Découverte de causalité pour séries temporelles utilisant un modèle constraint-based et la mesure d'information PMIME

Antonin Arsac, Aurore Lomet, Jean-Philippe Poli

Découverte de causalité pour séries temporelles utilisant un modèle *constraint-based* et la mesure d'information PMIME

Antonin Arsac [*,1] & Aurore Lomet [*,2] & Jean-Philippe Poli [*,3]

[*] *Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France*
[1] *antonin.arsac@cea.fr,* [2] *aurore.lomet@cea.fr,* [3] *jean-philippe.poli@cea.fr*

Résumé. La causalité est une notion qui définit les relations entre les causes et leurs effets. Dans le cadre des séries temporelles multivariées, cette notion permet de caractériser les liens entre plusieurs séries temporelles en considérant possiblement les décalages temporels. Par exemple, dans l'industrie elle détecte notamment les causes d'une anomalie dans un système complexe. En pratique, les systèmes complexes mesurés reposent peu souvent sur les hypothèses de données gaussiennes ou de linéarité entre les liens requises dans de nombreuses méthodes d'apprentissage automatique. Pour pallier ce problème, nous présentons dans ce papier une nouvelle approche pour établir les liens causaux entre séries temporelles qui combine un algorithme de découverte causale avec une mesure basée sur la théorie de l'information. Ainsi, l'approche proposée permet d'identifier des liens linéaires et non-linéaires et de construire le graphe causal sous-jacent. La méthode est évaluée sur différentes bases de données simulées issues de la littérature et montre des performances intéressantes.

Mots-clés. Causalité, Découverte causale *constraint-based*, Théorie de l'information, Séries temporelles

Abstract. Causality defines the relationship between cause and effect. In multivariate time series field, this notion allows to characterize the links between several time series considering temporal lags. These phenomena are particularly important in medicine to analyze the effect of a drug for example, in manufacturing to detect the causes of an anomaly in a complex system Most of the time, those complex systems do not rely on the assumption of linearity required in many machine learning methods. To circumvent this problem, we present in this paper a new approach for discovering causality in time series data that combines a causal discovery algorithm with an information theoretic-based measure. Hence the proposed method allows inferring both linear and non-linear relationships and building the underlying causal graph. We evaluate the performance of our approach on several simulated datasets, showing promising results.

Keywords. Causality, Consraint-based causal discovery, Information theory, Time series

1 Introduction

La causalité joue un rôle clé dans la compréhension humaine du monde. Ces deux dernières décennies ont vu se développer la formalisation des modèles causaux tels que ceux proposés par Halpern et Pearl (2008) qui permettent de faire la distinction entre causalité et

corrélation. L'étude de la causalité pour séries temporelles se retrouve dans de nombreuses applications comme l'étude d'EEG, les analyses de marchés (Brodersen et al. (2015)) ou encore dans l'étude du climat (Runge et al. (2019)). Les méthodes proposées se basent généralement sur les capacités des algorithmes à faire des liens entre les données au cours du temps. Parmi ces méthodes, plusieurs en font la représentation par des graphes causaux pour séries temporelles multivariées. Chaque noeud de ces graphes correspond à une variable et la présence d'une arête entre deux noeuds indique une dépendance directe, alors que l'absence d'arête représente une indépendance ou une indépendance conditionnelle, selon Spirtes et al. (2000) ou Pearl (2009).

Une fois les relations établies entre les variables d'un système, raisonner sur la base de ces liens devient possible. Cependant, un grand nombre de ces modèles repose sur des hypothèses peu réalistes dans de nombreuses applications comme des hypothèses de linéarité ou une distribution connue *a priori* des données.

Pour résoudre ces problèmes, les travaux présentés dans ce papier visent à inférer des relations causales entre des séries temporelles multivariées avec peu d'hypothèses. Ainsi, l'objectif est de construire automatiquement un graphe causal prenant compte les décalages temporels, considérant des liens linéaires et non-linéaires et ne supposant pas de distribution sur les données observées. Cette méthode consiste à fusionner un algorithme de découverte causale, l'algorithme PC, de Spirtes et al. (2000), avec une mesure de causalité, la *Partial Mutual Information from Mixed Embedding* (PMIME) développée par Kugiumtzis (2013). Basée sur la théorie de l'information, la mesure PMIME permet de limiter les hypothèses sur les données et sur les liens entre ces dernières. Muni de cette mesure, l'algorithme PC explicite les relations causales entre des séries temporelles multivariées, en représentant la causalité à l'aide de graphes dirigés acycliques (*DAGs*).

