Génération de modèles graphiques

Sophie ACHARD[1], Irène GANNAZ[2], Kévin POLISANO[3]

[1]Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK, France

[2]Univ. Lyon, INSA Lyon, UJM, UCBL, ECL, ICJ, UMR5208, 69621 Villeurbanne, France

[3]Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, France

sophie.achard@univ-grenoble-alpes.fr, irene.gannaz@insa-lyon.fr,
kevin.polisano@univ-grenoble-alpes.fr

Résumé – Nous étudions la répartition des valeurs de corrélation obtenues à partir de l'algorithme de Córdoba et al. pour différentes structures de graphes. Nous mettons en lumière un biais possible de l'algorithme et la difficulté à simuler des valeurs similaires à des données réelles.

Abstract – The distributions of correlation values obtained from the algorithm of Córdoba et al. for different graph structures are compared. The study highlights a possible bias of the algorithm and the difficulty in simulating values similar to real data.

1 Introduction

Les matrices de corrélation, ou de covariances, permettent de modéliser la structure de dépendance entre des variables aléatoires. Si elles sont très présentes dans les modèles statistiques, peu de procédures de simulation sont disponibles. Simuler une matrice de corrélation n'est pas aisé en grande dimension, car il faut que la matrice créée soit définie positive. Parmi les procédures de simulation proposées dans la littérature, on peut citer [6], qui propose une caractérisation de la loi uniforme sur l'espace des matrices de corrélations ainsi qu'un algorithme de simulation. Il existe aussi des simulations dites de *vines* et *onion* [11]. [8] utilise une paramétrisation de la décomposition de Cholesky d'une matrice de corrélation. On pourra se référer à cet article pour avoir un aperçu bibliographique des méthodes existantes, ainsi qu'à la thèse de Córdoba [4].

La motivation pour générer des matrices de corrélations est de permettre de simuler des jeux de données complexes afin de vérifier les performances de méthodes statistiques par Monte-Carlo. Par exemple, une application est la détection de clusters, ou l'inférence de modèles graphiques. Dans ces deux exemples, la matrice générée doit respecter certaines contraintes. Principalement, elle doit être associée à une structure de graphe sous-jacente, que les procédures essaient d'estimer. De nombreux travaux, confrontés à ce problème, utilisent alors des structures très parcimonieuses, ce qui peut biaiser les résultats.

Ainsi, nous sommes motivés par la génération de matrices de corrélation associées à des graphes. Récemment [3] a proposé un algorithme de simulations de matrices de corrélation sous la contrainte que certains coefficients de corrélation partielle sont nuls. Nous souhaitons étudier la répartition des valeurs de corrélation selon la structure sous-jacente de graphe à l'aide de cette approche. Nous mettons aussi en lumière les limitations de l'algorithme.

2 Corrélation partielle et Cholesky

Dans la suite, nous noterons $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ la matrice de covariance et $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ la matrice de précision, ou de corrélation partielle. Il est équivalent de générer une matrice de corrélation ou une matrice de précision ; nous considérerons ici la génération de la matrice de précision. Le modèle graphique associé à $\mathbf{\Omega}$ est le graphe avec p nœuds tels qu'il y a une arête entre les nœuds i et j, $i \sim j$, si $\omega_{ij} \neq 0$. L'ensemble des arêtes est noté E, $E = \{(i,j),\, i \sim j\}$. Étant donné $E \subset \mathbb{R}^2$, l'objectif est de construire $\mathbf{\Omega}^E$ telle que pour tout $(i,j) \notin E$, $\omega^E_{ij} = 0$.

2.1 Cholesky

La principale difficulté pour simuler une matrice $\mathbf{\Omega}^E$ est de faire en sorte qu'elle soit définie positive. Afin de s'en assurer, nous pouvons utiliser la décomposition de Cholesky. Soit $\mathbf{U}$ une matrice triangulaire supérieure de diagonale égale à 1, alors $\mathbf{U}\mathbf{U}^\top$ est définie positive. Ainsi pour générer $\mathbf{\Omega}^E$, nous allons générer $\mathbf{U}$ puis définir $\mathbf{\Omega}^E$ par $\mathbf{\Omega}^E = \mathbf{U}\mathbf{U}^\top$.

