

Génération de Matrices de Corrélation avec des Structures de Graphe par Optimisation Convexe

Ali FAHKAR¹ Kévin POLISANO¹ Irène GANNAZ² Sophie ACHARD¹

¹Univ. Grenoble Alpes, CNRS, Grenoble INP, Inria, LJK, F-38000 Grenoble, France

²Univ. Grenoble Alpes, CNRS, Grenoble INP, G-SCOP, 38000 Grenoble, France

Résumé – Ce travail porte sur la génération de matrices de corrélation théoriques présentant des motifs de parcimonie spécifiques, associés à des structures de graphe. Nous présentons une nouvelle approche basée sur l’optimisation convexe, offrant une plus grande flexibilité par rapport aux techniques existantes, notamment en contrôlant la moyenne de la distribution des entrées dans les matrices de corrélation générées. Cela permet de produire des matrices de corrélation représentant plus fidèlement des données réalistes et pouvant être utilisées pour l’évaluation comparative des méthodes statistiques d’inférence de graphes.

Abstract – This work deals with the generation of theoretical correlation matrices with specific sparsity patterns, associated to graph structures. We present a novel approach based on convex optimization, offering greater flexibility compared to existing techniques, notably by controlling the mean of the entry distribution in the generated correlation matrices. This allows for the generation of correlation matrices that better represent realistic data and can be used to benchmark statistical methods for graph inference.

1 Introduction

Les modèles graphiques permettent de représenter les dépendances entre variables aléatoires. Ce domaine a suscité un grand intérêt ces dernières années, voir par exemple [18,19,23], [12, Chapitre 7] et les références associées. Le champ d’application est vaste, incluant la génétique [13], l’étude des protéines [3], la caractérisation des maladies [5], la connectivité fonctionnelle du cerveau [15] ou encore la gestion des risques [16]. L’idée principale consiste à inférer une structure de graphe associée à la matrice de corrélation ou à la matrice de précision (inverse de la matrice de corrélation). Pour évaluer la qualité des procédures d’estimation, il est essentiel de pouvoir générer des matrices de corrélation ou de précision (théoriques) associées à une structure de graphe donnée, ce qui implique d’imposer des zéros particuliers dans la matrice tout en garantissant sa positivité, ce qui est généralement non trivial. L’objectif de cet article est de présenter une méthode permettant de générer de telles matrices.

Parmi les méthodes de génération de matrices de corrélation proposées dans la littérature, on peut citer *vines* et *onion* [20], basées sur la distribution Beta établie par [17]. Une autre approche consiste à utiliser la décomposition de Cholesky, comme dans [9, 24]. Nous renvoyons à ces articles pour un aperçu bibliographique des méthodes existantes. Dans [2], il a été observé que la distribution des corrélations de connectivité cérébrale était centrée autour de valeurs positives. Cependant, les méthodes existantes génèrent toutes des matrices de corrélation dont la distribution des entrées est centrée autour de zéro. Notre objectif est de proposer une nouvelle approche basée sur l’optimisation convexe permettant de contrôler cette distribution, en particulier sa moyenne.

Cet article¹ est organisé comme suit. La Section 2 introduit les principales définitions et notations. La Section 3 passe en

revue les travaux connexes. La Section 4 décrit l’approche proposée, et la Section 5 présente les résultats ainsi qu’une comparaison avec d’autres approches.