Cet article est structuré comme suit : la partie II introduit les notions nécessaires pour traiter la causalité au sein des séries temporelles ainsi que les travaux connexes ; la partie III se concentre sur la méthode que nous proposons et nous montrons des résultats dans la partie IV. Finalement, nous concluons et discutons notre approche et ses résultats dans la partie V.

2 Causalité pour les séries temporelles

Dans ce papier, $X^0, ..., X^g$ représentent des variables aléatoires, X^i est la $i^{\text{ème}}$ série temporelle et X_t^i est la $i^{\text{ème}}$ au temps t. $X,\ Y,\ Z$ sont des variables aléatoires où X est généralement la cause potentielle étudiée, Y est son effet et Z, l'ensemble de conditionnement (potentiellement multivarié). Un processus multivarié est écrit $\mathbf{X} = (\mathbf{X}_t, \mathbf{X}_{t-1}, ...)$. Soit n le nombre d'observations d'un processus multivarié de taille g au temps t, noté $\mathbf{X}_t = \{x_t^0, x_t^1, ..., x_t^g\}$ et $\mathbf{X}_t^- = (\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, ...)$ un processus à tout temps avant t. Un décalage est noté τ et le décalage maximal considéré est τ_{max}. Enfin, si $\mathcal{G}$ est un graphe causal, $X - Y$ dénote un lien entre X et Y et X est un voisin de Y dans $\mathcal{G}$ ($X \in adj(Y, \mathcal{G})$). S'il y a une flèche de X vers Y ($X \to Y$), alors X est un parent de Y dans $\mathcal{G}$ ($X \in Par(Y, \mathcal{G})$).

La causalité entre séries temporelles peut être représentée par des réseaux causaux Bayésiens (CBN), une classe de modèles de graphe permettant une représentation probabiliste de va-

riables aléatoires. Le principe de priorité temporelle, qui stipule qu'une cause précède ses effets induit l'asymétrie de la causalité dans le temps. Elle permet ainsi d'orienter les relations causales dans un graphe lorsqu'une cause est déjà connue. Si une variable prise à un temps $t - \tau$ cause une autre variable à un temps t, la relation causale est définie comme une relation causale décalée. Si la cause apparaît en même temps que son effet, la relation est dite instantanée (ou *contemporaine*).

Il existe plusieurs types de graphes pour représenter la causalité au sein des séries temporelles. Le *Full-time causal graph* montre toutes les dépendances entre chaque variable du graphe pour tout temps t. Par définition, un tel graphe est difficile à représenter. Sous l'hypothèse de stationnarité causale (Runge (2018)), indiquant que toutes les relations causales

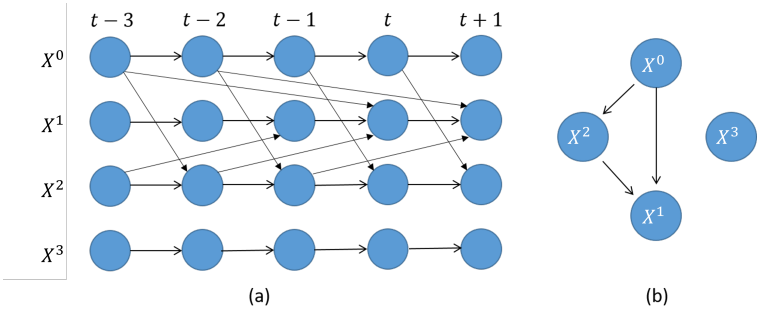

FIGURE 1 – Sur la gauche (a), un window causal graph et le graphe causal *summary* correspondant (b).

restent constantes en direction au cours du temps, ce type de graphe peut se réduire à un *window causal graph* (WCG). Les WCG montrent une section temporelle du graphe entier, après avoir sélectionné un décalage maximal en arrière τ_{max} et un temps maximal en avant. Le dernier type largement utilisé est le *summary causal graph* (SCG) dans lequel seulement les liens entre les variables sont représentés. Ainsi, le décalage entre une cause X et son effet Y n'est pas illustré. La Figure 1 présente un WCG et le graphe *summary* correspondant.

Sous les hypothèses décrites dans Spirtes et al. (2000), ces graphes encodent des indépendances conditionnelles, qui peuvent conduire à des DAGs. Dans un DAG, une flèche qui connecte deux noeuds traduit une dépendance directe. Une absence de flèche montre soit une indépendance soit une indépendance conditionnelle. Néanmoins, il peut arriver que deux DAGs différents encodent les mêmes dépendances. Dans ce cas, ces graphes appartiennent à la même classe équivalente de Markov (plus de détails dans Spirtes et al. (2000)).

