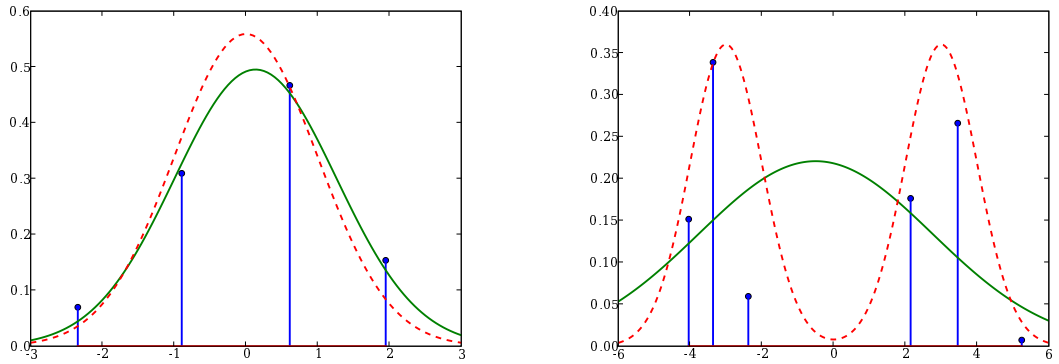

Figure 8.3 – *Exemples d'ajustements gaussiens et non-paramétriques par maximum de vraisemblance dans le modèle à effets mixtes, quand la vraie distribution de l'effet est gaussienne (à gauche) et bi-modale (à droite). Dans les deux cas, des échantillons de taille 20 ont été tirés et contaminés par un bruit hétéroscédastique de variances tirées dans une distribution Gamma* $\Gamma(3, \frac{1}{6})$. *Chacun des deux cadres montre la vraie distribution (en rouge pointillé), l'ajustement gaussien (en vert) et l'ajustement non-paramétrique (en bleu).*

8.2.3 Test unilatéral

Le rapport de vraisemblance (8.2) est adapté à un test bilatéral, étant insensible au signe de l'effet moyen. Afin de pouvoir effectuer des tests unilatéraux, nous définissons une variante signée du rapport de vraisemblance exactement comme au chapitre 7 (équation 7.4) :

$$T = \text{sign}(\hat{\mu})\sqrt{-2\log R} \tag{8.4}$$

où $\hat{\mu}$ est la moyenne de la densité maximisant la vraisemblance, et approchée en pratique par l'algorithme EM adapté au modèle considéré (gaussien ou non-paramétrique). Nous pouvons noter que $\hat{\mu}$ s'interprète également comme l'estimateur par maximum de vraisemblance de l'effet moyen au sens de la vraisemblance profile [Roche *et al.*, 2007].

Rappelons que la formule ci-dessus est motivée par le phénomène de Wilks,

dont on admettra sans démonstration qu'il reste vrai dans les conditions du modèle hiérarchique, et qui implique que T suit une loi normale standard dans la limite des grands échantillons. Nos simulations indiquent néanmoins qu'il faut des échantillons de plusieurs dizaines de sujets pour que cette approximation soit correcte, ce qui est bien supérieur aux échantillons dont nous disposons en pratique.

Dans le cas gaussien, nous montrons aisément que *si le bruit d'observation est homoscédastique*, alors la statistique unilatérale définie par l'équation 8.4 se calcule analytiquement et s'avère être une fonction strictement croissante de la statistique de Student T_s :

$$T = \mathrm{sign}(T_s)\sqrt{n \log\left(1 + \frac{T_s^2}{n-1}\right)}$$

Ainsi, dans ce contexte particulier, T et T_s sont strictement équivalentes en tant que statistiques de décision. Cette observation suggère que la statistique de Student n'est optimale que dans le cas particulier d'un bruit homoscédastique.

Dans le cas non-paramétrique, nous retrouvons la version unilatérale du rapport de vraisemblances empiriques définie en 7.1.3 dans le cas particulier où le bruit d'observation est négligeable devant la variabilité inter-sujets ($\hat{\sigma}_1 = \ldots = \hat{\sigma}_n = 0$).

8.3 Calibration statistique par permutations

Afin de calibrer finement les rapports de vraisemblance définis plus haut, nous proposons d'utiliser le même mécanisme de permutations de signes qu'au chapitre 7 plutôt qu'un résultat de théorie asymptotique. Dans Mériaux *et al.* [2006b], nous formulons trois conditions suffisantes pour que les tests de permutations de signe soient valides dans le contexte des effets mixtes, l'existence d'une erreur d'observation rendant leur application plus délicate que dans la situation classique d'observations supposées exactes. Mise à part éventuellement la troisième, ces hypothèses sont plus générales que celles du modèle hiérarchique défini au paragraphe 8.1 pour guider la construction des statistiques de décision :

(C_1) Les résumés statistiques individuels (effet estimé & variance) sont mutuellement indépendants d'un point de vue statistique. Cette hypothèse revient

à supposer, d'une part, que les sujets sont choisis de manière indépendante dans la population d'intérêt et, d'autre part, que les erreurs de mesure du signal BOLD induites par le scanner ne sont pas reproductibles à travers les sessions.

(C_2) Les résumés statistiques sont équivariants par translation et invariants par changements d'échelle, propriété naturelle garantie par les hypothèses du modèle linéaire utilisé pour les analyses intra-sujets.

(C_3) La distribution de l'effet dans la population est symétrique par rapport à l'effet moyen. Bien que cette condition soit très difficile à valider expérimentalement, les résultats obtenus par Romano [1990] sous des hypothèses plus restrictives indiquent que la symétrie n'est pas critique à condition de disposer d'un nombre suffisant d'observations.

Ces trois hypothèses impliquent que les signes des effets estimés sont échangeables sous l'hypothèse nulle globale H_0 que tous les voxels ont un effet moyen nul. Le couple d'images statistiques de chaque sujet peut ainsi être modifié de la manière suivante : $(\hat{\beta}_i, \hat{\sigma}_i) \rightarrow (-\hat{\beta}_i, \hat{\sigma}_i)$, sans modifier la distribution jointe de toutes les images statistiques des sujets. Cette propriété permet de généraliser le cadre classique des permutations développé par Holmes *et al.* [1996] ; Nichols et Holmes [2002] au cas où les variances du premier niveau sont non-nulles.

8.4 Résultats illustrant l'intérêt des statistiques à effets mixtes

L'intérêt des statistiques de décision à effets mixtes est illustré sur le jeu de données **G** décrit en annexe A. Ce jeu de données a été choisi car l'analyse des variances du premier niveau à travers les sujets a révélé une forte hétérogénéité (voir la figure 8.1).

8.4.1 Traitements des données

L'analyse intra-sujet des données d'IRMf acquises pour chaque sujet a été réalisée avec le logiciel SPM2. Les pré-traitements suivants ont été appliqués aux

données brutes d'IRMf, comme décrit en 2.1 : correction du mouvement, correction du délai d'acquisition entre les coupes, normalisation spatiale à l'aide du template du MNI et lissage spatial par un filtre gaussien de largeur à mi-hauteur $5 \times 5 \times 5$ mm^3. Ensuite, comme décrit en 2.2, un modèle linéaire généralisé a été appliqué aux données pré-traitées, en construisant une matrice de dessin expérimental constituée de quatre conditions convoluées par la fonction de réponse hémodynamique canonique : *Phrases identiques - Locuteurs identiques (SStSSp)*, *Phrases identiques - Locuteurs différents (SStDSp)*, *Phrases différentes - Locuteurs identiques (DStSSp)* et *Phrases différentes - Locuteurs différents (DStDSp)*. La première phrase de chaque bloc de stimulation a été retirée pour les quatre conditions précédentes et modélisée comme une cinquième condition. Les dérives basse-fréquence ont été retirées par application d'un filtre temporel passe-haut (fréquence de coupure 1/128 Hz) et le bruit a été modélisé par un processus autorégressif d'ordre 1. L'estimation des paramètres du modèle linéaire généralisé a permis d'obtenir pour chaque sujet une carte d'effet estimé et une carte de variance résiduelle, correspondant au contraste d'intérêt « Interaction Phrase $\times$ Locuteur » défini par l'expérimentateur.

Les résultats d'analyse de groupe rapportés ici ont été obtenus en utilisant cinq méthodes différentes : le test t paramétrique standard tel qu'il est mis en œuvre dans le logiciel SPM, un test t calibré par permutations à l'aide du logiciel *DISTANCE*, le test t-empirique présenté en 7.1.3, le test de rapport de vraisemblances gaussiennes à effets mixes présenté en 8.2.1 (GLR MFX) et le test de rapport de vraisemblances empiriques à effets mixes présenté en 8.2.2 (ELR MFX) également calibrés par permutations à l'aide de *DISTANCE*. Toutes les analyses de groupe ont été restreintes à un masque de 2920 voxels englobant les aires périsylviennes des deux hémisphères, afin de se concentrer sur les zones corticales connues pour être activées lors d'une tâche d'écoute passive.

Pour le contraste d'intérêt choisi, les cartes statistiques de groupe ont été seuillées pour un risque de première espèce de 1 % non corrigé des comparaisons multiples ($P_{non\ corrigée} \leq 0,01$), puis pour les voxels dépassant le seuil, les « p-values » corrigées au niveau voxel et au niveau cluster ont été calculées. Pour le test t paramétrique, le seuil correspondant au risque spécifié a été obtenu à partir de la distribution de Student à $n-1$ degrés de liberté (voir figure 3.3), avec

$n = 15$ sujets participant à l'étude, et les « p-values » corrigées au niveau voxel et au niveau cluster ont été calculées en utilisant la théorie des champs aléatoires gaussiens mise en œuvre dans SPM. En ce qui concerne les tests calibrés par permutations, les calculs du seuil correspondant au risque spécifié et des « p-values » corrigées au niveau voxel et au niveau cluster ont été effectués en générant aléatoirement $N = 10000$ permutations indépendantes (au lieu de prendre l'ensemble des permutations possibles, soit $2^{15} = 32768$ permutations), mais en conservant tous les voxels du masque d'analyse. Pour les statistiques calibrées par permutations, les temps de calcul varient entre 2 mn pour les statistiques t et t-empirique, 6 mn pour la statistique GLR MFX et environ 8 heures pour la statistique ELR MFX (calculs effectués par un PC standard avec un processeur de $2,80$ GHz tournant sous Linux). Ces temps de calcul restent tout à fait raisonnable au regard du temps nécessaire à la mise en place d'une étude d'IRMf et à l'acquisition des données. De plus, les calculs de tests de permutations sont facilement parallélisables, ce qui devrait permettre de réduire significativement ces temps de calcul.

Pour comparer les différents méthodes d'analyse de groupe entre elles, nous nous restreignons dans la présentation des résultats aux clusters dont la « p-value » corrigée est inférieure à 5 % pour au moins une des méthodes utilisées. Quant aux cartes d'activation du groupe présentées, elles correspondent à la projection du maximum d'intensité (MIP, *Maximum Intensity Projection*) des cartes statistiques seuillées à $P_{non\ corrigée} \leq 0,01$ (pas de seuil sur l'extension spatiale des clusters).

8.4.2 Contraste « Interaction Phrase × Locuteur »

La figure 8.4 et la table 8.1 résument les résultats obtenus pour le contraste d'intérêt « Interaction Phrase × Locuteur ». Seul un cluster situé dans le sulcus temporal supérieur droit (STS) est détecté (en rouge), pour les seuils considérés, en utilisant les statistiques à effets mixtes (GLR MFX et ELR MFX). Notons que les cartes d'activation du groupe obtenues en utilisant le test t paramétrique standard et le test t calibré par permutations étant très proches de celle obtenue en utilisant le test t-empirique, seule cette dernière est représentée sur la figure 8.4.

