Recherche de la connectivité de réseaux complexes. Application en IRMf.

Jérôme LEMOINE[1], Cédric GOUY-PAILLER[1], Sophie ACHARD[1], Pierre-Olivier AMBLARD[1]

[1]GIPSA-lab, Dépt. Images et Signal (UMR CNRS 5083)
ENSE3-BP 46 38402 Saint Martin d'Hères Cedex
prénom.nom@gipsa-lab.inpg.fr

Résumé – Nous considérons des processus aléatoires indexés par des graphes complexes. De la mesure des processus en chaque nœud du graphe, nous souhaitons identifier la connectivité du graphe. Nous avons recours à une analyse utilisant dans un premier temps les corrélations par paire pour déterminer les nœuds échangeant de l'information. Dans un deuxième temps, les paires corrélées sont examinées à la loupe des corrélations partielles pour éliminer les paires liées par des tiers. De plus, dans les applications concernées, les signaux mesurés présentent la propriété de longue mémoire. Pour lutter contre, nous utilisons une décomposition en ondelettes orthogonales. Enfin, cette méthodologie est appliquée à des données d'imagerie par résonance magnétique fonctionnelle.

Abstract – We consider random processes indexed by complex networks. A signal is recorded at each node of the network, and the problem adressed here is the recovery of the connectivity of the network from the signals measured. A first analysis consists in studying pairwise correlations to determine the nodes that share information. In a second step, correlated nodes are examined using partial correlation in order to eliminate cofounders. Moreover, all the analysis is performed on the wavelet coefficients for each band. The wavelet decomposition is adopted to eliminate the long memory property that characterizes the signals we study. To end, we apply the methodology to the analysis of data issued from functional magnetic resonance imaging of the brain.

1 Motivations

La plupart des grands réseaux étudiés aujourd'hui (informatiques, biologiques, sociaux, ...) présentent des caractéristiques communes qui les différencient des graphes déterministes simples (grilles) et des graphes aléatoires du type Erdös-Rényi [1, 2]. Ils constituent des intermédiaires, présentant des propriétés des graphes aléatoires (effet petit-monde par exemple) et des propriétés des grilles (fort clustering par exemple). Ces réseaux ont reçu la dénomination de réseaux complexes par les anglo-saxons (complex networks), traduit en graphes de terrain par la communauté francophone.

Les réseaux complexes font partie de la famille des systèmes complexes : systèmes comportant un grand nombre d'individus en interaction, robustes à l'élimination d'une partie, avec émergence de propriétés d'ensemble absentes chez les individus (auto-organisation) [3]. Des archétypes de ces systèmes sont les systèmes d'insectes (fourmis, abeilles,...), les grands réseaux informatiques, des systèmes physiques tels les tas de sable, ... Lorsque des mesures sont prises pour observer ces systèmes, il est courant d'obtenir des signaux non stationnaires, ayant des propriétés de longue mémoire, présentant des distributions à queues lourdes. Tel est le cas par exemple dans des mesures de résonance magnétique fonctionnelle (IRMf) ayant motivé ce travail. Nous disposons en effet de mesures de 90 zones du cerveau, chaque mesure étant une série temporelle de 2048 échantillons, chaque échantillon étant représentatif de l'activité de la zone cérébrale à un instant donné. Dans [4], la connectivité entre les 90 zones a été étudiée en utilisant la corrélation entre les séries temporelles. Précisément, ces séries présentant de la mémoire longue, la décomposition en ondelettes est utilisée pour détruire la longue dépendance. Les corrélations entre les séries sont alors calculées échelle par échelle, et un critère permet de décider de l'existence de connexions entre zones à partir de ces corrélations. Un graphe peut alors être construit, les nœuds correspondant aux aires cérébrales mesurées, et les arêtes entre nœuds représentant des connexions entre ces aires. Notons qu'une connexion existe si la corrélation entre les deux nœuds est forte. Or, deux zones cérébrales peuvent avoir des activités corrélées sans que des liens physiques existent entre elles, la corrélation existant à cause d'une interaction commune de ces deux nœuds avec un nœud tiers.

Dans ce travail, nous proposons d'étendre la technique en considérant non plus les corrélations, mais les corrélations partielles échelle par échelle entre les nœuds. Notons que l'utilisation de la corrélation partielle sur des transformées de signaux n'est pas nouvelle, bien qu'assez récente (voir par exemple [5, 6] pour l'utilisation des cohérences partielles). Toutefois et à notre connaissance, l'extension à l'analyse en ondelette n'existe pas. L'intérêt de l'analyse en ondelette est double. Premièrement, le caractère longue mémoire des données à disposition est anihilé si l'ondelette est bien choisie. Deuxièmement, le caractère bande passante de l'analyse permet d'étudier les in-

teractions dans certaines bandes de fréquence pertinentes pour les applications en neuroscience.