Muni de ces graphes, la recherche de relations causales à partir de données peut se traduire par une estimation statistique de paramètres décrivant la structure causale. La découverte de causalité, *causal discovery*, a pour objectif d'utiliser des données observationnelles pour analyser et identifier les propriétés d'un système. Lorsque le système varie au cours du temps, la causalité est souvent associée à la causalité au sens de Granger (1969). On dit qu'une variable X Granger-cause une variable Y si la prédiction du futur de Y est améliorée lorsque la connaissance du passé de X devient disponible.

Une restriction majeure de ce modèle est qu'il se concentre sur les relations linéaires. De plus, la causalité de Granger, par définition, exclue les liens instantanés. Plusieurs méthodes et

extensions ont été développées pour exprimer cette notion sur la base d'observations de séries temporelles comme la *Pairwise Granger Causality* (PWGC) ou ses extensions multivariées (Barrett et al. (2010)) ou pour considérer les cas non-linéaires. D'autres méthodes ont été développées pour faire de la découverte de causalité, et sont regroupées dans trois familles : les *Functional causal models* (FCMs), les approches *score-based* et les approches *constraint-based*.

Les FCMs reposent sur les modèles d'équations structurelles (Pearl (2009)) et ont pour objectif de faire une correspondance entre un graphe $\mathcal{G}$ et un système d'équations. Un algorithme populaire dans ce cadre est l'algorithme VarLiNGAM développé par Hyvärinen et al. (2010). Les méthodes *score-based* essayent de chercher des modèles d'équations parcimonieux qui maximisent une mesure de qualité d'ajustement du graphe aux données. DYNOTEARS, est un algorithme *score-based* récemment développé par Pamfil et al. (2020) pour traiter des séries temporelles. Pour finir, les approches *constraint-based* recherchent des graphes appartenant à une classe d'équivalence de Markov, qui respectent le mieux l'ensemble des indépendances conditionnelles détectées dans les données. Les deux algorithmes principaux de cette famille de méthodes sont l'algorithme PC (Peter-Clark) et l'algorithme *Fast Causal Inference* (FCI), tous deux introduits par Spirtes et al. (2000). Le premier émet l'hypothèse de suffisance causale, indiquant que toutes les causes sont observées, contrairement au second. L'algorithme PC a pour objectif de retourner un DAG, illustrant les liens causaux, qui peut éventuellement être partiellement complet dans le cas où certains liens sont restés non-orientés. Malgré cela, il y a une garantie qu'un tel graphe appartienne à une classe d'équivalence de Markov.

Pour les séries temporelles, plusieurs méthodes basées sur l'algorithme PC ont été proposées, telles que PCMCI et ses extensions (Runge et al. (2019)). Ces méthodes utilisent notamment l'Information Mutuelle (Conditionnelle) ((C)MI) pour mesurer les dépendances entre les variables en limitant les hypothèses sur la distribution des données et sur la nature des relations entre ces dernières. Dans ce cadre, des mesures basées sur la théorie de l'information ont été développées spécifiquement pour traiter des variables temporelles telles que l'Entropie de Transfert (TE) de Schreiber (2000), sa généralisation en Entropie de transfert partielle (PTE) ou encore la mesure PMIME.

La mesure PMIME est asymétrique, non-paramétrique et conçue pour détecter des couplages directs au sein de séries temporelles. Elle est dérivée d'un processus *d'embedding* basé sur un critère de sélection, la CMI. Dans le cas multivarié, pour évaluer si une variable X a une influence sur une variable Y, conditionnellement à un ensemble de variables $\mathbf{Z} = (Z^0, ..., Z^{g-2})$, le processus construit itérativement un vecteur d'*embedding* $\mathbf{w}$ à partir des composantes décalées extraites de $(X, Y, \mathbf{Z})$ qui expliquent le mieux le futur de Y, noté $Y_t^T = (Y_{t+1}, ..., Y_{t+T})$. Chaque itération définit un cycle d'*embedding* et utilise un critère d'arrêt pour accepter ou refuser une composante. Une composante est acceptée si l'information qu'elle apporte améliore strictement l'information déjà contenue dans le vecteur d'*embedding*.

