SHGR: A Generalized Maximal Correlation Coefficient

Samuel Stocksieker[1], Denys Pommeret[1]

[1] CNRS, I2M, Aix-Marseille Université

samuel.stocksieker@univ-amu.fr

1 Résumé

1.1 Introduction de SHGR : une nouvelle mesure de corrélation non linéaire

La compréhension des dépendances entre variables constitue un enjeu central en apprentissage automatique, en statistique et en science des données, avec des applications essentielles telles que la sélection de variables, la réduction de dimension, l'évaluation de l'équité, l'inférence causale ou l'apprentissage multimodal. Les mesures classiques, en particulier les corrélations de Pearson et de Spearman, sont largement utilisées, mais se limitent aux relations linéaires/monotones et binaires, échouant souvent à saisir des dépendances plus complexes ou d'ordre supérieur. Pour pallier ces limitations, plusieurs mesures de dépendance généralisées ont été proposées. Parmi elles, la corrélation maximale de Hirschfeld-Gebelein-Rényi (HGR) se distingue comme un outil théoriquement fondé pour quantifier la dépendance non linéaire entre variables aléatoires, qu'elles soient univariées ou multivariées. Introduite par Hirschfeld ([2]), étendue par Gebelein ([1]) et formalisée par Rényi ([5]), la mesure HGR définit la corrélation comme la corrélation linéaire maximale entre des versions transformées des variables. Malgré son attrait théorique, l'estimation du coefficient HGR demeure un défi, en raison de la difficulté d'identifier les transformations optimales. D'autres méthodes présentent des inconvénients similaires, tels qu'une interprétabilité limitée, une complexité computationnelle élevée ou une difficulté d'extension à des configurations multivariées. De plus, la sensibilité des estimateurs HGR neuronaux aux valeurs aberrantes peut conduire à des comportements de sur-apprentissage et à des scores de corrélation artificiellement gonflés.

Pour remédier à ces limitations, nous proposons une extension naturelle du coefficient HGR. Notre approche s'inspire du coefficient de Spearman et des méthodes basées sur les copules, qui reposent sur les fonctions de répartition et sont intrinsèquement robustes aux valeurs extrêmes et aux relations non linéaires. Cette substitution offre deux avantages principaux : (i) elle améliore la robustesse aux valeurs extrêmes, ii) elle présente une invariance par transformation monotone et (iii) elle étend la portée de la dépendance mesurable pour inclure des relations monotones mais non linéaires. Il est important de noter qu'elle reste cohérente avec le principe HGR de maximisation de la corrélation entre les transformations non linéaires des variables, mais recadre la notion de dépendance en termes de monotonicité basée sur les rangs plutôt que de linéarité brute. Cela pourrait aider à réduire les efforts de transformation et à optimiser la calibration pour éviter de rechercher une corrélation linéaire, mais simplement une corrélation monotone. Nous rappelons que le coefficient de corrélation de Spearman entre deux échantillons appariés iid $u = (u_1, \cdots, u_n)$ et $v = (v_1, \cdots, v_n)$ issus de U et V, noté $\rho(u, v)$, est défini comme le coefficient de corrélation de Pearson basé sur les rangs :

$$\rho(u,v) = r\left(n\widehat{F}_U(u), n\widehat{F}_V(v)\right) = r\left(\widehat{F}_U(u), \widehat{F}_V(v)\right)$$
$$\text{avec } \widehat{F}_U(u) = (\widehat{F}_U(u_1), \cdots, \widehat{F}_U(u_n))$$

DEFINITION 1. (Coefficient Spearman-HGR (SHGR)). Soient U et V deux variables aléatoires continues appariées prenant leurs valeurs dans $\mathcal{U}$ et $\mathcal{V}$, respectivement. Soit $\mathcal{E}(\mathcal{U})$ (resp. $\mathcal{E}(\mathcal{V})$) l'ensemble des fonctions mesurables de $\mathcal{U}$ (resp. $\mathcal{V}$) vers $\mathbb{R}$. Le coefficient Spearman-HGR (SHGR) associé à (U, V) est défini par $SHGR(U, V) :=$

$$\max_{\substack{f_u\in\mathcal{E}(\mathcal{U}), f_v\in\mathcal{E}(\mathcal{V})\\ \mathbb{E}(f_u(U))=0, \mathbb{E}(f_v(V))=0\\ \mathbb{E}(f_u^2(U))=1, \mathbb{E}(f_v^2(V))=1}} r(F_{f_u(U)}(f_u(U)), F_{f_v(V)}(f_v(V))).$$

