BLAIN Alexandre
SEGERIE Charbel-Raphael

MVA 2020-2021

All-Resolutions Inference for Brain Imaging

Jonathan D. Rosenblatt, Livio Finos, Wouter D. Weeda, Aldo Solari, Jelle J. Goeman

Table des matières

1 Introduction au problème posé

La cartographie du cerveau humain, qui consiste à associer à une fonction cognitive ou action une ou plusieurs régions du cerveau, est une question essentielle en neurosciences mais aussi dans la médecine moderne : dès la deuxième moitié du 20ème siècle, des chercheurs se sont intéressés à la corrélation entre l'oxygénation du sang dans certaines régions du cerveau et les comportements humains. Les progrès techniques en imagerie réalisés dans les années 70 ont permis de grandement préciser la cartographie du cerveau humain.

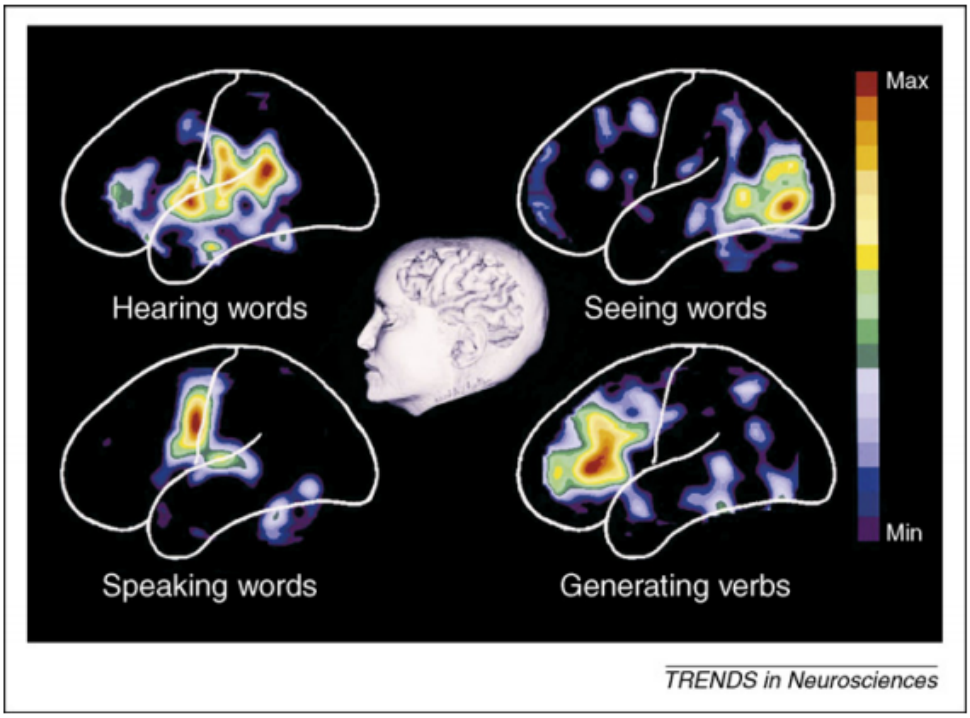

FIGURE 1 – Un exemple d'imagerie cérébrale dans l'étude du langage (1988)

La figure ci-dessus représente un exemple historique d'imagerie cérébrale obtenue par PET scan (émission de positrons). L'étude en question concerne l'activation de différentes zones cérébrales en réponse à la vue, la diction, la génération et l'audition de mots. Les quatre figures correspondent respectivement à ces quatre fonctions et montrent les différentes zones activées par les différentes fonctions. À partir de telles imageries, on a pu établir une cartographie cérébrale générale, illustrée figure 2. Il est important de noter que cette cartographie ne donne pas une image précise du cerveau de chaque individu : il existe des variabilités importantes entre les individus.

