Etude de la variabilité inter-individuelle de données de connectivités intrinsèques : détection de réseaux instables et de sous-populations dans un tableau tridimensionnel

Loïc Labache [1,2,a], Marc Joliot [1,b], Jérôme Saracco [2,c] & Nathalie Tzourio-Mazoyer [1,d]

[1] *Groupe d'Imagerie Neurofonctionnelle CEA & IMN, UMR 5293, 146 rue Léo Saignat, 33000 Bordeaux, France*
[2] *CQFD team - Inria Bordeaux Sud Ouest & IMB, UMR 5251, ENSC - Bordeaux INP, 109 Avenue Roul, 33400 Talence, France*

[a] *loic.labache@u-bordeaux.fr*
[b] *marc.joliot@u-bordeaux.fr*
[c] *jerome.saracco@inria.fr*
[d] *nathalie.tzourio-mazoyer@u-bordeaux.fr*

Résumé. Nous proposons deux méthodologies permettant de mieux comprendre la variabilité inter-individuelle de données cérébrales d'Imagerie par Résonance Magnétique (IRM) fonctionnelle. Il s'agit de quantifier si le dendrogramme "moyen" est bien représentatif de la population initiale et d'identifier ses éventuelles sources d'instabilité. La première méthode permet d'identifier les réseaux pouvant conduire à des partitions instables du dendrogramme "moyen". La seconde approche permet d'identifier des sous-populations homogènes de sujets pour lesquelles leurs dendrogrammes "moyen" associés est plus stable que celui de la population initiale. Ces deux méthodes seront illustrées sur des données simulées à partir de données cérébrales de connectivités intrinsèques obtenues en IRM fonctionnelle. Les deux approches proposées pour détecter un réseau instable ou bien la présence de sous-populations ont montré un bon comportement numérique lorsque le niveau de bruit ne masque pas la structure des données.

Mots-clés. Classification hiérarchique, variabilité inter-individuelle, détection de sources d'instabilité.

Abstract. We propose two methodologies to better understand the inter-individual variability of functional Magnetic Resonance Imaging brain data. The aim is to quantify whether the "average" dendrogram is representative of the initial population and to identify its possible sources of instability. The first method identifies networks that can lead to unstable partitions of the "average" dendrogram. The second approach identifies homogeneous sub-populations of subjects for whom their associated "average" dendrograms are more stable than that of the original population. These two methods will be illustrated on simulated data from intrinsic connectivity data obtained by functional MRI. The two suggested approaches to detect an unstable network or the presence of sub-populations

have shown good numerical behavior when the noise level does not mask the structure of the data.

Keywords. Hierarchical clustering, inter-individual variability, detection of sources of instability.

1 Introduction

L'étude de l'organisation fonctionnelle cérébrale à l'état de repos chez l'humain consiste en l'étude de la synchronicité entre les signaux BOLD (Blood Oxygen Level Dependent) de différents réseaux cérébrales. L'étude de la synchronicité est réalisée à l'aide du calcul des coefficients de corrélation de Pearson entre les signaux BOLD de toutes les paires de réseaux, résultant en des matrices de corrélation de taille $K \times K$ pour K réseaux cérébrales. Il existe des bases de données (de plusieurs centaines d'individus) regroupant des structures 3D (de dimension $S \times K \times K$) qui rassemblent l'ensemble des matrices de corrélation $M_s,\ s = 1, ..., S$ des S individus de la base.

Pour une base fixée (BIL&GIN, voir Doucet et al., 2011), un premier travail a consisté en une classification (ascendante) hiérarchique des K réseaux cérébrales, puis d'identifier une partition en un nombre optimal de classes reflétant au mieux l'organisation cérébrale à l'état de repos. Plus précisément, la méthode consiste à moyenner l'ensemble des transformées de Fisher des matrices M_s de la manière suivante :

$$M_m = \tanh\left(\frac{\sum_{s=1}^{S} \operatorname{arctanh}(M_s)}{S}\right).$$

Cette matrice est transformée en matrice de dissimilarités D_m : $D_m = (1 - M_m)/2$. L'agrégation des K réseaux est ensuite réalisée via un algorithme de classification ascendante hiérarchique (méthode de Ward) basée sur D_m.