2 Notations

Une matrice réelle symétrique \mathbf{A} de dimension $p \times p$ est semi-définie positive (SDP) si $\mathbf{x}^\top \mathbf{A} \mathbf{x} \geq 0$ pour tout $\mathbf{x} \in \mathbb{R}^p$. Soit $\mathbf{x} \in \mathbb{R}^p$ un vecteur aléatoire de matrice de covariance $\Sigma = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\top]$. La matrice de corrélation associée $\mathbf{C} \in \mathcal{C}$ est définie par $\mathbf{C} = (\text{diag}(\Sigma))^{-\frac{1}{2}} \Sigma (\text{diag}(\Sigma))^{-\frac{1}{2}}$. L’ensemble \mathcal{C} des matrices de corrélation satisfait :

$$\forall \mathbf{C} \in \mathcal{C}, \text{diag}(\mathbf{C}) = 1, \forall i, j \in 1, \dots, p, -1 \leq c_{ij} \leq 1. \quad (1)$$

Générer une matrice de corrélation revient à construire une matrice symétrique SDP vérifiant (1) [14, Problème 7.1.]. Dans le cas qui nous intéresse, nous cherchons une matrice de corrélation \mathbf{C} associée à un graphe $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, c’est-à-dire vérifiant $c_{ij} = 0$ si $(i, j) \notin \mathcal{E}$ et $c_{ij} \neq 0$ sinon, pour tout $(i, j) \in \mathcal{V} \times \mathcal{V}$ avec $i \neq j$. Les poids des arêtes dans le graphe correspondent aux valeurs de la matrice de corrélation. Nous définissons $\bar{\mathcal{E}}$ comme l’ensemble des non-arêtes, correspondant aux entrées nulles de la matrice \mathbf{C} . Ainsi, notre objectif est de générer une matrice de corrélation avec un ensemble prescrit $\bar{\mathcal{E}}$ d’entrées nulles. Ce problème peut être vu comme un problème de complétion de matrice [25, Chapitre 10].

Dans la suite, nous notons $\mathcal{C}(\mathcal{G})$ l’ensemble des matrices de corrélation associées à un graphe donné \mathcal{G} satisfaisant :

$$\mathbf{C} = (c_{ij}) \text{ est SDP, satisfait (1), et } c_{ij} = 0, (i, j) \notin \mathcal{E}. \quad (2)$$

Nous considérons différentes structures de graphe, à savoir les graphes aléatoires d’*Erdős-Rényi*, de *Barabási-Albert*, et de *Watts-Strogatz*, [4], ainsi que des *Modèles à blocs stochastiques* [1] et des graphes *cordaux* où tout cycle de longueur

¹Ce travail a été soutenu par l’Agence Nationale de la Recherche dans le cadre du programme France 2030, référence ANR-23-IACL-0006.

supérieure à trois possède une arête reliant deux sommets non adjacents dans le cycle. En pratique, nous générons un graphe cordal à partir d'un graphe de Barabási-Albert en ajoutant des arêtes si nécessaire pour satisfaire cette propriété.

Une caractéristique clé de la structure du graphe est sa densité, qui est le rapport entre le nombre d'arêtes et le nombre maximal d'arêtes possibles, $d = \frac{2|\mathcal{E}|}{p(p-1)}$.

3 Travaux connexes

Dans cette section, nous présentons brièvement les méthodes qui, à notre connaissance, peuvent générer des matrices de corrélation associées à un graphe donné \mathcal{G} . La première approche repose sur la décomposition de Cholesky. Nous définissons $\mathcal{U}(\mathcal{G}) \subset \mathcal{U}$ comme le sous-ensemble des matrices triangulaires supérieures avec des éléments diagonaux positifs et lignes normalisées à 1, où $u_{ij} = 0$ pour tous les $(i, j) \in \bar{\mathcal{E}}$. Si le graphe est cordal [25, Chapitre 4], il est possible de générer le facteur de Cholesky \mathbf{U} dans $\mathcal{U}(\mathcal{G})$ en imposant $u_{ij} = 0, \forall (i, j) \in \bar{\mathcal{E}}$. On a alors $\mathbf{C} = \mathbf{U}\mathbf{U}^\top \in \mathcal{C}(\mathcal{G})$.