Ainsi, $\mathbf{w}$ est formé de g' variables décalées dans le temps sélectionnées par la CMI et peut se décomposer $\mathbf{w}_t = (w_t^x, w_t^y, w_t^{\mathbf{Z}})$, où w_t^x sont les composantes de X sélectionnées dans le procédé, w_t^y, celles de Y et les composantes restantes sont regroupées sous $w_t^{\mathbf{Z}}$. On quantifie

alors l'effet causal de X vers Y conditionnellement à $\mathbf{Z}$ par :

$$R_{X \to Y|\mathbf{Z}} = \frac{I(Y_t^T, \mathbf{w}_t^x | \mathbf{w}_t^y, \mathbf{w}_t^{\mathbf{Z}})}{I(Y_t^T; \mathbf{w}_t)}.$$

La description détaillée de la construction du vecteur d'*embedding* est donnée dans Kugiumtzis (2013). Si w_t^x est vide cela signifie que X n'a aucune influence sur Y, ce qui se transcrit sur la mesure R : si w_t^x est vide, alors R est nul. De plus, la mesure est bornée entre 0 et 1, 0 signifie indépendance et 1 signifie que le futur de Y est totalement déterminé par X.

3 Approche proposée : PC-PMIME

Dans ce travail, nous considérons un système modélisé par des séries temporelles dont les probabilités jointes sont générées selon un modèle linéaire ou non-linéaire. Aucune hypothèse sur la distribution *a priori* des données n'est émise. Enfin, nous supposons que ces dernières respectent la propriété de suffisance causale.

Dans l'étude récente de Assaad et al. (2022), comparant de nombreuses méthodes de l'état de l'art, les approches adressant au mieux ces hypothèses sont les algorithmes PCMCI et ses dérivés. Ces méthodes *constraint-based* requièrent une mesure d'inférence causale adaptée. Une limite est que PCMCI utilise des mesures linéaires telles que la *Partial Correlation* ou bien la CMI comme mesure non-linéaire, mais nécessitant des paramétrages. Un autre problème est que PCMCI utilise un WCG, qui peut s'avérer sensible au bruit et coûteux du fait du besoin d'identifier toutes les relations causales à tous les temps considérés dans la fenêtre.

Pour répondre à ces problèmes, nous proposons de combiner l'algorithme PC, une méthode *constraint-based* avec la mesure PMIME, nommé PC-PMIME, pour inférer la causalité au sein de séries temporelles. L'avantage de la mesure PMIME est qu'elle quantifie les liens temporels entre variables ayant des relations décalées sur une fenêtre de temps. Elle ne suppose pas de distribution de probabilité particulière, peut traiter des relations linéaires ou non-linéaires et a peu de paramètres à ajuster. De plus, Papana et al. (2013) ont montré que PMIME est plus efficace que la PTE dans le cadre de systèmes non-linéaires. Enfin, le fait que PMIME soit bornée simplifie l'interprétabilité du résultat et évite d'ajouter un test de significance statistique.

Cependant, du fait de l'utilisation de l'algorithme PC, la méthode proposée dans ce papier se contraint à l'hypothèse de suffisance causale. De plus, PMIME nécessite que les séries temporelles soient stationnaires. Enfin, nous utilisons une estimation par plus proches voisins de la CMI, qui requiert des séries suffisamment grandes.

Pour trouver les structures causales dans des séries temporelles, l'algorithme PC est fusionné avec la mesure non-paramétrique PMIME. Seule la première phase de l'algorithme PC est utilisée : elle commence avec un graphe complet et essaye de trouver le squelette du graphe en testant successivement chaque arête entre chaque noeud. Par exemple, une arête entre X et Y est enlevée si $R_{X \to Y} = 0$, où R est la mesure PMIME. Lorsque toutes les arêtes ont été testées et que certaines ont été retirées, les tests se poursuivent pour celles qui

Algorithme 1 : *PC-PMIME*

Entrées : n observations de $X^0, \ldots, X^g$, paramètres de PMIME
Sorties : $\mathcal{G}$, le graphe estimé