Afin de prendre en compte la variabilité inter-individuelle dans Doucet et al. (2011), nous avons adapté la procédure du package R `pvclust` de Suzuki et al. (2006), qui permet d'évaluer l'incertitude des différentes partitions issues de la classification hiérarchique à l'aide d'une p-value obtenue par rééchantillonnage. `pvclust` fournit une bootstrap probability p-value (BP-value) et après correction, une Approximately Unbiased p-value (AU-value). Les AU-value sont calculés en utilisant les valeurs de bootstrap préconisées dans Suzuki et al. (2006) ; soit en utilisant de 50% à 140% de l'échantillon. Ces p-values indiquent à quel point les différentes partitions sont supportées par les données.

Nous avons notamment retenu, au sein d'une population de 439 sujets, une partition "optimale" en 3 classes ayant respectivement des AU et BP-values de 100% (voir Figure 1). Suzuki et al. (2006) recommandant d'utiliser les AU-values, les 3 classes retenues sont alors parfaitement représentées par nos données. Cependant, les AU-values représentent

la stabilité des classes basées sur M_m et non pas la fréquence empirique d'apparition de ces classes à travers les S sujets, appelées fréquences individuelles ci-après. La Figure 1 illustre cet aspect. Nous pouvons voir l'évolution des BP-value en fonction de la proportion de sujets utilisés dans l'échantillonnage, ainsi que la valeur réelle des fréquences individuelles d'apparition d'une classe. Par exemple, pour une partition en 3 classes, nous constatons que les BP-values sont toutes les trois de 100% alors que les fréquences individuelles d'apparition sont respectivement de 5%, 4% et 12% pour ces 3 classes.