Dans [24], les auteurs proposent une paramétrisation polaire des entrées du facteur de Cholesky $\mathbf{U} \in \mathcal{U}$ et établissent la distribution de probabilité à adopter de telle sorte que l'échantillonnage des matrices résultantes soit uniforme sur l'ensemble \mathcal{C} . En utilisant cette paramétrisation polaire, il est facile d'incorporer la contrainte $u_{ij} = 0$ pour tous $(i, j) \in \bar{\mathcal{E}}$ pour obtenir $\mathbf{U} \in \mathcal{U}(\mathcal{G})$. Ceci assure que $\mathbf{C} = \mathbf{U}\mathbf{U}^\top \in \mathcal{C}(\mathcal{G})$ mais pour les graphes cordaux uniquement. Dans [9], la méthode de génération proposée est basée sur l'algorithme Metropolis-Hastings [8] et donne alors une distribution uniforme sur $\mathcal{C}(\mathcal{G})$, mais n'est applicable là aussi que pour les graphes cordaux.

À notre connaissance, seules deux méthodes ont été proposées dans la littérature pour générer des matrices de corrélation associées à un graphe donné \mathcal{G} sans exiger une structure cordale : la dominance diagonale et l'orthogonalisation partielle, toutes deux présentées dans [9]. Un inconvénient majeur de la dominance diagonale est qu'elle donne des matrices de corrélation avec des valeurs hors-diagonales très faibles. La méthode d'orthogonalisation partielle ne souffre pas de cette limitation, néanmoins notre méthode présentée à la Section 4 est moins sensible à la matrice initiale utilisée (en particulier parce que l'orthogonalisation partielle dépend de l'ordonnement des nœuds) et permet de s'affranchir de la symétrie autour de zéro.

4 Approche proposée

L'objectif de notre travail est de générer des matrices de corrélation dans $\mathcal{C}(\mathcal{G})$, c'est-à-dire satisfaisant les contraintes (2). Des contraintes supplémentaires peuvent être ajoutées, en fonction du contexte. L'une de nos motivations est de construire des matrices de corrélation avec une distribution qui ressemble aux données réelles, typiquement en neurosciences, où celles-ci sont décalées vers des valeurs positives [2]. Pour refléter cette propriété, nous imposons cette contrainte supplémentaire sur la moyenne : pour $b \geq -1$,

$$\frac{1}{2|\mathcal{E}|} \sum_{i \neq j} c_{ij} \geq b. \quad (3)$$

Prendre $b \leq -1$ revient à ne pas imposer de contrainte.

Nous cherchons à résoudre le problème d'optimisation :

$$\underset{\mathbf{C}}{\text{minimiser}} \quad \frac{1}{2} \|\mathbf{C} - \bar{\mathbf{C}}\|_F^2, \quad \text{sous contraintes (2) et (3),} \quad (4)$$

où $\bar{\mathbf{C}}$ est une matrice donnée arbitraire. Avec des données réelles, cela peut être la matrice de corrélation empirique. Il convient de noter que la résolution de (4) garantit que la moyenne des éléments non diagonaux est au moins égale à b . Puisque la fonction objectif dans (4) est convexe, une solution existe chaque fois que les contraintes sont réalisables. Par ailleurs, la matrice identité satisfait les contraintes (2), garantissant ainsi la faisabilité en l'absence de la contrainte supplémentaire (3). Dans la Section 5, nous examinons l'impact de cette contrainte supplémentaire sur l'existence de solutions.