Dans la suite, nous détaillons la technique, montrons son comportement face à des données simulées, puis l'illustrons en montrant des premiers résultats sur les données réelles.

2 Etude de la connectivité par corrélation partielle des coefficients en ondelettes.

On considère un graphe $G = (V, E)$, V étant l'ensemble des nœuds et E l'ensemble des arêtes. Chaque nœud γ porte un processus aléatoire $x_\gamma(t)$. Si l'on note $|V|$ le cardinal de V, l'ensemble des $|V|$ processus constitue un processus multidimensionnel. On suppose que la structure de dépendance entre les processus $x_\gamma(t)$ est reflétée par des liens entre les nœuds du graphe. Ceci définit un modèle graphique [7, 8]. Deux nœuds γ et μ du graphe ne sont pas liés par une arête si les signaux x_γ et x_μ sont indépendants conditionnellement aux autres signaux $x_\beta(t), \beta \in V\backslash(\gamma, \mu)$. A défaut de tester l'indépendance conditionnelle, nous nous contentons ici d'étudier la décorrélation partielle.

La démarche suivie débute par le calcul d'une décomposition en ondelette de chaque signal $x_\gamma(t)$. On obtient à chaque échelle j de décomposition, une suite de coefficients $d_\gamma(j, k)$, où k indexe le temps. La corrélation partielle entre deux signaux est définie comme la corrélation entre les résidus des régressions linéaires des deux signaux sur les autres signaux. Soit $\mathcal{H}^{j,k}_{V\backslash(\gamma,\mu)}$ l'espace vectoriel engendré par $d_\beta(j, k), \beta \in V\backslash(\gamma, \mu)$. Le résidu de la régression linéaire de $d_\mu(j, k)$ sur $\mathcal{H}^{j,k}_{V\backslash(\gamma,\mu)}$ est $\varepsilon_\mu(j, k) = d_\mu(j, k) - \mathcal{P}\left(d_\mu(j, k)\middle|\mathcal{H}^{j,k}_{V\backslash(\gamma,\mu)}\right)$ où $\mathcal{P}\left(X\middle|\mathcal{H}\right)$ est la projection orthogonale de la variable X sur l'espace $\mathcal{H}$. La corrélation partielle entre les séries μ et γ à l'échelle j est alors définie par

$$C_{j,k}(\mu, \gamma) = \text{Cov}\left[\varepsilon_\mu(j, k), \varepsilon_\gamma(j, k)\right]$$

Un résultat fondamental de l'analyse des statistiques multivariées montre que les corrélations partielles sont données par l'opposé des éléments extra-diagonaux de la matrice de corrélation inverse [7]. Deux stratégies de calcul sont donc à disposition : évaluation des corrélations des résidus dans les régressions linéaires ou estimation de l'inverse de la matrice de corrélation. En pratique, les corrélations usuelles et partielles sont normalisées et prennent leurs valeurs dans $[-1; 1]$.

Pour étudier la connectivité entre des signaux aléatoires, on commence par mesurer les corrélations entre les signaux. Chaque signal est décomposé en coefficients d'ondelette. Echelle par échelle, la matrice de corrélation $\Gamma^j_{\mu,\gamma} = \text{Cov}\left[d_\mu(j, k), d_\gamma(j, k)\right]$ est estimée (le moyennage est effectué en sommant sur le paramètre temporel k des coefficients), et une corrélation forte à une échelle entre les indices μ et γ est interprétée comme un lien entre les nœuds correspondant du graphe (pour la notion de corrélation forte, nous renvoyons à [4] où les tests statistiques sont explicités.) La procédure se poursuit alors en calculant les corrélations partielles entre les nœuds liés par la corrélation. Un test sur la nullité de la corrélation partielle est effectué et permet de conserver ou non le lien entre les deux nœuds considérés.

