A propos de transformations vers des distributions uniformes et de corrélations de quantiles

Guillaume Jean-Paul Claude Becq

A propos de transformations vers des distributions uniformes et de corrélations de quantiles

Guillaume BECQ

Laboratoire Gipsa-lab
11 rue des mathématiques, BP 46, 38402 Saint Martin d'Hères cedex, France
guillaume.becq@gipsa-lab.grenoble-inp.fr

Résumé – Des corrélations de quantiles sur des variables bidimensionnelles sont évaluées dans cette étude. Des transformations vers des distributions uniformes sont utilisées pour calculer ces coefficients. L'approche est évaluée sur des variables gaussiennes liées, des variables uniformes avec transformations non linéaires induisant de fortes queues de distributions, sur des signaux sinusoïdaux, sur des signaux d'épilepsie et sur des enregistrements électriques d'un matériau inerte imbibé d'un électrolyte conducteur. La démarche utilisée met en évidence son rapprochement avec l'utilisation de copules qui en étend son domaine d'utilisation et indique que la corrélation de quantiles proposée est équivalente à une corrélation de rang de Spearman.

Abstract – A quantile correlation on bivariate is proposed in this paper. Transformations towards uniform distribution are used to compute this coefficient. The approach is evaluated on gaussian covariate, uniform covariate with non linear transformations inducing large queues on distributions, on sinusoidal signals, on epileptic signals and on electrical recordings from an inert material soaked with conductive electrolyte. The approach puts evidence for a link with copula that extends its application and indicates that the proposed correlation is equivalent to a Spearman's rank correlation.

1 Introduction

Une pratique courante pour la recherche de dépendences entre variables consiste à calculer des corrélations linéaires entre ces variables avec prise en compte ou non de délais temporels en utilisant des fonctions d'intercorrelation [5, 7]. Cependant la linéarité entre variables est souvent mise en défaut et l'utilisation de cette corrélation n'est plus justifiée même si les données sont liées entre elles. Dans cet article, nous nous intéressons à une corrélation entre variables après transformation non linéaire. Cette dernière étape constitue un prétraitement des données souvent qualifiée de normalisation [3] et peut être qualifé, dans notre cas de normalisation quantile [2, 1] qui permet de travailler dans un espace plus approprié pour la représentation des données et le calcul de distances. Nous nous intéressons ici à des transformations vers des distributions uniformes [5]. Dans un premier temps, les fonctions de répartition empiriques des variables aléatoires en jeu sont estimées. On regarde ensuite la corrélation des quantiles entre ces variables aléatoires. Cette approche permet de bien déterminer les corrélations entre données pour des distributions à longues queues et pour des variables liées par des fonctions monotones. Une approche par voisinages liées telles que développées dans [7] peut aussi être mise en évidence par cette approche. L'article est organisé de cette façon. Dans un premier temps, les différentes étapes pour évaluer la corrélation de quantiles sont décrites et leurs propriétés analysées. Les résultats entre signaux avec différentes normalisations sont ensuite présentés. Une application est proposée pour des données réelles telles que des enregistrements électroencéphalographiques épileptiques ou des enregistrements électriques à travers une éponge imbibée d'un électrolyte conducteur.

2 Méthodes

2.1 Transformation vers des distributions uniformes

Soit F la fonction de répartition de la variable aléatoire stationnaire X définie de $\mathbf{R}$ dans $[0, 1]$ telle que :

$$F(x) = Pr(X < x)$$

Ceci correspond à la définition de Saporta [6] [1]. Soit $u = F(x)$ la valeur associée à la valeur x. On dit que x est le $u \times 100$ ème percentile de la variable aléatoire X ou le q ème quantile. La fonction inverse : $x = F^{-1}(u)$ est appelée fonction quantile. La variable aléatoire U associée à u suit une loi uniforme. En effet, on peut transformer la v.a. X en une variable aléatoire Y qui vérifie une fonction de repartition G autre qu'une distribution uniforme, telle que $u = F(x) = G(y)$ on utilise alors une transformation ϕ définie par :

$$y = \phi(x) = G^{-1}(F(x)) \quad (1)$$

1. Les résultats sont comparables pour une définition telle que celle présentée par Wasserman [9] : $F(x) = Pr(X \leq x)$.