On remarque que les deux figures sont bien cohérentes (bien que la figure 1 date

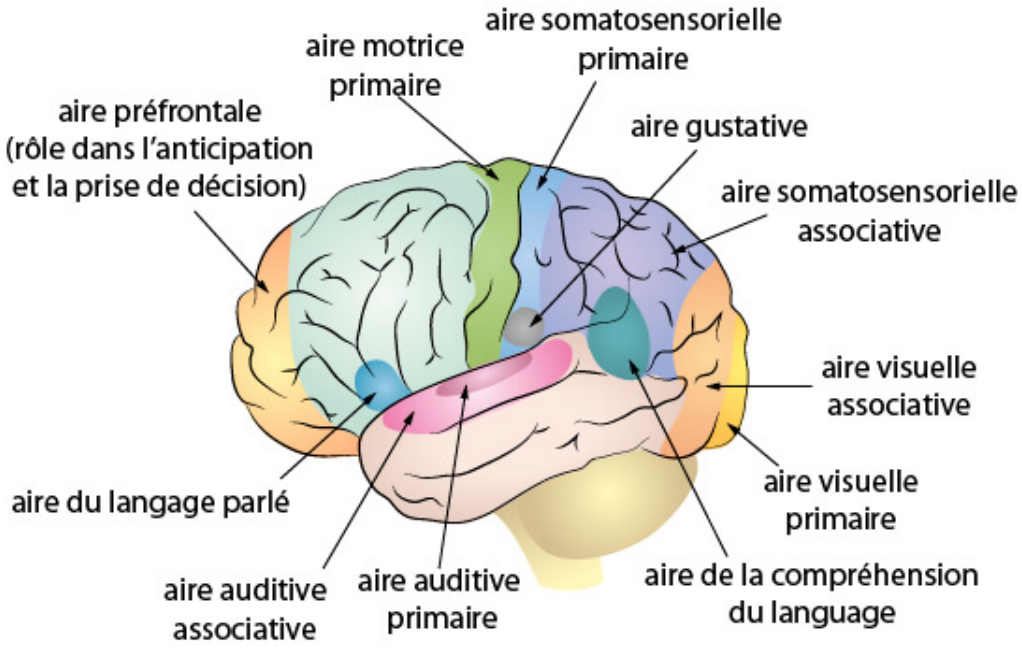

FIGURE 2 – Cartographie cérébrale et fonctions cognitives

de 1988, ce qui est relativement tôt dans la chronologie de la cartographie cérébrale) : lorsque l'on fait entendre des mots au patient, on observe bien une augmentation du flux sanguin au centre de la figure, qui correspond aux aires auditives primaire et associative sur la figure 2.

Savoir quelles zones du cerveau remplissent quelles fonctions est essentiel pour étudier la cognition humaine et plus généralement les comportements humains. Cette connaissance est cruciale dans l'étude de certaines maladies qui endommagent gravement le cerveau à court ou long terme, en particulier les maladies neurodégénératives comme la maladie d'Alzheimer ou de Parkinson mais aussi dans l'étude des AVC.

Pour établir une telle cartographie, ou en tout cas détecter l'activation d'une zone donnée du cerveau pour une tâche, on se place dans le cadre suivant : le cerveau d'un patient est analysé par IRM dans deux états distincts, "au repos" ou bien "exécutant une action". L'action en question peut être physique, comme prendre un objet par exemple, ou simplement mentale, penser à une certaine idée ou objet par exemple. On va ici chercher à détecter les zones qui s'activent différemment dans les deux états, signe de l'implication dans la tâche réalisée. Après avoir réalisé cette étude pour un nombre N de patient, typiquement dans l'ordre d'une à plusieurs dizaines, on rassemble tous ces résultats en une liste de p_i p-valeurs, les différents i correspondant à différents voxels. Ainsi, nous ne nous intéressons qu'à la dernière étape de l'analyse de la pipeline des

fMRI, celle qui analyse les p-valeurs associées à chacun des voxels.

2 La méthode ARI

La méthode All-Resolutions Inference proposée par Rosenblatt et al. permet d'étudier le problème de l'activation cérébrale. Formellement, on segmente le cerveau en un ensemble d'unités de volume appelées voxels, et on va chercher à tester les hypothèses $\mathbf{H_{0,i}}$: "voxel *i* inactif pour cette action" contre $\mathbf{H_{1,i}}$: "voxel *i* actif pour cette action".

2.1 Multiplicité et circularité

Les approches utilisées pour résoudre ce problème doivent obligatoirement tenir compte de la multiplicité du problème : un test simple classique sur chacun des voxels pour mesurer l'activité ne suffit pas. En effet, dans une image typique par fMRI, on dénombre 100 000 voxels. Même dans le cas où aucun voxel est activé, c'est à lire que l'hypothèse nulle $\mathbf{H_{0,i}}$ est vraie pour tous les voxels *i*, on s'attend à obtenir 5000 voxels associés à un test faux positifs. C'est un problème classique de multiplicité de test statistique. Par ailleurs, une approche se restreignant à l'étude individuelle de chacun des voxels ne serait pas cohérente avec la réalité neurologique. Comme expliqué plus haut, l'activation du cerveau se fait par aires. Les aires du cerveau peuvent comporter quelques centaines voire quelques milliers de voxels ce qui entraîne une multiplicité forte. Il faut donc prendre en compte le fait que nous étudions un nombre élevé de voxels simultanément (problème de multiplicité statistique) ainsi que le fait que l'activation se fait par aire cérébrale (a priori des aires cérébrales).