Pour effectuer pratiquement le test, nous avons calculé un intervalle de confiance pour l'estimateur de la corrélation partielle des coefficient en ondelette. La détermination suit les développements effectués dans [9] pour la corrélation. Nous utilisons la transformation de Fisher. Cette transformation permet de se rapprocher de l'hypothèse gaussienne pour des petites tailles d'échantillon. Sous l'hypothèse que les coefficients d'ondelettes forment un processus multivarié gaussien stationnaire avec un spectre de carré intégrable, à l'échelle j, on montre que l'intervalle de confiance estimé à $100(1 - 2p)\%$ de $C_{j,k}(\mu, \gamma)$ s'écrit,

$$\left[\tanh\left\{h[C_{j,k}(\mu,\gamma)] - \frac{\Phi^{-1}(1-p)}{\sqrt{N_j - L'_j - 3}}\right\},\ \tanh\left\{h[C_{j,k}(\mu,\gamma)] + \frac{\Phi^{-1}(1-p)}{\sqrt{N_j - L'_j - 3}}\right\}\right]$$

où $h(x) = \tanh^{-1}(x)$, N_j est le nombre de coefficient d'ondelettes à l'échelle j et $L'_j = \lceil (L-2)(1-2^{-j})\rceil$, L est la longueur du filtre de la transformée en ondelette utilisé. Φ représente la fonction erreur.

Pratiquement, l'hypothèse nulle est la nullité de la corrélation partielle. Si l'on fixe le taux de fausse alarme à $p\%$, l'alternative est acceptée lorsque l'estimateur sort de l'intervalle précédemment défini, avec $C_{j,k}(\mu, \gamma) = 0$.

Nous illustrons maintenant sur des graphes simulés le fonctionnement de cette méthodologie.

3 Recherche de connectivité dans des graphes aléatoires simulés.

Trois processus liés. Nous avons mené deux simulations de trois processus à différence finie corrélés entre eux par les matrices de corrélation suivantes

$$A = \begin{bmatrix} 1 & -0.5 & 0.5 \\ -0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{bmatrix} B = \begin{bmatrix} 1 & 0.57 & -0.57 \\ 0.57 & 1 & -0.33 \\ -0.57 & -0.33 & 1 \end{bmatrix}$$

Notons que ces matrices sont inverses l'une de l'autre de sorte que l'une constitue la matrice des opposées des corrélations partielles de l'autre.

Ces processus simulés présentent la propriété de longue mémoire et les coefficients d'ondelettes forment un processus multivarié gaussien stationnaire avec un spectre de carré intégrable, [10]. Nous avons simulés 1000 réalisations de ces réseaux de trois processus pour dix tailles de signaux différentes, $N = 1024, 2048, 3072, 4096, 5120, 6144, 7168, 8192, 12288, 16384$.

Chacune des séries est décomposée en ondelette, et nous avons testé la méthodologie sur l'échelle 3. Les résultats empiriques et théoriques sont représentés dans la figure 3. On superpose à l'analyse statistique l'intervalle de confiance à 99.5 et 50 % : l'accord est remarquable et valide l'utilisation de l'intervalle de confiance pour le choix du seuil à probabilité de fausse alarme donnée.

Deux clusters liés. Afin de comprendre le comportement de la corrélation partielle dans le cas de réseaux complexes de taille plus importantes, on simule à présent deux clusters indépendants de trois variables chacun. Les signaux des points appartenant au premier cluster suivent une distribution normale multivariée de moyenne nulle et de matrice de covariance Σ_x telle qu'indiquée dans la figure (2). De même, les signaux du deuxième cluster suivent une loi normale multivariée de moyenne nulle et de variance Σ_y (*cf.* figure 2). Enfin, on introduit une variable z telle que $z(t) = x_1(t) + y_1(t) + 0,2 \times n(t)$ où $n \sim \mathcal{N}(0;1)$. De même que précédemment, tous les signaux sont blancs. On estime alors les liens mesurés dans ce réseau à l'aide de la corrélation et de la corrélation partielle. La figure (2) rend compte des valeurs des liens mesurés (épaisseur des traits) ainsi que du signe de la mesure. Alors que la corrélation rend bien compte des liens à l'intérieur de chaque cluster, l'introduction de z s'accompagne d'un ensemble de liens indésirables entre z et les nœuds fortement corrélés à x_1 ou y_1. L'estimation des liens entre les nœuds par corrélation partielle est quant à elle moins perturbée par l'apparition de z dans le réseau. Notons cependant l'apparition indésirable d'une corrélation partielle forte entre x_1 et y_1 due à la relation existant entre ces trois variables : en effet, bien que marginalement indépendantes, les variables x_1 et y_1 ne sont pas conditionnellement indépendantes puisque liées par z.

L'utilisation de la corrélation en conjonction avec la corrélation partielle permet donc de s'affranchir d'un ensemble de liens indésirables dans les réseaux complexes. On doit donc dorénavant envisager une procédure en deux étapes :

1. sélection des liens pour lesquels la corrélation est forte ;
2. rejet des liens sélectionnés dont la corrélation partielle est faible.

Avec cette démarche, les liens indépendants marginalement ne sont pas considérés dans la deuxième étape, et n'apparaissent pas sur le graphe final. Notons que ce graphe n'est donc pas en général un modèle graphique pour les données.