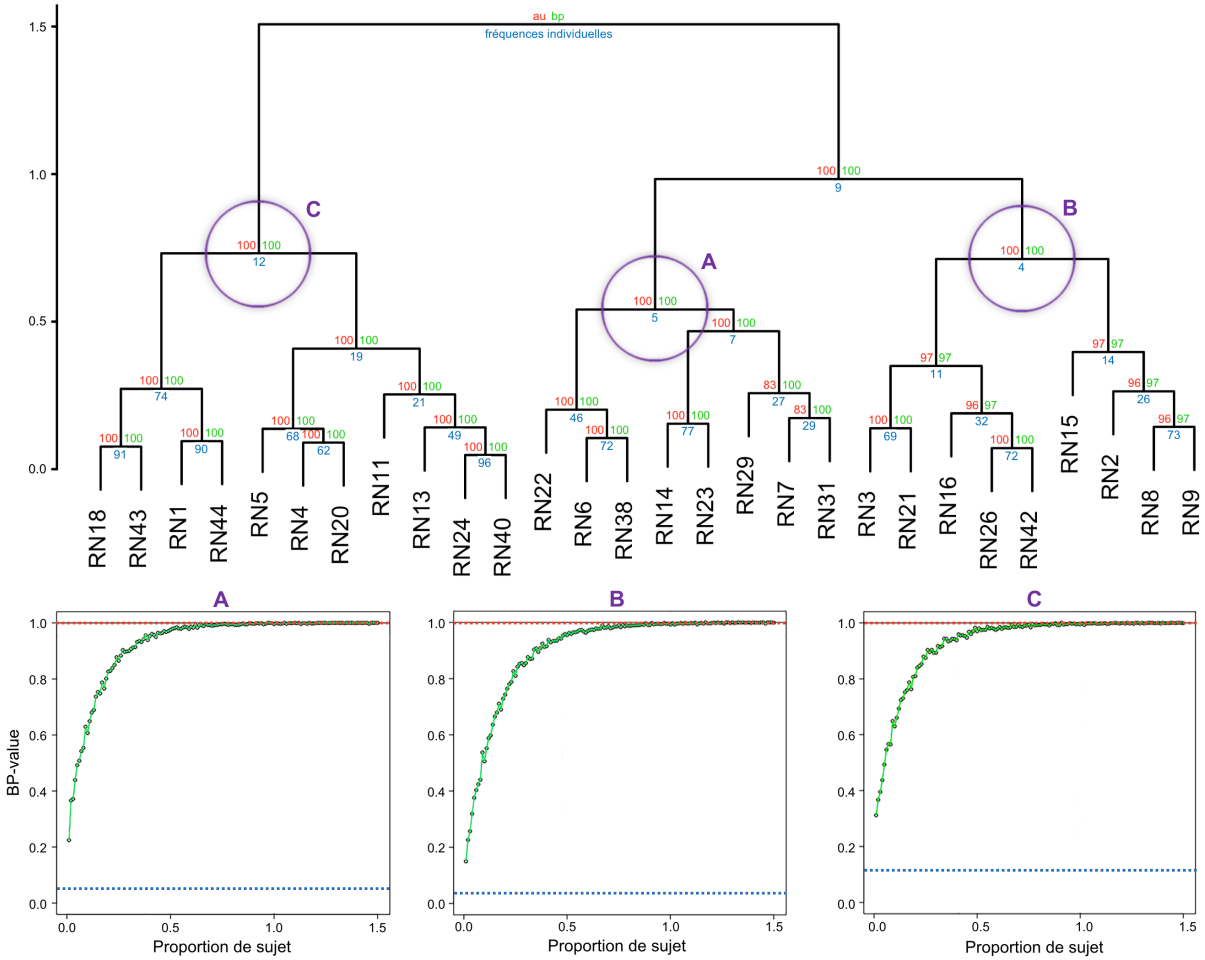

Figure 1: Classification ascendante hiérarchique issue de M_m (en haut). Evolution des BP-value en fonction de la proportion d'individus échantillonnés (en bas). En bleu, la fréquence individuelle des "classes". En vert, le BP-values. En rouge, les AU-values.

Nous proposons deux méthodologies permettant de mieux comprendre la variabilité inter-individuelle des partitions construites à partir de M_m :

- la première permettant d'identifier les **réseaux** pouvant conduire à des partitions issues de M_m instables ;
- la seconde permettant d'identifier, à travers la population, des sous-populations homogènes de **sujets** maximisant la fréquence individuelle d'apparition des classes issues des partitions construites sur leurs matrices M_m associées.

2 Présentation des approches proposées

Considérons les S matrices individuelles $K \times K$ de corrélation $M_s,\ s = 1, ..., S$.

Dans l'introduction, nous avons indiqué comment obtenir la classification ascendante hiérarchique à partir de la matrice "moyenne" M_m. On peut ainsi en déduire K partitions emboîtées $P_m^j,\ j = 1, \ldots, K$ que l'on peut résumer par des vecteurs de taille K définis de la manière suivante. Pour une partition en j classes, la $k^{\text{ème}}$ composante de P_m^j vaut 1 si le réseau k apparaît dans la nouvelle classe créée en passant d'une partition en $(j+1)$ classes à une partition en j classes. Dans la suite, on notera

$$P_m = \{P_m^1, \ldots, P_m^j, \ldots, P_m^K\}$$

avec $P_m^1 = (1, \ldots, 1)$ et $P_m^K = (0, \ldots, 0)$ par convention.

De la même manière, pour chaque sujet s, on peut obtenir à partir des matrices M_s les partitions

$$P_s = \{P_s^1, \ldots, P_s^j, \ldots, P_s^K\} \ \text{ pour } s = 1, \ldots, S.$$

Pour chaque partition non triviale[1] issues de M_m, nous allons chercher, parmi les $(K-2)$ partitions non triviales de chaque sujet s, l'élément de P_s le plus proche de P_m^j au sens de l'indice de Sorensen-Dice défini comme suit :

$$C_s^j = 1 - \min_{l=2,\ldots,K-1} \frac{2(P_m^j)' P_s^l}{(P_m^j)' P_m^j + (P_s^l)' P_s^l},$$

où la notation v' désigne la transposée du vecteur v. Un indice C_s^j nul (resp. égal à 1) indique que les vecteurs P_m^j et P_s^j égaux (resp. ne partagent aucun réseau commune).