L'inférence en neuro-imagerie se concentrant sur des aires d'intérêt susceptibles de s'activer, les méthodes d'inférence sont généralement *cluster-based*, c'est-à-dire que l'on sélectionne un cluster, un ensemble de voxels connexes du cerveau, sur lequel on veut tester simultanément pour chaque voxel $\mathbf{H_{0,i}}$ contre $\mathbf{H_{1,i}}$. Pour contrôler l'erreur de type 1, on peut adopter différentes approches, comme le contrôle du Family Wise Error Rate (FWER) par la correction de Bonferroni, ou encore le contrôle du False Discovery Rate (FDR) par la correction de Benjamini-Hochberg.

Afin de donner une idée de l'esprit de ce genre de correction, on peut expliquer le principe de la correction de Bonferroni : si l'on considère les $|C|$ p-valeurs p_i associées au cluster C. Alors, on rejette H_0 au niveau α dès que l'une des p-valeurs est inférieur

à $\alpha/|C|$.

Dans le cadre de ces méthodes, on extrait des clusters d'intérêt à partir des données puis on effectue un test, une inférence, sur ces clusters. Mais si plusieurs tests d'activation sont effectués sur la région déjà sélectionnée, cela peut induire un problème de dépendance aussi appelé un problème de circularité statistique. Par exemple, si après avoir sélectionné un cluster, la correction de Bonferroni indique qu'il y a au moins un voxel i activé, on ne peut pas conduire un nouveau test en sélectionnant maintenant comme cluster les voxels voisins du voxel i, car cela crée un problème de circularité qui est donc le deuxième problème principal (après la multiplicité) que les méthodes utilisées en neuro-imagerie doivent résoudre.

Autrement dit, l'utilisation des mêmes données pour la sélection et pour l'inférence sur les données sélectionnées pose problème. Selon l'article [4], on peut expliquer ces biais par le fait que le critère de sélection est proche de la garantie attendue sur les données sélectionnées. Ceci entraîne une forme de circularité dans l'inférence réalisée, qui entraîne une inflation du nombre de faux positifs.

Les méthodes de Bonferroni et de Benjamini-Hochberg permettent de contrôler l'erreur de type I soit sous la forme du FWER ou du FDR mais ne permettent pas l'inférence hiérarchique par drill-down successifs. L'objectif est donc de préserver les garanties qui permettent une approche par drill-down successifs tout en gardant une puissance de test maximale. Nous allons étudier la méthode ARI et vérifier expérimentalement que la puissance de cette méthode est proche de l'état de l'art.

2.2 Nécessité de bornes post-hoc

Dans l'article que nous présentons [6] publié en 2018 dans la revue NeuroImage, Rosenblatt et ses co-auteurs proposent une approche répondant à ces deux problèmes en construisant des bornes post-hoc. Leur approche consiste à utiliser l'inégalité de Simes pour contrôler le nombre de faux positifs dans tout sous-ensemble d'hypothèses sélectionné, éventuellement après avoir vu les données.

La méthode ARI permet de répondre au problème de multiplicité et de circularité de l'inférence en construisant des bornes post-hoc, c'est-à-dire des bornes qui fournissent des garanties quel que soit le sous-ensemble d'hypothèses sélectionnées, éventuellement choisi par l'utilisateur en utilisant les données. On peut formellement com-

parer comme suit l'écriture d'une borne classique avec une borne post-hoc :

Une borne inférieure classique $V(S)$ du nombre de vrais positifs dans la région S :

$$\forall P \in \mathscr{P}\ \mathbb{P}_{X\sim P}\left(a(S) \geq V(S)\right) \geq 1-\alpha$$

Une borne post-hoc correspond à une fonction $V(S)$ vérifiant la propriété suivante :

$$\forall P \in \mathscr{P}\ \mathbb{P}_{X\sim P}\left(\forall S \subset \mathbb{N}_m, a(S) \geq V(S)\right) \geq 1-\alpha$$

Crucialement, les méthodes post-hoc permettent d'inclure le "pour tout" dans la probabilité. On peut donc étudier n'importe quel sous ensemble de voxels dans l'ordre voulu.

