Sigcor – un paquet Python pour le calcul de seuils de coefficients de corrélation significatifs pour des petits et grands échantillons

Guillaume BECQ

University Grenoble Alpes, Laboratoire Gipsa-Lab, UMR 5216 CNRS, Grenoble-INP, Grenoble, France

Résumé – Dans cet article, le paquet Python Sigcor est présenté. Il propose des fonctions pour le calcul des seuils de coefficients de corrélation significatifs entre signaux. En particulier, il permet d'obtenir des valeurs précises lorsque le nombre d'échantillons des signaux est faible ou lorsque les signaux sont filtrés. Les fonctions du paquet sont présentées et leurs intérêts pratiques sont évalués.

Abstract – In this article, the Python package Sigcor is presented. Its aim is to propose functions for computing thresholds of significant correlation coefficients between signals. Computations are not using approximated equations, thus enabling to obtain precise values for signals with small samples or when signals are filtered. Functions of the package are presented putting in evidence their practical interests.

1 Introduction

Lors de l'étude du cerveau par imagerie par résonnance magnétique (IRM) nucléaire, une modalité permet d'enregistrer la variation au cours du temps de l'activité dans différentes régions du cerveau. Les signaux obtenus varient en fonction de la fonction réalisée par le cerveau. On parle alors d'IRM fonctionnelle. Il est alors courant d'analyser les connectivités fonctionnelles (functional connectivities – FC) entre régions du cerveau, celles-ci étant généralement obtenues à partir des corrélations empiriques entre signaux de différentes régions [1]. Cependant, si les acquisitions sur un cerveau entier permettent d'obtenir une bonne résolution spatiale, elles sont obtenues au détriment d'une fréquence d'acquisition faible, souvent poussée à la limite technologique des machines, et sur de courtes durées afin de respecter le confort des sujets étudiés et ne pas modifier la fonction étudiée. Les signaux contiennent donc finalement peu d'échantillons temporels. Ils sont aussi filtrés pour permettre de regarder dans des bandes de fréquence d'intérêt. On se demande alors quelles sont les valeurs de corrélations qui sont significatives, au sens où elles ne sont pas obtenues par hasard et générées par le bruit important contenu dans ces enregistrements [3].

L'étude des coefficients de corrélation a été inroduite par Galton, Pearson et Fisher à la fin du XIXème siècle et au début du XXème siècle, les lois théoriques des distributions des corrélations empiriques pour des variables gaussiennes ont été formulées, et des tables pour des petits échantillons ont été proposées par David [5]. Plus récemment, Muirhead pointe sur ces résultats et en regroupe de plus généraux dans son ouvrage « Aspects of multivariate statistical theory » [7]. Lors du dernier Gretsi, nous avons présenté l'effet du filtrage sur ces corrélations et présenté des courbes issues de simulation [2]. Nous sommes parti des résultats de Witcher, qui utilise des ondelettes et propose des bornes qui tiennent compte des coefficients d'échelles [8, 9]. Cependant, celui-ci utilise des transformations de Fisher, pour déterminer des seuils de corrélations d'ondelettes significatives et donc une approximation des lois théoriques. Le paquet Python `sigcor` que nous présentons ici permet de combler plusieurs besoins : il propose de calculer les seuils de corrélations significatives en tenant compte des petits échantillons ; il permet d'utiliser des valeurs issues des lois théoriques ; il prend en compte l'effet du filtrage dans une bande de fréquence donnée. La version 1.0 du paquet est accessible sur zenodo à l'adresse `https://doi.org/10.5281/zenodo.15115356`.

Dans un premier temps, les notations utilisées sont introduites et la fonction permettant d'obtenir les distributions empiriques est présentée. Ensuite, les fonctions permettant d'obtenir les valeurs seuils en fonctions des échantillons sont ensuite étudiées, en tenant compte ou non de l'effet du filtrage. Enfin une comparaison aux tables obtenues par David [5] est proposée.