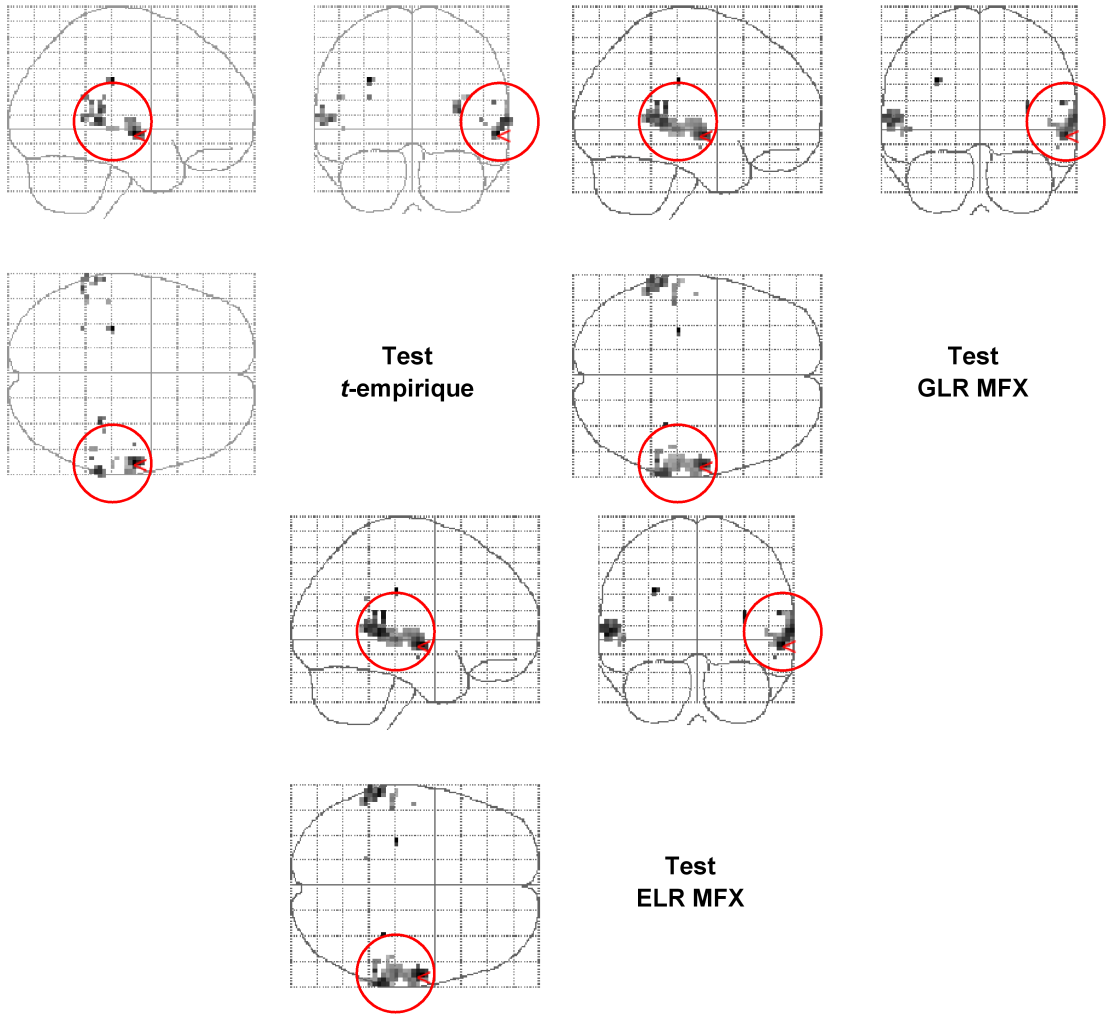

Figure 8.4 – *Cartes d'activation du groupe obtenues par différentes procédures de test statistique – Contraste « Interaction Phrase × Locuteur » – Jeu de données* ***G*** *décrit en annexe A*

Ces résultats confirment que l'utilisation d'une information supplémentaire (la variance intra-sujet) permet d'améliorer la puissance de détection des tests fondés sur des statistiques de décision à effets mixtes, en ce sens que les clusters détectés sont plus étendus et présentent des « p-values » corrigées plus faibles. De plus, nous pouvons remarquer que la version non-paramétrique de la statistique de rapport de vraisemblances à effets mixtes (ELR MFX) affiche les mêmes performances statistiques que la version gaussienne (GLR MFX). Cette constatation s'est renouvelée

pour tous les autres contrastes d'intérêt que nous avons testés pour ce jeu de données. Ainsi, il semblerait que la distribution des effets estimés à travers les sujets soit à peu près normale. Nous pouvons prévoir que des différences plus significatives pourraient être observées entre les performances statistiques de l'ELR MFX et du GLR MFX pour des distributions d'effets estimés fortement hétérogènes, comme cela peut se produire pour des contrastes moteurs par exemple.

Position anatomique du cluster	Procédure de test statistique	$P_{corrigée}$ au niveau cluster	Étendue du cluster (en voxels)	$P_{corrigée}$ au niveau voxel (maximum)	Position du maximum (en mm)		
					x	y	z
Sulcus temporal supérieur droit (STS)	Test t paramétrique	0.13	24	0.70	60	-12	-3
	Test t permutations	0.09	25	0.28	60	-12	-3
	Test t-empirique	0.13	20	0.44	60	-9	-3
	Test GLR MFX	**0.02**	97	0.04	60	-12	-3
	Test ELR MFX	**0.01**	111	0.04	60	-12	-3

Table 8.1 – *Résultats de l'analyse de groupe pour différentes procédures de test statistique – Contraste « Interaction Phrase × Locuteur » – Jeu de données* ***G*** *décrit en annexe A*

D'un point de vue cognitif, les résultats obtenus pour le contraste d'intérêt « Interaction Phrase × Locuteur » confirme que la région du sulcus temporal supérieur droit (STS) ne semble pas sensible à la répétition du contenu linguistique de la voix lorsque des variations acoustiques sont introduites. Comme le montrent certaines études précédentes [Belin *et al.*, 2000 ; Belin et Zatorre, 2003 ; von Kriegstein *et al.*, 2003], cette région pourrait être impliquée dans les processus de reconnaissance des aspects paralinguitiques de la voix, comme par exemple la détection des émotions ou des changements de locuteur, ainsi que le montrent les résultats présentés ici.

8.5 Conclusion

Dans ce chapitre, nous proposons de nouvelles statistiques à effets mixtes permettant d'utiliser l'information de variance intra-sujet disponible à l'issue des analyses du premier niveau. Cette information permet de « pondérer » les sujets en fonction de l'incertitude associée à leur effet estimé : un sujet aura un « poids » d'autant plus élevé dans le calcul de la statistique de décision à effets mixtes que l'estimation de son effet est fiable.

Cette approche permet une pondération locale (*a priori* différente en chaque voxel) des sujets tout en préservant leur indépendance statistique, contrairement à l'approche naïve consistant à retirer un sujet jugé atypique après une analyse *conjointe* des effets. Une telle approche reviendrait formellement à une analyse à effets mixtes dans laquelle une incertitude infinie serait associée au sujet atypique, mais la mesure d'incertitude étant alors dépendante des autres sujets, la condition d'indépendance statistique serait violée.

Néanmoins, les statistiques de test introduites dans ce chapitre pourraient être calculées à partir de mesures plus fines de l'incertitude du premier niveau, à condition qu'elles restent statistiquement indépendantes ; ces mesures pourraient incorporer, par exemple, l'erreur de recalage, ou l'erreur de normalisation spatiale,...

Chapitre 9

Conclusion

Comme nous l'avons dit à plusieurs reprises, l'analyse de groupe en IRMf pose à la fois la question de la normalisation spatiale, à savoir le positionnement des cerveaux des sujets dans un même espace de référence, et la question de l'inférence statistique dont l'objectif est de résumer les résultats de l'analyse intra-sujet au niveau du groupe.

Durant cette thèse, nous nous sommes concentrés sur la question de l'inférence statistique. Le premier objectif était de développer des méthodes de vérification de l'hypothèse forte, faite par l'analyse de groupe « standard », de normalité des effets à travers les sujets. Le deuxième objectif était de proposer des statistiques de décision alternatives pour les cas où cette hypothèse serait invalide. L'ensemble des développements méthodologiques associés à la réalisation de ces deux objectifs ont été intégrés dans le logiciel *DISTANCE*.

9.1 Objectifs de la thèse

L'analyse de groupe « standard », telle qu'elle est mise en œuvre dans le logiciel SPM, repose sur une hypothèse très forte de normalité des effets estimés à travers les sujets. Jusqu'à présent, aucune méthode n'était à notre connaissance spécifiquement dédiée à la validation de cette hypothèse, sans doute parce qu'elle

est difficile à vérifier étant donné le petit nombre de sujets participant aux études d'IRMf. Cependant, sa mise en défaut a pour conséquence de biaiser la spécificité du test de Student paramétrique utilisé pour tester la significativité de l'effet de groupe, mais également de diminuer sa sensibilité. Il nous a donc semblé utile, dans un premier temps, de développer des outils de diagnostique capables de nous renseigner sur l'homogénéité des effets estimés. Afin d'évaluer la fréquence des jeux de données hétérogènes en pratique, nous avons appliqué ces outils de diagnostique à un grand nombre de jeux de données couvrant un large éventail de paradigmes expérimentaux. La fréquence élevée que nous avons observée semble contredire l'hypothèse de normalité.

Dans un deuxième temps, nous avons relâché l'hypothèse de normalité et avons donc cherché à proposer des méthodes d'inférence alternatives à la méthode « standard ». Ces méthodes doivent être capables de corriger l'éventuel manque de sensibilité due à l'utilisation de la statistique de Student pour tester l'effet moyen mais également le biais de spécificité introduit par la calibration paramétrique de la statistique de décision. C'est pourquoi il nous a semblé intéressant de travailler dans le cadre de la méthode de tests par permutations introduite par Holmes *et al.* [1996] ; Nichols et Holmes [2002], mais en proposant de nouvelles statistiques de décision présentant des niveaux de sensibilité stables à travers un large éventail de distributions d'effets estimés, mais également capables de prendre en compte l'incertitude du premier niveau sur chaque effet estimé pour pondérer les sujets en conséquence.

Enfin, un dernier point nous a semblé important : afin de pouvoir valider nos nouvelles méthodes dédiées à l'analyse de groupe, celles-ci doivent être accessibles à la communauté de neuro-imagerie et facilement utilisables dans le cadre des traitements classiques de données d'IRMf.

9.2 Principaux résultats de la thèse

Dans un premier temps, il convenait donc de disposer de procédures permettant de contrôler l'homogénéité des données afin de pouvoir raisonnablement valider l'hypothèse de normalité faite par l'analyse de groupe « standard ». Pour cela, nous proposons une méthode multivariée de diagnostique d'influence (chapitre 4),

fondée sur le calcul d'une matrice de distances entre les sujets à partir de leurs cartes d'effets estimés. Cette matrice peut révéler la présence de sous-groupes de sujets ou de sujets atypiques lorsqu'elle est visualisée et le calcul de la distance de Cook permet d'estimer l'influence de chaque sujet (via sa distance aux autres) sur la moyenne globale des distances [Kherif *et al.*, 2004]. Notre approche ne constitue pas pour l'instant une méthode de test d'homogénéité permettant de valider (ou d'infirmer) directement les hypothèses de normalité de l'approche « standard » mais elle peut servir de diagnostique d'homogénéité en repérant les éventuels sujets atypiques qui présentent une trop grande influence sur la moyenne globale.

Nous complétons cet outil diagnostique par le test statistique de Grubbs (chapitre 5), qui permet réellement de tester de manière univariée l'hypothèse de distribution normale. Le calcul du seuil par simulations de Monte-Carlo révèle que la formule analytique proposée dans la littérature n'est vérifiée qu'asymptotiquement ce qui rendait en pratique le test de Grubbs anti-conservatif pour les petits échantillons de données. Cette approche présente l'intérêt de donner une information locale quant à la présence éventuelle de données atypiques. En effet, l'hétérogénéité à travers les sujets peut ne se révéler que localement, par exemple à cause d'erreurs de recalage ou de normalisation dans certaines régions. Ainsi les résultats obtenus par application du test de Grubbs nous amène à favoriser des méthodes capables de pondérer les sujets localement, et non globalement, en fonction de leurs effets estimés.

Afin de déterminer si la présence de données atypiques dans les jeux de données d'IRMf est une situation fréquente ou exceptionnelle, nous avons testé notre méthode de diagnostique d'influence et le test de Grubbs sur un ensemble de vingt jeux de données, collectés au cours de la thèse. Cette étude, présentée au chapitre 6, révèle que 30 % des jeux de données présentent au moins un sujet atypique. Cette observation nous semble cruciale, car d'après nos simulations, ce pourcentage indique que le phénomène n'est pas marginal. Il devient vraiment nécessaire de le prendre en compte et il peut en plus rendre plus délicate l'interprétation des résultats déjà publiés avec les techniques standards.

Ce résultat nous conforte dans l'idée que les hypothèses de distribution normale des effets à travers les sujets sont sans doute fausses et nous proposons donc comme alternative à l'analyse de groupe « standard » l'utilisation de statistiques de dé-

cision robustes, la robustesse se définissant dans le contexte du test d'hypothèses comme la stabilité de la sensibilité de la statistique à travers les distributions possibles des effets (chapitre 7). Cette approche se justifie aussi par le petit nombre de sujets participant aux études d'IRMf, qui rend difficile l'estimation de la distribution vraie des effets à travers les sujets. Les trois statistiques robustes utilisées (statistique du signe, statistique du rang signé et statistique du rapport de vraisemblances empiriques) sont par ailleurs calibrées par permutations afin d'améliorer la spécificité du test statistique, ce qui permet de limiter les hypothèses sur la distribution des données à une hypothèse de symétrie, beaucoup plus faible que la normalité. Cette nouvelle approche combinant des statistiques robustes calibrées par permutations permet d'améliorer la sensibilité des résultats de l'analyse de groupe pour un jeu de données *a priori* très hétérogène, puisque issu d'une étude chez des nourrissons de 11 à 17 semaines, période au cours de laquelle le développement du cortex cérébral est très rapide [Mériaux *et al.*, 2006c ; Dehaene-Lambertz *et al.*, 2006b]. Remarquons toutefois que pour d'autres jeux de données, les statistiques de décision robustes n'ont pas permis d'améliorer la sensibilité des résultats de l'analyse de groupe « standard », voire ont donné des résultats moins significatifs. En effet, pour des distributions à peu près homogènes d'effets estimés à travers les sujets, la statistique de Student reste la plus sensible. En revanche, quelque soit le jeu de données testés, les résultats confirment que l'approche par permutations améliore la spécificité du test et permet une correction exacte du problème des comparaisons multiples (résultats déjà indiqués dans Hayasaka et Nichols [2003]).