2.3 Inférence par cluster et drill-down

En pratique, ce type de méthode sélectionne des clusters à partir d'un premier seuil, puis sélectionne des clusters plus fins au sein de ces clusters, afin d'aboutir à une inférence sur une région d'intérêt. Typiquement, le résultat de la méthode est un nombre minimum de vraies détections (au seuil de risque α) dans une région. Cette approche appelée *drill-down* dans l'article est possible car les bornes post-hoc permettent de conserver des garanties sur le nombre de faux positifs dans les régions sélectionnées hiérarchiquement.

On pourrait même théoriquement aller jusqu'à choisir des clusters singletons ne contenant qu'un seul voxel, mais ceci n'est pas conseillé car la méthode ARI a une bien meilleure puissance statistique lorsqu'on l'utilise sur un grand cluster. En pratique on réalisera un drill-down de la même manière que l'on épluche la peau extérieure d'un oignon, procédure qui est réalisée empiriquement automatiquement en sélectionnant comme nouveau cluster l'intersection de l'ancien cluster et de l'ensemble des voxels ayant une p-valeur inférieure à un nouveau seuil. Ceci est possible car le caractère post-hoc des bornes utilisées permet de conserver la garantie voulue même après plusieurs sélections utilisant les données.

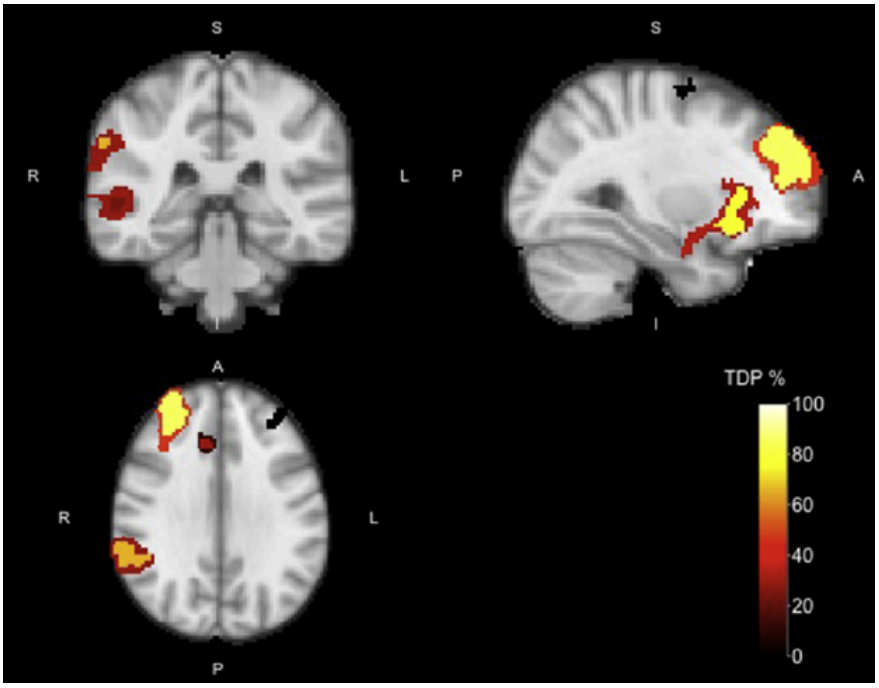

FIGURE 3 – Un exemple d'inférence tiré de l'article de Rosenblatt et al., 2018

La figure ci-dessus représente un exemple d'inférence du papier de Rosenblatt et al : ces figures sont le résultat de la superposition de deux seuillages par une statistique de test. Le premier seuillage sert à isoler les premiers clusters, en sélectionnant tous les voxels ayant une p-valeur inférieure à un seuil et en conduisant un test par partie connexe qui sont alors autant de clusters. Chacun des tests donne une borne inférieure du nombre de voxels activé par cluster. Le deuxième seuillage permet d'extraire des régions d'intérêt au sein des premiers clusters (*drill-down* hiérarchique). Le taux minimum de vraies détections est représenté par une échelle de couleur.

