

Information dirigée et graphes de causalité au sens de Granger

Pierre-Olivier AMBLARD^{1,2}, Olivier MICHEL¹

¹GIPSA-lab, Dépt. Images et Signal (UMR CNRS 5083)
ENSE3-BP 46 38402 Saint Martin d'Hères Cedex

²The University of Melbourne, Dept. of Math&Stat
Parkville, VIC 3010, Australie

bidou.amblard@gipsa-lab.inpg.fr, olivier.michel@gipsa-lab.inpg.fr

Résumé – L'objet du papier est d'établir les liens entre les graphes de causalité de Granger proposés par Dalhaus et Eichler, et la théorie de l'information dirigée initiée par Massey. Nous montrons comment l'information dirigée se décompose en une somme de mesures permettant d'inférer les graphes de causalité. Les mesures envisagées doivent être utilisées dans leurs versions conditionnellement causale. L'aspect pratique est évoqué à travers la présentation d'un estimateur reposant sur l'estimation de l'entropie utilisant les plus proches voisins. Deux exemples simulés illustrent à la fois le conditionnement causal et les problèmes liés à l'estimation des mesures d'information.

Abstract – This paper is devoted to studying the links between Granger causality graphs initially proposed by Dalhaus and Eichler, and directed information theory developed after Massey. We show how directed information decomposes into a sum of measures that allow to infer causality graphs. The measures considered must be used in their causal conditional form. Some practical considerations are examined. We detail an estimator based on nearest-neighbor entropy estimators. We illustrate causal conditioning and the difficulties of estimation on two examples.

1 Graphes de causalité

Les graphes de causalité de Granger ont été introduits par Eichler et Dalhaus pour compléter l'arsenal des modèles graphiques dans le cas de processus aléatoires multivariés [1]. Considérons deux processus x_t et y_t , et un ensemble de mesures complémentaires dénoté m_t ne contenant ni x ni y . x est une cause (au sens de Granger) de y relativement à m si la qualité de la prédiction de y_t à partir de son passé et de celui de m_t est améliorée pas la prise en compte du passé x_t . La prédiction peut être réalisée de différentes manières. Il peut s'agir par exemple de prédiction linéaire. Dans un cadre général, la prédiction est encodée par les mesures de probabilité conditionnelle. Considérons pour un processus (ou un ensemble) W la filtration \mathcal{W}_t résumant le passé jusqu'à la date t comprise. On a alors la définition suivante :

Définition 1 : x ne cause pas y relativement à $m \Leftrightarrow P(y_{t+1}|\mathcal{Y}_t, \mathcal{X}_t, \mathcal{M}_t) = P(y_{t+1}|\mathcal{Y}_t, \mathcal{M}_t)$

Autrement dit, conditionnellement à son passé et au passé des autres mesures m , le processus y est indépendant du passé du processus x . Cette définition exclut la possibilité (courante en pratique) d'un lien instantané entre deux signaux (dû par exemple à des bruits de dynamique corrélés). Une deuxième définition doit alors être considérée :

Définition 2 : x n'est pas une cause instantanée de y relativement à $m \Leftrightarrow P(y_{t+1}|\mathcal{Y}_t, \mathcal{X}_{t+1}, \mathcal{M}_{t+1}) = P(y_{t+1}|\mathcal{Y}_t, \mathcal{X}_t, \mathcal{M}_{t+1})$

Notons que cette deuxième définition est symétrique. Notons

également que le conditionnement par rapport aux processus supplémentaires se fait jusqu'à la date $t + 1$. Une version de cette définition dans laquelle le conditionnement s'arrête à la date t est possible, mais la notion d'indépendance (instantanée) n'est alors plus conditionnelle, mais inconditionnelle. Pratiquement, on identifiera dans la suite la filtration \mathcal{W}_t à l'ensemble des observations passées, noté $w_{1:t} = (w_1, \dots, w_t)$

Nous avons étudié dans [2] une version plus forte de ces définitions. Le conditionnement adopté dans la définition 1 est affaibli puisque l'on ne conditionne que sur \mathcal{M}_t et non pas \mathcal{M}_{t+1} comme dans [2]. Ceci rend l'étude des liens avec l'information dirigée plus difficile.

Eichler a proposé de synthétiser les dépendances sous forme d'un graphe mixte de causalité. Chaque nœud du graphe représente un processus ou une composante d'un processus multivarié. Un lien dirigé du nœud i vers le nœud j signifie que la composante x_i est une cause de la composante y_j relativement aux autres composantes. Un lien non dirigé entre les nœuds i et j signifie que les composantes associées sont des causes instantanées l'une de l'autre, relativement aux autres composantes.

