

GÉNÉRATION DE MODÈLES GRAPHIQUES PAR OPTIMISATION CONVEXE

Ali Fakhar¹ & Sophie Achard² & Irène Gannaz³ & Kevin Polisano⁴

¹ *Univ. Grenoble Alpes, LJK, F-38000 Grenoble, France, ali.fakhar@univ-grenoble-alpes.fr*

² *Univ. Grenoble Alpes, CNRS, LJK, F-38000 Grenoble, France
sophie.achard@univ-grenoble-alpes.fr*

³ *Univ. Grenoble Alpes, CNRS, Grenoble INP, G-SCOP, 38000 Grenoble, France,
irene.gannaz@grenoble-inp.fr*

⁴ *Univ. Grenoble Alpes, CNRS, LJK, F-38000 Grenoble, France,
kevin.polisano@univ-grenoble-alpes.fr*

Résumé. Ce travail concerne la génération de matrices de corrélation théoriques avec des motifs de parcimonie spécifiques, associés à des structures de graphes. Nous présentons une nouvelle approche basée sur l'optimisation convexe, offrant une plus grande flexibilité par rapport aux techniques existantes. Notre méthode permet de générer des matrices de corrélation qui peuvent être utilisées pour comparer les méthodes d'inférence de graphes.

Mots-clés. Matrices de corrélation, modèles graphiques, structures de parcimonie

Abstract. This work deals with the generation of theoretical correlation matrices, with fixed sparsity patterns. We propose an optimization approach. It offers some flexibility, in particular by the addition of some constraints. The generated correlation matrices may be used for comparisons of statistical inferences in graphical models.

Keywords. Correlation matrices, graphical models, sparsity patterns

1 Introduction

Les matrices de corrélation 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. 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. Des exemples d'application sont 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.

2 Notations

Nous considérons dans la suite une dimension p fixée. Une matrice symétrique réelle $\mathbf{A} \in \mathbb{R}^{p \times p}$ est semi-définie positive (PSD) si $\mathbf{x}^\top \mathbf{A} \mathbf{x} \geq 0$ pour tout $\mathbf{x} \in \mathbb{R}^p$. Elle est définie positive (PD) si $\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$ pour tout $\mathbf{x} \in \mathbb{R}^p$ non nul.

Soit $\mathbf{x} \in \mathbb{R}^p$ un vecteur aléatoire de matrice de covariance Σ , défini comme : $\Sigma = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\top] = (\sigma_{ij})_{i,j=1,\dots,p}$. La matrice de corrélation correspondante $\mathbf{C} \in \mathcal{C}$ est définie par $c_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$ pour tout $(i, j) \in \{1, \dots, p\}$, ou de manière équivalente $\mathbf{C} = (\text{diag}(\Sigma))^{-\frac{1}{2}} \Sigma (\text{diag}(\Sigma))^{-\frac{1}{2}}$. L'ensemble $\mathcal{C}$ de matrices de corrélation satisfait

$$\forall \mathbf{C} \in \mathcal{C} \quad \text{diag}(\mathbf{C}) = 1, \quad \forall i, j \in \{1, 2, \dots, p\} \quad -1 \leq c_{ij} \leq 1. \quad (1)$$

D'un point de vue génératif, travailler avec des matrices de corrélation et des matrices de précision est similaire. Par conséquent, nous nous concentrons sur les matrices de corrélation. La génération d'une matrice de corrélation implique la construction d'une matrice PSD symétrique qui satisfait la condition (1) [5, Problème 7.1].

Un graphe $\mathcal{G}$ est défini comme $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, où $\mathcal{V} = \{v_1, v_2, \dots, v_p\}$ est l'ensemble des sommets, et $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ est l'ensemble des arêtes. Une arête est une paire de sommets connectés par le graphe. La densité du graphe est alors $d = \frac{|\mathcal{E}|}{p(p-1)}$.

Nous désignons par $\mathcal{C}(\mathcal{G})$ l'ensemble des matrices de corrélation associées à un graphe donné $\mathcal{G}$. Une matrice $\mathbf{C} \in \mathcal{C}(\mathcal{G})$ est telle que

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

Nous recherchons à générer une matrice de corrélation dans $\mathcal{C}(\mathcal{G})$. En d'autres termes, nous imposons la parcimonie à certaines relations. Nous souhaitons générer des matrices dans $\mathcal{C}(\mathcal{G})$. Ce problème peut être considéré comme un problème de complétion de matrice [9, Chapitre 10].