2.4 Condition de Simes et borne ARI

La condition principale d'obtention de la borne ARI est la condition PRDS, qui est une condition de régularité des données sous l'hypothèse nulle. Sous cette condition, l'inégalité de Simes est vérifiée. Pour m_0 le nombre d'hypothèses nulles vraies et α un niveau de risque, on a :

$$\mathbb{P}\left(\forall k \in \{1,\dots,m_0\} : p_{(k:\mathscr{H}_0)} \geq \alpha k/m_0\right) \geq 1-\alpha$$

Pour tout ensemble S de voxels, p_S est définie comme la p-valeur correspondant au test de Simes dont l'hypothèse nulle est le fait qu'aucun voxel de S est activé, avec :

$$p_S = \min_{1 \leqslant i \leq |S|} \frac{|S|}{i} p_{i:S}$$

et $p_{i:S}$ est la i-ème p-valeur la plus petite parmi S. Sous l'hypothèse PRDS, ce test est bien de niveau α. L'hypothèse PRDS est bien connue dans la littérature des tests multiples puisqu'elle est nécessaire à l'obtention de Benjamini-Hochberg [2]. Dans le cas des données fMRI, cette hypothèse est communément acceptée (voir [5]).

Pour un cerveau représenté par un ensemble de voxels B, un ensemble d'hypothèses S donné, un niveau α de risque et la famille de p-valeurs ordonnées $(p_{(i:B)})_i$, le taux de vraies détections représenté sur la figure est basé sur la borne de Simes :

$$\bar{a}(S) = \min\left\{0 \leq k \leq |S| : \min_{1\leq i\leq |S|-k} \frac{h(\alpha)}{i} p_{(i+k:S)} > \alpha\right\}$$

Avec

$$h(\alpha) = \max\left\{i \in \{0,\ldots,m\} : i p_{(m-i+j:B)} > j\alpha, \quad \text{for } j = 1,\ldots,i\right\}$$

On peut interpréter la grandeur h comme un majorant du nombre de voxels inactifs dans le cerveau pour la tache d'intérêt.

3 Implémentation et résultats

3.1 Outils

Pour implémenter la méthode décrite dans le papier, on utilise le package Nilearn (voir [1]) sur Python. Ce package permet non seulement de récupérer les données de plusieurs datasets de neuro-imagerie mais aussi de traiter ces données pour les préparer à l'application de méthodes statistiques. Dans notre cas, on utilise d'abord Nilearn pour récupérer le dataset *Haxby* [3] qui étudie des tâches de vision. Dans la figure ci-dessous, on représente un exemple de régions d'intérêt (le Gyrus Cingulaire). Il est à noter que cette visualisation permet juste d'avoir une idée de la localisation anatomique de la région et ne représente pas l'activation ou les p-valeurs à l'étude.

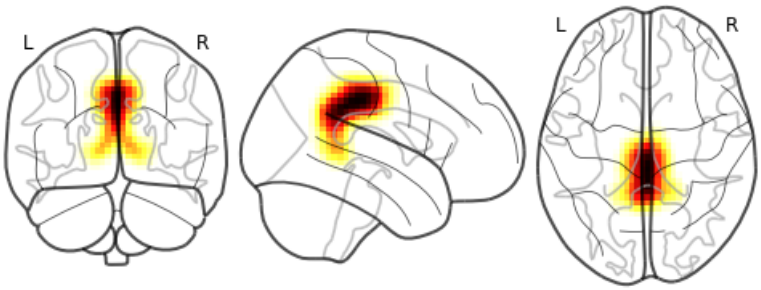

FIGURE 4 – Exemple de visualisation du Gyrus Cingulaire

Plus précisément, on s'intéresse aux deux états "au repos" contre "regarder un visage" tirés de ce dataset. Ensuite, on utilise le package Nilearn pour préparer les données à l'application de la méthode ARI. Finalement, on obtient une liste de p-valeurs de longueur 39912 (nombre de voxels). Ensuite, on utilise Nilearn pour extraire un atlas (ici, l'atlas Harvard-Oxford), qui va nous permettre de nous intéresser à des régions d'intérêt pour la tâche à l'étude.