L'analyse de groupe « standard » n'utilise pas l'information de variance intra-sujet disponible à l'issue des analyses du premier niveau, la considérant comme constante à travers les sujets. Or les résultats obtenus par application du test de Grubbs nous amène à favoriser des méthodes capables de pondérer les sujets localement, et notamment en utilisant cette mesure d'incertitude associée à l'effet estimé. En partant du cadre très général des rapports de vraisemblances maximales, nous proposons deux nouvelles statistiques de décision à effets mixtes, l'une conservant l'hypothèse de distribution normale des effets à travers les sujets [Mériaux *et al.*, 2006b] et dénommée GLR MFX, l'autre non-paramétrique [Roche *et al.*, 2007] et dénommée ELR MFX (chapitre 8). Chacune de ces deux approches

permet d'améliorer la sensibilité des résultats d'analyse de groupe pour le jeu de données du FIAC 2005, confirmant l'intérêt de disposer d'une mesure d'incertitude supplémentaire capable de pondérer les sujets.

9.3 Perspectives

Tout d'abord, comme nous l'avons fait remarquer dans le chapitre 2, les différentes étapes de pré-traitements et de modélisation permettant d'obtenir les cartes d'effet estimé et de variance résiduelle pour chaque sujet constituent autant de sources d'erreur potentielles, avec des répercussions plus ou moins importantes sur les résultats de l'analyse de groupe. Par exemple, des erreurs de recalage ou de normalisation spatiale peuvent introduire une incertitude sur la localisation de l'effet estimé alors que des erreurs de modélisation de la HRF peuvent biaiser l'estimation de l'amplitude de l'effet.

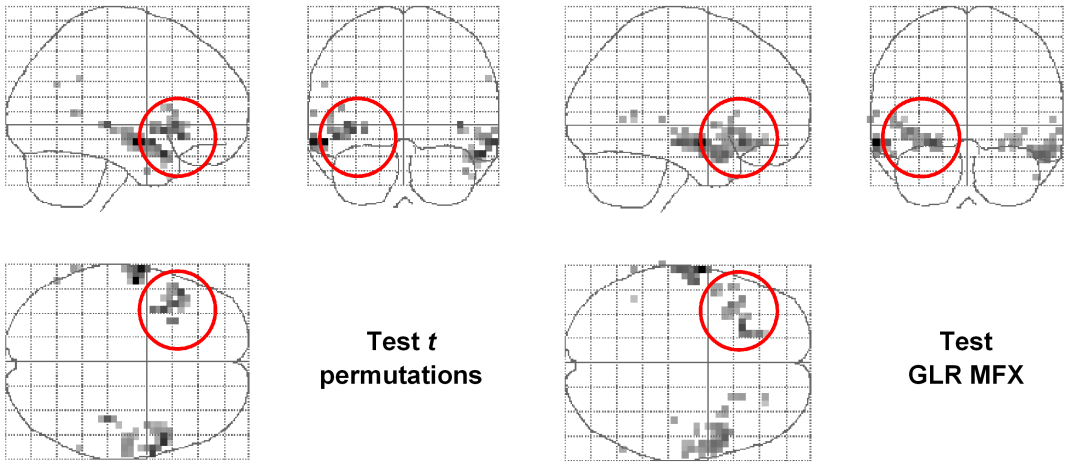

Figure 9.1 – *Cartes d'activation du groupe obtenues par différentes procédures de test statistique – Contraste « Activation à la parole » – Jeu de données* ***J*** *décrit en annexe A*

La figure 9.1 compare les cartes d'activation (affichage de la projection du maximum d'intensité, seuil de $P_{non\ corrigée} \leq 0,01$) obtenues pour la statistique de Student et la statistique GLR MFX introduite au chapitre 8, toutes les deux

calibrées par permutations. Nous pouvons remarquer que l'approche à effets mixtes ne permet pas dans ce cas d'améliorer la sensibilité des résultats de l'analyse de groupe, au contraire ils apparaissent moins significatifs [Mériaux *et al.*, 2006a]. Par exemple, la « p-value » corrigée pour le cluster cerclé de rouge est plus élevée en utilisant la statistique GLR MFX.

Une explication possible de ce résultat est que le modèle de premier niveau est inexact et notamment les erreurs de modélisation de la HRF induisent des erreurs importantes non seulement sur l'estimation de l'effet, mais également *sur l'estimation de son incertitude*. En effet, une partie du signal BOLD peut se retrouver rejeté dans le terme d'erreur du modèle linéaire si la HRF est mal modélisée, ce qui fait augmenter de façon artificielle la variance de l'erreur. On voit ainsi que plus un effet est fort, plus sa variance l'est aussi et donc l'approche à effets mixtes a tendance à attribuer un point faible aux meilleurs activateurs. Lorsque nous affichons la carte de corrélation entre l'effet estimé et la variance résiduelle (voir l'affichage de la projection du maximum d'intensité sur la figure 9.2, seuil de $P_{non\ corrigée} \leq 0,01$), nous remarquons en effet que la zone déjà identifiée sur la figure 9.1, dans laquelle l'utilisation de la statistique à effets mixtes provoque une baisse de sensibilité (en rouge), correspond bien à une zone de forte corrélation positive entre effet et variance.

Une des causes de l'observation de données atypiques peut être une mauvaise normalisation spatiale dans le repère de Talairach (espace du template du MNI). Il est donc intéressant d'utiliser l'information sur l'erreur de recalage spatial dans la construction de la statistique de décision à effets mixtes. C'est un des objectifs principaux de la thèse en cours de Merlin Keller à Neurospin, qui l'ont conduit à étendre les techniques présentées dans ce manuscrit pour modéliser explicitement l'incertitude de localisation spatiale (publication en préparation).

D'autres résultats obtenus par Anne-Laure Fouque (travail original non publié pour le moment) montrent qu'il est possible d'estimer une hémodynamique régionale selon des critères d'homogénéité fonctionnelle, permettant de rendre le modèle de régression pour l'analyse intra-sujet plus précis et ainsi de réduire la variance au premier niveau. Une des conséquences observées pour l'analyse de groupe « standard » est l'obtention d'une meilleure sensibilité de détection ; de plus ces résultats ont été obtenus sur des données non lissées spatialement ce qui permet de

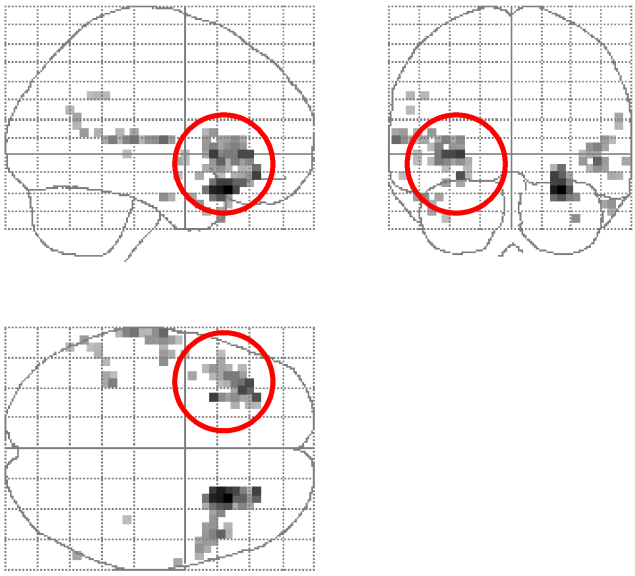

Figure 9.2 – *Carte de corrélation entre l'effet estimé et la variance résiduelle – Contraste « Activation à la parole » – Jeu de données* ***J*** *décrit en annexe A*

réconcilier les efforts des RMNistes en termes de résolution spatiale et la volonté des cognitivistes d'obtenir des résultats les plus précis possibles (à l'échelle d'une macro-colonne de neurones).

Annexe A

Description des jeux de données collectés

Nous présentons ici un court descriptif de la problématique cognitive, du paradigme expérimental et des éventuels résultats obtenus, pour chacune des études d'IRMf correspondant aux jeux de données collectés au cours de cette thèse et dont les caractéristiques sont résumées dans la table 6.1.

L'étude en cours correspondant au jeu de données **A** cherche à identifier les régions cérébrales impliquées spécifiquement dans le traitement syntaxique de la parole. Pour cela, les sujets scannés effectuent une tâche de reconnaissance de phrases partageant le même contenu syntaxique, le même contenu sémantique ou des contenus différents. Les premiers résultats n'ont permis pour l'instant que de détecter certaines régions, en particulier temporales antérieures, s'intéressant au contenu sémantique des phrases, dans la mesure où elles sont sensibles à la répétition du même contenu par des structures syntaxiques différentes.

L'étude correspondant au jeu de données **B** s'est intéressée à la plasticité cérébrale associée à l'acquisition d'une nouvelle langue. Huit sujets, de langue maternelle le coréen avant d'être adoptés en France pendant leur l'enfance, ont été comparés à huit sujets de langue maternelle le français dans une expérience d'écoute passive de phrases lues en français, en coréen, en japonais et en polonais. L'analyse des résultats a montré que les mêmes régions étaient activées pour les deux groupes de sujets lors de l'écoute de la langue parlée (le français) mais aussi des autres langues, indiquant une probable réversibilité de la plasticité cérébrale associée à l'acquisition d'une langue dans l'enfance [Pallier *et al.*, 2003].

Le jeu de données **C** correspond à une étude des processus cérébraux impliqués

dans une tâche de déduction logique. Pour cela, les seize sujets scannés devaient alternativement découvrir une combinaison de quatre boutons par essai-erreur et répéter une séquence connue de quatre boutons. La comparaison entre les tâches de recherche et de répétition a mis en évidence l'activation d'un large réseau cérébral impliquant des aires pariétales bilatérales, préfrontales, cingulaires et striatales [Landmann *et al.*, 2007].

Le jeu de données **D** correspond à une étude préliminaire d'IRM fonctionnelle sur le traitement subliminal de l'information et l'accès à la conscience, à l'aide d'une tâche de « clignement attentionnel » (*attentional blink*). Cette tâche consiste à présenter un premier stimulus, activement traité par le sujet, qui rend invisible, pendant une brève période de temps, la présentation d'un second stimulus. L'étude principale, utilisant les potentiels évoqués précoces (P1 et N1), a permis de révéler précisément la séquence d'activité cérébrale associée à l'accès à la conscience, impliquant un réseau distribué comprenant notamment les régions temporales latérales et antérieures, ainsi que de vastes secteurs du cortex pré-frontal et cingulaire [Sergent et Dehaene, 2004 ; Sergent *et al.*, 2005].

Le jeu de données **E** correspond à une étude comparative de l'activité cérébrale mesurée par IRM fonctionnelle dans le cortex cingulaire antérieur, en présence ou en absence de sillon paracingulaire, chez des sujets sains et des patients schizophrènes. Quinze sujets atteints de schizophrénie et seize sujets sains ont été scannés alors qu'ils effectuaient une tâche de comparaison de nombres avec présentation d'un distracteur. Cette tâche avait déjà été utilisée pour mettre en évidence une altération de l'activité cérébrale dans le cortex cingulaire antérieure chez les patients schizophrènes [Dehaene *et al.*, 2003]. Cette nouvelle étude [Artiges *et al.*, 2006] a permis de montrer que l'altération se produit principalement chez les patients schizophrènes sans sillon paracingulaire.

L'étude correspondant au jeu de données **F** s'est intéressée aux subdivisions fonctionnelles du cortex pariétal. Les données d'IRM fonctionnelle de neuf sujets ont été acquises alors qu'ils effectuaient six tâches différentes : une tâche de saisie d'objet (*grasping*), une tâche de pointage d'objet (*pointing*), une tâche de saccades visuelles (*saccades*), une tâche d'attention visuelle (*attention*), une tâche de soustraction mentale (*calculation*) et une tâche de détection de phonèmes (*language*). La comparaison des activations détectées lors de l'exécution de ces différentes

tâches a révélé une organisation structurée de l'activité cérébrale dans le cortex pariétal, principalement des zones antérieures vers les zones postérieures, ainsi que certaines analogies avec l'organisation observée chez le singe [Simon *et al.*, 2002].

Le jeu de données **G** correspond à une étude sur les aires du langage cherchant à distinguer les zones sensibles spécifiquement au contenu linguistique de la parole de celles sensibles spécifiquement au locuteur (contenu vocal). Les quinze sujets scannés ont été soumis à une tâche d'écoute passive de différentes phrases tirées d'une histoire pour enfant (« *Les Trois Petits Cochons* »), et lues par différentes personnes. Cette étude a mis en évidence un effet significatif du locuteur dans la région du sillon temporal supérieur droit alors que la région homologue située à gauche apparaît comme sensible au contenu linguistique (Dehaene-Lambertz *et al.* [2006a], voir les résultats de l'expérience 2).