2 Présentations des fonctions de `sigcor`

2.1 Etude de la distribution des corrélations empiriques

Soit N le nombre d'échantillons de deux signaux X et Y. On note $n = N - 1$. La correlation empirique r entre X et Y est donnée par :

$$r = \frac{\sum_{i=1}^{N}(X_i - \bar{X})\,(Y_i - \bar{Y})}{\left[\sum_{i=1}^{N}(X_i - \bar{X})^2 \sum_{i=1}^{N}(Y_i - \bar{Y})^2\right]^{1/2}} \qquad (1)$$

avec $\bar{X}$ et $\bar{Y}$ les moyennes empiriques de X et Y.

Lorsque les échantillons sont issus des variables aléatoires indépendantes X et Y de densité jointe $f(x,y) = \frac{1}{2\,\pi\,\sigma_1\,\sigma_2\,\sqrt{(1-\rho^2)}}\,\exp\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\,\rho(x-\mu_1)\,(y-\mu_2)}{\sigma_1\,\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2})\right)$, X et Y suivant les lois normales $f(x) = \frac{1}{\sqrt{2\,\pi}\,\sigma_1}\,\exp\left(-\frac{1}{2}\,\frac{(x-\mu_1)^2}{\sigma_1^2}\right)$ et $f(y) = \frac{1}{\sqrt{2\,\pi}\,\sigma_2}\,\exp\left(-\frac{1}{2}\,\frac{(y-\mu_2)^2}{\sigma_2^2}\right)$, la densité de fonction de r

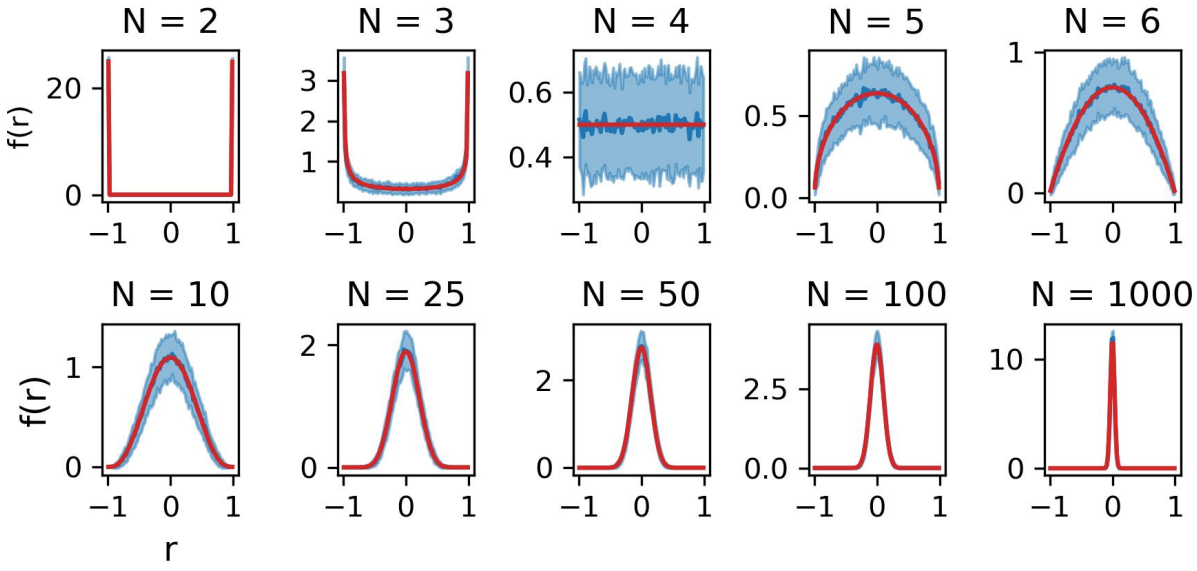

FIGURE 1 : Densité de probabilité du coefficient de corrélation empirique r pour des variables indépendantes et non corrélées qui suivent une distribution normale, en fonction du nombre d'échantillons disponibles N pour estimer r : en rouge, courbes théoriques, en bleu, histogrammes empiriques, moyennes et écart-types réalisés sur 100 histogrammes obtenus à partir de 1000 tirages aléatoires. Les histogrammes sont réalisés sur 100 points dans l'intervalle $[-1, 1]$.