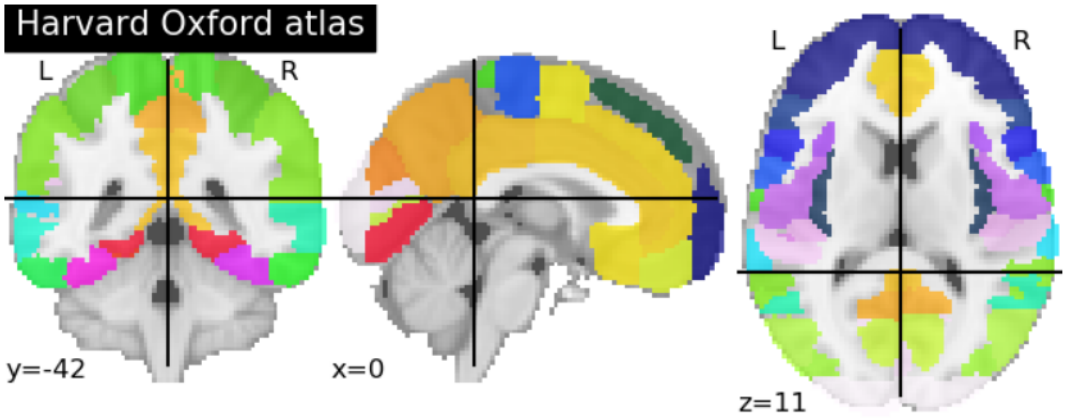

FIGURE 5 – Atlas Harvard-Oxford sous Nilearn

3.2 Résultats et comparaisons

Après avoir implémenté la borne de l'article, nous avons choisi de comparer la méthode ARI à deux autres méthodes. La première méthode est la correction de Bonferroni, procédé le plus simple de correction en tests multiples. La seconde méthode, proche de l'état de l'art, utilise un test non-paramétrique et une correction de type max-type. Cette méthode est notamment proposée dans des exemples d'inférence avec Nilearn.

Le résultat de notre implémentation est une fonction qui pour des données, un ni-

veau de risque α, et une région anatomique d'un atlas renvoie un plot de la région d'intérêt et les détections données par chacune des méthodes. Voici les résultats obtenus pour deux régions d'intérêt liées à la vision, le Précuneus et le Gyrus Cingulaire :