Ceci implique que G soit inversible, hypothèse qui est vérifiée par définition de G, fonction de répartition de Y. La fonction ϕ réalise une anamorphose [6]. Lorsque G est la fonction identité caractéristique d'une distribution uniforme sur $[0,1]$, $U = Y \sim \text{Unif}(0,1)$

Soit ω_x une séquence de n réalisations de X :

$$\omega_x = x(1)x(2)\cdots x(n)$$

Cette séquence est utilisée pour estimer la fonction de répartition de X. Pour cela, on trouve $x_{(1)}x_{(2)}\cdots x_{(N)}$ les N valeurs distinctes ordonnées associées à cette réalisation, i.e. les réalisations des N premières statistiques d'ordre. On note $\rho_i = n_i/n$ une estimation de la fréquence des réalisations de ces valeurs ordonnées avec n_i le nombre de réalisations de $x_{(i)}$. Soit $\hat{F}$ une approximation de F définie par :

$$\hat{F}(x) = \begin{cases} 0, \text{ pour } x \leq x_{(1)} \\ \sum_{k=1}^{i-1} \rho_i, \text{ pour } x_{(i)} < x \leq x_{(i+1)} \text{ i < N} \\ 1, \text{ pour } x_{(N)} < x \end{cases} \quad (2)$$

$\hat{F}$ correspond à une fonction de répartition empirique de X.

On peut former une nouvelle séquence associée aux réalisations de ω_x représentée par les successions des valeurs de u :

$$\omega_u = u(1)u(2)\cdots u(n)$$

en notant $u(i) = \hat{F}(x(i))$

2.2 Corrélations de quantiles

Soit U la v. a. définie par la transformation vers une distribution uniforme de X. On définit r_q, la corrélation quantile de deux séquences contenant le même nombre d'observations soit n réalisations des v. a. X_1 et X_2 par la correlation de leurs séquences des variables U_1 et U_2 associées à leurs quantiles : $\omega_{u_1} = u_1(1)u_1(2)\cdots u_1(n)$ et $\omega_{u_2} = u_2(1)u_2(2)\cdots u_2(n)$.

$$r_q = \frac{E\left[(U_1 - \mu_{U_1})(U_2 - \mu_{U_2})\right]}{\sigma_{U_1}\sigma_{U_2}} \quad (3)$$

avec $E[X]$ espérance de X, $\mu_{U_i} = 1/2$ la moyenne et $\sigma_{U_i} = \sqrt{1/12}$ l'écart-type des variables qui sont uniformes. Un calcul simple conduit à :

$$r_q = 12\,E\left[(U_1\,U_2)\right] - 3 \quad (4)$$

Pratiquement, Soit $X = (X_1, X_2)$ un couple de 2 variables aléatoires. Soit ω_x une séquence de réalisations de X. On travaille sur chaque variable séparément pour obtenir des fonctions de répartitions empiriques des 2 variables X_1 et X_2 qui conduisent à $U = (U_1, U_2)$ et la séquence des réalisations ω_u. On utilise ensuite un estimateur empirique pour calculer r_q.

Plus généralement, l'estimation de la fonction de répartition du couple de variables $U = (U_1, U_2)$ pourrait être réalisée à partir des valeurs utilisées pour calculer la corrélation. Ceci conduirait à une fonction de répartition bivariée dont les marginales suivraient des lois uniformes. En d'autres termes, l'utilisation de cette normalisation met en évidence les réalisations

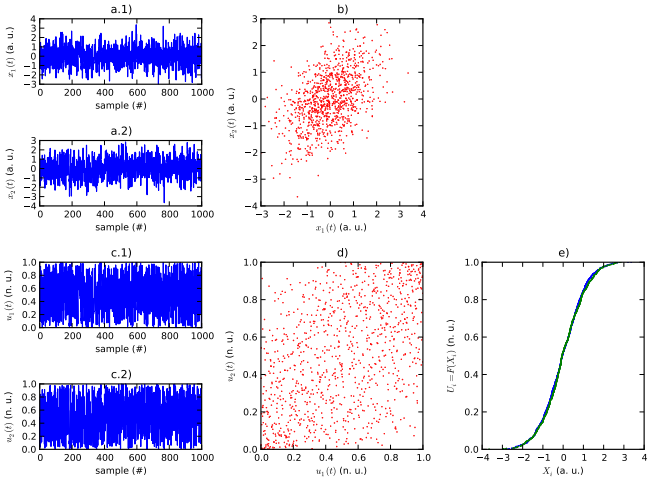

FIGURE 1 – Exemple de variables bivariées gaussiennes liées et résultats de transformations vers des distributions uniformes. a) Signaux centrés réduits : a.1) x_1, a.2) x_2. b) Réalisations de $X = (X_1, X_2)$. c) Signaux transformés vers une distribution uniforme $U_i \sim$ Unif(0, 1) : c.1) u_1, c.2 : u_2 d) Réalisations de $U = (U_1, U_2)$. e) Fonctions de répartition empiriques : $u_i = F_i(x_i)$. Ces fonctions correspondent aux transformations de x_i vers u_i.