5 Résultats et discussions

Dans nos simulations, nous considérons $\mathbf{C} \in \mathbb{R}^{51 \times 51}$ et $\bar{\mathbf{C}}$ une matrice de même taille dont les entrées suivent une loi uniforme sur l'intervalle $[-1, 1]$. Pour le motif \mathcal{E} , nous utilisons les différents modèles de graphes aléatoires mentionnés à la Section 2. Le problème d'optimisation (4) est résolu en utilisant le solveur CVXOPT de la bibliothèque Python CVXPY [10] qui implémente une méthode de point intérieur. Nous l'avons appliqué à ces modèles de graphes sur 50 exécutions². Dans certains cas, la résolution de (4) aboutit numériquement à une matrice dont la valeur propre minimale est proche de zéro, mais négative, ce qui indique que la matrice n'est pas strictement SDP. Pour y remédier, nous considérons la matrice $\tilde{\mathbf{C}}_\epsilon = \tilde{\mathbf{C}} + \epsilon \mathbf{I}$ (avec $\epsilon = 10^{-8}$) que nous renormalisons par $\mathbf{C} = \frac{1}{1+\epsilon} \tilde{\mathbf{C}}_\epsilon$. La matrice \mathbf{C} n'est pas le minimiseur de la fonction objectif, mais elle est une matrice de corrélation qui satisfait les contraintes (4)³.

5.1 Comparaison avec d'autres approches

Pour la comparaison, nous considérons un graphe avec 51 nœuds, c'est-à-dire $\mathbf{C} \in \mathbb{R}^{51 \times 51}$. La Figure 1 montre la densité des éléments non diagonaux non nuls. Nous comparons notre méthode avec deux autres approches : la dominance diagonale et l'orthogonalisation partielle. Plus précisément, nous générons 50 graphes d'*Erdős-Rényi* en utilisant la dominance diagonale, l'orthogonalisation partielle, et la méthode proposée. Pour cette dernière, nous fixons la densité cible du graphe à 0.5. La distribution des valeurs de corrélation est représentée par les lignes rouge, verte et violette, respectivement - la ligne orange est liée aux données réelles et expliquée ci-dessous. Nous fixons le paramètre $b = -1$ dans notre algorithme pour faciliter la comparaison avec d'autres algorithmes qui n'utilisent pas de seuil. Dans notre algorithme, à la fois $\bar{\mathbf{C}}$ et le point initial pour la méthode de dominance diagonale sont des réalisations d'une distribution uniforme, puisque nous visons à générer des matrices de corrélation aléatoires. Il convient de noter que, lorsqu'on utilise la dominance diagonale, aucune

²Les simulations ont été réalisées à l'aide de l'infrastructure GRICAD, soutenue par la communauté de recherche grenobloise. Le code pour reproduire les expériences est disponible en ligne [11].

³Pour être plus précis, avec cette étape de post-traitement, la valeur moyenne change, et donc la contrainte (3) peut ne pas être satisfaite. Augmenter b à $b(1 + \epsilon)$ permet d'atteindre notre objectif.

perturbation positive n'est appliquée pour garantir que la matrice est SDP. Comme mentionné précédemment, les densités obtenues avec la dominance diagonale sont concentrées autour de faibles valeurs (ligne rouge). Notre approche (ligne violette) donne des entrées plus élevées dans la matrice de corrélation. Nous disposons ainsi d'un modèle de génération de matrices de corrélations synthétiques à graphe fixé, dont les valeurs de corrélations significatives ne se confondent pas avec le bruit, nous permettant ainsi dans [7] d'effectuer un *benchmark* de différentes techniques d'inférence de graphes.

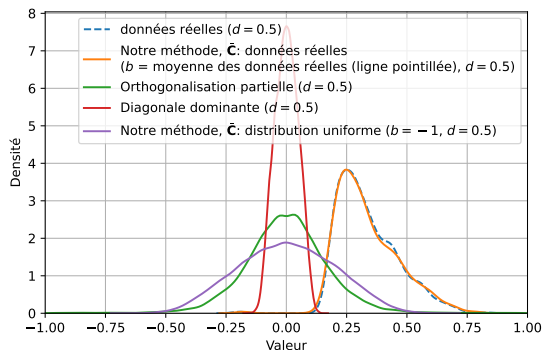


FIGURE 1 : Densité des éléments non diagonaux non nuls dans les matrices de corrélation générées par différentes méthodes (dominance diagonale, orthogonalisation partielle et notre méthode) comparées à la matrice de corrélation obtenue à partir des données fMRI de rats.