Notons par la suite $\mathcal{U}_1^p$ l'ensemble des matrices triangulaires supérieures de dimension $p\times p$, à diagonale positive et de lignes normalisées à 1 ; $\mathcal{U}_1^p(G) \subset \mathcal{U}_1^p$ le sous-ensemble où $u_{ij} = 0$ pour tout $(i,j) \notin E$; $\mathcal{E}^p$ l'elliptope de dimension p des matrices symétriques définies positives de diagonale unitaire ; et

$$\mathcal{E}^p(G) = \left\{\mathbf{U}\mathbf{U}^\top \mid \mathbf{U} \in \mathcal{U}_1^p \text{ et } u_{ij} = 0 \text{ si } (i,j) \notin E\right\} \subset \mathcal{E}^p.$$

2.2 Corrélation partielle sur graphe ordonné

Soient $\{X_i,\, i = 1,\dots,p\}$ des variables *i.i.d* gaussiennes, centrées, de variance $\mathbf{\Sigma}$, avec $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$. Alors

$$X_i = \sum_{\substack{j=1,\dots,p\\ j\neq i}} \beta_{ij} X_j + \varepsilon_i,$$

avec $\beta_{ij} = -\frac{\omega_{ij}}{\omega_{ii}}$ et $\varepsilon_i \sim \mathcal{N}(0, \omega_{ii}^{-1})$.

Chaque nœud i va ainsi dépendre de parents notés $\text{pa}(i) = \{j \,:\, \omega_{ij} \neq 0\}$. Supposons de plus que le graphe est ordonné, c'est-à-dire que l'on peut mettre une relation d'ordre sur les nœuds $1 \prec 2 \prec \cdots \prec p$ avec $\{i \prec j\} \subseteq \text{pa}(i)$. La relation de dépendance s'exprime alors comme suit :

$$X_i = \sum_{j \in \text{pa}(i)} \beta_{ij} X_j + \varepsilon_i = \sum_{j \succ i} \beta_{ij} X_j + \varepsilon_i,$$

où on note $j \succ i$ si $j \in \text{pa}(i)$. En ce cas [7] fait remarquer que $\beta_{ij} = -\frac{u_{ij}}{u_{ii}}$. Nous avons ainsi

$$\omega_{ij} = 0 \iff u_{ij} = 0.$$

Ainsi, dans le cadre d'un graphe ordonné, il est possible de générer la matrice de Cholesky $\mathbf{U}$ dans $\mathcal{U}_1^p(G)$ en imposant $u_{ij} = 0$ pour tout $(i,j) \notin E$, ceci afin d'obtenir une matrice $\mathbf{\Omega}^E = \mathbf{U}\mathbf{U}^\top \in \mathcal{E}^p(G)$ telle que $\omega_{ij}^E = 0$ pour tout $(i,j) \notin E$. Ceci nécessite donc d'avoir des graphes ordonnés, ce qui n'est pas le cas en général.

3 Algorithme de simulation

Une caractérisation de la distribution uniforme sur l'espace engendré par l'ensemble des matrices de corrélation partielle a été formulé par [6]. Avec un changement de variable, il est possible d'en déduire une distribution pour les éléments non nuls de la matrice triangulaire supérieure $\mathbf{U}$.

En s'appuyant sur les résultats ci-dessus, [3] propose un algorithme de simulation d'une matrice $\mathbf{\Omega}^E$ pour $E \subset \{1, \ldots, p\}^2$. Son algorithme est réalisé en plusieurs étapes :

1. Modification de E en $\widetilde{E}$ de sorte que le graphe $\widetilde{G}$ associé à $\widetilde{E}$ puisse être ordonné. Le graphe $\widetilde{G}$ est un graphe *cordal* obtenu en "triangulant" le graphe G par ajout d'arêtes [1] ;
2. Génération de $\mathbf{U} \in \mathcal{U}_1^p(\widetilde{G})$ avec $u_{ij} = 0$ si $(i,j) \notin \widetilde{E}$;
3. Simulation de $\mathbf{\Omega}^{\widetilde{E}}$ en utilisant la décomposition de Cholesky via le changement de variable suivant :
$$\Phi \quad : \quad \begin{array}{ccc} \mathcal{U}_1^p(\widetilde{G}) & \longrightarrow & \mathcal{E}^p(\widetilde{G}) \\ \mathbf{U} & \longmapsto & \mathbf{U}\mathbf{U}^\top \end{array} ;$$
4. Construction de la matrice $\mathbf{U}$ finale associée à E en fixant $u_{ij} = 0$ si $(i,j) \in \widetilde{E} \setminus E$. Pour tout $i = 1, \ldots, p$, orthogonalisation et renormalisation de l'ensemble $\{(u_{ij})_j,\ (i,j) \notin E,\ i \succ j\}$.