Le jeu de données **H** correspond à une étude en cours sur la perception des émotions sollicitant un jugement moral. Le paradigme expérimental consiste à demander aux sujets scannés de se représenter mentalement les scènes décrites par écrit sur un écran. L'IRM fonctionnelle est alors utilisée pour tenter d'identifier les zones cérébrales impliquées dans le jugement de stimuli « moraux », c'est-à-dire la condamnation (ou la validation) de situations émotionnelles précises, mais aussi pour étudier si un lien existe entre les zones activées et le profil sociopsychologique du sujet.

Le jeu de données **I** correspond à une étude sur l'organisation fonctionnelle du réseau d'aires périsylviennes associées au langage. Le paradigme expérimental auquel ont été soumis les dix sujets scannés était le même que celui utilisé pour l'étude de Dehaene-Lambertz et Houston [1998] chez les enfants. Il consistait en l'écoute passive de phrases tirées d'une histoire pour enfant (« *Les Trois Petits Cochons* »), chaque phrase étant répétée de deux à quatre fois de manière à induire un effet de répétition-suppression. L'étude de la dynamique des réponses BOLD a révélé une hiérarchie temporelle dans les régions temporales supérieures, avec les réponses les plus rapides dans le cortex auditif primaire et les plus lentes dans les aires associatives, comme le gyrus temporal frontal gauche (Dehaene-Lambertz *et al.* [2006a], voir les résultats de l'expérience 1).

Le jeu de données **J** correspond à la même étude que celle du jeu de données **I** sur le réseau d'aires périsylviennes associées au langage, avec le même paradigme

expérimental [Dehaene-Lambertz et Houston, 1998], mais à la différence que les dix sujets scannés étaient des nourrissons de quelques mois. Une organisation fonctionnelle des régions temporales supérieures similaire à celle observée chez l'adulte a été mise en évidence, ainsi qu'une activation de l'aire de Broca impliquée dans la perception et la production du langage [Dehaene-Lambertz *et al.*, 2006b].

Le jeu de données **K** correspond à une étude sur les effets de la répétition lors de tests de mémorisation de visages. Les douze sujets scannés devaient réaliser deux tâches de reconnaissance de visages : pour la tâche « implicite », le sujet devait indiquer si le visage présenté était comme connu ou inconnu, et pour la tâche « explicite », il devait indiquer si le visage était présenté pour la première ou la deuxième fois. L'analyse des résultats a montré que seule la tâche « implicite » induit un effet de la répétition, situé dans la région fusiforme latérale droite lors de la présentation de visages connus et dans la région occipitale inférieure gauche quel que soit le visage présenté [Henson *et al.*, 2002]. Notons que les résultats présentés dans les chapitres 4 et 5 (figures 4.1, 4.2, 4.3, 4.4, 4.7 et 5.2) ont été obtenus en utilisant les cartes d'effet estimé pour la tâche « implicite ».

L'étude correspondant au jeu de données **L** s'est intéressée à l'activité cérébrale mesurée dans le cortex visuel primaire (aire V1) lors de la production d'images mentales. Pour chaque essai du paradigme expérimental, les dix sujets scannés devaient visualiser mentalement un animal, puis déterminer si une caractéristique concrète ou abstraite lui correspondait ou non. L'étude a montré que la production d'images mentales provoque une activité transitoire dans le sillon calcarin, même lorsque la tâche à effectuer ne nécessite pas de visualiser de nombreux détails [Klein *et al.*, 2000].

Le jeu de données **M** correspond à une étude sur les fluctuations de l'activité cérébrale spontanée au repos. Pour cela, l'activité cérébrale de quinze sujets a été mesurée par électroencéphalographie (EEG) couplée à l'IRM fonctionnelle, avec pour seule instruction de rester allongé les yeux fermés sans s'endormir. L'analyse conjointe des activations détectées en IRM et des signaux EEG a révélé l'activation d'un réseau distribué d'aires cérébrales pendant le repos, chaque composante de ce réseau étant corrélé à une bande de puissance spécifique de l'EEG. Ainsi, l'activité mesurée dans les cortex pariétal et fronto-latéral semble correspondre à une activité de base du cerveau (corrélation négative avec la bande alpha de l'EEG) alors

que celle détectée dans les cortex rétrosplénial, temporo-pariétal et préfrontal dorsomédial semble indiquer l'exécution d'opérations cognitives spontanées pendant l'état de repos (corrélation positive avec la bande beta de l'EEG) [Laufs *et al.*, 2003].

Le jeu de données **N** correspond à une étude en cours de localisation fonctionnelle de certains processus cognitifs et sensori-moteurs, impliquant déjà plus d'une centaine de sujets. Le paradigme expérimental est une tâche de « *Localizer* » comprenant une dizaine de conditions, comme l'écoute ou la lecture de phrases, le calcul d'une soustraction, des clics droit ou gauche de bouton, la vision passive de damiers. L'objectif de cette étude est de pouvoir quantifier la variabilité fonctionnelle entre les sujets et de disposer d'une large base de données pour tester les méthodes d'inférence statistique sur le groupe (voir notamment Thirion *et al.* [2005, 2007]).

L'étude correspondant au jeu de données **O** s'est intéressée aux processus de production syntaxique chez des sujets bilingues. Pour chaque essai du paradigme expérimental, une série de trois à cinq mots français ou anglais était présentée et les douze sujets scannés devaient soit les lire (condition de base), soit les utiliser pour former une phrase (condition expérimentale). L'étude a montré que la tâche de production de phrases active des régions incluant le gyrus frontal inférieur gauche ainsi que l'aire motrice supplémentaire. De plus, l'analyse des activations détectées dans le gyrus frontal inférieur gauche a révélé une plus forte activité lorsque les mots présentés étaient en anglais par rapport aux mots français (langue maternelle des sujets scannés), suggérant qu'un plus grand effort cognitif est nécessaire pour la production syntaxique dans une langue acquise plus tardivement [Golestani *et al.*, 2006].

Le jeu de données **P** correspond à une étude sur les bases représentationnelles et neurales sous-jacentes à l'acte de désignation des parties du corps. Les stimuli utilisés dans le paradigme expérimental consistaient en des représentations de complexité croissante (dessin, photographie ou vidéo) soit de corps humains, soit de chiens. Les quinze sujets participant à l'étude devaient pointer, pour chaque stimulus présenté, une partie spécifique du corps à l'aide d'un curseur. Les résultats obtenus ont montré que l'analyse visuelle d'un corps humain sur une représentation dynamique (vidéo) s'accompagne d'une activation de la partie médiane du

gyrus occipital moyen bilatéralement, des gyri temporaux supérieur et moyen à droite, mais également du lobule pariétal inférieur à gauche qui semble impliqué dans l'analyse du corps d'autrui [Felician *et al.*, 2005].

Le jeu de données **Q** correspond à une étude préliminaire d'habituation linguistique. Les dix sujets scannés ont été soumis à une tâche d'écoute passive de phrases répétées quatre fois toutes les 14 secondes. L'analyse des résultats a montré que parmi le réseau d'aires périsylviennes associées au langage, certaines régions situées le long du sillon temporal supérieur sont sensibles à la répétition des phrases. De plus, un effet d'adaptation a été observé : l'activation diminue à la seconde présentation d'une phrase. D'autres études ont ensuite complété les résultats obtenus en cherchant à distinguer les régions sensibles au contenu syntaxique, au contenu sémantique ou au contenu vocal (voir notamment les études correspondant aux jeux de données **A** et **G**).

Le jeu de données **R** correspond à une étude sur les relations entre les processus de perception acoustique et phonologique. Pour cela, les huit sujets scannés ont été soumis à deux tâches : l'une consistait à vérifier le contenu sémantique de phrases présentées avec des taux variables de compression temporelle, et l'autre consistait à comparer le contenu phonologique de mots présentés par paire. La comparaison des activations détectées pour les deux tâches a permis de mettre en évidence un sous-ensemble de régions frontales inférieures gauches impliquées dans la perception phonologique de la parole mais également sensibles à son contenu acoustique [Poldrack *et al.*, 2001].

L'étude correspondant au jeu de données **S** s'est intéressée à l'effet placebo dans l'anticipation et la perception de la douleur. Pour cela, le paradigme expérimental consistait à provoquer une sensation de brûlure (ou un choc) sur l'avant-bras gauche (ou droit) des vingt-quatre sujets scannés. Les sujets étaient informés que la zone touchée étaient préalablement traitée avec une crème anesthésiante (condition placebo) ou une crème inerte (condition contrôle), alors qu'en réalité la crème inerte était appliquée dans tous les cas. L'étude a montré que l'effet placebo altère la perception de la douleur : il provoque à la fois une diminution de l'activité cérébrale dans les régions sensibles à la douleur, comme le thalamus, l'insula et le cortex cingulaire antérieur, mais également une augmentation de l'activité cérébrale dans le cortex préfrontal lors de l'anticipation de la douleur [Wager *et al.*,

2004].

Le jeu de données **T** correspond à une étude préliminaire sur la représentation des nombres. Les onze sujets ont été scannés alors qu'ils lisaient de courtes histoires contenant un ou plusieurs nombres, avec lesquels ils devaient effectuer différentes opérations (comparaison, multiplication,...). L'analyse des résultats a révélé l'activation du sulcus intrapariétal et de certaines aires des cortex précentral et préfrontal inférieur, indiquant leur possible rôle dans la représentation quantitative des nombres et le calcul mental. Cette étude s'est poursuivie en modifiant le paradigme expérimental et les résultats publiés ont confirmé certaines des observations faites lors de l'étude préliminaire [Piazza *et al.*, 2004].

Annexe B

EM algorithm for a Gaussian population

In this case, $\mathcal{F}$ is the Gaussian family so that the random effect's PDF f is parameterized by its mean μ_{f} and standard deviation σ_{f}. The algorithm iteratively refines initial estimates $\hat{\mu}_{\mathrm{f}}$ and $\hat{\sigma}_{\mathrm{f}}$ by alternating two steps, the E-step (expectation) and the M-step (maximization). In our implementation, the initial estimates are respectively taken as the classical sample mean and sample standard deviation of the observed effects $(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_n)$.

E-step. Assume current estimates are exact, and compute the posterior joint PDF of all subject's true effects. Since the subjects are conditionally independent, this reduces to computing each subject's posterior, $p(\beta_i|\hat{\beta}_i, \sigma_i, \mu_{\mathrm{f}}, \sigma_{\mathrm{f}})$ which is a Gaussian with parameters (m_i, s_i) :

$$m_i = \frac{\hat{\sigma}_{\mathrm{f}}^2}{\sigma_i^2 + \hat{\sigma}_{\mathrm{f}}^2}\hat{\beta}_i + \frac{\sigma_i^2}{\sigma_i^2 + \hat{\sigma}_{\mathrm{f}}^2}\hat{\mu}_{\mathrm{f}}, \qquad s_i = \frac{\sigma_i \hat{\sigma}_{\mathrm{f}}}{\sqrt{\sigma_i^2 + \hat{\sigma}_{\mathrm{f}}^2}}$$

M-step. Update $(\hat{\mu}_{\mathrm{f}}, \hat{\sigma}_{\mathrm{f}})$ by maximizing the expected log-likelihood of the complete data :

$$\mathcal{Q}(\mu_{\mathrm{f}}, \sigma_{\mathrm{f}}) = n \log \sqrt{2\pi}\sigma_{\mathrm{f}} + \frac{1}{2\sigma_{\mathrm{f}}^2} \sum_i \left[s_i^2 + (\mu_{\mathrm{f}} - m_i)^2 \right],$$

yielding :

$$\hat{\mu}_{\mathrm{f}} = \frac{1}{n} \sum_i m_i, \qquad \hat{\sigma}_{\mathrm{f}}^2 = \frac{1}{n} \sum_i [s_i^2 + (\hat{\mu}_{\mathrm{f}} - m_i)^2]$$

In the constrained problem of maximizing likelihood subject to $\mu_{\mathrm{f}} = 0$, the algorithm is identical except that $\hat{\mu}_{\mathrm{f}}$ is frozen to zero in the M-step. In practice, we perform five EM iterations, which proves generally sufficient to achieve good precision on the ensuing test statistics.

Annexe C

EM algorithm for a general population

In this case, $\mathcal{F}$ is the space of all PDFs on $\mathbb{R}$. It is then proved in Lindsay [1983] that the maximum likelihood PDF is necessarily a mixture of n or less Dirac masses :

$$f(\beta) = \sum_{k=1}^{n} w_k \delta(\beta - z_k),$$

which amounts to saying that each observation $\hat{\beta}_i$ is drawn from an unobserved class $k(i)$. In practice, we initialize the EM algorithm with uniform mixing proportions and support points coinciding with the observations (corresponding to the maximum likelihood PDF under exact observations).