Dans la suite, on appelle "pattern alternatif" à P_m^j, noté $\tilde{P}^j$, le (ou les) pattern(s) de P_s le plus proche de P_m^j au sens de l'indice C_s^j (si ce pattern n'est pas égal au pattern moyen P_m^j).

A partir de l'ensemble des $\tilde{P}^j$, il est alors possible de calculer un score Q_k de participation pour chaque réseau R_k. Q_k représente la fréquence empirique (exprimée en %) que le réseau R_k soit constitutif d'un pattern alternatif quel que soit le pattern moyen. Pour un pattern moyen P_m^j, on définit F^j, comme étant le nombre de sujet unique exprimant un

[1] La partition en une classe (resp. en K classe) est considérée comme triviale.

$\tilde{P}^j$, et F_k^j le nombre de sujet unique exprimant un $\tilde{P}^j$ contenant le réseau k. La fréquence empirique d'un réseau R_k est alors défini comme suit :

$$Q_k = \frac{\sum_{j=1}^{K-2} F_k^j}{\sum_{j=1}^{K-2} F^j}$$

Par la suite, la fréquence individuelle d'un pattern (moyen et alternatif) sera représentée comme à la Figure 2. Pour un pattern moyen P_m^j, il est possible de voir les patterns alternatifs $\tilde{P}^j$ associés, ainsi que leurs fréquences d'apparition à travers les S sujets.

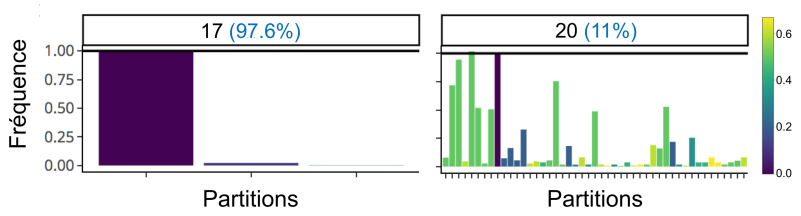

Figure 2: Exemple de représentation de la fréquence individuelle des pattern moyens et alternatifs. A gauche, P_m^{17} et ses 2 patterns alternatifs $\tilde{P}^{17}$. A droite, P_m^{20} et l'ensemble de ses patterns alternatifs $\tilde{P}^{20}$. Pour des raisons de lisibilité, la fréquence individuelle d'un pattern est relative à la fréquence individuelle du pattern moyen. La couleur reprénte la dissimilarité entre un pattern alternatif et le pattern moyen associé. Les P_m^j apparaissent en violet et ont une fréquence individuelle de 1. Le pourcentage en bleu dans le bandeau de chaque graphe indique la fréquence individuelle du pattern moyen à travers les S sujets.

Remarque. Un pattern alternatif de P_m^j ne peut jamais être égal à un pattern moyen P_m^l d'un autre niveau que $l \neq j$.

2.1 Méthode d'identification d'un réseau instable

A partir du score de participation Q_k de chaque réseau R_k, il est possible d'identifier parmi les K réseaux celui qui sera dit "instable", i.e. un réseau qui ne sera pas stable vis à vis de P_m à travers les sujets. Typiquement, un réseau R_{k^*} qui se "promème" sur le dendrogramme associé à P_m, sans avoir de rattachement fixe au sein des S dendrogrammes individuels P_s, conduit à un arbre moyen instable. Le réseau montrant l'instabilité la plus marquée est obtenu par :

$$k^* = \arg \max_{k=1,\ldots,K} Q_k.$$

Par convention, si $Q_1 = \cdots = Q_K$, on considère qu'il n'y a pas de réseaux instables.

2.2 Méthode d'identification de sous-populations de sujets

Définissons tout d'abord les fréquences individuelles des patterns moyen P_m^j du dendrogramme "moyen" P_m notées F_m^j comme étant le nombre de sujet unique exprimant P_m^j.

La méthodologie consiste à sélectionner de manière itérative les sujets appartenant au pattern $P_m^{j^*}$ tel que $j^* = \arg\min_{j=2,\ldots,K-1} F_j$. A chaque itération, les sujets sélectionnés forment alors une nouvelle sous-population, et la nouvelle population courante est composée des sujets non sélectionnés. Le critère d'arrêt est l'obtention d'un arbre dans la sous-population courante telle que $\forall j, F_m^j = 1$.