3 Méthodes existantes

La génération d'une matrice de corrélation (théorique) a été proposée par exemple dans [6] avec une caractérisation de la distribution uniforme sur l'espace des matrices de corrélation. Les approches courantes incluent les *vines* et *onion* [7]. Pour une revue plus large des techniques disponibles, nous renvoyons aux études bibliographiques dans [8, 4]. La plupart des méthodes existantes ne peuvent cependant pas être étendues pour générer des matrices de corrélation avec un ensemble de parcimonie $\bar{\mathcal{E}}$ contraint. Ci-dessous, nous présentons les méthodes qui, à notre connaissance, peuvent générer de telles matrices.

3.1 Graphes chordaux et décomposition de Cholesky

La première approche repose sur la décomposition de Cholesky. Soit $\mathcal{U} \in \mathbb{R}^{p \times p}$ l'ensemble des matrices triangulaires supérieures, à éléments diagonaux positifs et à lignes normalisées

à 1. On définit $\mathcal{U}(\mathcal{G}) \subset \mathcal{U}$ comme le sous-ensemble où $u_{ij} = 0$ pour tout $(i, j) \in \bar{\mathcal{E}}$. Si le graphe est ordonné (Perfect Elimination Order, PEO), il est possible de générer le facteur de Cholesky $\mathbf{U}$ dans $\mathcal{U}(\mathcal{G})$ en imposant $u_{ij} = 0$ pour tout $(i, j) \in \bar{\mathcal{E}}$ pour obtenir une matrice $\mathbf{C} = \mathbf{U}\mathbf{U}^\top \in \mathcal{C}(\mathcal{G})$ telle que $m_{ij} = 0$ pour tout $(i, j) \notin \mathcal{E}$. Pour plus de détails, nous nous référons à [4].

Dans [8], les auteurs proposent une écriture en coordonnées polaires des éléments de la matrice de Cholesky $\mathbf{U} \in \mathcal{U}$. Ils établissent la distribution de probabilité des quantités induites, telle que la distribution des matrices obtenues soit uniforme sur l'ensemble $\mathcal{C}$. En utilisant cette paramétrisation en coordonnées polaires, il est aisé d'incorporer la contrainte $u_{ij} = 0$ pour tout $(i, j) \in \bar{\mathcal{E}}$, pour obtenir $\mathbf{U} \in \mathcal{U}(\mathcal{G})$.

Dans [4] la première contrainte est déjà prise en compte. La méthode de génération proposée est basée sur l'algorithme de Metropolis-Hastings et donne une distribution uniforme sur $\mathcal{C}(\mathcal{G})$.

Comme mentionné ci-dessus, ces méthodes supposent que les graphes sont ordonnés (PEO), ce qui n'est généralement pas le cas. En fait, seuls les graphes dits cordaux peuvent être parfaitement ordonnés [9, Chapitre 4]. Un graphe est cordal s'il ne contient pas de cycle induit de longueur 4 ou plus. Cette approche n'est ainsi utilisable que pour les graphes cordaux, ce qui est une limitation forte.

3.2 Dominance diagonale et orthogonalisation partielle

A notre connaissance, deux méthodes ont été proposées dans la littérature pour générer des matrices de corrélation associées à un graphe donné $\mathcal{G}$ sans nécessiter de structure cordale : la dominance diagonale et l'orthogonalisation partielle.

La dominance diagonale telle que proposée dans [4] peut être utilisée pour générer une matrice PD. L'idée est de construire une matrice symétrique $\tilde{\mathbf{C}}$, où les éléments $\tilde{c}_{ij}$ sont choisis uniformément au hasard dans l'intervalle $[-1, 1]$ si $(i, j) \notin \bar{\mathcal{E}}$, et $\tilde{c}_{ij} = 0$ sinon. Ensuite, on applique la transformation

$$\forall i \in \{1, \dots, p\} \quad \text{do } \tilde{c}_{ii} \leftarrow \sum_{\substack{j=1, \dots, p \\ i \neq j}} |\tilde{c}_{ij}| + \text{perturbation positive aléatoire.}$$

Ensuite, on définit $\mathbf{C} = \text{diag}(\tilde{\mathbf{C}})^{-1/2} \tilde{\mathbf{C}} \text{diag}(\tilde{\mathbf{C}})^{-1/2}$ pour récupérer une matrice dans $\mathcal{C}(\mathcal{G})$.