E-step. Given the PDF parameters, we compute the posterior probability q_{ik} of class label k for subject i, yielding :

$$q_{ik} = \frac{\hat{w}_k g_i(\hat{\beta}_i - \hat{z}_k)}{\sum_{k'} \hat{w}_{k'} g_i(\hat{\beta}_i - \hat{z}_{k'})}$$

M-step. Given the posterior probabilities q_{ik}, we form the negated expected complete-data log-likelihood :

$$\mathcal{Q}(w, z) = \sum_{i,k} q_{ik} \Big[\log \sqrt{2\pi}\sigma_i + \frac{(\hat{\beta}_i - z_k)^2}{2\sigma_i^2} - \log w_k \Big]$$

This criterion is to be minimized subject to the constraint that the mixing proportions sum up to one, and possibly that the mean population effect vanishes. We therefore consider the Lagrangian :

$$\mathcal{L}(w, z, \lambda_0, \lambda_1) = \mathcal{Q}(w, z) + \lambda_0(\sum_k w_k - 1) + \lambda_1(\sum_k w_k z_k),$$

whose derivatives read :

$$\frac{\partial \mathcal{L}}{\partial w_k} = -\frac{1}{w_k}\sum_i q_{ik} + \lambda_0 + \lambda_1 z_k, \qquad \frac{\partial \mathcal{L}}{\partial z_k} = \sum_i \frac{q_{ik}}{\sigma_i^2}(z_k - \hat{\beta}_i) + \lambda_1 w_k$$

When no mean constraint is applied, so that $\lambda_1 = 0$, the M-step yields an explicit updating rule. We easily get $\lambda_0 = n$, then :

$$\hat{w}_k = \frac{1}{n}\sum_i q_{ik}, \qquad \hat{z}_k = \frac{1}{S_k}\sum_i \frac{q_{ik}}{\sigma_i^2}\hat{\beta}_i \qquad \text{with} \quad S_k = \sum_i \frac{q_{ik}}{\sigma_i^2}$$

Constrained M-step. When maximizing likelihood subject to the zero mean constraint, the Lagrange multiplier λ_1 becomes a free parameter. In this case, there is no explicit solution to the M-step. A possibility is to recast the joint constrained minimization w.r.t. w and z as sequential constrained minimizations :

- Along w (at fixed z). We, again, find that $\lambda_0 = n$, and :

$$\hat{w}_k = \frac{\frac{1}{n}\sum_i q_{ik}}{1 + \lambda_1 z_k}$$

  We then solve for λ_1 by writing the zero-mean constraint, leading to a weighted version of the standard empirical likelihood equation [Owen, 2001], whose solution generally exists uniquely, and can be found using a Newton algorithm. The only exception is when all the support points have the same sign, in which case no solution exists and we simply leave the mixing proportions unchanged until the next iteration.
- Along z (at fixed w). We easily get the following implicit equation :

$$\hat{z}_k = \frac{1}{S_k}\Big(\sum_i \frac{q_{ik}}{\sigma_i^2}\hat{\beta}_i - \lambda_1 w_k\Big),$$

(with S_k like in the unconstrained case) which becomes explicit after expressing the zero-mean constraint and solving the resulting linear equation for λ_1.

In practice, we do not iterate the alternate minimization, meaning that, in the constrained case, our algorithm is actually an EM variant known as the expectation conditional maximization (ECM) algorithm [Meng et Rubin, 1993], which maintains the essential property that the likelihood value increases on each iteration.

Bibliographie

[Aguirre *et al.*, 1997] G. K. Aguirre, E. Zarahn et M. D'Esposito. Empirical analyses of BOLD fMRI statistics. II Spatially smoothed data collected under null-hypothesis and experimental conditions. *NeuroImage*, 5 (3) : 199–212, avril 1997.

[Almeida et Ledberg, 2001] R. Almeida et A. Ledberg. A spatially constrained clustering algorithm with no prior knowledge of the number of clusters. In *Proc. 7th HBM*, Brighton, Royaume-Uni, 10-14 juin 2001.

[Amunts *et al.*, 2000] K. Amunts, L. Jäncke, H. Mohlberg, H. Steinmetz et K. Zilles. Interhemispheric asymmetry of the human motor cortex related to handedness and gender. *Neuropsychologia*, 38 (3) : 304–312, mars 2000.

[Anderson et Darling, 1952] T. W. Anderson et D. A. Darling. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann Math Stat*, 23 (2) : 193–212, juin 1952.

[Andersson *et al.*, 1999] A. H. Andersson, D. M. Grash et M. J. Avison. Principal component analysis of the dynamic response measured by fMRI : a generalized linear systems framework. *Magn Reson Imaging*, 17 (6) : 795–815, juillet 1999.

[Andersson *et al.*, 2001] J. L. R. Andersson, C. Hutton, J. Ashburner, R. Turner et K. J. Friston. Modeling geometric deformations in EPI time series. *NeuroImage*, 13 (5) : 903–919, mai 2001.

[Andrade, 2001] A. Andrade. *Surface Brain Mapping.* Thèse de doctorat, Université de Lisbonne, Portugal, septembre 2001.

[Andrade *et al.*, 2001] A. Andrade, F. Kherif, J.-F. Mangin, K. J. Worsley, A.-L. Paradis, O. Simon, S. Dehaene, D. LeBihan et J.-B. Poline. Detection of fMRI activation using cortical surface mapping. *Hum Brain Mapp*, 12 (2) : 79–93, février 2001.

[Arbuthnott, 1710] J. Arbuthnott. An argument for divine providence, taken from the constant regularity observed in the births of both sexes. *Philosophical Transactions of the Royal Society of London*, 27 : 186–190, 1710.

[Artiges *et al.*, 2006] E. Artiges, C. Martelli, L. Naccache, D. Bartrés-Faz, J.-B. Leprovost, A. Viard, M.-L. Paillère-Martinot, S. Dehaene et J.-L. Martinot. Paracingulate sulcus morphology and fMRI activation detection in schizophrenia patients. *Schizophr Res*, 82 (2-3) : 143–151, février 2006.

[Ashburner et Friston, 1999] J. Ashburner et K. J. Friston. Nonlinear spatial normalization using basis functions. *Hum Brain Mapp*, 7 (4) : 254–266, juin 1999.

[Ashburner et Friston, 2005] J. Ashburner et K. J. Friston. Unified segmentation. *NeuroImage*, 26 (3) : 839–851, juillet 2005.

[Ashburner *et al.*, 1997] J. Ashburner, P. Neelin, D. L. Collins, A. C. Evans et K. J. Friston. Incorporating prior knowledge into image registration. *NeuroImage*, 6 (4) : 344–352, novembre 1997.

[Baillet *et al.*, 2001] S. Baillet, J. J. Riera, G. Marin, J.-F. Mangin, J. Aubert et L. Garnero. Evaluation of inverse methods and head models for EEG source localization using a human skull phantom. *Phys Med Biol*, 46 (1) : 77–96, janvier 2001.

[Bandettini *et al.*, 1997] P. A. Bandettini, K. K. Kwong, T. L. Davis, R. B. H. Tootell, E. C. Wong, P. T. Fox, J. W. Belliveau, R. M. Weisskoff et B. R. Rosen. Characterization of cerebral blood oxygenation and flow changes during prolonged brain activation. *Hum Brain Mapp*, 5 (2) : 93–109, février 1997.

[Bandettini *et al.*, 1992] P. A. Bandettini, E. C. Wong, R. S. Hinks, R. S. Tikofsky et J. S. Hyde. Time course EPI of human brain function during task activation. *Magn Reson Med*, 25 (2) : 390–397, juin 1992.

[Barnett et Lewis, 1994] V. Barnett et T. Lewis. *Outliers in statistical data.* Wiley & Sons, New York, USA, 3ème edition, 1994.

[Beckmann *et al.*, 2003] C. F. Beckmann, M. Jenkinson et S. M. Smith. General multilevel linear modeling for group analysis in fMRI. *NeuroImage*, 20 (2) : 1052–1063, octobre 2003.

[Beckmann et Smith, 2004] C. F. Beckmann et S. M. Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans Med Imag*, 23 (2) : 137–152, février 2004.

[Belin et Zatorre, 2003] P. Belin et R. J. Zatorre. Adaptation to speaker's voice in right anterior temporal lobe. *NeuroReport*, 14 (16) : 2105–2109, novembre 2003.

[Belin *et al.*, 2000] P. Belin, R. J. Zatorre, P. Lafaille, P. Ahad et B. Pike. Voice-selective areas in human auditory cortex. *Nature*, 403 (6767) : 309–312, janvier 2000.

[Belsley *et al.*, 1980] D. A. Belsley, E. Kuh et R. E. Welsch. *Regression Diagnostics : identifying influential data and sources of collinearity.* Wiley, New York, USA, 1980.

[Bollen et Jackman, 1990] K. Bollen et R. Jackman. Regression diagnostics : an expository treatment of outliers and influential cases. In J. Fox et J. S. Long, éditeurs, *Modern Methods of Data Analysis*, pages 257–291. Sage Publications, Newbury Park, USA, 1990.

[Brammer *et al.*, 1997] M. J. Brammer, E. T. Bullmore, A. Simmons, S. C. R. Williams, P. M. Grasby, R. J. Howard, P. W. R. Woodruff et S. Rabe-Hesketh. Generic brain activation mapping in functional magnetic resonance imaging : a nonparametric approach. *Magn Reson Imaging*, 15 (7) : 763–770, mai 1997.

[Brett *et al.*, 2002] M. Brett, I. S. Johnsrude et A. M. Owen. The problem of functional localization in the human brain. *Nat Rev Neurosci*, 3 (3) : 243–249, mars 2002.

[Bullmore *et al.*, 1996] E. T. Bullmore, M. J. Brammer, S. C. R. Williams, S. Rabe-Hesketh, N. Janot, A. S. David, J. D. C. Mellers, R. Howard et P. Sham. Statistical methods of estimation and inference for functional MR image analysis. *Magn Reson Med*, 35 (2) : 261–277, février 1996.

[Bullmore *et al.*, 2001] E. T. Bullmore, C. Long, J. Suckling, J. Fadili, G. Calvert, F. Zelaya, T. A. Carpenter et M. J. Brammer. Colored noise and computational inference in neurophysiological (fMRI) time series analysis : resampling methods in time and wavelet domains. *Hum Brain Mapp*, 12 (2) : 61–78, février 2001.

[Bullmore *et al.*, 1999] E. T. Bullmore, J. Suckling, S. Overmeyer, S. Rabe-Hesketh, E. Taylor et M. J. Brammer. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Trans Med Imaging*, 18 (1) : 32–42, janvier 1999.

[Burock *et al.*, 1998] M. A. Burock, R. L. Buckner, M. G. Woldorff, B. R. Rosen et A. M. Dale. Randomized event-related experimental designs allow for extremly rapid presentation rates using functional MRI. *Neuroreport*, 9 (16) : 3735–3739, novembre 1998.

[Cachier *et al.*, 2001] P. Cachier, J.-F. Mangin, X. Pennec, D. Rivière, D. Papadopoulos-Orfanos, J. Régis et N. Ayache. Multisubject non-rigid registration of brain MRI using intensity and geometric features. In *Proc. 4th MICCAI*, pages 734–742, Utrecht, Pays-Bas, 14-17 octobre 2001.

[Calhoun *et al.*, 2001a] V. D. Calhoun, T. Adali, G. D. Pearlson et J. J. Pekar. A method for making group inferences from functional MRI data using independent component analysis. *Hum Brain Mapp*, 14 (3) : 140–151, novembre 2001.

[Calhoun *et al.*, 2001b] V. D. Calhoun, T. Adali, G. D. Pearlson et J. J. Pekar. Spatial and temporal independent component analysis of functional MRI data

containing a pair of task-related waveforms. *Hum Brain Mapp*, 13 (1) : 43–53, mai 2001.

[Carroll, 1968] J. D. Carroll. Generalisation of canonical analysis to three or more sets of variables. *Proc. 76th Amer. Psych. Assoc.*, 3 : 227–228, 1968.

[Chatfield et Collins, 1992] C. Chatfield et A. J. Collins. *Introduction to multivariate analysis.*, pages 198–200. Chapman and Hall, Londres, Royaume-Uni, 1992.

[Cochran, 1950] W. G. Cochran. The comparison of percentages in matched samples. *Biometrika*, 37 (3-4) : 256–266, décembre 1950.

[Cook, 1977] R. D. Cook. Detection of influential observation in linear regression. *Technometrics*, 19 (1) : 15–18, février 1977.

[Cook et Weisberg, 1982] R. D. Cook et S. Weisberg. *Residuals and influence in regression.*, page 116. Chapman and Hall, Londres, Royaume-Uni, 1982.

[Corouge *et al.*, 2003] I. Corouge, P. Hellier, B. Gibaud et C. Barillot. Interindividual functional mapping : a nonlinear local approach. *NeuroImage*, 19 (4) : 1337–1348, août 2003.

[Dehaene *et al.*, 2003] S. Dehaene, E. Artiges, L. Naccache, C. Martelli, A. Viard, F. Schürhoff, C. Recasens, M.-L. Paillre-Martinot, M. Leboyer et J.-L. Martinot. Conscious and subliminal conflicts in normal subjects and patients with schizophrenia : the role of the anterior cingulate. *Proc Natl Acad Sci*, 100 (23) : 13722–13727, novembre 2003.

[Dehaene-Lambertz *et al.*, 2006a] G. Dehaene-Lambertz, S. Dehaene, J.-L. Anton, A. Campagne, P. Ciuciu, G. Dehaene, I. Denghien, A. Jobert, D. LeBihan, M. Sigman, C. Pallier et J.-B. Poline. Functional segregation of cortical language areas by sentence repetition. *Hum Brain Mapp*, 27 (5) : 360–371, mai 2006.