Créer un graphe complet $\mathcal{G}$ avec g noeuds
Initialiser rm_e la liste des arêtes retirées
pour *chaque permutation* (X^i, X^j) *de noeuds* $\mathcal{G}$ **faire**
    Calculer la mesure PMIME R avec X^j et X^i sans conditionnement
    Affecter R à l'arête $X^j \to X^i$
    si $R \approx 0$ **alors** ajouter l'arête entre X^i et X^j dans rm_e
Retirer toutes les arêtes de $\mathcal{G}$ contenues dans rm_e
Initialiser $l = 1$, *process* = Vrai
tant que *process est Vrai* **faire**
    Définir *process* sur Faux
    Réinitialiser $rm\_e = [\,]$
    pour *chaque permutation* (X^i, X^j) *de noeuds de* $\mathcal{G}$ **faire**
        $adj\_set = Par(X^i, \mathcal{G}) \backslash X^j$ la liste des prédécesseurs de X^i, sans X^j
        si $Card(adj\_set) \geq l$ **alors**
            pour *chaque combinaison* $\mathbf{Z}$ *de noeuds de adj_set de taille* l **faire**
                Calculer R, la mesure PMIME de $X^j \to X^i | \mathbf{Z}$
                Affecter R à l'arête entre X^j et X^i
                si R *est proche de 0* **alors**
                    Ajouter $X^j \to X^i$ dans rm_e
                    Sortir de la boucle
            Définir *process* comme Vrai
    Retirer toutes les arêtes de $\mathcal{G}$ contenues dans rm_e
Orienter $X^j \to X^i$, $\forall (X^i, X^j) \in \mathcal{G}$ si $R_{X^j \to X^i} > 0$

restent en regardant si deux noeuds sont conditionnellement indépendants. L'ensemble de conditionnement se compose dans un premier temps d'une seule variable, connectée à X ou Y, puis sa taille augmente incrémentalement jusqu'à ce qu'une indépendance conditionnelle soit détectée ou que toutes les arêtes liées à X et Y ont été testées. Comme la PMIME est asymétrique, l'algorithme teste les deux directions : de X vers Y et de Y vers X. Dans notre implémentation, une arête n'est pas retirée dès que $R = 0$, mais lorsque l'algorithme a testé toutes les arêtes pour une même taille de l'ensemble de conditionnement. Cette propriété définie comme PC stable d'après Colombo & Maathuis et al. (2014), permet d'éviter la dépendance de l'algorithme PC à l'ordre des variables testées.

L'algorithme 1, PC-PMIME, considère en entrées les données (g séries temporelles de taille n), un décalage maximal τ_{max}, le nombre de plus proches voisins k pour l'estimation de la CMI et A, le critère d'arrêt lors de la construction du vecteur d'embedding. L'algorithme retourne $\mathcal{G}$, le graphe causal orienté estimé. On observe qu'en général, R n'est pas égal à 0 en cas d'indépendance, mais est proche de 10^{-15}, à cause des erreurs d'estimation. Ainsi, nous considérons qu'il y a indépendance lorsque R est proche de 0 (on note $R \approx 0 \Leftrightarrow R < 10^{-10}$).

Bien que l'algorithme PC-PMIME n'est pas adapté pour rechercher des variables latentes, son implémentation est faite de telle sorte que si $R_{X^j \to X^i} > 0$ et $R_{X^i \to X^j} > 0$, alors cela mène à une arête orientée des deux côtés ($X^j \leftrightarrow X^i$). Cela apporte l'information que ces deux

variables sont mutuellement corrélées et qu'il y a potentiellement une cause commune cachée derrière ces deux variables.

4 Expérimentations

La méthode présentée dans la section précédente, PC-PMIME, est évaluée sur des données simulées issues de l'étude récente de Assaad et al (2022). Les auteurs ont simulé des structures causales basiques souvent rencontrées en recherche causale. Ces données contiennent un total de 5 structures différentes, toutes simulées 10 fois, sur 4000 observations. De ces 5 structures, nous en étudions 4, ne contenant pas de variable latente : la *fork* (une cause commune de deux effets), la v-structure, le *Mediator* (un *collider* dont l'un des parents cause l'autre) et le *Diamond* (une cause commune menant à une v-structure), représentées respectivement en figure 2. Ces données sont simulées avec des relations linéaires d'auto-corrélation et non-linéaires entre les variables.