Afin d'échantillonner uniformément sur $\mathcal{E}^p(\widetilde{G})$ une matrice $\mathbf{\Omega}^{\widetilde{E}}$ au travers de la paramétrisation Φ, il est nécessaire à l'étape 2 d'échantillonner les matrices $\mathbf{U} \in \mathcal{U}_1^p(\widetilde{G})$ à partir d'une densité proportionnelle au déterminant du jacobien de Φ, dont l'expression est :

$$\det(J\Phi(\mathbf{U})) = 2^p \prod_{i=1}^{p} u_{ii}^{\text{pa}(i)+1}.$$

1. Nous passons sous silence les détails techniques de cette étape afin de ne pas obscurcir le propos général. Pour comprendre le procédé d'ordonnancement le lecteur pourra se référer à la notion de *graphe cordal* qui est la pierre angulaire à l'intersection des réseaux de Markov et des réseaux bayésiens gaussiens.

La factorisation à travers les lignes de $\mathbf{U}$ permet de les échantillonner de manière indépendante, grâce à l'algorithme de Metropolis-Hastings, à partir d'une densité $f(\boldsymbol{u}_i) \propto u_{ii}^i$.

4 Résultats : influence de la structure de graphes

L'objectif est d'étudier l'influence de la structure de graphe sous-jacente sur la distribution des valeurs de corrélations des matrices échantillonnées. Les différentes classes de graphes aléatoires considérées sont les suivantes : le modèle d'Erdős–Rényi, de Barabási–Albert, du "petit monde" (*small-world*) et des blocs stochastiques à 2 communautés [12, 13, 1]. Ainsi deux sources d'aléatoires entrent en jeu : d'une part la génération de matrices d'adjacence correspondant à un graphe G issu de la classe de graphe aléatoire considérée ; et d'autre part l'échantillonnage d'une matrice de corrélation $\mathbf{\Omega}^{\widetilde{E}}$ à partir du graphe $\widetilde{G}$ associé, *via* la méthode de Córdoba et al. exposée à la section 3.

4.1 Caractérisation des structures utilisées

L'étape 1 dans l'algorithme de [3] transformant G en $\widetilde{G}$ est en pratique cruciale. En effet la modification du graphe pour obtenir un graphe ordonné peut nécessiter d'ajouter un grand nombre d'arêtes. Nous comparons en Figure 1 les distances de Hamming entre les graphes initiaux G et les graphes associés $\widetilde{G}$ de cette étape, pour différentes structures de graphes. La distance de Hamming entre deux graphes est la somme du nombre d'arêtes qui diffèrent. Nous observons que les graphes de Erdős–Rényi sont les graphes initiaux G les plus éloignés de ceux $\widetilde{G}$ générés lors de l'étape 1 et les graphes de Barabási–Albert les moins éloignés. Les autres structures considérées ici, petit monde et à blocs stochastiques, sont intermédiaires. Ainsi le caractère uniforme de la génération des matrices de corrélation est *a priori* mieux préservé pour les graphes ayant une structure de type Barabási–Albert.

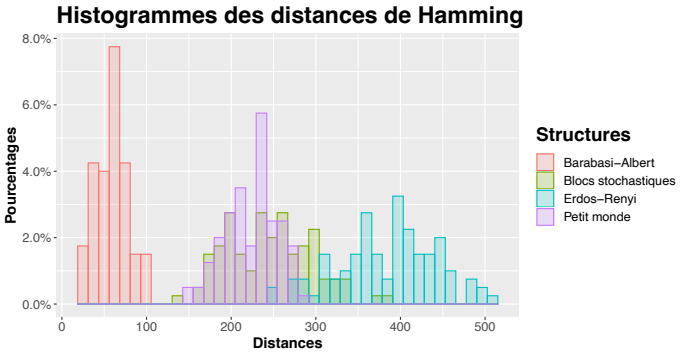

FIGURE 1 – Histogrammes des distances de Hamming entre les graphes G et les graphes $\widetilde{G}$ définis à l'étape 1 en section 3. Les graphes ont 51 nœuds et 128 arêtes en moyenne (degré moyen $d = 10\%$).