Il est important de noter que la transformation par copule préserve la dépendance entre les vecteurs originaux U et V (voir par exemple, [4]). En utilisant l'estimateur empirique $\widehat{F} : \widehat{F}_z(z_i) := \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{z_j \le z_i}$ de F, nous obtenons un estimateur du SHGR $\widehat{SHGR} =$:

$$\max_{\substack{f_u\in\mathcal{E}(\mathcal{U}), f_v\in\mathcal{E}(\mathcal{V})\\ \mathbb{E}(f_u(U))=0, \mathbb{E}(f_v(V))=0\\ \mathbb{E}(f_u^2(U))=1, \mathbb{E}(f_v^2(V))=1}} r(\widehat{F}_{f_u(U)}(f_u(U)), \widehat{F}_{f_v(V)}(f_v(V))),$$

et sa version neuronale $SHGR_\Theta(u, v) =$:

$$= \max_{f_{\theta_u}, f_{\theta_v}\in\Theta} \rho(f_{\theta_u}(u), f_{\theta_v}(v))$$
$$= \max_{f_{\theta_u}, f_{\theta_v}\in\Theta} r(\widehat{F}_{f_{\theta_u}(U)}(f_{\theta_u}(u)), \widehat{F}_{f_{\theta_v}(V)}(f_{\theta_v}(v)))$$

Notons que la transformation par rang est appliquée uniquement au calcul de la corrélation, et non aux variables d'entrée elles-mêmes. Cela préserve l'information des entrées originales tout en bénéficiant de la robustesse des objectifs basés sur les rangs. Typiquement, pour l'estimation bivariée, l'algorithme considère p variables d'entrée $u_1, \ldots, u_p$

et vise à optimiser les entrées correspondantes de la matrice de corrélation. La conception à encodeurs empilés permet des transformations marginales capturant des dépendances complexes et non linéaires entre les variables. La fonction objectif pour les estimations bivariées (corrélations par paires) est définie comme suit :

$$\mathcal{L}_{SHGR}(\boldsymbol{u}) := -\sum_{i,j=1\ i\neq j}^{p}\left[\rho^2\left(f_{\theta_{u_i}}(u_i), f_{\theta_{u_j}}(u_j)\right)^{1/2}\right]^{\alpha}$$

with $\alpha > 0$

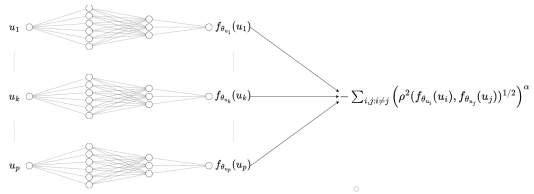
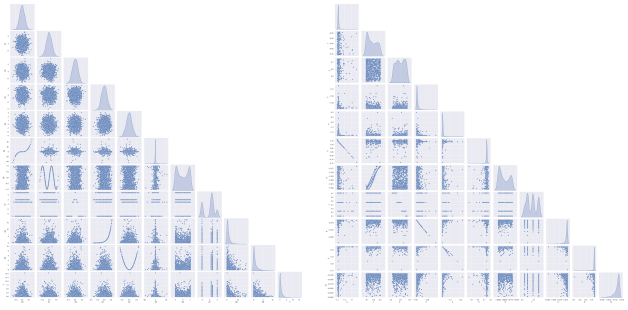


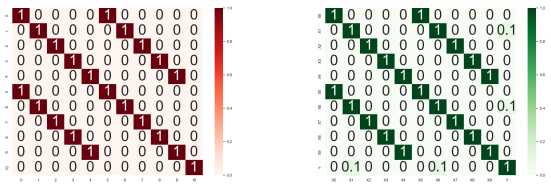
FIGURE 1 – Cross-Encoder Architecture

Ce nouveau coefficient estime des matrices de corrélation par paires, multivariées (p-vs-1 variables) et par groupes (p-vs-q variables). Cette extension s'appuie sur deux composantes clés : (i) une approximation neuronale conçue pour l'estimation simultanée de corrélations (Figure 1), et (ii) une formulation à base de copules qui améliore la robustesse et la stabilité en réduisant la sensibilité aux valeurs extrêmes.
La figure 2 présente une illustration de l'approche : l'algorithme transforme les données pour que les corrélations non linéaires deviennent linéaires.