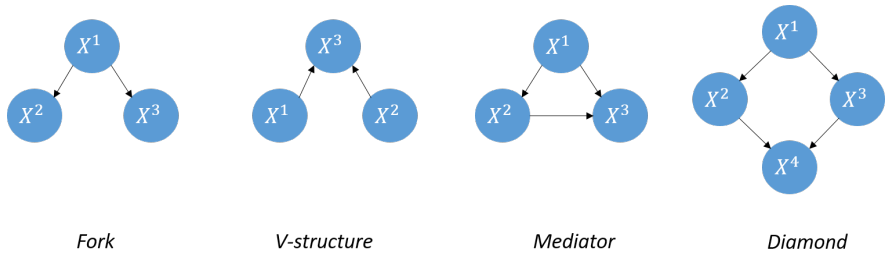

FIGURE 2 – Structures simulées de graphes causaux élémentaires

Sur chaque structure, PC-PMIME et quatre autres méthodes issues de l'état de l'art sont testées : PWGC, VarLiNGAM, DYNOTEARS et PCMCI. Pour chaque algorithme, le décalage maximal considéré est $\tau_{max} = 3$, selectionné empiriquement sur ces jeux de données. PWGC est utilisé avec un F-test et un seuil de signification $\alpha = 0.03$. VarLiNGAM utilise une pénalisation Lasso paramétrée par le Critère d'Information Bayesien (BIC). La paramétrisation de DYNOTEARS est celle recommandée par Palmfil et al. (2020), avec $\lambda_w = 0.05 = \lambda_a$ et le seuil $\tau_w = 0.01$. Pour finir, pour PCMCI, la mesure utilisée est la *Partial Correlation* avec un seuil de signification fixé à $\alpha = 0.03$. Le nombre de voisins utilisé pour l'estimation de la CMI dans PC-PMIME est $k = 0.01n$, selon Frenzel et Pompe (2007). Après plusieurs tests, il apparaît que si le critère d'arrêt A est proche de 0 ($A \leq 0.01$), il est trop permissif, à l'inverse, si $A \geq 0.05$, il est trop conservateur. Le critère d'arrêt est donc fixé à $A = 0.03$.

Pour comparer notre algorithme aux différentes méthodes, le F1-score est utilisé. Dans ce cadre, nous considérons qu'un vrai positif apparaît lorsqu'un lien dans le graphe estimé est aussi dans le graphe simulé. Aussi, l'auto-corrélation n'est pas mesurée car PC-PMIME ne l'intègre pas.

Pour chaque structure, nous faisons tourner les algorithmes sur dix jeux de données différents, avec des longueurs différentes de séries, $n = \{125, 250, 500, 1000, 2000, 4000\}$. Nous calculons ensuite la moyenne et la variance du F1-score pour chaque valeur de n. Les scores illustrés en figure 3 diffèrent de ceux obtenus par Assaad et al. (2022) car ces derniers prennent

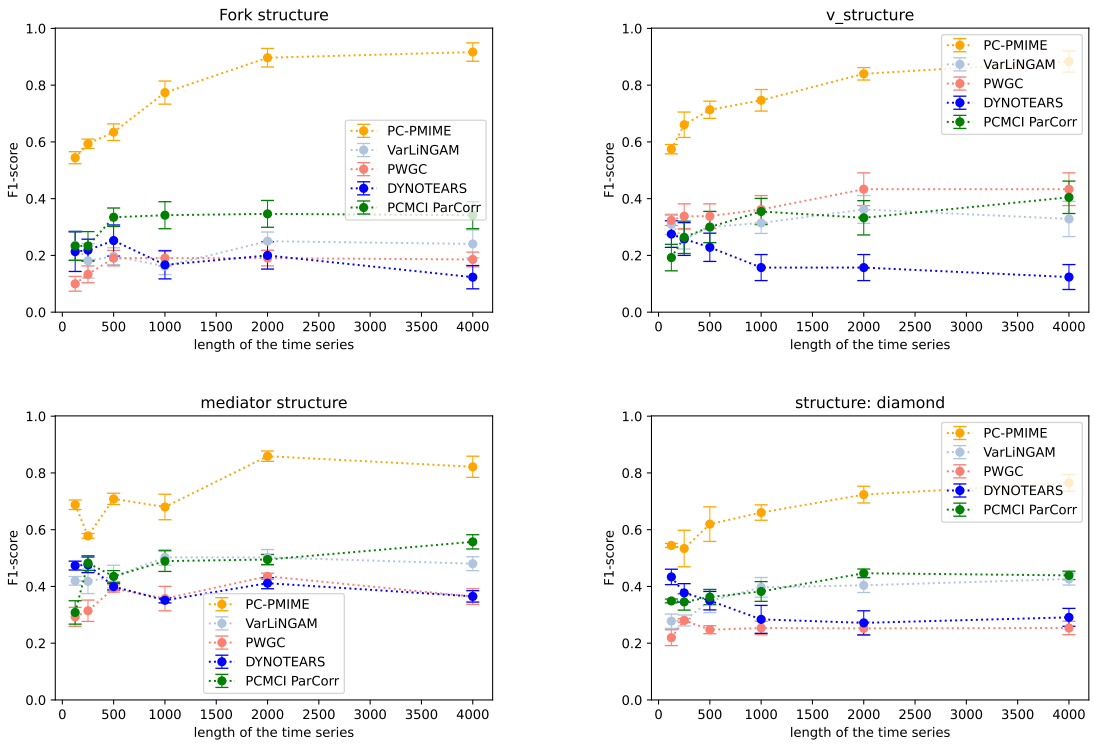

FIGURE 3 – F1-score des 5 méthodes en fonction de n. Chaque graphe correspond à une structure causale de base.