4 Données réelles, enregistrement d'IRMf

L'imagerie par résonance magnétique présente l'intérêt d'enregistrer l'activité cérébrale (par l'intermédiaire du signal BOLD, Blood Oxygen Level Dependent) de petits volumes (voxels) du cerveau. Après avoir regroupé en 90 régions anatomiques les voxels, on extrait 90 séries temporelles décrivant l'activité cérébrale de chacune de ces régions (voir [4] pour une description plus complète). Les données réelles utilisées dans ce papier sont téléchargeables sur le site :
http ://fs2.psychiatry.cam.ac.uk/∼js369/nihrest
Ce sont des données acquises avec une machine IRMf de 3 Tesla. Les volontaires avaient pour consignes de rester calmes avec les yeux fermés. Dans cette étude nous avons analysé un jeu de donnée provenant d'un seul individu. Ces données comprennent 2048 images de 64x64x20 volumes. Après un prétraitement comprenant la correction des mouvements et la normalisation par le volume référence MNI, on extrait les 90 séries temporelles étudiées. Sur la base de ces séries, on calcule tout d'abord la mesure de corrélation par ondelettes entre paires de séries temporelles pour sélectionner les paires de régions possédant une connexion significative. Puis, pour chacune de ces paires, on effectue un test sur la corrélation partielle. La figure (3) présente pour la bande de fréquence 0.06-0.11Hz, les 71 corrélations les plus fortes calculées sur les données réelles. Ceci correspond à prendre les valeurs absolues des corrélations significativement supérieures à 0.55. Parmi ces 70 corrélations, 13 ont été détectées comme possédant une corrélation partielle nulle. Pour effectuer ce test, nous avons utilisé la méthodologie précédemment développée, avec un taux de fausse alarme de 0.1. Notons que les liens rejetés appartiennent à une zone fortement connectée. L'utilisation de la corrélation partielle permet de mettre en évidence des causes communes au sein de ces régions denses.

Références

[1] A. BARRAT, M. BARTHELEMY, and A. VESPIGNIANI. *Dynamics on complex networks*. Cambridge University Press, 2008.

[2] M. E. J. NEWMAN. The structure and function of complex networks. *SIAM reviews*, 45 :167–256, 2003.

[3] N. BOCCARA. *Modeling complex systems*. Springer, 2004.

[4] S. ACHARD, R. SALAVADOR, B. WHITCHER, J. SUCKLING, and E. BULLMORE. A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *J. Neurosci.*, 26 :63–72, 2006.

[5] R. DAHLHAUS, M. EICHLER, and J. SANDKUHLER. Identification of synaptic connections in neural ensembles by graphical models. *Journal of neuroscience methods*, 77 :93–107, 1997.

[6] R. DAHLHAUS. Graphical interaction models for multivariate time series. *Metrika*, 51 :157–172, 2000.

[7] J. WHITTAKER. *Graphical models in applied multivariate statistics*. Wiley&Sons, 1989.

[8] S. LAURITZEN. *Graphical models*. Oxford University Press, 1996.

[9] B. WHITCHER, P. GUTTORP, and D. B. PERCIVAL. Wavelet analysis of covariance with application to atmospheric time series. *Journal of Geophysical Research*, 14 :941–962, 2000.

[10] M. J. CHAMBERS. The simulation of random vector time series with given spectrum. *Mathematical and Computer Modelling*, 22 :1–6, 1995.