est donnée pour $\rho = 0$ par (voir [7] p.147) :

$$\begin{aligned} f(r) &= \frac{\Gamma(\frac{1}{2}n)}{\sqrt{\pi}\,\Gamma(\frac{1}{2}(n-1))}\,(1-r^2)^{\frac{(n-3)}{2}} \\ &= \frac{\Gamma(\frac{1}{2}n)}{\Gamma(\frac{1}{2})\,\Gamma(\frac{1}{2}(n-1))}\,(1-r^2)^{\frac{(n-3)}{2}} \\ &= \frac{1}{\mathrm{B}(\frac{1}{2},\frac{1}{2}(n-1))}\,(1-r^2)^{\frac{(n-3)}{2}} \end{aligned} \tag{2}$$

avec $\mathrm{B}(z_1, z_2)$ la fonction beta et $\Gamma(z)$ la fonction gamma, ou intégrales d'Euler de première et seconde espèces.

La fonction `sigcor.get_fr(r, N)` permet d'obtenir la valeur de la densité de probabilité pour une valeur de r et N échantillons en utilisant l'équation 2, i.e. pour $\rho = 0$. Elle permet d'obtenir plusieurs distributions qui sont proposés Fig. 1. Une comparaison avec les tracés expérimentaux obtenus à partir de simulations réalisées sur des signaux gaussiens non correlés est aussi proposée. Sur ces exemples, les estimations empiriques sont réalisées sur N=1000 échantillons, 100 réalisations, et les histogrammes calculés sur 100 intervalles entre -1 et 1. Le calcul des valeurs théoriques pour $N = 3$, $n = 2$ a été approximé au voisinage de $r = -1$ et $r = 1$, en remarquant que :

$$\begin{aligned} f(r) &= \frac{1}{\mathrm{B}(\frac{1}{2},\frac{1}{2})}\,\frac{1}{(1-r^2)^{\frac{1}{2}}} \\ f(r) &= \frac{1}{\pi}\,\frac{1}{(1-r^2)^{\frac{1}{2}}} \end{aligned} \tag{3}$$

En prenant $r = -1 + \epsilon$ ou $r = 1 - \epsilon$, on approxime par :

$$\frac{1}{(1-r^2)^{1/2}} = \frac{1}{((\epsilon\,(1-\epsilon))^{1/2}} \approx \frac{1}{(\epsilon)^{1/2}} \tag{4}$$

dont l'intégration conduit à $P_\epsilon = P(1-\epsilon < X \leq 1) = P(-1 < X \leq -1+\epsilon) = \frac{1}{\pi}\,2\epsilon^{1/2}$ soit en prenant $\epsilon = dx/2$, $P_{dx/2} = \frac{1}{\pi}\sqrt{2}\,\sqrt{dx}$

2.2 Calcul des seuils pour les coefficients de corrélation significatifs

Soit H0 l'hypothèse suivante : la valeur du coefficient de corrélation empirique entre deux signaux contenant N échantillons est issue d'une distribution de variables aléatoires indépendantes non corrélées. On rejette l'hypothèse au seuil α, pour un test bilatéral, lorsque $p = P(|R| \geq |r||H0) \leq \alpha/2$. La valeur de r miminale pour laquelle cette condition est vérifiée est notée r^*. Pour un test unilateral, on rejette l'hypothèse si $p = P(R \geq r|H0) \leq \alpha$ ou $p = P(R \leq r|H0) \leq \alpha$. L'équation donnant r^* est la suivante (données dans [7] p.147) :

$$r^* = \frac{t^*_{n-1}(p)}{(n-1+t^{*2}_{n-1}(p))^{1/2}} \tag{5}$$

avec $t^*_\nu(p)$ le quantile d'ordre p de la distribution t de Student de paramètre ν vérifiant la limite du test.