en compte l'auto-corrélation qui augmente la valeur du score. Par exemple, si l'on considère peu de noeuds et donc peu d'arêtes, détecter l'auto-corrélation compte comme un vrai positif et augmente significativement le score. L'auto-corrélation est calculée mais non prise en compte dans le calcul du F1-score. Ainsi, nous nous concentrons sur la découverte du graphe.

Parmi les scores obtenus par les différentes méthodes, montrés dans la figure 3, PC-PMIME obtient de performances comparables voire supérieures à celles des autres méthodes. En effet, le F1-score atteint une moyenne de 0.9 pour n assez grand, sur chaque structure exceptée le *Diamond*, alors que les scores des autres méthodes varient entre 0.1 et 0.5. Une faible variance des scores obtenus est également observée dans les expérimentations.Cependant, notre méthode n'est pas stable pour des petites tailles de séries temporelles (sous $n = 1000$). Cela est dû à l'estimation de la CMI dans PMIME. En fait, l'estimation par k-nn de l'entropie est robuste asymptotiquement. Lorsque la taille des séries temporelles est petite, détecter des indépendances (conditionnelles) est plus difficile et le nombre d'arêtes est alors généralement sur-estimé. La figure 4 illustre ce phénomène : dans le graphe donné par PC-PMIME pour $n = 125$, on constate que l'algorithme n'a pas su détecter l'indépendance entre X^2 et X^3 par exemple, ou encore le sens de l'influence entre X^1 et X^2. Cela rend alors une mauvaise interprétation causale : selon ce graphe, il y aurait une influence directe de X^3 vers X^2, ce qui est faux au regard du graphe simulé présenté figure 2. En revanche, le graphe estimé par PC-PMIME pour $n = 2000$ est exactement le même que le graphe simulé dans lequel X^1 est une cause commune de X^2 et X^3, qui causent à leur tour X^4.

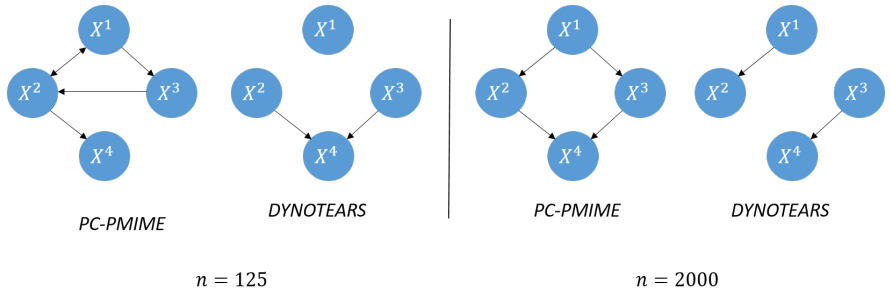

FIGURE 4 – Exemple de graphes causaux estimés sur le *Diamond* par la méthode PC-PMIME et DYNOTEARS pour deux tailles d'échantillons $n = 125$ à gauche et $n = 2000$ à droite.

5 Conclusion et futurs travaux

Dans ce papier, nous présentons une méthode pour inférer des relations causales entre séries temporelles multivariées, en faisant peu d'hypothèses sur les données. Ainsi, l'approche PC-PMIME permet de construire un graphe causal à partir de séries temporelles linéaires non-linéaires. La méthode produit des résultats intéressants sur des données simulées sans variable latente. Cependant, la phase d'orientation des arêtes, la mesure de l'auto-corrélation ou encore la prise en compte de causes cachées peuvent être améliorées. Enfin, un test sur des données réelles permettrait d'évaluer la méthode sur des données complexes. Par exemple, en manufacturing, l'objectif serait de déterminer les causes d'une défaillance au cours du temps dans un système monitoré complexe.