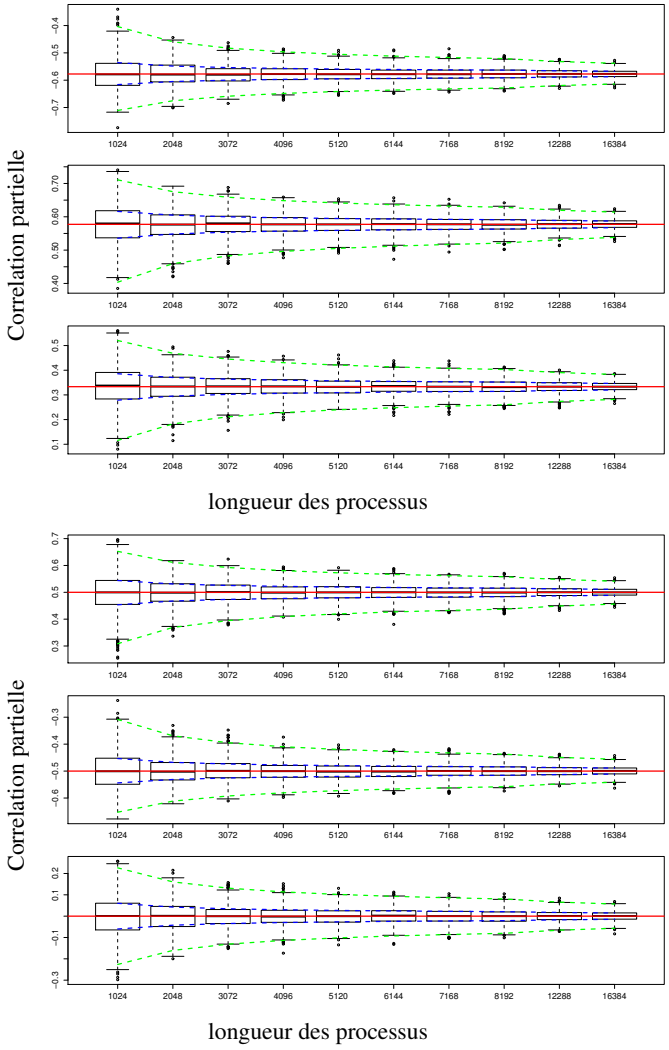

FIG. 1 – Représentation de la corrélation partielle entre 3 processus à différence finie simulés. En noir, les boites à moustache représentent la dispersion de la corrélation partielle obtenue pour 1000 réalisations. Le trait en gras correspond à la médiane, les traits en continus correpondent à 25% et 75% de l'échantillon, et les barres horizontales correspondent à 0.5% et 99.5% de l'échantillon. La droite rouge correpond à la valeur théorique de la corrélation partielle. Les droites bleues correpondent à un intervalle de confiance de 50% et les droites vertes correspondent à un intervalle de confiance de 99%. La figure supérieure correspond au couplage par la matrice de corrélation A, voir le texte, alors que la figure du bas correspond au couplage par la matrice de corrélation B.

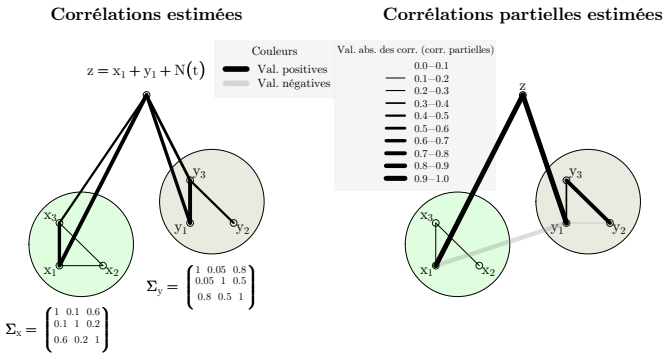

FIG. 2 – Recherche de liens entre deux clusters des noeuds corrélés. A gauche, l'épaisseur des liens correspond à la corrélation entre les nœuds considérés. A droite, l'épaisseur correspond à la corrélation partielle entre les nœuds considérés. On note l'apparition d'un lien à gauche entre x_1 et y_1 dû à la non indépendance conditionnelle de ces deux variables. On constate également la disparition des liens dûs à des causes communes.

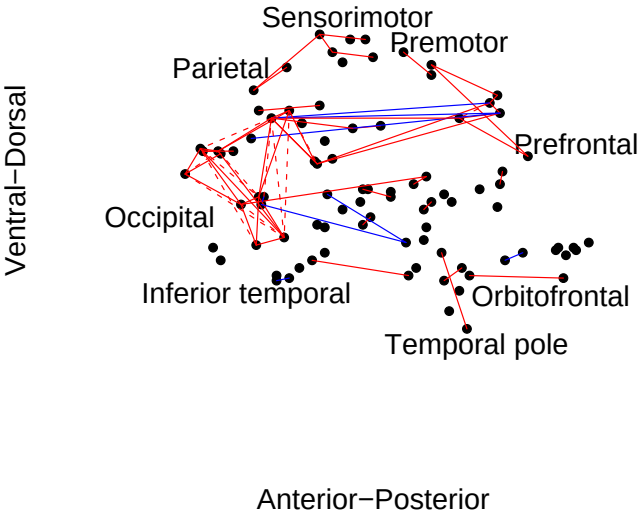

FIG. 3 – Illustration de liens trouvés par corrélation mais rejetés par corrélation partielle dans des données de fMRI en utilisant au préalable une décomposition en ondelettes. Les résultats présentés concernent la bande fréquentielle 0.06-0.11 Hz. Les liens rejetés apparaissent en pointillés sur ce graphe. Ils appartiennent à des zones fortement connectées.