La fonction `sigcor.get_rs(N, alpha, twosided=True)` permet d'obtenir le seuil de significativité en fonction du nombre d'échantillons des signaux. Par défaut le calcul se fait sur un test d'hypothèse symétrique mais la valeur à `False` permet d'utiliser un test unilatéral. La figure Fig.2.a propose les évolutions de r^* en fonction du seuil de significativité α et du nombre d'échantillons obtenu par cette fonction. La figure Fig.2.b permet de vérifier que les valeurs expérimentales obtenues par simulation sont pour la plupart confondues avec les valeurs théoriques calculées par `get_rs`. On utilise les mêmes paramètres que précédemment pour les simulations. La figure Fig.2.c permet de comparer les résultats théoriques aux valeurs obtenues en utilisant des lois normales et des transformations de Fisher, qui sont utilisées pour des valeurs de N grandes et des seuils de significativé grand. Les différences sont en effet faibles. Ces valeurs peuvent être obtenues à partir de `sigcor.get_rs_fisher(N, alpha,`

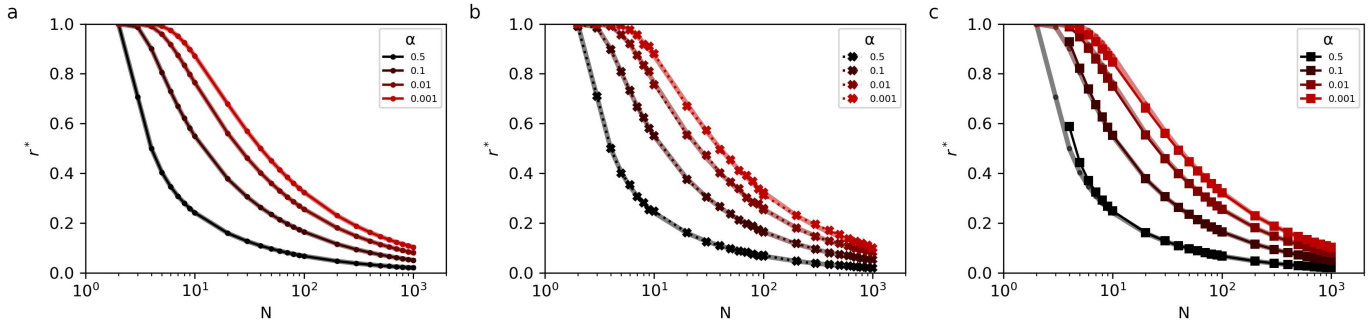

FIGURE 2 : Valeurs du seuil de significativité r^* en fonction du nombre d'échantillons des signaux N : a) courbes théoriques paramétrées par α le seuil de significativité ; b) courbes expérimentales en pointillés obtenue par simulation et comparées au tracé théorique ; c) courbes théoriques en utilisant l'approximation de Fisher en ligne fine, comparée aux courbes théoriques.

`twosided=true`).

Lorsque les signaux sont filtrés, une correction doit être réalisée sur le nombre d'échantillons utilisé dans l'équation 2 [8, 9, 2]. Au lieu d'utiliser $n = N - 1$ directement dans l'équation 2 ou 5, on doit remplacer N par N_f :

$$N_f = \lfloor 2\,B\,T \rfloor = \lfloor \kappa N \rfloor \tag{6}$$

avec B la bande passante du signal, T la durée d'observation des signaux, soit, si F_s est la fréquence d'acquisition du signal, $T = N\,\frac{1}{Fs}$, $B = \kappa\,\frac{F_s}{2}$ avec $\kappa \in [0, 1]$. Si on utilise des ondelettes en notant a le facteur d'échelle, on a, pour des ondelettes dyadiques telles que $a = 2^{k-1}$ et $k \in \mathbb{N}^*$, $\kappa = \frac{1}{a} = \frac{1}{2^{k-1}}$, l'équation 6 donne

$$N_f = \lfloor 2^{-(k-1)}\,N \rfloor \tag{7}$$

La fonction qui permet de prendre en compte l'effet du filtrage est `sigcor.get_rs_filtered(N, alpha, B, Fs, twosided=True)`. L'effet de la prise en compte du filtrage est proposé figure Fig 3 pour deux valeurs de seuils significatifs.

On remarque que la prise en compte du filtrage augmente les seuils $r*$ en fonction de la largeur de la bande passante du filtre réalisé : plus la largeur est petite, plus la valeur seuil pour les corrélations significative est grande. Pour les deux seuils α, on remarque que les largeurs des créneaux sont identiques et ne dépendent que de la largeur de bande, de la fréquence d'échantillonnage et du temps d'observation des signaux, ce qui agit au niveau des fractions entières de $\kappa\,N$.