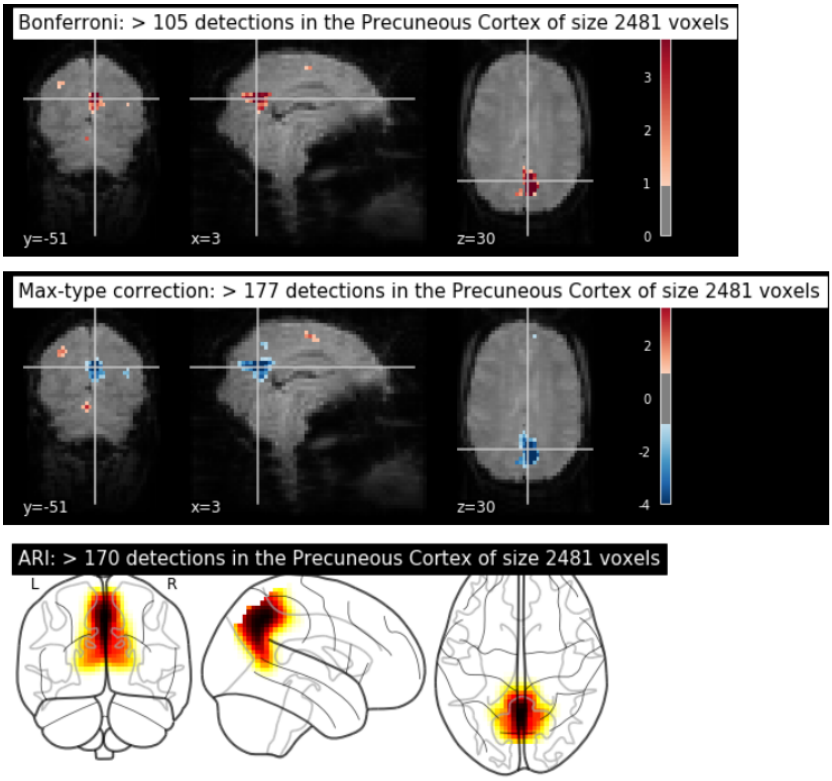

FIGURE 6 – Résultats sur le Précuneus

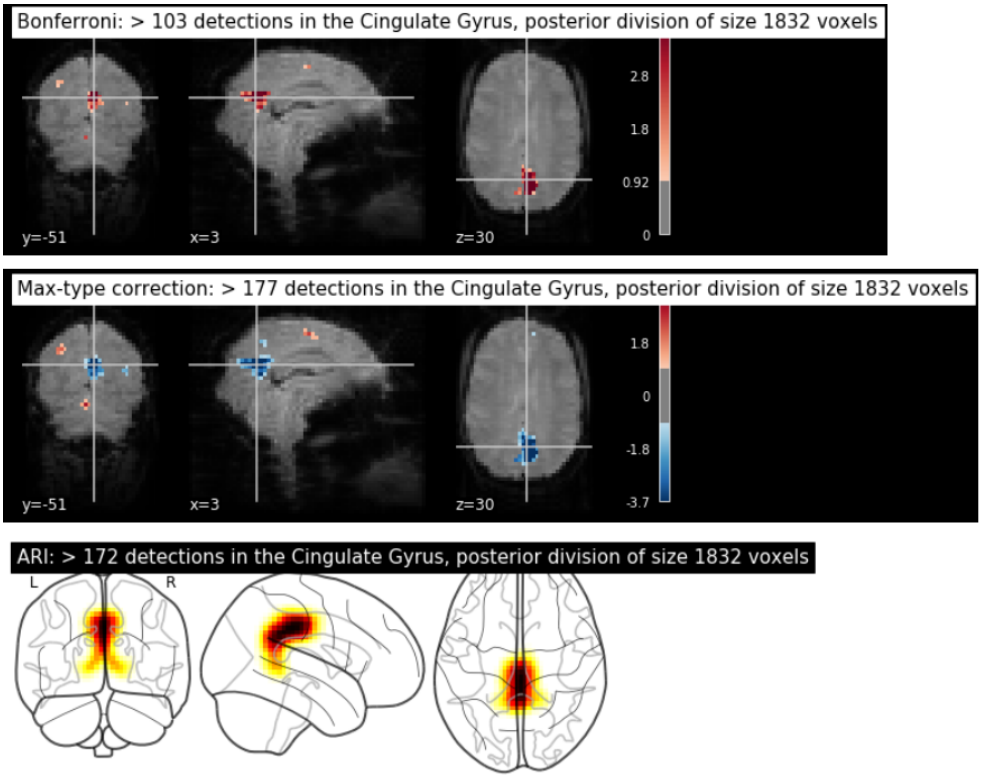

FIGURE 7 – Résultats sur le Gyrus Cingulaire

Les résultats d'ARI sur le dataset Haxby sont proches de la méthode "état de l'art" proposée dans Nilearn, et bien meilleurs que ceux de la correction de Bonferroni. Pour le Gyrus Cingulaire par exemple, la correction de Bonferroni donne au minimum 103 détections et la correction de l'état de l'art donne au minimum 177 détections. ARI donne au minimum 172 détections.

4 Limites de la méthode

Dans cette partie, nous allons exposer quelques limites de la méthode ARI et quelques points qui nous ont paru gênants dans l'article.

4.1 Limite des hypothèses et des a priori de la modélisation

Nous avons longtemps été gêné par le fait que le modèle statistique sous-jacent qui génère les p-valeurs utilisées n'est pas précisé : on suppose en début d'article qu'on dispose d'une p-valeur par voxel et il n'y a aucune mention de comment ces p-valeurs sont obtenues, ou encore de l'impact du modèle choisi sur les résultats. Ce point est important car selon le modèle choisi, le sens de l'inférence que l'on fait par la suite

peut être complètement différent. Après documentation, il s'avère que ce modèle est fait une combinaison de modèles englobant le traitement des séries temporelles de chaque voxels, la correction du mouvement, le co-recalage, la régularisation spatiale, et l'agrégation des résultats pour les différents individus. Toutes ces étapes sont à la fois cruciales mais également relativement bien standardisées. Nous ne tenons donc plus autant rigueur à la méthode ARI de ne pas s'intéresser aux étapes précédentes, même si nous pensons que tout scientifique s'intéressant à la méthode devrait prendre le temps de comprendre au moins la philosophie du processus de formation des p-valeurs. Mais il demeure que la méthode ARI, comme toutes les autres méthodes statistiques actuellement utilisées, n'est pas en mesure d'apporter des garanties sur le bon fonctionnement des étapes précédentes de la pipeline.

De plus, la méthode ARI contrôle uniquement l'erreur de type 1 et l'erreur de type 2 n'est jamais mentionnée. Peut-être que celle-ci est négligeable par rapport à l'erreur de type 1 dans les données de neuro-imagerie, mais ce point n'est jamais mentionné dans l'article. Contrôler l'erreur de type 1 a pour conséquence naturelle de faire un compromis sur l'erreur de type 2. Il nous semble alors qu'il serait bon d'utiliser une courbe ROC afin de vérifier la puissance du test pour différents seuils α. On pourrait également utiliser une fonction de perte asymétrique (loss function) afin quantifier la gravité relative des erreurs de type 1 et de type 2 dans l'optique d'optimiser α afin de minimiser la fonction de perte.