5.2 Influence de la contrainte sur la moyenne

Dans la section précédente, nous avons montré que la méthode proposée génère des corrélations non négligeables et dont la distribution est symétrique autour de zéro. En revanche dans certains jeux de données réelles, la distribution des corrélations empiriques est décalée vers les valeurs positives. Or à notre connaissance peu de travaux ont concerné la simulation de matrices de corrélation semblables aux données réelles, à l'exception de [22] utilisant des réseaux antagonistes génératifs (GANs), mais qui nécessitent un grand ensemble de données de matrices de corrélation observées, ce qui peut ne pas toujours être disponible en pratique. Nous examinons maintenant l'impact de la contrainte (3), qui modifie le centre de la distribution des valeurs de corrélation. En particulier, cela permet de générer des matrices de corrélation plus proches des données réelles et cela peut être utile pour contrôler le rapport signal/bruit dans les simulations.

Dans notre contexte, nous sommes motivés par une application en neurosciences impliquant des données d'IRM fonctionnelle acquises sur des rats. Les données sont décrites et en accès libre [2, 6]. Dans la Figure 1, la ligne bleue représente la distribution des corrélation (empirique) des données IRM d'un rat, tandis que la ligne orange montre la distribution des entrées de la matrice de corrélation (théorique) générée à l'aide de notre méthode. Pour les données réelles, nous calculons la densité du graphe en sélectionnant les 50% des entrées avec les valeurs absolues les plus élevées dans la matrice de corrélation. Lors de la génération de la matrice synthétique avec notre méthode, nous fixons $d = 50\%$ et le paramètre b égal à la moyenne des entrées de la matrice de corrélation des données réelles correspondant au graphe calculé ($d = 50\%$). La matrice

initiale \bar{C} est ici égale à la corrélation empirique des données réelles. La Figure 1 montre que la distribution des données simulées non nulles est en effet proche de celle des données réelles décalées vers le positif, avec en sus la contrainte des zéros satisfaite, ce qui n'aurait pas été réalisable avec les méthodes présentées à la Section 3. À noter qu'un choix aléatoire pour \bar{C} eut été également possible pour approcher cette distribution une fois la moyenne b adéquatement fixée, nous avons choisi ici d'illustrer la projection de la matrices de corrélation empirique sur les contraintes de faisabilité.

5.3 Influence de la structure du graphe

L'ajout de la contrainte (3) peut parfois aboutir à un problème d'optimisation sans solution. La Figure 2 illustre la proportion de cas où une matrice de corrélation valide $C \in \mathbb{R}^{51 \times 51}$ est trouvée pour différentes densités de graphes d et valeurs de moyenne b dans (4), et différentes structures de graphes.

Le coût computationnel de la résolution du problème d'optimisation est significativement plus élevé que pour des méthodes comme la dominance diagonale ou l'orthogonalisation partielle qui n'excèdent pas la dizaine de secondes. L'augmentation de la dimension p de C entraîne généralement une augmentation du temps d'exécution. La Figure 3 compare les temps d'exécution pour $p = 51$ avec différentes densités de graphes et modèles. Globalement, le temps d'exécution diminue à mesure que la densité d du graphe augmente. Seul le cas $b = -1$ est représenté, mais des résultats similaires sont obtenus pour différentes valeurs de b .

Conclusion et perspectives

La méthode proposée pour générer une matrice de corrélation correspondant à un graphe présente plusieurs avantages : elle ne repose pas sur une structure de type cordale, elle évite de générer des valeurs trop faibles en contrôlant la moyenne, et enfin elle peut approcher une matrice de corrélation empirique. Cependant, le coût de cette approche augmente avec la dimension. Puisque nous projetons à court terme d'étudier comment l'augmentation du nombre de nœuds affecte l'influence des structures de graphes, l'utilisation de l'algorithme QSDPNAL [21] en MATLAB pourrait être envisagé pour résoudre le problème d'optimisation en plus grande dimension. Enfin une perspective consistera à étudier l'échantillonnage de \bar{C} pouvant conduire à une distribution uniforme sur $\mathcal{C}(\mathcal{G})$.