de la copule qui lie les variables. Dans [4], un calcul du coefficient ρ de Spearman à partir des copules est donné par $\rho(C) = 12\int_{[0,1]^2} C(u,v)\,du\,dv - 3 = 12\int_{[0,1]^2} u\,v\,dC(u,v) - 3$ et qui correspond à la formule 4 trouvée plus haut en se référant à la définition de l'espérance. La corrélation de quantiles est donc identique à la corrélation de Spearman donnée par cette formule.

3 Exemples sur des signaux stationnaires

3.1 Variables gaussiennes liées

Soit $X = (X_1, X_2)$ une variable aléatoire gaussienne bidimensionnelle dont la matrice de variance-covariance, avec ρ un coefficient de liaison entre X_1 et X_2, est donnée par $M = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Une séquence de $N = 1000$ réalisations pour un tel couple de variables en prenant $\rho = 0.5$ est représenté Fig. 1. Une transformation des données ainsi que le tracé des couples obtenus sont aussi proposés sur cette figure. On remarque que le 'nuage' gaussien observé en (b) se transforme en un 'nuage' de forme caractéristique en (d) centré autour de la fonction identité $y = x$. Différentes valeurs de correlations linéaires avec ou sans transformations pour cette séquence sont données Tab.1. On retrouve des valeurs similaires avec ces deux approches ce qui indique que la corrélation des quantiles est identique à la corrélation linéaire dans ce cas.

TABLE 1 – Valeurs de correlations pour une v.a. gaussienne multidimensionnelle. ρ valeur du modèle, , r_l correlation linéaire, r_q corrélation des quantiles. Valeurs obtenues avec $N = 1000$ réalisations.

ρ	0.01	0.1	0.5	0.9	0.99
r_l	0.02	0.06	0.49	0.90	0.99
r_q	0.01	0.06	0.46	0.89	0.99

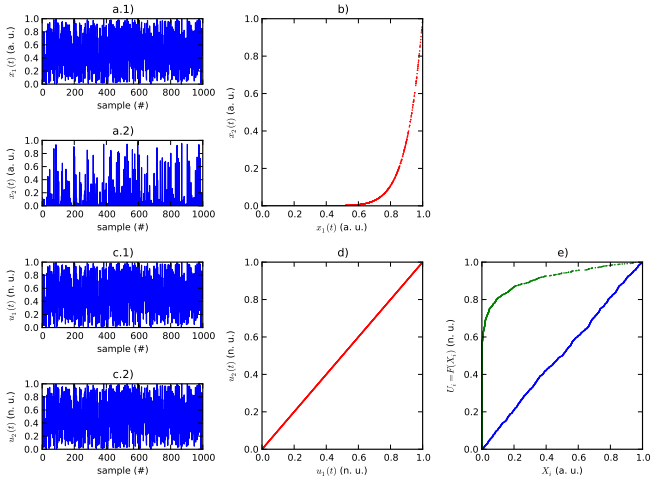

FIGURE 2 – Exemple de variables liées par une transformation non linéaire. Mêmes notations que dans Fig. 1.

3.2 Transformations non linéaires monotones

On part d'une variable aléatoire uniforme $X_1 \sim \text{Unif}(0, 1)$. A partir de X_1, on génère une variables aléatoires X_2 telle que $X_2 = r(X_1) = X_1^{100}$. Cette transformation a pour effet de condenser les valeurs vers 0 et de générer une grande queue jusqu'à 1. La relation liant les variables n'est plus linéaire et le calcul du coefficient de corrélation linéaire fournira une valeur non exploitable. En revanche les distributions des données sont liées par les formules classiques disponibles dans [6, 9]. Une représentation des valeurs prises par U_1 et U_2 pour le calcul de la corrélation de quantiles définies précédemment et reliant les mêmes quantiles des variables sont données Fig. 2. La liaison des quantiles représentées Fig. 2 (d) est linéaire. Cet effet est la conséquence de la monotonicité de la fonction r reliant X_2 à X_1. Le coefficient de corrélation de quantiles vaut alors 1 pour ces données.