4.2 Expériences numériques

Pour chacune des classes, différents nombres de noeuds p avec différents degrés moyens $d = |E|/p$, nous avons simulé $N = 50$ graphes aléatoires composés de p noeuds, puis généré $K = 10$ matrices de corrélations à partir de ces graphes grâce à l'algorithme de Córdoba et al. [3]. Des exemples de distributions des valeurs obtenues sont donnés en Figure 2. Par souci de lisibilité des résultats, nous écartons les entrées strictement nulles ($\omega_{ij}^{\widetilde{E}} = 0 \Leftrightarrow (i,j) \notin \widetilde{E}$) ou celles égales à 1 (situées sur la diagonale).

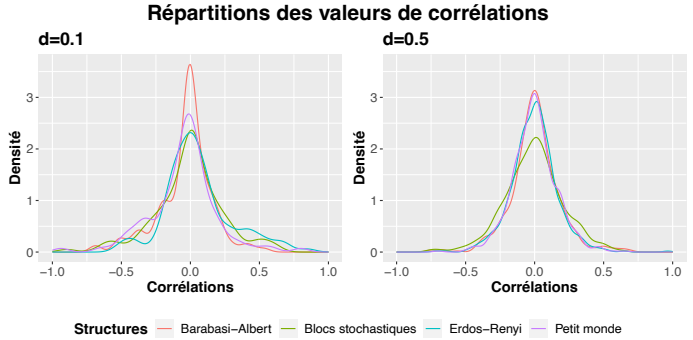

FIGURE 2 – Densités des valeurs non nulles des corrélations pour une simulation de graphe de différentes structures. Les graphes ont $p = 51$ nœuds et un degré moyen $d = 10\%$ (à gauche) et $d = 50\%$ (à droite).

Nous souhaitons comparer les distributions des corrélations obtenues pour ces différents graphes. Nous nous intéressons alors au comportement de la moyenne des valeurs absolues $\widehat{M}$ des entrées non nulles de ces matrices. Plus précisément pour chaque couple (p, d) nous observons la moyenne de ces estimations sur les $N \times K$ matrices générées.

Les paramètres qui influencent principalement la distribution sont la classe du graphe, le nombre de nœuds p et le degré moyen d. Des études (non présentées ici) montrent que le nombre d'arêtes dans le graphe n'est pas une variable pertinente pour expliquer les répartitions des valeurs de $\widehat{M}$, d'où cette paramétrisation par p et d.

La Figure 5 représente l'évolution des moyennes de $\widehat{M}$. Pour chaque classe de graphe, nous observons que ces valeurs diminuent quand le nombre de nœuds augmente et quand le degré moyen augmente. Ainsi plus il y a d'arêtes dans un graphe, plus les valeurs de corrélation sont faibles. Cependant l'évolution n'est pas régulière en fonction du nombre d'arêtes. On constate également qu'à paramètres (p, d) égaux, les graphes de Erdős–Rényi et de blocs stochastiques ont les valeurs les plus élevées. Inversement, les valeurs observées pour les graphes de Barabási–Albert sont significativement plus faibles. Une explication possible à cette différence est l'étape de normalisation 4 de l'algorithme, décrite en section 3. En effet, plus le graphe $\widetilde{G}$ est éloigné du graphe initial G, plus cette normalisation va augmenter les valeurs initialement générées. Or comme vu en section 4.1 les graphes de Barabási–Albert sont ceux pour lesquels le graphe $\widetilde{G}$ est le plus proche du graphe initial.

4.3 Limitations et comparaisons avec des données réelles

Nous observons sur la Figure 2 que les distributions sont toujours symétriques autour de zéro, traduisant une génération de variables de corrélations très faibles, qui ne correspond donc pas nécessairement aux observations sur données réelles.