Références

- [1] Emmanuel ABBE : Community detection and stochastic block models : recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.
- [2] Sophie ACHARD, Irène GANNAZ et Kévin POLISANO : Génération de modèles graphiques. In *GRETSI 2022-XXVIIIème Colloque francophone de traitement du signal et des images*, pages 1–3, 2022.
- [3] Rehan AKBANI, Patrick Kwok Shing NG, Henrica MJ WERNER, Maria SHAHMORADGOLI, Fan ZHANG, Zhenlin JU, Wenbin LIU, Ji-Yeon YANG, Kosuke YOSHIHARA, Jun LI *et al.* : A pan-cancer proteomic perspective on the Cancer Genome Atlas. *Nature communications*, 5(1):3887, 2014.

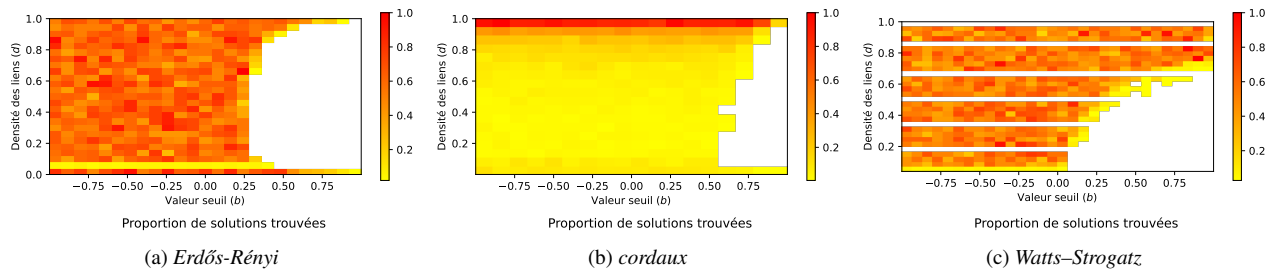


FIGURE 2 : Proportion de réussites pour trouver une matrice de corrélation $\mathbf{C} \in \mathbb{R}^{51 \times 51}$ par bin en fonction des paramètres (b, d) , utilisant les modèles (a) *Erdős-Rényi*, (b) *cordaux* et (c) *Watts-Strogatz*. Les zones blanches correspondent à une absence de solution.

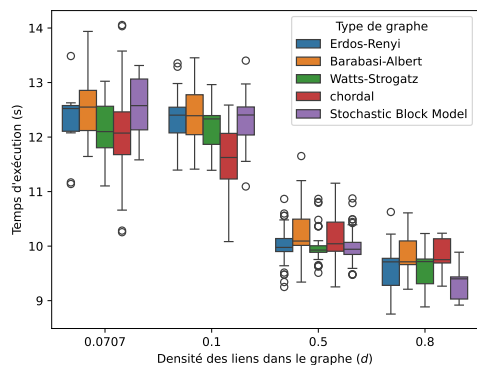


FIGURE 3 : Temps d'exécution (en secondes) de notre méthode pour calculer la matrice de corrélation $\mathbf{C} \in \mathbb{R}^{51 \times 51}$, moyenné sur 50 exécutions pour chaque boîte à moustaches (représentant un type de graphe différent) en fonction de la densité des éléments non nuls et non diagonaux dans la matrice de corrélation. Le paramètre b est fixé à -1 .

[4] Réka ALBERT et Albert-László BARABÁSI : Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[5] Cherie ARMOUR, Eiko I FRIED, Marie K DESERNO, Jack TSAI et Robert H PIETRZAK : A network analysis of DSM-5 posttraumatic stress disorder symptoms and correlates in US military veterans. *Journal of anxiety disorders*, 45:49–59, 2017.