[Dehaene-Lambertz *et al.*, 2002] G. Dehaene-Lambertz, S. Dehaene et L. Hertz-Pannier. Functional neuroimaging of speech perception in infants. *Science*, 298 (5600) : 2013–2015, décembre 2002.

[Dehaene-Lambertz *et al.*, 2006b] G. Dehaene-Lambertz, L. Hertz-Pannier, J. Dubois, S. Mériaux, A. Roche, M. Sigman et S. Dehaene. Functional organization of perisylvian activation during presentation of sentences in preverbal infants. *Proc Natl Acad Sci*, 103 (38) : 14240–14245, septembre 2006.

[Dehaene-Lambertz et Houston, 1998] G. Dehaene-Lambertz et D. Houston. Faster orientation latency toward native language in two-month-old infants. *Lang Speech*, 41 (1) : 21–43, 1998.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird et D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B (Methodological)*, 39 (1) : 1–38, 1977.

[Desmond et Glover, 2002] J. E. Desmond et G. H. Glover. Estimating sample size in functional MRI (fMRI) neuroimaging studies : statistical power analyses. *J Neurosci Methods*, 118 (2) : 115–128, août 2002.

[D'Esposito *et al.*, 1999] M. D'Esposito, E. Zarahn et G. K. Aguirre. Event-related functional MRI : implications for cognitive psychology. *Psychol Bull*, 125 (1) : 155–164, janvier 1999.

[Dixon, 1950] W. J. Dixon. Analysis of extreme values. *Ann Math Stat*, 21 (4) : 488–506, décembre 1950.

[Escoufier, 1973] Y. Escoufier. Le traitement des variables vectorielles. *Biometrics*, 29 (4) : 751–760, décembre 1973.

[Esposito *et al.*, 2005] F. Esposito, T. Scarabino, A. Hyvarinen, J. Himberg, E. Formisano, S. Comani, G. Tedeschi, R. Goebel, E. Seifritz et F. D. Salle. Independent component analysis of fMRI group studies by self-organizing clustering. *NeuroImage*, 25 (1) : 193–205, mars 2005.

[Felician *et al.*, 2005] O. Felician, P. Romaiguère, J.-L. Anton, S. Mériaux, A. Roche, B. Nazarian, M. Roth et J.-P. Roll. Neural correlates of seeing someone else's body. In *IX International Conference on Cognitive Neuroscience*, La Havane, Cuba, 5-10 septembre 2005.

[Fiebach *et al.*, 2002] C. J. Fiebach, A. D. Friederici, K. Müller et D. Y. von Cramon. fMRI evidence for dual routes to the mental lexicon in visual word recognition. *J Cogn Neurosci*, 14 (1) : 11–23, janvier 2002.

[Fischl *et al.*, 1999a] B. Fischl, M. I. Sereno et A. M. Dale. Cortical surface-based analysis. II : Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9 (2) : 195–207, février 1999.

[Fischl *et al.*, 1999b] B. Fischl, M. I. Sereno, R. B. H. Tootell et A. M. Dale. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum Brain Mapp*, 8 (4) : 272–284, novembre 1999.

[Fisher, 1925] R. A. Fisher. *Statistical methods for research workers.* Oliver & Boyd, Édimbourg, Royaume-Uni, 1925.

[Fisher, 1935] R. A. Fisher. *The design of experiment.* Oliver & Boyd, Édimbourg, Royaume-Uni, 1935.

[Flandin *et al.*, 2002a] G. Flandin, F. Kherif, X. Pennec, G. Malandain, N. Ayache et J.-B. Poline. Improved detection sensitivity of functional MRI data using a brain parcellation technique. In *Proc. 5th MICCAI*, Tokyo, Japon, 25-28 septembre 2002.

[Flandin *et al.*, 2002b] G. Flandin, F. Kherif, X. Pennec, D. Rivière, N. Ayache et J.-B. Poline. Parcellation of brain images with anatomical and functional constraints for fMRI data analysis. In *Proc. 1st ISBI*, pages 907–910, Washington, USA, 7-10 july 2002.

[Fox, 1991] J. Fox. Regression diagnostics : provides a thorough review of methods of testing the assumptions of regression models. In M. S. Lewis-Beck, éditeur, *Quantitative Applications in the Social Sciences, Series No. 79*, page 34. Sage Publications, Thousand Oaks, USA, 1991.

[Fox *et al.*, 1999] P. T. Fox, A. Y. Huang, L. M. Parsons, J. H. Xiong, L. Rainey et J. L. Lancaster. Functional volumes modeling : scaling for group size in averaged images. *Hum Brain Mapp*, 8 (2-3) : 143–150, juillet 1999.

[Fox *et al.*, 1988] P. T. Fox, M. E. Raichle, M. A. Mintun et C. Dence. Nonoxidative glucose consumption during focal physiologic neural activity. *Science*, 241 (4864) : 462–464, juillet 1988.

[Frackowiak *et al.*, 1997] R. S. J. Frackowiak, K. J. Friston, C. Frith, R. Dolan et J. C. Mazziotta, éditeurs. *Human Brain Function*. Academic Press, USA, 1997.

[Freire et Mangin, 2001] L. Freire et J.-F. Mangin. Motion correction algorithms may create spurious brain activations in the absence of subject motion. *NeuroImage*, 14 (3) : 709–722, septembre 2001.

[Freire *et al.*, 2002] L. Freire, A. Roche et J.-F. Mangin. What is the best similarity measure for motion correction in fMRI time series ? *IEEE Trans Med Imag*, 21 (5) : 470–484, mai 2002.

[Friman *et al.*, 2001] O. Friman, J. Carlsson, P. Lundberg, M. Borga et H. Knutsson. Detection of neural activity in functional MRI using canonical correlation analysis. *Magn Reson Med*, 45 (2) : 323–330, février 2001.

[Friston *et al.*, 1995a] K. J. Friston, J. Ashburner, C. D. Frith, J.-B. Poline, J. D. Heather et R. S. J. Frackowiak. Spatial registration and normalisation of images. *Hum Brain Mapp*, 3 (3) : 165–189, février 1995.

[Friston *et al.*, 1993] K. J. Friston, C. D. Frith, P. F. Liddle et R. S. J. Frackowiak. Functional connectivity : the Principal-Component Analysis of large PET data sets. *J Cereb Blood Flow Metab*, 13 (1) : 5–14, janvier 1993.

[Friston *et al.*, 1995b] K. J. Friston, A. P. Holmes, J.-B. Poline, P. J. Grasby, S. C. R. Williams, R. S. J. Frackowiak et R. Turner. Analysis of fMRI time-series revisited. *NeuroImage*, 2 (1) : 45–53, mars 1995.

[Friston *et al.*, 1999] K. J. Friston, A. P. Holmes et K. J. Worsley. How many subjects constitute a study ? *NeuroImage*, 10 (1) : 1–5, juillet 1999.

[Friston *et al.*, 1995c] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-B. Poline, C. D. Frith et R. S. J. Frackowiak. Statistical parametric maps in functional

imaging : a general linear approach. *Hum Brain Mapp*, 2 (4) : 189–210, avril 1995.

[Friston *et al.*, 2000] K. J. Friston, A. Mechelli, R. Turner et C. J. Price. Nonlinear responses in fMRI : the Balloon model, Volterra kernels, and other hemodynamics. *NeuroImage*, 12 (4) : 466–477, octobre 2000.

[Friston *et al.*, 2002] K. J. Friston, W. Penny, C. Phillips, S. J. Kiebel, G. Hinton et J. Ashburner. Classical and bayesian inference in neuroimaging : theory. *NeuroImage*, 16 ((2)) : 465–483, juin 2002.

[Friston *et al.*, 2005] K. J. Friston, K. E. Stephan, T. E. Lund, A. Morcom et S. J. Kiebel. Mixed-effects and fMRI studies. *NeuroImage*, 24 (1) : 244–252, janvier 2005.

[Genovese *et al.*, 1997] C. R. Genovese, D. C. Noll et W. F. Eddy. Estimating test-retest reliability in functional MR imaging. I : Statistical methodology. *Magn Reson Med*, 38 (3) : 497–507, septembre 1997.

[Golestani *et al.*, 2006] N. Golestani, F. X. Alario, S. Mériaux, D. LeBihan, S. Dehaene et C. Pallier. Syntax production in bilinguals. *Neuropsychologia*, 44 (7) : 1029–1040, janvier 2006.

[Gower, 1984] J. C. Gower. Multidimensional scaling displays. In H. G. Law, C. W. Snyder, J. A. Hattie et R. P. McDonald, éditeurs, *Research methods for multimode data analysis*, pages 470–517. Praeger, New York, USA, 1984.

[Gray, 1993] J. B. Gray. Approximating the internal norm influence measure in linear regression. *Commun stat, Simul comput*, 22 (1) : 117–135, 1993.

[Grubbs, 1950] F. E. Grubbs. Sample criteria for testing outlying observations. *Ann Math Stat*, 21 (1) : 27–58, mars 1950.

[Grubbs, 1969] F. E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11 (1) : 1–21, février 1969.

[Hansen *et al.*, 1999] L. K. Hansen, J. Larsen, F. A. Nielsen, S. C. Strother, E. Rostrup, R. Savoy, N. Lange, J. J. Sidtis, C. Svarer et O. B. Paulson. Generalizable patterns in neuroimaging : how many principal components ? *NeuroImage*, 9 (5) : 534–544, mai 1999.

[Hayasaka et Nichols, 2003] S. Hayasaka et T. E. Nichols. Validating cluster size inference : random field and permutation methods. *NeuroImage*, 20 (4) : 2343–2356, décembre 2003.

[Henson *et al.*, 1999] R. N. A. Henson, C. Büchel, O. Josephs et K. J. Friston. The slice-timing problem in event-related fMRI. In *Proc. 5th HBM CD-Rom, NeuroImage vol.9(1)*, Düsseldorf, Allemagne, 22-26 juin 1999.

[Henson *et al.*, 2002] R. N. A. Henson, T. Shallice, M.-L. Gorno-Tempini et R. J. Dolan. Face repetition effects in implicit and explicit memory tests as measured by fMRI. *Cereb Cortex*, 12 (2) : 178–186, février 2002.

[Hollander et Wolfe, 1999] M. Hollander et D. A. Wolfe. *Nonparametric statistical methods.* Wiley & Sons, New York, USA, 2ème edition, 1999.

[Holmes *et al.*, 1996] A. P. Holmes, R. C. Blair, J. D. G. Watson et I. Ford. Nonparametric analysis of statistic images from functional mapping experiments. *J Cereb Blood Flow Metab*, 16 (1) : 7–22, janvier 1996.

[Hotelling, 1936] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28 (3/4) : 321–377, décembre 1936.

[Houdé *et al.*, 2002] O. Houdé, B. Mazoyer et N. Tzourio-Mazoyer. *Cerveau et psychologie : introduction à l'imagerie cérébrale anatomique et fonctionnelle.* Presses Universitaires de France, Paris, France, 2002.

[Jenkinson *et al.*, 2002] M. Jenkinson, P. R. Bannister, J. M. Brady et S. M. Smith. Improved optimisation for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17 (2) : 825–841, octobre 2002.

[Jensen et Ramirez, 1998] D. R. Jensen et D. E. Ramirez. Some exact properties of Cook's D_I. In N. Balakrishnan et C. R. Rao, éditeurs, *Handbook of Statistics*, volume 16, pages 387–402. Elsevier Science Publishers, Amsterdam, Pays-Bas, 1998.

[Jezzard et Balaban, 1995] P. Jezzard et R. S. Balaban. Correction for geometric distortion in echo planar images from B0 field variations. *Magn Reson Med*, 34 (1) : 65–73, juillet 1995.

[Jezzard et Clare, 1999] P. Jezzard et S. Clare. Sources of distortion in functional MRI data. *Hum Brain Mapp*, 8 (2-3) : 80–85, septembre 1999.

[Josephs et Henson, 1999] . Josephs et R. N. Henson. Event-related functional magnetic resonance imaging : modelling, inference and optimization. *Philosophical Transactions : Biological Sciences*, 354 (1387) : 1215–1228, juillet 1999.

[Kettenring, 1971] J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58 (3) : 433–451, décembre 1971.

[Kherif *et al.*, 2003] F. Kherif, M. Brett, S. Mériaux, H. Benali et J.-B. Poline. Inter-individual differences through a battery of tasks. In *Proc. 9th HBM CD-Rom, NeuroImage vol.19(2)*, New-York, USA, 18-22 juin 2003.

[Kherif *et al.*, 2002] F. Kherif, G. Flandin, P. Ciuciu, H. Benali, O. Simon et J.-B. Poline. Model based spatial and temporal similarity measures between series of functional magnetic resonance images. In *Proc. 5th MICCAI*, Tokyo, Japon, 25-28 septembre 2002.

[Kherif *et al.*, 2004] F. Kherif, J.-B. Poline, S. Mériaux, H. Benali, G. Flandin et M. Brett. Group analysis in functional neuroimaging : selecting subjects using similarity measures. *NeuroImage*, 20 (4) : 2197–2208, janvier 2004.