3.3 Cas de signaux sinusoïdaux de fréquences différentes

On peut observer ce qui se passe dans le cas de signaux déterministes tels que deux signaux sinusoïdaux définis par $x1(t) = a_1\, sin(2\,\pi\, f_1\, t)$ et $x_2(t) = a_2\, sin(2\,\pi\, f_2\, t + \phi)$. Un tel exemple est représenté Fig. 3 pour $f_1 = 2$ et $f_2 = 10$, $a_1 = 1$, $a_2 = 1$ et $\phi = 0$. La courbe de Lissajoux observée en (b) se trans-

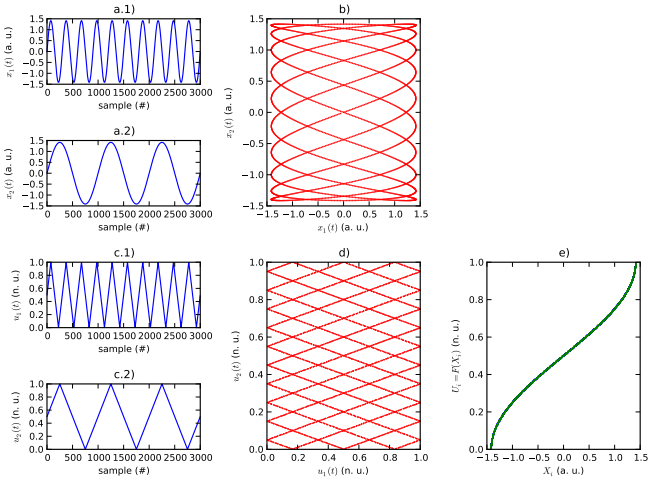

FIGURE 3 – Exemple de sinusoïdes. Mêmes notations que dans Fig. 1.

forme en un grillage en (d). On constate que les relations par morceaux entre ces signaux sont linéaires et conduirait à une correlation conditionnelle valable dans les différents segments entre intersections du tracé (d) pour des relations de variables tels que U_1 et U_2 ayant des caractères monotones communs tels que des couplages de croissances ou de décroissances. On remarque que le nombre de points de rebroussement des courbes U_2 par rapport à U_1 et dans un rapport 10/3, qui correspond au rapport de fréquences des deux sinusoïdes. Enfin, la modification de phase entre signaux induit des effets de dissymétries non présentés dans cette étude.

3.4 Cas de signaux réels

3.4.1 EEG d'épilepsie

Les effets des transformations vers des distributions uniformes pour deux signaux électroencéphalographiques issus d'électrodes positionnées sur le cortex d'un rat épileptique sont représentés Fig.4. Les données sélectionnées correspondent à un épisode de crise. On remarque des zones de corrélations en (d) pour différents intervalles de valeurs qui sont moins évidents pour les variables non transformées en (b), ainsi que des zones de transitions pour les hautes et basses valeurs des variables. On remarque que certaines valeurs hautes de x_1 correspondent à des valeurs basses sur x_2. Les différents segments observés en (d) pourrait correspondre à différents déphasages entre signaux. Les transformations de variables représentées en (e) permettent de normaliser les données de façon appréciables et ressemblent aux fonctions sigmoidales utilisées en entrée de réseaux de neurones artificiels pour la classification de données EEG.

3.4.2 EEG d'éponges

L'effet des transformations vers des distributions uniformes pour deux signaux électroencéphalographiques issus d'électro-

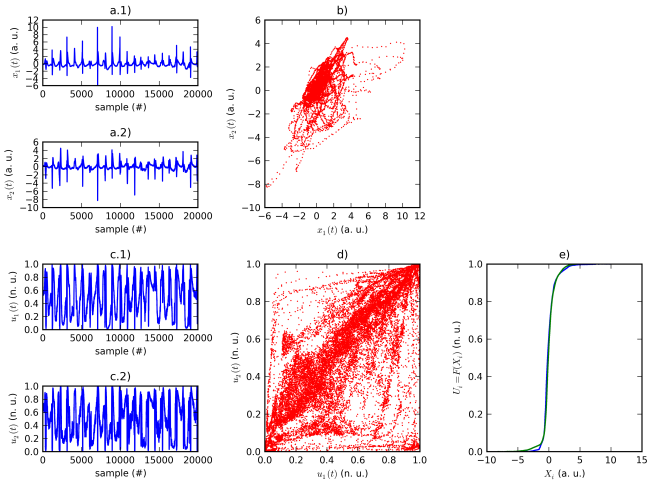

FIGURE 4 – Exemple de signaux d'EEG au cours d'une crise d'épilepsie. Mêmes notations que dans Fig. 1.