Pour illustrer ce constat, nous considérons des données d'IRM fonctionnelle acquises sur des rats. Les données sont décrites et en accès libre [2]. La durée d'enregistrement est de 30 minutes avec un temps de répétition de 0,5 seconde, et 3600 points de temps sont ainsi disponibles en fin d'expérience. Après le prétraitement expliqué dans [2], des séries chronologiques de 51 régions cérébrales pour chaque rat ont été extraites. Nous calculons alors la transformée en ondelette avec l'ondelette de Daubechies d'ordre 8 des 51 signaux. Nous nous intéressons ici à l'échelle d'ondelettes 4, correspondant à l'intervalle de fréquence [0.06 ; 0,12] Hz. Il y alors 122 coefficients d'ondelettes disponibles pour chacune des 51 régions. Les distributions des corrélations paires à paires entre les coefficients d'ondelettes des régions sont représentées en Figure 3.

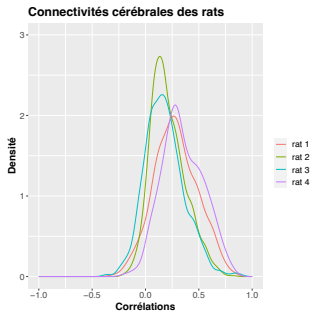

FIGURE 3 – Densités des corrélations entre les coefficients d'ondelettes d'ordre 4 des signaux d'IRMf des 51 régions cérébrales. Chaque courbe correspond à un rat donné.

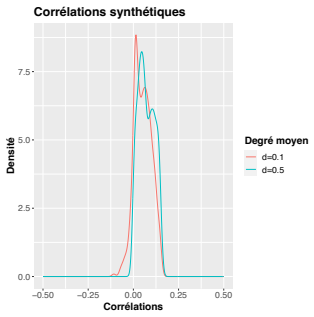

FIGURE 4 – Exemples de corrélations non simulés mais construits *ad hoc*. Leurs déterminants sont égaux à 0.0108.

Nous pouvons constater que les distributions obtenues ne sont pas symétriques, et que les valeurs théoriques associées ne peuvent non plus être symétriques. [9] obtient un constat similaire sur des données de finance. Outre la complexité algorithmique de la méthode de génération utilisée [3], celle-ci ne permet donc pas de simuler des matrices de corrélations "réalistes" permettant la validation de procédures d'inférence statistique sur des données réelles comparables. Remarquons qu'il est néanmoins possible de générer des matrices de corrélations avec des distributions non symétriques comme illustré en Figure 4. Cela peut être réalisé par des méthodes utilisant des GAN [9, 10], lesquels sont en mesure de reproduire ces distributions dont la moyenne est déplacée vers les valeurs positives. Cependant, le caractère aléatoire de l'échantillonnage n'est alors plus aussi bien maitrisé que dans [3].

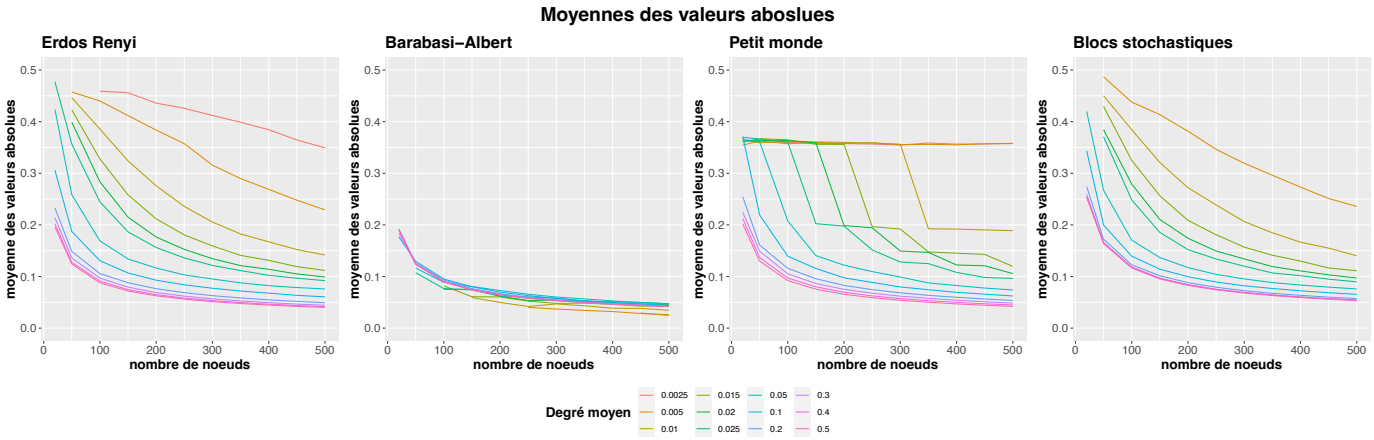

FIGURE 5 – Moyennes $\widehat{M}$ des valeurs absolues des valeurs non nulles des matrices de corrélation, en fonction de p le nombre de nœuds dans les graphes, pour différents degrés moyens d. Notons que toutes les valeurs de (p, d) ne peuvent être simulées pour certaines classes de graphe.