Cette procédure conduit ainsi à extraire itérativement les sujets les plus stables du dendrogramme "moyen" courant $P_m^{courant}$. Il est à noter que par construction, $P_m^{courant}$ va varier en terme d'agencement et de composition de ses partitions. Finalement, nous parviendrons à identifier l'existence de sous-population de sujets ayant un pattern qui diffère de manière constante en son sein, ou l'existence d'une sous-population n'ayant pas de structure forte en terme de dendrogramme.

3 Etude du comportement numérique

Nous allons illustrer le bon comportement numérique des deux méthodologies proposées sur des données simulées de "type IRM".

- Dans la première simulation, nous considérons une population de S sujets dont les matrices de corrélation $M_s,\ s = 1, \ldots, S$ sont "proches" (en terme de dendrogramme "moyen"), exceptées pour un réseau R_{k^*} qui est permutée aléatoirement pour chaque sujet. L'objectif est ici de détecter ce réseau instable qui rend le dendrogramme "moyen" P_m peu représentatif de la population des S sujets.

- Dans la seconde simulation, la population des S sujets est composée de deux sous-populations de sujets homogènes (en terme de dendrogramme "moyen"). L'objectif est ici d'identifier ces deux sous-populations.

Simulation des données. Afin de créer des sujets "proches" dans la population (ou sous-population), on se fixe une matrice de corrélation M que l'on va "bruiter" avec un bruit gaussien ε centré d'écart-type σ de la manière suivante : pour $1 \leq i < l \leq K$,

$$M_{(i,l)} = \tanh(\text{arctanh}(M_{(i,l)}) + e_{i,l}) \quad \text{et} \quad M_{(l,i)} = M_{(i,l)},$$

où $M_{(i,l}$ désigne l'élément (i, j) de la matrice M et $e_{i,l}$ est une réalisation du bruit ε.

Différents niveaux σ de bruit ont été considérés dans les simulations. Naturellement, plus le niveau de bruit est élevé, moins la population (ou sous-population) générée est homogène. Pour que l'étude par simulation ait de l'intérêt, nous avons choisi de considérer des valeurs de σ où le dendrogramme "moyen" est similaire au dendrogramme "moyen" non bruité.

3.1 Simulation 1 : identification du réseau instable

Dans cette partie, nous considérons tout d'abord un niveau de bruit $\sigma = 0.03$. Nous avons généré une population de $S = 500$ sujets à partir de la matrice de corrélation moyenne M_m présentée dans l'introduction et permuté aléatoirement[2] le réseau `RN14`. L'objectif est de détecter `RN14` comme étant le réseau instable à l'aide de la méthodologie proposée.

Dans le dendrogramme des données initiales (non bruitées et sans permutation aléatoire de le réseau `RN14`), le réseau `RN14` est agrégée avec le réseau `RN23`. Nous pouvons voir sur la Figure 3 (à gauche) que la présence des permutations (et du bruit) conduit le réseau `RN14` à changer de classe et se retrouve agrégée avec le réseau `RN15`. La Figure 3 (au centre) nous indique quelle partition contient le plus de patterns alternatifs. Nous pouvons voir que les patterns P^{20}, P^{22}, P^{23} et P^{24} sont les plus affectés, ce qui n'est pas surprenant puisque ces patterns sont ceux qui sont liés directement à `RN14` avant (pour P^{23}) ou après les permutations (pour les autres patterns). Enfin, la Figure 3 (à droite) montre que le réseau ayant le plus grand score Q_k est bien la région instable (permutée) `RN14` avec un score de 16.33. Notre méthodologie permet donc bien de détecter le réseau instable conduisant à l'instabilité du dendrogramme moyen. En sortant le réseau `RN14`, le nouveau dendrogramme moyen est alors bien plus stable avec seulement quelques patterns alternatifs "non dominants" (dus à la présence du bruit). La fréquence individuelle minimum pour un pattern moyen étant atteint pour $P_m^{20} = 58.2\%$. Par comparaison, la fréquence individuelle minimum dans les données initiale était de $P_m^{22} = 10.4\%$. La stabilité du dendrogramme moyen augmente alors d'un facteur 5.