[Kim et Ugurbil, 1997] S.-G. Kim et K. Ugurbil. Comparison of blood oxygenation and cerebral blood flow effects in fMRI : estimation of relative oxygen consumption change. *Magn Reson Med*, 38 (1) : 59–65, juillet 1997.

[Klein *et al.*, 2000] I. Klein, A.-L. Paradis, J.-B. Poline, S. M. Kosslyn et D. LeBihan. Transient activity in the human calcarine cortex during visual-mental imagery : an event-related fMRI study. *J Cogn Neurosci*, 12 (Suppl 2) : 15–23, novembre 2000.

[Kruggel et von Cramon, 1999] F. Kruggel et D. Y. von Cramon. Temporal properties of the hemodynamic response in functional MRI. *Hum Brain Mapp*, 8 (4) : 259–271, novembre 1999.

[Landmann *et al.*, 2007] C. Landmann, S. Dehaene, S. Pappata, A. Jobert, M. Bottlaender, D. Roumenov et D. LeBihan. Dynamics of prefrontal and cingulate activity during a reward-based logical deduction task. *Cereb Cortex*, 17 (4) : 749–759, avril 2007.

[Laufs *et al.*, 2003] H. Laufs, K. Krakow, P. Sterzer, E. Eger, A. Beyerle, A. Salek-Haddadi et A. Kleinschmidt. Electroencephalographic signatures of attentional and cognitive default modes in spontaneous brain activity fluctuations at rest. *Proc Natl Acad Sci*, 100 (19) : 11053–11058, septembre 2003.

[Lauterbur, 1973] P. C. Lauterbur. Image formation by induced local interactions : examples employing nuclear magnetic resonance. *Nature*, 242 (5394) : 190–191, mars 1973.

[Lazar *et al.*, 2002] N. A. Lazar, B. Luna, J. A. Sweeney et W. F. Eddy. Combining brains : a survey of methods for statistical pooling of information. *NeuroImage*, 16 (4) : 538–550, décembre 2002.

[Lazraq *et al.*, 1992] A. Lazraq, R. Cléroux et H. A. L. Kiers. Mesures de liaison vectorielle et généralisation de l'analyse canonique. *Revue de Statistique Appliquée*, 40 (1) : 23–35, 1992.

[Lehmann, 1986] E. Lehmann. *Testing statistical hypotheses.* Springer, New York, USA, 1986.

[Lilliefors, 1967] H. W. Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *J Am Stat Ass*, 62 (318) : 399–402, juin 1967.

[Lindsay, 1983] B. G. Lindsay. The geometry of mixture likelihoods : a general theory. *The Annals of Statistics*, 11 (1) : 86–94, mars 1983.

[Lingoes et Schönemann, 1974] J. C. Lingoes et P. H. Schönemann. Alternative measures of fit for Schönemann-Caroll matrix fitting algorithm. *Psychometrika*, 39 (4) : 423–427, décembre 1974.

[Mahalanobis, 1936] P. C. Mahalanobis. On the generalized distance in statistics. In *Proc. Natl. Inst. Sci. India*, pages 49–55, 1936.

[Maitra *et al.*, 2002] R. Maitra, S. R. Roys et R. P. Gullapalli. Test-retest reliability estimation of functional MRI data. *Magn Reson Med*, 48 (1) : 62–70, juillet 2002.

[Mangin *et al.*, 2004] J.-F. Mangin, D. Rivière, O. Coulon, C. Poupon, A. Cachia, Y. Cointepas, J.-B. Poline, D. LeBihan, J. Régis et D. Papadopoulos-Orfanos. Coordinate-based versus structural approaches to brain image analysis. *Artif Intell Med*, 30 (2) : 177–197, février 2004.

[Mansfield, 1977] P. Mansfield. Multi-planar image formation using NMR spin echoes. *J Phys C : Solid State Phys*, 10 (3) : L55–L58, 1977.

[Mazziotta *et al.*, 1995] J. C. Mazziotta, A. W. Toga, A. C. Evans, P. Fox et J. Lancaster. A probabilistic atlas of the humain brain : theory and rationale for its development. the International Consortium for Brain Mapping (ICBM). *NeuroImage*, 2 (2) : 89–101, juin 1995.

[McIntosh *et al.*, 1996] A. R. McIntosh, F. L. Bookstein, J. V. Haxby et C. L. Grady. Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage*, 3 (3 Pt 1) : 143–157, juin 1996.

[McKeown, 2000] M. J. McKeown. Detection of consitently task-related activation in fMRI data with hybrid independent component analysis. *NeuroImag*, 11 (1) : 24–35, janvier 2000.

[McKeown *et al.*, 1998] M. J. McKeown, S. Makeig, G. G. Brown, T.-P. Jung, S. S. Kindermann, A. J. Bell et T. J. Sejnowski. Analysis of fMRI data by

blind separation into independent spatial components. *Hum Brain Mapp*, 6 (3) : 160–188, janvier 1998.

[Meng et Rubin, 1993] X.-L. Meng et D. B. Rubin. Maximum likelihood estimation via the ECM algorithm : a general framework. *Biometrika*, 80 (2) : 267–278, juin 1993.

[Mériaux *et al.*, 2003] S. Mériaux, F. Kherif, C. Pallier, V. Izard, M. Brett, L. Garnero et J.-B. Poline. Assessing the homogeneity of subject's samples for random effect analyses in fMRI. In *Proc. 9th HBM CD-Rom, NeuroImage vol.19(2)*, New-York, USA, 18-22 juin 2003.

[Mériaux *et al.*, 2004] S. Mériaux, F. Kherif, A. Roche, M. Brett, L. Garnero et J.-B. Poline. How frequently do we sample inhomogeneous group of subjects in fMRI studies ? In *Proc. 10th HBM CD-Rom, NeuroImage vol.22(1)*, Budapest, Hongrie, 13-17 juin 2004.

[Mériaux *et al.*, 2006a] S. Mériaux, A. Roche, G. Dehaene-Lambertz et J.-B. Poline. When do mixed-effect models fail to improve detection sensitivity in fMRI group activation maps ? In *Proc. 12th HBM CD-Rom, NeuroImage vol.31(1)*, Florence, Italie, 11-15 juin 2006.

[Mériaux *et al.*, 2006b] S. Mériaux, A. Roche, G. Dehaene-Lambertz, B. Thirion et J.-B. Poline. Combined permutation test and mixed-effect model for group average analysis in fMRI. *Hum Brain Mapp*, 27 (5) : 402–410, mai 2006.

[Mériaux *et al.*, 2006c] S. Mériaux, A. Roche, B. Thirion et G. Dehaene-Lambertz. Robust statistics for nonparametric group analysis in fMRI. In *Proc. 3rd ISBI*, pages 936–939, Arlington, USA, 6-9 avril 2006.

[Moeller et Strother, 1991] J. R. Moeller et S. C. Strother. A regional covariance approach to the analysis of functional patterns in positron emission tomographic data. *J Cereb Blood Flow Metab*, 11 (2) : A121–A135, mars 1991.

[Moritz *et al.*, 2005] C. H. Moritz, J. D. Carew, A. B. McMillan et M. E. Meyerand. Independent component analysis applied to self-paced functional MR imaging paradigms. *NeuroImage*, 25 (1) : 181–192, mars 2005.

[Neter *et al.*, 1985] J. Neter, M. H. Kutner, W. Wasserman et C. J. Nachtsheim. *Applied linear statistical models.* McGraw-Hill Education, Homewood, USA, 2ème edition, 1985.

[Neumann *et al.*, 2003] J. Neumann, G. Lohmann, S. Zysset et D. Y. von Cramon. Within-subject variability of BOLD response dynamics. *NeuroImage*, 19 (3) : 784–796, juillet 2003.

[Nichols et Holmes, 2002] T. E. Nichols et A. P. Holmes. Nonparametric permutation tests for functional neuroimaging : a primer with examples. *Hum Brain Mapp*, 15 (1) : 1–25, janvier 2002.

[Nieto-Castanon *et al.*, 2003] A. Nieto-Castanon, S. S. Ghosh, J. A. Tourville et F. H. Guenther. Region of interest based analysis of functional imaging data. *NeuroImage*, 19 (4) : 1303–1316, août 2003.

[Nybakken *et al.*, 2002] G. E. Nybakken, M. A. Quigley, C. H. Moritz, D. Cordes, V. M. Haughton et M. E. Meyerand. Test-retest precision of functional magnetic resonance imaging processed with independent component analysis. *Neuroradiology*, 44 (5) : 403–406, mai 2002.

[Ogawa *et al.*, 1990] S. Ogawa, T. M. Lee, A. R. Kay et D. W. Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc Natl Acad Sci*, 87 (24) : 9868–9872, décembre 1990.

[Owen, 2001] A. B. Owen. *Empirical likelihood.* CRC Press, New York, USA, 2001.

[Pallier *et al.*, 2003] C. Pallier, S. Dehaene, J.-B. Poline, D. LeBihan, A.-M. Argenti, E. Dupoux et J. Mehler. Brain imaging of language plasticity in adopted adults : can a second language replace the first ? *Cereb Cortex*, 13 (2) : 155–161, février 2003.

[Penny *et al.*, 2003a] W. D. Penny, A. P. Holmes et K. J. Friston. Random-effects analysis. In R. S. J. Frackowiak, K. J. Friston, C. Frith, R. Dolan, K. J. Friston, C. J. Price, S. Zeki, J. Ashburner et W. D. Penny, éditeurs, *Human Brain Function.* Academic Press, USA, 2nd edition, 2003.

[Penny *et al.*, 2003b] W. D. Penny, S. J. Kiebel et K. J. Friston. Variational Bayesian inference for fMRI time series. *NeuroImage*, 19 (3) : 727–741, juillet 2003.

[Penny *et al.*, 2005] W. D. Penny, N. J. Trujillo-Barreto et K. J. Friston. Variational Bayesian inference for fMRI time series. *NeuroImage*, 24 (2) : 350–362, janvier 2005.

[Piazza *et al.*, 2004] M. Piazza, V. Izard, P. Pinel, D. LeBihan et S. Dehaene. Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, 44 (3) : 547–555, octobre 2004.

[Pitman, 1937a] E. J. G. Pitman. Significance tests which may be applied to samples from any population. Part I. *Supplement to the Journal of the Royal Statistical Society*, 4 (2) : 119–130, 1937.

[Pitman, 1937b] E. J. G. Pitman. Significance tests which may be applied to samples from any population. Part II. The correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society*, 4 (2) : 225–232, 1937.

[Pitman, 1938] E. J. G. Pitman. Significance tests which may be applied to samples from any population. Part III. The analysis of variance test. *Biometrika*, 29 (3/4) : 322–335, février 1938.

[Poldrack *et al.*, 2001] R. A. Poldrack, E. Temple, A. Protopapas, S. Nagarajan, P. Tallal, M. Merzenich et J. D. Gabrieli. Relations between the neural bases of dynamic auditory processing and phonological processing : evidence from fMRI. *J Cogn Neurosci*, 13 (5) : 687–697, juillet 2001.

[Poline *et al.*, 2006] J.-B. Poline, S. C. Strother, G. Dehaene-Lambertz, G. F. Egan et J. L. Lancaster. Motivation and synthesis of the FIAC experiment : reproducibility of fMRI results across expert analyses. *Hum Brain Mapp*, 27 (5) : 351–359, mai 2006.

[Purdon *et al.*, 2001] P. L. Purdon, V. Solo, R. M. Weisskoff et E. N. Brown. Locally regularized spatiotemporal modeling and model comparison for functional MRI. *NeuroImage*, 14 (4) : 912–923, octobre 2001.

[Purdon et Weisskoff, 1998] P. L. Purdon et R. M. Weisskoff. Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI. *Hum Brain Mapp*, 6 (4) : 239–249, décembre 1998.

[Reber *et al.*, 1998] P. J. Reber, E. C. Wong, R. B. Buxton et L. R. Frank. Correction of off resonance-related distortion in echo-planar imaging using EPI-based field maps. *Magn Reson Med*, 39 (2) : 328–330, février 1998.

[Robert et Escoufier, 1976] P. Robert et Y. Escoufier. A unifying tool for linear multivariate statistical methods : the RV coefficient. *Appl Statist*, 25 (3) : 257–265, mai 1976.

[Roche, 2001] A. Roche. *Recalage d'images médicales par inférence statistique.* Thèse de doctorat, Université de Nice-Sophia Antipolis, France, février 2001.

[Roche *et al.*, 2007] A. Roche, S. Mériaux, M. Keller et B. Thirion. Mixed-effect statistics for group analysis in fMRI : a nonparametric maximum likelihood approach. *NeuroImage*, 38 (3) : 501–510, novembre 2007.

[Romano, 1990] J. P. Romano. On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85 (411) : 686–692, septembre 1990.

[Rorden *et al.*, 2007] C. Rorden, L. Bonilha et T. E. Nichols. Rank-order versus mean based statistics for neuroimaging. *NeuroImage*, 35 (4) : 1531–1537, mai 2007.

[Rosner, 1975] B. Rosner. On the detection of many outliers. *Technometrics*, 17 (2) : 221–227, mai 1975.

[Rousseeuw et Leroy, 1987] P. Rousseeuw et A. Leroy. *Robust regression and outlier detection.* Wiley & Sons, New York, USA, 1987.