En outre, on peut se questionner sur la terminologie utilisée : qu'est-ce qu'un voxel "activé" ? Cette modélisation est très binaire : soit un voxel est activé, soit il est désactivé. Mais un voxel ne représente pas quelque chose de physique : un voxel d'un millimètre cube est une agrégation de plusieurs centaines de milliers de neurones. Par exemple l'abeille est dotée d'un « mini-cerveau » d'un millimètre cube, composé d'environ 960 000 neurones. Il y a 4 km d'interconnexions de réseaux neuronaux emballés dans chaque millimètre cube de matière grise. Ainsi, il nous semble très simplificateur de considérer qu'un voxel est activé de manière binaire. Effectivement, il nous semble pertinent de ne remonter l'information que pour une aire cérébrale pour ne pas se noyer dans trop d'information, mais le modèle statistique sous-jacent devrait pouvoir prendre en compte des activations continues pour les différents voxels, par exemple avec un modèle statistique bayésien manipulant les différentes probabilités d'activation.

4.2 Limite de l'applicabilité

La méthode ARI donne la possibilité de "drill-down", mais n'indique pas en pratique combien de fois le faire, ni comment le faire de manière intéressante ou optimale. Nous pensons qu'il faudrait sans doute trouver un moyen de standardiser le drill-down car la littérature nous semble contenir une trop grande variété d'approche.

Une autre critique de la méthode ARI est le fait qu'elle ne donne qu'une valeur pour tout le cluster et aucune valeur pour les clusters non sélectionnés. Il faudrait donc l'appliquer itérativement sur une partition du cerveau afin d'obtenir une carte de l'activation sur l'ensemble du cerveau, mais même de cette manière la carte résultante est constante par morceau sur chacune des parties de la partition.

4.3 Critique de la validation des résultats

Enfin, la méthode de validation des résultats est discutable : pour 218 individus présents dans l'étude à la base, 33 sujets sont sélectionnés comme sujets "test" et 66 sont sélectionnés pour validation. On pourrait déjà se poser la question : pourquoi ne pas sélectionner tout le reste du dataset (les 218 - 33 individus) pour validation ? De plus, les TDP (True Discovery Proportion, TDP + FDP = 1) rapportés semblent très élevés comparés à ceux obtenus pour nôtre implémentation. Même dans des régions fortement activées, on obtient des TDP d'au maximum 10%. Par exemple dans notre implémentation, nous avons trouvé 172 détections pour le Gyrus Cingulaire, région contenant 1832 voxels soit 172/1882 soit 9.1%, contre des TDP parfois à plus de 90% dans l'article.

En réalité, le TDP rapporté dans l'article est simplement le pourcentage de voxels en commun entre les clusters calculés sur le dataset de test et de validation. Ceci revient à faire l'hypothèse que le dataset de validation manipulé avec la méthode ARI est une vérité terrain ('ground-truth'), car dans ces conditions la proportion de voxels en commun est effectivement le TDP, puisque le dataset de validation représente les vraies détections. Or ceci est forcément faux, car cela présuppose que l'inférence ARI est une vérité terrain. Plutôt que de se comparer à la même méthode sur des données différentes, il nous semble qu'il aurait été plus pertinent de se comparer à une autre méthode qui a déjà fait ses preuves. Même dans le cas où l'on admet qu'il est acceptable d'utiliser la méthode ARI pour la validation, pourquoi dans ce cas ne pas utiliser l'ensemble des 218 - 33 sujets disponibles ?

Références

[1] Alexandre ABRAHAM et al. “Machine learning for neuroimaging with scikit-learn”. In : *Frontiers in neuroinformatics* 8 (2014), p. 14.

[2] Yoav BENJAMINI et Yosef HOCHBERG. “Controlling the false discovery rate : a practical and powerful approach to multiple testing”. In : *Journal of the Royal statistical society : series B (Methodological)* 57.1 (1995), p. 289-300.

[3] James V HAXBY et al. “Distributed and overlapping representations of faces and objects in ventral temporal cortex”. In : *Science* 293.5539 (2001), p. 2425-2430.

[4] Nikolaus KRIEGESKORTE et al. “Circular analysis in systems neuroscience : the dangers of double dipping”. In : *Nature neuroscience* 12.5 (2009), p. 535.

[5] Thomas NICHOLS et Satoru HAYASAKA. “Controlling the familywise error rate in functional neuroimaging : a comparative review”. In : *Statistical methods in medical research* 12.5 (2003), p. 419-446.

[6] Jonathan D ROSENBLATT et al. “All-resolutions inference for brain imaging”. In : *Neuroimage* 181 (2018), p. 786-796.