La méthodologie proposée ayant pour but d'être appliquée à des données d'IRM, nous avons voulu voir, à l'aide de simulations, jusqu'à quel niveau σ de bruit l'approche était capable de détecter le réseau instable `RN14`. Nous avons alors considéré des données générées avec un niveau de bruit *sigma* variant entre 0 et 0.25. La Figure 4, montrant l'évolution des scores Q_k en fonction de σ, permet de voir que le réseau instable `RN14` a été correctement identifiée jusqu'à $\sigma = 0.13$. Par la suite, la méthodologie sélectionne le réseau `RN15` qui est le réseau partenaire de `RN14` suite à la permutation aléatoire et ce jusqu'à $\sigma = 0.19$. Finalement, pour $\sigma > 0.19$, nous pouvons voir qu'il y alternativement un quatuor de régions détectées comme instables par notre approche : `RN14`, `RN15`, `RN23` et `RN31`, l'approche proposée n'est naturellement plus capable d'identifier correctement `RN14` à cause d'un bruit trop important dans les données. Notons que lorsque σ devient grand (il n'y a plus d'information/de structure commune dans les données), tous les scores Q_k vont converger vers $1/k$, i.e. chaque réseau participe à de très nombreux patterns alternatifs.

3.2 Simulation 2 : identification de sous-populations homogènes

Dans cette partie, nous considérons un seul niveau de bruit $\sigma = 0.01$. Nous avons généré une population de $S = 500$ sujets répartis en deux sous-population A et B de 250 sujets

[2]i.e. permuter les valeurs de la ligne/colonne correspondante dans la matrice de corrélation

Régions	Q_k
RN13	0.00
RN24	0.00
RN40	0.00
RN40	0.02
RN1	0.02
RN18	0.02
RN43	0.02
RN44	0.02
RN26	1.57
RN42	1.57
RN4	1.59
RN5	1.59
RN20	1.69
RN7	1.69
RN38	2.02
RN6	2.02
RN2	3.68
RN22	3.68
RN16	4.02
RN29	4.37
RN8	5.26
RN23	5.26
RN11	5.75
RN31	6.06
RN21	6.56
RN3	6.62
RN9	7.90
RN15	10.66
RN14	16.33

Figure 3: A gauche : dendrogramme "moyen" (construit à partir de M_m) de la population des $S = 500$ sujets ayant eu le réseau `RN14` permutée aléatoirement. En violet, le numéro du pattern est indiqué. Au centre : graphique de la fréquence des "patterns alternatifs" pour chaque partition du dendrogramme "moyen". Une fréquence de 1 indique un nombre de sujets dans un "pattern alternatif" équivalent au nombre de sujets dans le "pattern moyen". La couleur indique la valeur de l'indice C. En bleu, la fréquence du pattern moyen. A droite : tableau des scores Q_k pour chaque réseau R_k.

chacune, ayant seulement quelques permutations de différence dans leurs dendrogrammes respectifs, voir Figure 4. Nous pouvons également noter l'impact du bruit sur les deux sous-population A et B : le bruit conduit à des partitions ayant une stabilité minimum de 73% (fréquence individuelle) pour la sous-population A et de 68% pour la sous-population B.

Avec l'ensemble des $S = 500$ sujets, la méthode proposée permet, au bout de 2 itérations, de retrouver convenablement les sous-populations A et B cachées dans les données simulées, voir la Figure 5.

- Le premier groupe identifié correspond à la sous-population A et contient 209 sujets, soit 84% de la sous-population initiale.
- Le second groupe, correspond à la sous-population B contient elle 171 sujets, soit 68% de la sous-population initiale.
- Les 120 sujets restants ne sont pas associés aux sous-populations A et B à cause du bruit introduit dans les matrices individuelles lors de la génération du jeu de données.