Une autre méthode est l'orthogonalisation partielle, proposée dans [4]. L'idée est de commencer avec une matrice initiale $\mathbf{C}$ dont les entrées nulles sont incluses dans le motif de parcimonie souhaité $\bar{\mathcal{E}}$. Les arêtes supplémentaires de $\mathbf{C}$ sont ensuite supprimées à l'aide d'un processus d'orthogonalisation partielle basé sur un algorithme de Gram-Schmidt modifié. La procédure repose sur une écriture de $\mathbf{C}$ sous la forme $\mathbf{C} = \mathbf{Q}\mathbf{Q}^\top$, et orthogonalise itérativement chaque ligne $\mathbf{q}_i$, par rapport à l'ensemble des lignes $\{\mathbf{q}_j, s.t. (i, j) \notin \mathcal{E} \text{ et } j < i\}$.

Dans [4], les auteurs suggèrent de commencer par trianguler le graphe $\mathcal{G}$ pour obtenir un graphe cordal, puis d'appliquer la procédure basée sur Cholesky de [3], comme décrit ci-dessus. La matrice obtenue est alors l'initialisation de l'algorithme d'orthogonalisation partielle.

4 Méthode 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). De plus, des contraintes supplémentaires peuvent être ajoutées, selon le contexte. Une de nos motivations est de construire des matrices de corrélation avec une distribution qui ressemble à des données du monde réel, notamment en neurosciences. Dans [1], la distribution des corrélations de connectivité cérébrale s’est avérée centrée autour de valeurs positives. 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 $b = -1$ équivaut à n’avoir aucune contrainte.

Nous cherchons à résoudre le problème d’optimisation suivant :

$$\begin{aligned} & \underset{\mathbf{C}}{\text{minimize}} && \frac{1}{2} \|\mathbf{C} - \bar{\mathbf{C}}\|_F^2, \\ & \text{sous réserve de} && \text{contraintes (2) et (3),} \end{aligned} \quad (4)$$

avec $\bar{\mathbf{C}}$ une matrice arbitraire donnée. Avec des données réelles, cela peut être la matrice de corrélation empirique.

5 Résultats

Dans nos simulations, nous considérons $p = 51$, $\mathbf{C} \in \mathbb{R}^{51 \times 51}$ et $\bar{\mathbf{C}}$ une matrice de même taille dont les entrées sont tirées d’une distribution uniforme sur l’intervalle $[-1, 1]$. Pour le motif $\mathcal{E}$, nous utilisons un graphe *Erdős-Rényi*. D’autres structures de graphes ont été considérées (*Barabási-Albert*, *Watts-Strogatz*, *Stochastic Block Model*...) mais ne sont pas présentées ici.

5.1 Comparaison avec les autres approches

La Figure 1 montre la densité des éléments non diagonaux et non nuls des matrices de corrélation générées. Nous comparons notre méthode avec deux autres approches : la dominance diagonale et l’orthogonalisation partielle. Nous avons considéré 50 graphes, avec une densité des graphes cibles à 0.5. Dans notre algorithme, la matrice initiale $\bar{\mathbf{C}}$ est obtenues avec des réalisations d’une distribution uniforme sur $[-1, 1]$. Les densités obtenues en utilisant la dominance diagonale sont concentrées autour de valeurs faibles. Notre approche donne des valeurs plus élevées dans la matrice de corrélation que les autres procédures.

5.2 Influence de la contrainte sur la moyenne

Nous examinons maintenant l’impact de la contrainte dans (3), qui modifie le centrage de la distribution des entrées de corrélation. Nous sommes motivés par une application en

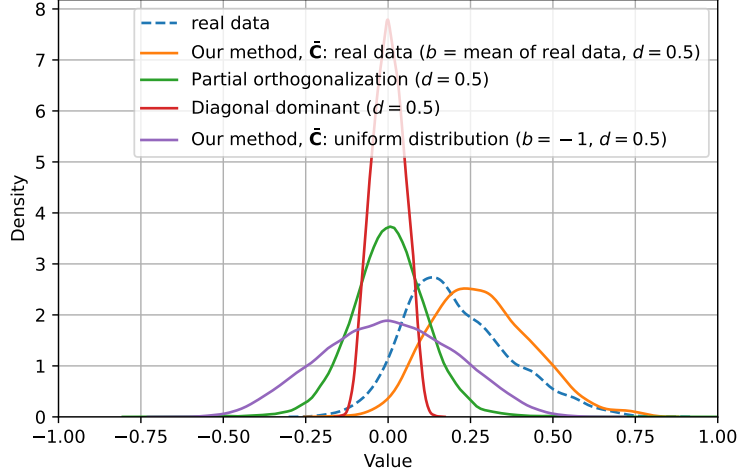


Figure 1: Densité des éléments non diagonaux et non nuls dans les matrices de corrélation générées à l’aide de différentes méthodes.