(a) Inputs Pairplot

(b) SHGR Transformations



(c) Reference Correlation

(d) SHGR Correlation

FIGURE 2 – Illustration of SHGR on bivariate correlations

1.2 Protocole d'évaluation numérique

Nous proposons d'évaluer le SHGR et ses concurrents au moyen d'un protocole complet suivant l'analyse ci-dessous (définissant la *puissance multivariée des mesures de corrélation*) :

- Performance : capacité à capturer des corrélations non linéaires complexes sans bruit ;
- Robustesse au bruit : capacité à identifier des corrélations non linéaires complexes en présence de bruit ;
- Robustesse à l'hallucination : corrélation nulle en cas d'indépendance ;
- Robustesse aux valeurs extrêmes : corrélation nulle en cas d'indépendance, en présence de valeurs extrêmes ;
- *Puissance bivariée d'une mesure de dépendance* telle que proposée par [3] ;
- Temps de calcul : rapidité d'estimation des corrélations ;
- Analyse du test de significativité : possibilité d'effectuer un test de significativité sur le coefficient et capacité à rejeter l'hypothèse nulle (de corrélation nulle) en présence de bruit.

Nous appliquons le protocole de *puissance multivariée des mesures de corrélation* dans trois configurations : (i) corrélations bivariées (par paires) (1-vs-1), (ii) corrélations multivariées (p-vs-1), et (iii) corrélations complètes (par groupes) (p-vs-q). Pour (i), nous générons six variables gaussiennes indépendantes et cinq autres qui en dépendent de manière non linéaire (Figure 2a). Pour (ii), nous simulons 20 variables, incluant des dépendances non linéaires impliquant plus de deux variables. Pour (iii), nous générons deux jeux de données présentant diverses structures de corrélation globale. Nous estimons les corrélations non linéaires à l'aide de SHGR et les comparons aux méthodes alternatives suivantes (certaines n'étant pas disponibles dans les configurations multivariées et complètes) :

- Le coefficient de corrélation de Pearson (Pearson) ;
- Le coefficient de corrélation de Spearman (Spearman) ;
- Le coefficient de corrélation de Kendall (Kendall) ;
- Le Randomized Dependence Coefficient (RDC) ;
- Le Mutual Information Criterion (MIC) ;
- La Distance Correlation (dCor) ;
- L'Analyse en Corrélations Canoniques (CCA) ;
- L'Analyse en Corrélations Canoniques à noyau ; : cette approche n'a pas été intégrée car instable et trop coûteuse en temps de calcul
- L'Alternating Conditional Expectations (ACE) ;
- L'estimation HGR par noyau (HGRkde) ;
- L'estimation HGR par réseau neuronal (HGRnn) ;
- L'estimation HGR par treillis (HGRlat) ;
- L'estimation HGR par double noyau (dk) ;
- L'estimation HGR par noyau simple (sk) ;
- Le Normalized Hilbert-Schmidt Independence Criterion (HSIC) ;

— La corrélation de Hellinger : Cette approche n'a pas été intégrée car inefficace, limitée aux corrélations par paires et trop coûteuse en temps de calcul.

Nous entraînons notre modèle SHGR pendant 100, 200 et 500 epochs, ainsi qu'avec arrêt anticipé. Une analyse de sensibilité de l'architecture et des hyperparamètres de SHGR a été réalisée. Une architecture et une configuration d'hyperparamètres uniques ont ensuite été utilisées de manière cohérente pour l'ensemble des illustrations et expériences. Dans les analyses ci-dessous, nous générons $K = 10$ jeux de données synthétiques et comparons les corrélations estimées par chaque méthode aux corrélations vraies, connues, appelées corrélations de référence. Nous évaluons l'écart à ces corrélations de référence via : la matrice de corrélation par paires pour la corrélation bivariée, le vecteur de corrélations multiples pour la corrélation multivariée, et le coefficient de corrélation pour l'analyse de corrélation complète.
Les figures ci-dessous présentent les résultats obtenus dans le cas de corrélations bivariées. On remarque que le SHGR est plus performant et plus robuste que les compétiteurs. En effet, la mesure est la distance à la matrice de corrélation de référence : plus bas est a valeur meilleur est le coefficient.

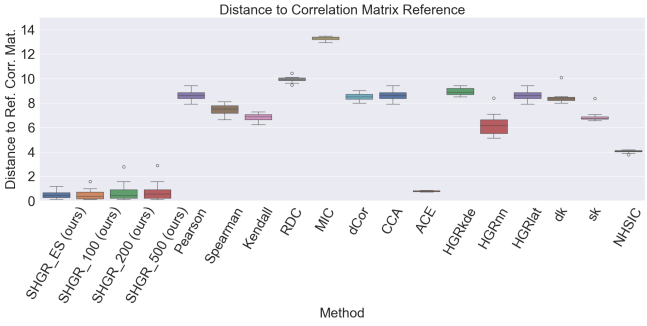


FIGURE 3 – Performance (Correlation bivariée)