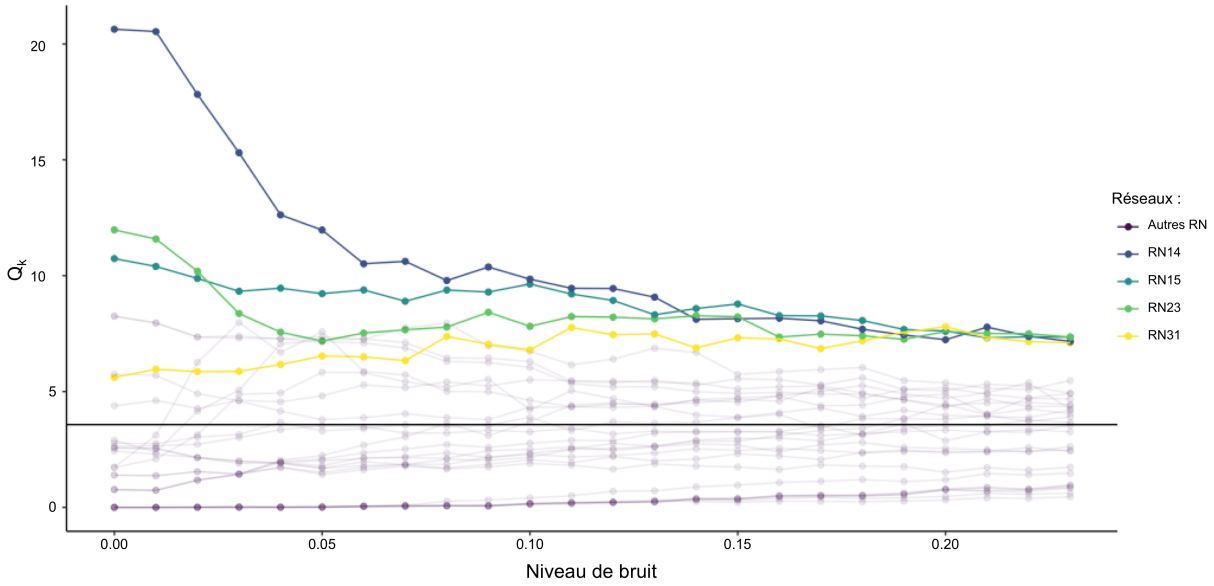

Figure 4: Evolution des scores Q_k pour chaque réseau R_k en fonction du niveau de bruit σ. La ligne horizontale noire représente le cas d'égalité des scores ($\forall k, Q_k = 1/K$).

Nous remarquerons que la plus petite fréquence individuelle pour un pattern moyen (voir Figure 5) est de 93% pour les sous-populations A et B, et de 0% pour le groupe des 120 sujets restants. Cet index nous donne une piste pour identifier le nombre de groupes (sous-populations) "cachés".

4 Brève conclusion et perspectives

Les deux approches proposées pour détecter un réseau instable ou bien la présence de sous-populations ont montré un bon comportement numérique sur données simulées (lorsque le niveau de bruit ne masque pas la structure des données). Une perspective nécessaire et naturelle est de combiner les deux approches afin de pouvoir détecter ces deux sources d'instabilité du dendrogramme "moyen" et ainsi de mieux comprendre la variabilité inter-individuelle dans les données cérébrales de connectivités intrinsèque obtenues en IRM fonctionnelle à la base de ce travail.