neurosciences impliquant des données d’IRM fonctionnelle acquises sur des rats. Les données sont accessibles en accès libre et décrites dans [2]. Après le prétraitement expliqué dans [2], des séries temporelles de 51 régions cérébrales pour chaque rat ont été extraites. Nous calculons ensuite la transformée en ondelettes 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 a alors 122 coefficients d’ondelettes disponibles pour chacune des 51 régions. Les distributions des corrélations deux à deux entre les coefficients d’ondelettes des régions pour un rat donné. Nous calculons la matrice de corrélation en sélectionnant les 50% d’entrées ayant les valeurs absolues les plus élevées dans la matrice de corrélation empirique, ce qui donne une densité du graphe $d = 50\%$. Les distributions obtenues sont présentées dans la Figure 1.

Nous générons ensuite la matrice synthétique avec notre méthode, avec $d = 50\%$ et le paramètre b égal à la moyenne des entrées dans la matrice de corrélation des données réelles. La matrice initiale $\tilde{\mathbf{C}}$ est ici égale à la corrélation empirique des données réelles. La Figure 1 montre que la distribution des données simulées est proche de la distribution réelle.

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

Conclusion

La méthode proposée offre plusieurs avantages. Elle ne repose pas sur une structure cordale, garantit la génération d’une matrice semi-définie positive et évite de générer des entrées proches de zéro dans la matrice de corrélation. De plus, elle permet de générer une matrice de corrélation qui correspond à un graphe lié à une matrice de corrélation empirique.

D’autres analyses, notamment sur l’influence de la structure du graphe sur la distribution

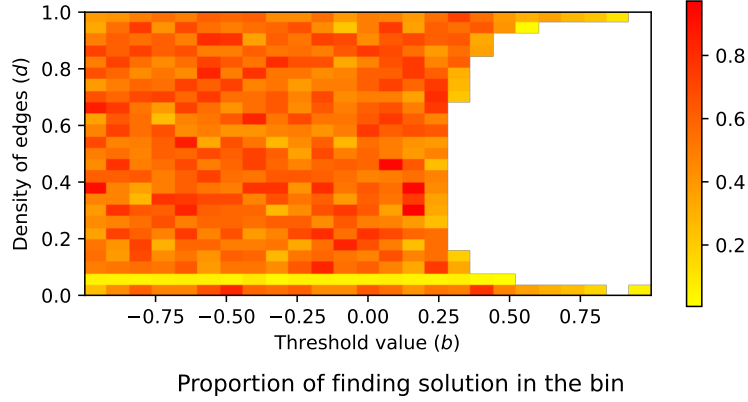


Figure 2: Proportion de réussite à trouver une matrice de corrélation $\mathbf{C} \in \mathbb{R}^{51 \times 51}$ en fonction du seuil b , et de la densité du graphe d . Les zones blanches indiquent les cas où aucune solution n’a été trouvée.

des corrélations, sont en cours. Dans la continuité de ce travail, nous souhaitons comparer les performances de divers procédures d’inférence de modèles graphiques sur données simulées.

Bibliographie

- [1] Achard, S., Gannaz, I., & Polissano, K. (2022, September). Génération de modèles graphiques. In GRETSI 2022-XXVIIIème Colloque francophone de traitement du signal et des images (pp. 1-3).
- [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., & Larrañaga, P. (2018). A fast Metropolis-Hastings method for generating random correlation matrices. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 117-124). Cham: Springer International Publishing.
- [4] Córdoba, I., Varando, G., Bielza, C., & Larrañaga, P. (2020). On generating random Gaussian graphical models. *International Journal of Approximate Reasoning*, 125, 240-250.
- [5] Horn, R. A., & Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.
- [6] Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97(10), 2177-2189.
- [7] Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9), 1989-2001.
- [8] Pourahmadi, M., & Wang, X. (2015). Distribution of random correlation matrices: Hyperspherical parameterization of the Cholesky factor. *Statistics & Probability Letters*, 106, 5-12.
- [9] Vandenberghe, L., & Andersen, M.S. Chordal graphs and semidefinite optimization. (2015). *Foundations and Trends in Optimization*, 2015, 1(4), 241-433.