5 Conclusion et perspectives

Nous cherchons à simuler des matrices de corrélation similaires à celles obtenues sur des données réelles afin de pouvoir simuler des jeux de données réalistes et d'étudier la qualité d'algorithmes d'analyse de données sur ces jeux de données. Nous avons ici étudié les résultats obtenus par l'algorithme proposé dans [3]. Cet algorithme permet de simuler des matrices de corrélations associées à des graphes. L'étude menée montre que les valeurs des corrélations générées sont faibles et symétriques, ce qui n'est pas nécessairement le cas avec des données réelles. Une adaptation de l'algorithme est envisagée afin de permettre des réalisations selon des lois plus complexes.

De plus, à partir de cet algorithme, nous avons étudié l'influence des structures de graphes sur les valeurs de corrélations générées, à nombres d'arêtes et de nœuds fixés. Nous observons des différences significatives selon les structures. En particulier les graphes de Barabási–Albert ont des valeurs de corrélations associées plus faibles que les autres structures. Cependant, il est possible que cette observation soit due à un biais de l'algorithme de [3]. Les autres algorithmes de génération de matrices de corrélation ne proposent pas de simulations contraintes à un graphe. Il serait intéressant de mettre en place cette contrainte sur certains algorithmes, tel celui de [8] (voir [5, section 4]). Ceci pourrait permettre de s'affranchir du biais potentiel lié à l'ordonnancement de graphe.

Remerciements. Les auteurs souhaitent remercier les ingénieurs en informatique du LJK et de l'ICJ, en particulier Roland Denis, qui ont contribué à la mise en place des simulations.

Références

[1] Abbe, E., Bandeira, A. S. et Hall, G. (2015). Exact recovery in the stochastic block model. *IEEE Transactions on information theory*, 62(1), 471-487.

[2] Becq, G. J-PC, Habet, T., Collomb, N., Faucher, M., Delon-Martin, C., Coizet, V., Achard, S. et Barbier, E.L. (2020). Functional connectivity is preserved but reorganized across several anesthetic regimes. *NeuroImage*, 219, 116945.

[3] Córdoba, I., Varando, G., Bielza, C. et Larrañaga, P. (2020). On generating random Gaussian graphical models. *International Journal of Approximate Reasoning*, 125, 240–250.

[4] Córdoba, I. (2020). Unifying methodologies for graphical models with Gaussian parametrization. *Doctoral dissertation, ETSI Informatica*, Universidad Politécnica de Madrid, Spain.

[5] Flórez, A.J., Abad, A.A., Molenberghs, G. et Van Der Elst, W. (2020). Generating random correlation matrices with fixed values : An application to the evaluation of multivariate surrogate endpoints. *Computational Statistics & Data Analysis*, 142, 106834.

[6] Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97(10), 2177-2189.

[7] Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data : Unconstrained parameterisation. *Biometrika*, 86(3), 677-690.

[8] Pourahmadi, M. et Wang, X. (2015) Distribution of random correlation matrices : Hyperspherical parameterization of the Cholesky factor. *Statistics & Probability Letters*, 106, 5–12.

[9] Marti, G. (2020). Corrgan : Sampling realistic financial correlation matrices using generative adversarial networks. *In ICASSP 2020, IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8459-8463).

[10] Marti, G., Goubet, V. et Nielsen, F. (2021). cCorrGAN : Conditional Correlation GAN for Learning Empirical Conditional Distributions in the Elliptope. In *International Conference on Geometric Science of Information* (pp. 613-620). Springer, Cham.

[11] Lewandowski, D., Kurowicka, D., et Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9), 1989-2001.

[12] Wang, X.F. et Chen, G. (2003). Complex networks : small-world, scale-free and beyond. *IEEE circuits and systems magazine*, 3(1), 6-20.

[13] Albert, R. et Barabási, A.L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1), 47.