2.3 Comparaison avec les tables historiques

Un des objectif du paquet `sigcor` est de s'affranchir de l'utilisation des tables proposées dans [5]. Les valeurs obtenues par `sigcor.get_rs` pour les valeurs de N proposées dans [5], pour $\rho = 0$, $\alpha = 0.05$ et $\alpha = 0.1$ sont tracés et superposées, après agrandissement et mise en transparence à 0.75 % aux abaques disponible dans [5]. Le résultat de ces superpositions est proposé Fig.4a-b. Les valeurs semblent identiques malgré l'imprecision de la méthode.

3 Discussion

Dans la version 1.0 de `sigcor`, seuls les calculs pour $\rho = 0$ dans la densité jointe des variables aléatoires indépendantes sont implémentées. La prochaine version devrait permettre de calculer les seuils significatifs pour d'autres valeurs de corrélations, en reprenant la loi proposée dans [5] et en travaillant sur celle-ci pour obtenir des valeurs approchées d'intégrales.

Lors de la comparaison des valeurs seuils des corrélations significatives r^*, on remarque que pour des valeurs de α faible, quelques petits écarts liés aux estimations réalisées sur les 1000 échantillons utilisés dans cette étude, tombent dans les valeurs critiques des seuils étudiés puisqu'on teste pour $\alpha = 0.001 = 1/1000$, et ceci même pour des valeurs de N proche de 2. Une augmentation du nombre de simulations permettrait certainement d'obtenir une meilleur estimation aux valeurs théoriques. On remarque aussi que les seuils obtenus avec transformations de Fisher sont proches des valeurs théoriques mais que pour N petit et un seuil α faible il existe néanmoins des différences. L'utilisation de `sigcor.get_rs` est donc conseillée dans ces conditions.

Dans un cadre d'application au calcul de connectivité fonctionnelle dynamique [6, 4], c'est à dire sur de courtes fenêtres d'analyse, il est important de prendre en compte la significativité des corrélations obtenues sur de petits échantillons. Si l'on prend en compte les corrections multiples liées à l'évaluation de signaux issus de plusieurs voxels ou régions du cerveau, les seuils α s'écroulent rapidement. Si l'on compare d régions, on a $d\,(d-1)/2$ tests deux à deux, soit si on a 100 régions à comparer, une correction de Bonferroni correspondant à une division du seuil α par ≈ 5000. L'utilisation du paquet `sigcor` facilite la prise en compte de cet aspect et son usage est donc conseillé.

4 Conclusion

Nous avons développé le paquet Python `sigcor` pour permettre d'améliorer le calcul des valeurs de seuils de corrélations significatives entre deux signaux lorsque le nombre d'échantillons était faible et lorsque les signaux étaient filtrés. Nous avons montré que la version 1.0 du paquet réalisait pleinement ces fonctions pour $\rho = 0$. La prise en compte d'autres lois devrait être prise en compte dans les futures versions du paquet.