[6] Guillaume J-PC BECQ, Tarik HABET, Nora COLLOMB, Margaux FAUCHER, Chantal DELON-MARTIN, Véronique COIZET, Sophie ACHARD et Emmanuel L BARBIER : Functional connectivity is preserved but reorganized across several anesthetic regimes. *NeuroImage*, 219:116945, 2020.

[7] Alice CHEVAUX, Ali FAHKAR, Kévin POLISANO, Irène GANNAZ et Sophie ACHARD : Benchmarking Brain Connectivity Graph Inference : A Novel Validation Approach. In *33rd European Signal Processing Conference (EUSIPCO 2025)*, Palerme, Italy, septembre 2025.

[8] Siddhartha CHIB et Edward GREENBERG : Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

[9] Irene CÓRDOBA, Gherardo VARANDO, Concha BIELZA et Pedro LARRAÑAGA : On generating random Gaussian graphical models. *International Journal of Approximate Reasoning*, 125:240–250, 2020.

[10] Steven DIAMOND et Stephen BOYD : CVXPY : A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

[11] Ali FAHKAR, Poliano KÉVIN, Irène GANNAZ et Sophie ACHARD : Code of the present paper, 2025. <https://gricad-gitlab.univ-grenoble-alpes.fr/polisank/generating-correlation-matrices-with-graph-structures-using-convex-optimization>.

[12] Christophe GIRAUD : *Introduction to high-dimensional statistics*. CRC Press, 2021.

[13] Maxim GRECHKIN, Maryam FAZEL, Daniela WITTEN et Su-In LEE : Pathway graphical lasso. *Proceedings of the AAAI conference on artificial intelligence*, 29(1), 2015.

[14] Roger A HORN et Charles R JOHNSON : *Matrix analysis*. Cambridge University Press, 2012.

[15] Shuai HUANG, Jing LI, Liang SUN, Jieping YE, Adam FLEISHER, Teresa WU, Kewei CHEN, Eric REIMAN et Alzheimer's Disease NeuroImaging INITIATIVE : Learning brain connectivity of Alzheimer's disease by sparse inverse covariance estimation. *NeuroImage*, 50(3):935–949, 2010.

[16] John HULL : *Risk management and financial institutions, + Web Site*, volume 733. John Wiley & Sons, 2012.

[17] Harry JOE : Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97(10): 2177–2189, 2006.

[18] Angelika KIMMIG, Lilyana MIHALKOVA et Lise GETOOR : Lifted graphical models : a survey. *Machine Learning*, 99:1–45, 2015.

[19] Daphne KOLLER et Nir FRIEDMAN : *Probabilistic graphical models : principles and techniques*. MIT press, 2009.

[20] Daniel LEWANDOWSKI, Dorota KUROWICKA et Harry JOE : Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9):1989–2001, 2009.

[21] Xudong LI, Defeng SUN et Kim-Chuan TOH : QSDPNAL : A two-phase augmented lagrangian method for convex quadratic semidefinite programming. *Mathematical Programming Computation*, 10:703–743, 2018.

[22] Gautier MARTI, Victor GOUBET et Frank NIELSEN : cCorrGAN : Conditional correlation gan for learning empirical conditional distributions in the ellipsope. In *International Conference on Geometric Science of Information*, pages 613–620. Springer, 2021.

[23] Judea PEARL : *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Elsevier, 2014.

[24] Mohsen POURAHMADI et Xiao WANG : Distribution of random correlation matrices : Hyperspherical parameterization of the Cholesky factor. *Statistics & Probability Letters*, 106:5–12, 2015.

[25] Lieven VANDENBERGHE et Martin S ANDERSEN : Chordal graphs and semidefinite optimization. *Foundations and Trends® in Optimization*, 1(4):241–433, 2015.