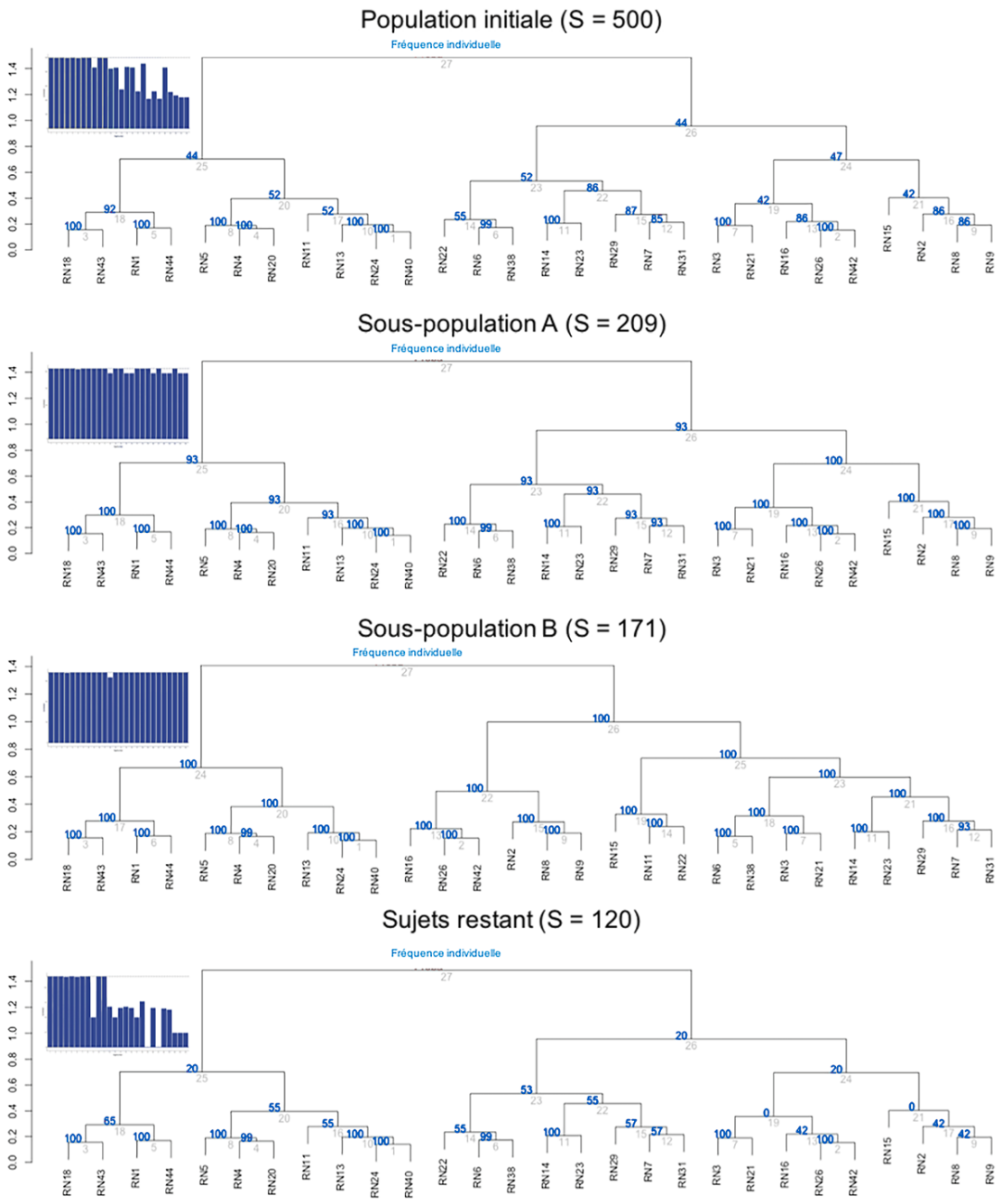

Figure 5: Les 2 dendrogrammes au centre correspondent aux sous-populations A et B identifiées par la méthodologie proposée, ayant respectivement 209 et 171 sujets sur les 250 initiaux. En bleu, la fréquence individuelle des patterns moyens pour un bruit $\sigma = 0.01$. En haut à gauche de chaque dendrogramme, le nombre de sujet pour chaque niveau de pattern. Les sous-populations A et B identifiées sont homogènes : tous les sujets expriment chaque pattern du dendrogramme, à l'inverse de la population initiale (en haut) et des 120 sujets restants (en bas).

Bibliographie

Doucet, G., Naveau, M., Petit, L., Delcroix, N., Zago, L., Crivello, F., Jobard, G., Tzourio-Mazoyer, N., Mazoyer, B., Mellet, E. and Joliot, M. (2011).Brain activity at rest: a multiscale hierarchical functional organization, *Journal of Neurophysiology*, 105(6), pp. 2753-2763.

Suzuki, R. and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering, *Bioinformatics*, 22(12), pp. 1540-1542.