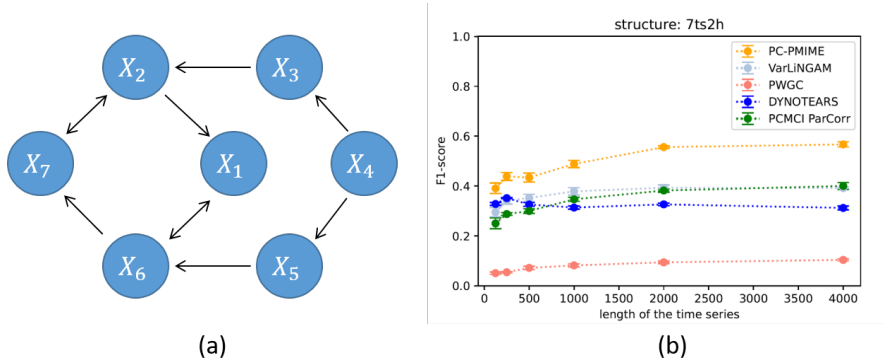

FIGURE 5 – Le graphe à gauche est le SCG de la structure simulée, incluant deux variables latentes marquées par des flèches à double-tête. Le second présente le F1-score des 5 méthodes en fonction de n.

Adapter les règles d'orientation de l'algorithme PC pour les inclure dans PC-PMIME améliorerait la méthode et éviterait certaines erreurs d'inférence causale. De plus, pour calculer l'auto-corrélation, on pourrait intégrer des *extended summary graphs* (Assaad et al. (2022)) et utiliser une mesure complémentaire à PMIME.

Enfin, la prise en compte des variables latentes pourrait s'avérer être un véritable enjeu pour améliorer notre approche. L'algorithme PC n'est pas défini dans ce cadre comme il nécessite l'hypothèse de suffisance causale. Pour confirmer cela, les différentes méthodes de

la section d'expérimentation sont testées sur le dernier jeu de données proposé dans Assaad et al. (2022), contenant deux causes latentes. La Figure 5 présente le graphe à inférer et les scores obtenus par chaque méthode. Comme attendu, les résultats se dégradent en présence de variables latentes.

Ainsi, la prise en compte de variable latentes nécessite d'utiliser un autre algorithme, tel que FCI par exemple, déjà adapté pour séries temporelles mais avec certaines limitations. Notre objectif serait alors de les réduire par une nouvelle approche.

Références

[1] Assaad, C. K., Devijver, E., and Gaussier, E. Discovery of extended summary graphs in time series. In *Uncertainty in Artificial Intelligence* (2022), PMLR, pp. 96–106.

[2] Assaad, C. K., Devijver, E., and Gaussier, E. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research 73* (2022), 767–819.

[3] Barrett, A. B., Barnett, L., and Seth, A. K. Multivariate granger causality and generalized variance. *Physical Review E 81*, 4 (2010), 041907.

[4] Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., and Scott, S. L. Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics 9*, 1 (2015), 247–274.

[5] Colombo, D., Maathuis, M. H., et al. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res. 15*, 1 (2014), 3741–3782.

[6] Frenzel, S., and Pompe, B. Partial mutual information for coupling analysis of multivariate time series. *Physical review letters 99*, 20 (2007), 204101.

[7] Granger, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica : journal of the Econometric Society* (1969), 424–438.

[8] Halpern, J. Y., and Pearl, J. Causes and explanations : A structural-model approach. part i : Causes. *The British journal for the philosophy of science* (2008).

[9] Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research 11*, 5 (2010).

[10] Kugiumtzis, D. Direct-coupling information measure from nonuniform embedding. *Physical Review E 87*, 6 (2013), 062918.

[11] Pamfil, R., Sriwattanaworachai, N., Desai, S., et al. Dynotears : Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics* (2020), PMLR, pp. 1595–1605.

[12] Papana, A., Kyrtsou, C., Kugiumtzis, D., and Diks, C. Simulation study of direct causality measures in multivariate time series. *Entropy 15*, 7 (2013), 2635–2661.

[13] Runge, J. Causal network reconstruction from time series : From theoretical assumptions to practical estimation. *Chaos : An Interdisciplinary Journal of Nonlinear Science 28*, 7 (2018), 075310.

[14] Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., et al. Inferring causation from time series in earth system sciences. *Nature communications 10*, 1 (2019), 1–13.

[15] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances 5*, 11 (2019), eaau4996.

[16] Schreiber, T. Measuring information transfer. *Physical review letters 85*, 2 (2000), 461.

[17] Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search*. MIT press, 2000.