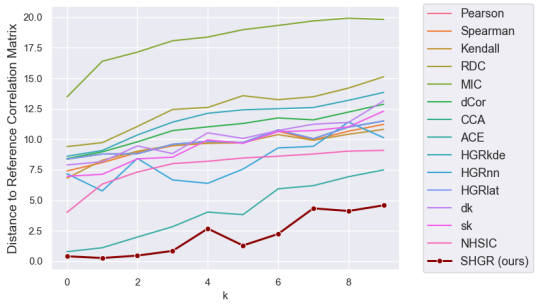


FIGURE 4 – Robustesse au bruit (Correlation bivariée)

Nos principales contributions peuvent être résumées comme suit :

— Nous revisitons la corrélation maximale HGR en proposant une extension basée sur Spearman : SHGR
— Nous présentons un premier estimateur de SHGR qui est i) ii) différentiable, iii) rapide, iv) efficace : détecte efficacement les liens non linéaires v) robuste au bruit :

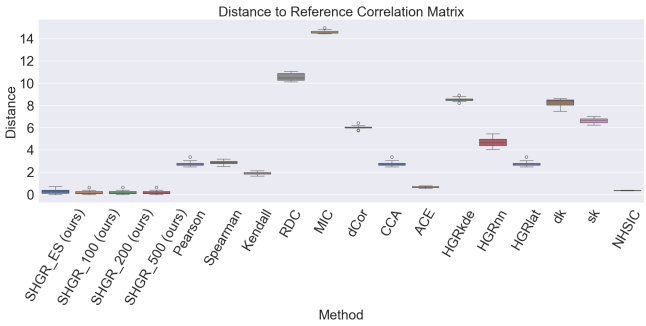


FIGURE 5 – Robustesse aux hallucinations (Correlation bivariée)

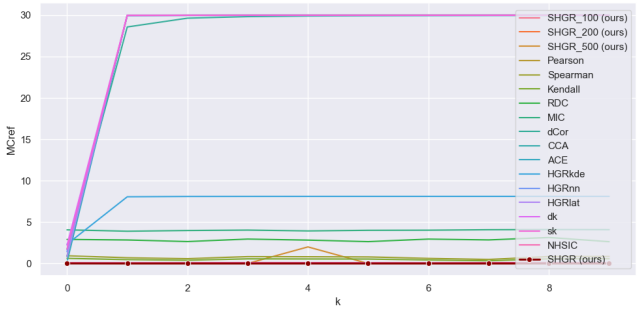


FIGURE 6 – Robustesse aux outliers (Correlation bivariée)

detecte les liens non linéaires en présence de bruit vi) robuste aux valeurs aberrantes : invariance du coefficient en présence de valeurs extrêmes et/ou abérantes et vii) robuste aux dépendances hallucinatoires : n'indique pas de corrélation signficatives lorsque les variales sont indépendantes. Contrairement à certaines méthodes antérieures, il récupère également les transformations de données et supporte les tests de significativité de la corrélation.

— Nous introduisons une architecture de cross-encodeur empilé spécifiquement conçue pour estimer simultanément (non itérative) plusieurs corrélations en contextes bivariés et multivariés.
— L'interprétabilité, généralement limitée pour les méthodes existantes, peut être partiellement améliorée via les transformations apprises accessibles dans SHGR, ouvrant des pistes d'analyse visuelle.
— Nous déduisons un protocole d'évaluation complet, la *Puissance multivariée de la mesure de corrélation*, pour évaluer les estimateurs de corrélation maximale en termes de performance, robustesse au bruit, hallucinations (respectant l'Axiome 4 de Rényi), valeurs extrêmes, ainsi que l'estimation de corrélations bivariées, multivariées et complètes, tests de significativité et efficacité computationnelle.
— Nous validons SHGR sur des ensembles de données tabulaires synthétiques et réels. Une évaluation sur des applications réelles de sélection de vairables est également réalisée, sur 9 datasets. Nous démontrons que SHGR surpasse les méthodes de l'état de l'art exis-

tantes en termes de performance et de robustesse. L'article original est disponible à l'adresse : https://neurips.cc/virtual/2025/loc/san-diego/poster/117128. SHGR constitue une base prometteuse pour de futurs travaux, notamment en haute dimension, en architectures à grande échelle, en apprentissage multimodal ou dans des contextes d'équité algorithmique.

Mots-clés

Corrélation non linéaire

Références

[1] Hans Gebelein. Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *Zammzeitschrift Fur Angewandte Mathematik Und Mechanik*, 21 :364–379, 1941.

[2] H. O. Hirschfeld. A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4) :520–524, 1935.

[3] David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. The randomized dependence coefficient. *Advances in neural information processing systems*, 26, 2013.

[4] Roger B. Nelsen. *An introduction to copulas.* Springer, New York, 2006.

[5] Alfréd Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10 :441–451, 1959.