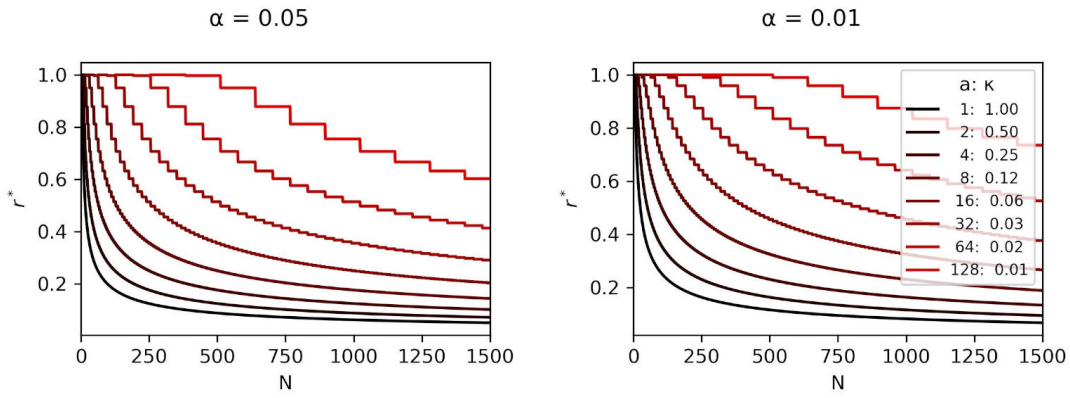

FIGURE 3 : Influence du filtrage sur les valeurs des seuils de corrélations significatives pour $\alpha = 0.05$ et $\alpha = 0.01$. Soit N le nombre d'échantillons des signaux. Lorsque les signaux sont filtrés dans la bande de fréquence $B = \kappa\, F_s$ on doit utiliser remplacer N par $N_f = \lfloor 2\, B\, T \rfloor = \lfloor \kappa\ N \rfloor$ et les courbes obtenues deviennent crénelées. On note a le facteur d'échelle dans le cas d'un filtrage par ondelette.

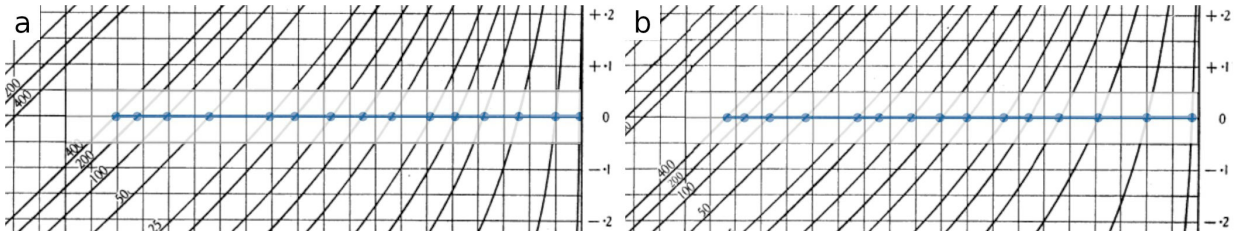

FIGURE 4 : Comparaison des valeurs seuils pour $\rho = 0$ avec les tracés historiques de [5] : a) au seuil $\alpha = 0.05$; b) au seuil $\alpha = 0.10$. Les valeurs sont calculées pour $N \in \{3, 4, 5, 6, 7, 8, 10, 12, 15, 20, 25, 50, 100, 200, 400\}$

Références

[1] GJC BECQ et AL. : Functional connectivity is preserved but reorganized across several anesthetic regimes. *NeuroImage*, page 116945, 2020.

[2] GJC BECQ et AL. : Effet du filtrage sur l'évaluation de la connectivité fonctionnelle dynamique et application sur des données d'irms fonctionnelles de rats. *In GRETSI 2022-XXVIIIème Colloque Francophone de Traitement du Signal et des Images*, 2022.

[3] GJC BECQ, EL BARBIER et S ACHARD : Brain networks of rats under anesthesia using resting-state fmri : comparison with dead rats, random noise and generative models of networks. *Journal of Neural Engineering*, 17(4):045012, 2020.

[4] VD CALHOUN et AL. : The chronnectome : time-varying connectivity networks as the next frontier in fmri data discovery. *Neuron*, 84(2):262–274, 2014.

[5] FN DAVID : *Tables of the ordinates and probability integral of the distribution of the correlation coefficient in small samples*. Cambridge University Press, 1938.

[6] X LIU et JH DUYN : Time-varying functional network information extracted from brief instances of spontaneous brain activity. *Proceedings of the National Academy of Sciences*, 110(11):4392–4397, 2013.

[7] RJ MUIRHEAD : *Aspects of multivariate statistical theory*. John Wiley & Sons, 2009.

[8] BJ WHITCHER : *Assessing nonstationary time series using wavelets*. Citeseer, 1998.

[9] BJ WHITCHER, P GUTTORP et DB PERCIVAL : Wavelet analysis of covariance with application to atmospheric time series. *Journal of Geophysical Research : Atmospheres*, 105(D11):14941–14962, 2000.