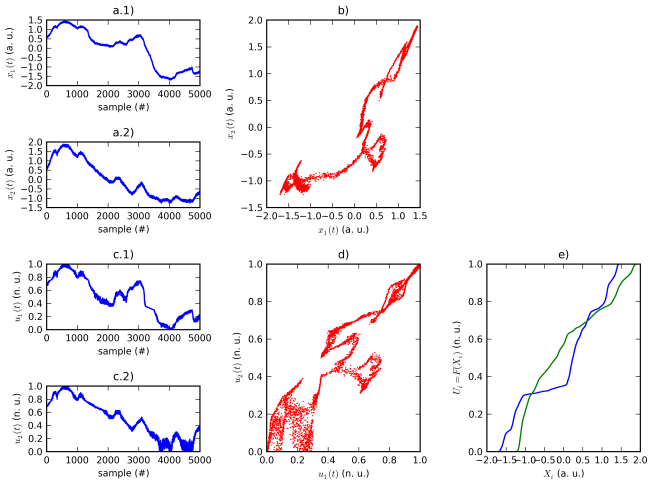

FIGURE 5 – Exemple de signaux d'activités électriques observées sur deux capteurs distants positionnés sur une éponge imbibé d'un électrolyte conducteur. Mêmes notations que dans Fig. 1.

des positionnées sur une éponge imbibée d'un électrolyte conducteur est donné Fig.5. Les signaux enregistrés semblent persistants et présentent des caractéristiques longues mémoires. Les électrodes sont positionnées à deux endroits éloignés.

Certaines zones de (d) mettent en évidence des corrélations de quantiles, mais aussi des régions où les variables sont indépendantes ce qui se traduit par des nuages de points rectangulaires sur la figure.

4 Discussion, Conclusion

Une technique de corrélation linéaire dite corrélation quantile a été introduite dans cette étude. Ces effets ont été comparés à une corrélation linéaire classique et nous avons montré que cette corrélation est équivalente à une corrélation de rang dite de Spearman telle qu'utilisée dans les études de copules. Cette corrélation peut être intéressante pour l'étude des dépendances entre variables multidimensionnelles en particulier pour les variables dont les queues de distributions sont longues ou pour l'étude de variables couplées de façon monotone. Pour des variables couplées de façon non monotone, quelques exemples ont été proposés pour se rapporter à une étude locale ou la structure du couplage reste monotone. Aussi pour des variables sinusoïdales, il a été montré que les courbes de Lissajoux se transforme en grillage dont le nombre de points de rebroussement indique les fréquences temporelles. Pour des données réelles, l'analyse semble plus complexe mais les transformations introduites pourraient permettre de mieux représenter les données pour la mise en évidence de différentes structures de couplage.

Remerciements

Ce projet a été financé par l'IXXI dans le cadre du projet Fracnets 2011-2013. Je tiens à remercier Steeve Zozor pour m'avoir indiqué l'approche par copules qui lui paraissait proche de ce que je faisais.

Références

[1] G. Becq et al., "Classification of Epileptic Motor Manifestations and Detection of Tonic-Clonic Seizures with Acceleration Norm Entropy", IEEE Trans. Biomed. Eng., in press 2013.

[2] B. M. Bolstad et al. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.", Bioinformatics, 19(2) :185-193, 2003.

[3] W. Meisel, "Computer-oriented approaches to pattern recognition", Academic Press Inc., 1972.

[4] R. B. Nelsen et M. Úbeda-Flores, "Directional dependence in multivariate distributions", Ann. Inst. Sta. Math., 64 :677-685, 2012.

[5] B. Pompe, "Measuring statistical dependences in a time series", J. Stat. Phys., 73(3/4) :587–610, 1993.

[6] G. Saporta, "Probabilités, analyse de données et statistiques", Editions TechniP, 2006.

[7] G. Sugihara et al., "Detecting causality in complex ecosystems", Science, 338 :496–500, 2012.

[8] P. K. Trivedi et D. M. Zimmer, "Copula modeling : an introduction for practitioners", Foundations and Trends in Econometrics, 1(1) :1–11, 2005.

[9] L. Wasserman, " All of parametric statistics", Springer Science + Business Media Inc., 2006.