[Roy et Sherrington, 1890] C. S. Roy et C. S. Sherrington. On the regulation of the blood supply of the brain. *Journal of Physiolohy*, 11 (1-2) : 85–108, janvier 1890.

[Royston, 1982] J. P. Royston. An extension of Shapiro and Wilk's W test for normality to large samples. *Appl Stat*, 31 (2) : 115–124, 1982.

[Savoy, 2006] R. L. Savoy. Using small numbers of subjects in fMRI-based research. *IEEE Eng Med Biol Mag*, 25 (2) : 52–59, avril 2006.

[Sergent *et al.*, 2005] C. Sergent, S. Baillet et S. Dehaene. Timing of the brain events underlying access to consciousness during the attentional blink. *Nat Neurosci*, 8 (10) : 1391–1400, octobre 2005.

[Sergent et Dehaene, 2004] C. Sergent et S. Dehaene. Is consciousness a gradual phenomenon ? Evidence for an all-or-none bifurcation during the attentional blink. *Psychol Sci*, 15 (11) : 720–728, novembre 2004.

[Shapiro *et al.*, 1968] S. S. Shapiro, M. B. Wilk et H. J. Chen. A comparative study of various tests for normality. *J Am Stat Assoc*, 63 (324) : 1343–1372, décembre 1968.

[Simon *et al.*, 2002] O. Simon, J.-F. Mangin, L. Cohen, D. LeBihan et S. Dehaene. Topographical layout of hand, eye, calculation and language-related areas in the human parietal lobe. *Neuron*, 33 (3) : 475–487, janvier 2002.

[Stark et Bradley, 1992] D. D. Stark et W. G. Bradley. *Magnetic resonance imaging.* Mosby-Year Book, Saint-Louis, USA, 1992.

[Stewart et Love, 1968] D. Stewart et W. Love. A general canonical correlation index. *Psychol. Bull.*, 70 (3) : 160–163, septembre 1968.

[Strasser et Weber, 1999] H. Strasser et C. Weber. On the asymptotic theory of permutation statistics. *Math Methods Statist*, 8 (2) : 220–250, juin 1999.

[Strother *et al.*, 2002] S. C. Strother, J. R. Anderson, L. K. Hansen, U. Kjems, R. Kustra, J. J. Sidtis, S. Frutiger, S. Muley, S. LaConte et D. A. Rottenberg. The quantitative evaluation of functional neuroimaging experiments : the NPAIRS data analysis framework. *NeuroImage*, 15 (4) : 747–771, avril 2002.

[Strother *et al.*, 1995a] S. C. Strother, J. R. Anderson, K. A. Schaper, J. J. Sidtis, J. S. Liow, R. P. Woods et D. A. Rottenberg. Principal component

analysis and the scaled subprofile model compared to intersubject averaging and statistical parametric mapping : I. "Functional connectivity" of the human motor system studied with [15O]water PET. *J Cereb Blood Flow Metab*, 15 (5) : 738–753, septembre 1995.

[Strother *et al.*, 1995b] S. C. Strother, I. Kanno et D. A. Rottenberg. Commentary and opinion : I. Principal component analysis, variance partitioning and "functional connectivity". *J Cereb Blood Flow Metab*, 15 (3) : 353–360, mai 1995.

[Tegeler *et al.*, 1999] C. Tegeler, S. C. Strother, J. R. Anderson et S. G. Kim. Reproducibility of BOLD-based functional MRI obtained at 4T. *Hum Brain Mapp*, 7 (4) : 267–283, février 1999.

[Thévenaz *et al.*, 2000] P. Thévenaz, T. Blu et M. Unser. Interpolation revisited. *IEEE Trans Med Imaging*, 19 (7) : 739–758, juillet 2000.

[Thirion et Faugeras, 2003] B. Thirion et O. Faugeras. Dynamical components analysis of fMRI data through kernel PCA. *NeuroImage*, 20 (1) : 34–49, septembre 2003.

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

[Thirion *et al.*, 2007] B. Thirion, P. Pinel, S. Mériaux, A. Roche, S. Dehaene et J.-B. Poline. Analysis of a large fMRI cohort : Statistical and methodological issues for group analyses. *NeuroImage*, 35 (1) : 105–120, mars 2007.

[Thirion *et al.*, 2005] B. Thirion, P. Pinel et J.-B. Poline. Finding landmarks in the functional brain : detection and use for group characterization. In *Proc. MICCAI 2005*, Palm Springs, USA, 26-29 octobre 2005.

[Thirion *et al.*, 2006b] B. Thirion, P. Pinel et J.-B. Poline. Making reliable Random Effects Analyses : How many subjects ? In *Proc. 12th HBM CD-Rom, NeuroImage vol.31(1)*, Florence, Italie, 11-15 juin 2006.

[Thode, 2002] H. C. Thode. Testing for normality. In M. Dekker, éditeur, *STATISTICS : textbooks and monographs*, volume 164. New-York, USA, 2002.

[Tietjen et Moore, 1972] G. L. Tietjen et R. H. Moore. Some Grubbs-type statistics for the detection of several outliers. *Technometrics*, 14 (3) : 583–597, août 1972.

[Tucker, 1958] L. R. Tucker. An inter-battery method of factor analysis. *Psychometrika*, 23 (2) : 111–136, juin 1958.

[Turner *et al.*, 1998] R. Turner, A. Howseman, G. E. Rees, O. Josephs et K. J. Friston. Functional magnetic resonance imaging of the human brain : data acquisition and analysis. *Exp Brain Res*, 123 (1-2) : 5–12, octobre 1998.

[Unser, 1999] M. Unser. Splines : a perfect fit for signal/image processing. *IEEE Signal Processing Magazine*, 16 (6) : 22–38, novembre 1999.

[van de Moortele, 1999] P.-F. van de Moortele. *IRM fonctionnelle à 3 Tesla : développements méthodologiques.* Thèse de doctorat, Université Paris XI, France, 1999.

[van der Vaart, 1998] A. W. van der Vaart. *Asymptotic statistics.* Cambridge University Press, New York, USA, 1998.

[von Kriegstein *et al.*, 2003] K. von Kriegstein, E. Eger, A. Kleinschmidt et A.-L. Giraud. Modulation of neural responses to speech by directing attention to voices or verbal content. *Brain Res Cogn Brain Res*, 17 (1) : 48–55, juin 2003.

[Wager *et al.*, 2005] T. D. Wager, M. C. Keller, S. C. Lacey et J. Jonides. Increased sensitivity in neuroimaging analyses using robust regression. *NeuroImage*, 26 (1) : 99–113, mai 2005.

[Wager *et al.*, 2004] T. D. Wager, J. K. Rilling, E. E. Smith, A. Sokolik, K. L. Casey, R. J. Davidson, S. M. Kosslyn, R. M. Rose et J. D. Cohen. Placebo-induced changes in fMRI in the anticipation and experience of pain. *Science*, 303 (5661) : 1162–1167, février 2004.

[Welchew *et al.*, 2002] D. E. Welchew, G. D. Honey, T. Sharma, T. W. Robbins et E. T. Bullmore. Multidimensional scaling of integrated neurocognitive function and schizophrenia as a disconnexion disorder. *NeuroImage*, 17 (3) : 1227–1239, novembre 2002.

[Welsch, 1982] R. E. Welsch. Influence functions and regression diagnostics. In R. L. Launer et A. F. Siegel, éditeurs, *Modern Data Analysis*, pages 149–169. Academic Press, New York, USA, 1982.

[Welsch et Kuh, 1977] R. E. Welsch et E. Kuh. Linear regression diagnostics. Rapport interne, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, USA, 1977.

[Wilcoxon, 1945] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1 (6) : 80–83, décembre 1945.

[Woods, 1996] R. P. Woods. Modeling for intergroup comparisons of imaging data. *NeuroImage*, 4 (3 Pt 3) : S84–S94, décembre 1996.

[Woolrich *et al.*, 2004] M. W. Woolrich, T. E. J. Behrens, C. F. Beckmann, M. Jenkinson et S. M. Smith. Multilevel linear modelling for fMRI group analysis using Bayesian inference. *NeuroImage*, 21 (4) : 1732–1747, avril 2004.

[Woolrich *et al.*, 2001] M. W. Woolrich, B. D. Ripley, M. Brady et S. M. Smith. Temporal autocorrelation in univariate linear modeling of fMRI data. *NeuroImage*, 14 (6) : 1370–1386, décembre 2001.

[Worsley, 1994] K. J. Worsley. Local maxima and the expected Euler characteristic of excursion sets of χ^2, F and t fields. *Advances in Applied Probability*, 26 (1) : 13–42, mars 1994.

[Worsley *et al.*, 2002] K. J. Worsley, C. H. Liao, J. Aston, V. Petre, G. H. Duncan, F. Morales et A. C. Evans. A general statistical analysis for fMRI data. *NeuroImage*, 15 (1) : 1–15, janvier 2002.

[Worsley *et al.*, 1997] K. J. Worsley, J.-B. Poline, K. J. Friston et A. C. Evans. Characterizing the response of PET and fMRI data using multivariate linear models. *NeuroImage*, 6 (4) : 305–319, novembre 1997.

[Xiong *et al.*, 2000] J. H. Xiong, S. Rao, P. Jerabek, F. Zamarripa, M. Woldorff, J. L. Lancaster et P. T. Fox. Intersubject variability in cortical activations during a complex language task. *NeuroImage*, 12 (3) : 326–339, septembre 2000.

[Zarahn *et al.*, 1997] E. Zarahn, G. K. Aguirre et M. D'Esposito. Empirical analyses of BOLD fMRI statistics. I Spatially unsmoothed data collected under null-hypothesis conditions. *NeuroImage*, 5 (3) : 179–197, avril 1997.

Table des matières

Liste des figures

Liste des tableaux

Nom : MÉRIAUX **Prénom :** Sébastien **Laboratoire :** CEA/I^{2}BM - Neurospin, France

---

Titre : Diagnostique d'homogénéité et inférence non-paramétrique pour l'analyse de groupe en imagerie par résonance magnétique fonctionnelle

Résumé : L'un des objectifs principaux de l'imagerie par résonance magnétique fonctionnelle (IRMf) est la localisation *in vivo* et de manière non invasive des zones cérébrales associées à certaines fonctions cognitives. Le cerveau présentant une très grande variabilité anatomo-fonctionnelle inter-individuelle, les études d'IRMf incluent généralement plusieurs sujets et une analyse de groupe permet de résumer les résultats intra-sujets en une carte d'activation du groupe représentative de la population d'intérêt. L'analyse de groupe « standard » repose sur une hypothèse forte d'homogénéité des effets estimés à travers les sujets. Dans un premier temps, nous étudions la validité de cette hypothèse par une méthode multivariée diagnostique et un test de normalité univarié (le test de Grubbs). L'application de ces méthodes sur une vingtaine de jeux de données révèle la présence fréquente de données atypiques qui peuvent invalider l'hypothèse d'homogénéité. Nous proposons alors d'utiliser des statistiques de décision robustes calibrées par permutations afin d'améliorer la spécificité et la sensibilité des tests statistiques pour l'analyse de groupe. Puis nous introduisons de nouvelles statistiques de décision à effets mixtes fondées sur le rapport de vraisemblances maximales, permettant de pondérer les sujets en fonction de l'incertitude sur l'estimation de leurs effets. Nous confirmons sur des jeux de données que ces nouvelles méthodes d'inférence permettent un gain en sensibilité significatif, et nous fournissons l'ensemble des outils développés lors de cette thèse à la communauté de neuro-imagerie dans le logiciel *DISTANCE*.

Mots clés : IRM fonctionnelle, analyse de groupe, diagnostique d'homogénéité, coefficient RV, distance de Cook, test de Grubbs, statistiques robustes, modèles à effets mixtes, tests de permutations

---

Title: Homogeneity diagnosis and non-parametric inference for group analysis using functional magnetic resonance imaging

Abstract: One of the most challenging purposes to reach for the functional magnetic resonance imaging (fMRI) is the *in vivo* and non invasive localization of cerebral areas involved in some cognitive functions. Due to the high degree of anatomo-functional variability observed for human brains, fMRI studies generally involve several subjects, which results are summarized into a group activation map representing the population of interest through a group analysis procedure. The « standard » procedure for group analysis inference relies on the strong assumption that the estimated effects are normally distributed across subjects. Our first concern is to study the validity of this assumption using both a multivariate diagnosis approach and a univariate normality test (the Grubbs test). These methods are tested on twenty datasets revealing that the homogeneity assumption may be violated by the frequent presence of atypical data. To enhance both sensibility and sensitivity of statistical tests used for group analysis, we then propose to use robust decision statistics calibrated through permutation testing methods. We also introduce new mixed effects statistics based on maximum likelihoods ratio, which allow to re-weight the subjects according to the reliability of their respective effect estimates. The results obtained on several datasets confirm a significant enhancement of sensibility in group activations maps. We therefore propose all our group analysis methods to the neuro-imaging community through our *DISTANCE* software.

Keywords: Functional MRI, group analysis, homogeneity diagnosis, RV-coefficient, Cook distance, Grubbs test, robust statistics, mixed effects models, permutation tests