A titre d'exemple, soient $\rho_{1,2,3} \in (-1, 1)$ et soit

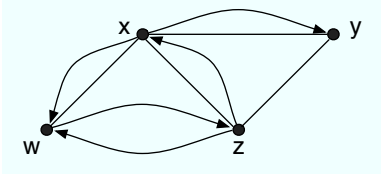
$$\Gamma_\varepsilon = \begin{pmatrix} 1 & \rho_1 & 0 & \rho_1\rho_2 \\ \rho_1 & 1 & 0 & \rho_2 \\ 0 & 0 & 1 & \rho_3 \\ \rho_1\rho_2 & \rho_2 & \rho_3 & 1 \end{pmatrix}$$

la matrice de covariance de la séquence multivariée i.i.d. gaus-

sienne $(\varepsilon_{w,t}, \varepsilon_{x,t}, \varepsilon_{y,t}, \varepsilon_{z,t})^t$. Le système suivant

$$\begin{cases} w_t = f_w(w_{t-1}, x_{t-1}, z_{t-1}) + \varepsilon_{w,t} \\ x_t = f_x(x_{t-1}, z_{t-1}) + \varepsilon_{x,t} \\ y_t = f_y(x_{t-1}, y_{t-1}) + \varepsilon_{y,t} \\ z_t = f_z(w_{t-1}, z_{t-1}) + \varepsilon_{z,t} \end{cases} \quad (1)$$

a pour graphe de causalité le graphe



Les graphes de causalité permettent d'une manière synthétique de représenter ce que l'intuition dicte, en éliminant de la représentation tout l'aspect dynamique, et en ne conservant que des liens de dépendances. Insistons sur le fait que la présence de liens repose sur des propriétés conditionnelles, que ce soit dans le cas dirigé comme dans le cas instantané. Ainsi dans l'exemple, bien que w et z soient corrélés instantanément (terme $\rho_1\rho_2$ dans Γ_ε), ils ne sont pas instantanément cause l'un de l'autre au sens de la définition 2 : il n'y a donc pas de lien non dirigé entre les deux. Ceci traduit le fait que la corrélation instantanée entre les deux est due à une autre cause commune. A l'opposé, x et y sont décorrélés, mais un lien non dirigé existe entre les deux. Ce paradoxe provient encore une fois de la définition adoptée, qui repose sur l'indépendance conditionnelle. Dans l'exemple, x et y sont corrélés conditionnellement à z , mais ne le sont pas inconditionnellement (cela peut être vérifié en examinant la matrice de précision Γ_ε^{-1} , dont les termes extra-diagonaux sont proportionnels aux coefficients de corrélation partielle). Cette situation apparaît car z est fonction de x et y : le lien fonctionnel relie les deux variables de manière conditionnelle. Ce phénomène est bien connu en théorie des graphes acycliques dirigés et est révélé par la notion de graphe moral : deux parents d'un même nœud doivent être connectés !

2 Information dirigée

Dans [3], Massey propose qu'en présence de feedback, la mesure correcte d'information entre deux signaux n'est plus l'information mutuelle mais l'information dirigée, définie par

$$I(x_{1:t} \rightarrow y_{1:t}) = \sum_{k=1}^t I(x_{1:k}; y_k | y_{1:k-1})$$

où $I(x_{1:k}; y_k | y_{1:k-1})$ est l'information mutuelle conditionnelle usuelle [4]. Une définition alternative repose sur la notion de conditionnement causal introduit par Kramers [5]. L'entropie causale conditionnelle est définie par

$$H(y_{1:n} | x_{1:n}) = \sum_{i=1}^n H(y_i | x_{1:i}, y_{1:i-1}).$$

La différence avec le conditionnement usuel apparaît quand on utilise les règles de chaînage pour écrire

$H(y_{1:n} | x_{1:n}) = \sum_{i=1}^n H(y_i | x_{1:n}, y_{1:i-1})$. Le terme dans la somme considère $x_{1:i}$ jusqu'à l'indice courant dans le conditionnement causal, alors qu'il considère tout le vecteur $x_{1:n}$ dans le conditionnement usuel. Avec cette définition, il est aisé de montrer que l'information dirigée s'écrit $I(x_{1:t} \rightarrow y_{1:t}) = H(y_{1:n}) - H(y_{1:n} | x_{1:n})$. Le conditionnement causal peut s'appliquer à l'information dirigée elle-même lorsque l'on a affaire à des données supplémentaires m_t . L'information dirigée conditionnellement causale est définie par

$$\begin{aligned} I(x_{1:t} \rightarrow y_{1:t} | m_{1:t}) &= H(y_{1:t} | m_{1:t}) - H(y_{1:t} | x_{1:t}, m_{1:t}) \\ &= \sum_{i=1}^t I(x_{1:i}; y_i | y_{1:i-1}, m_{1:i}). \end{aligned} \quad (2)$$

On peut alors montrer deux résultats [3]. Le premier

$$\begin{aligned} I(x_{1:t} \rightarrow y_{1:t}) + I(y_{1:t} \rightarrow x_{1:t}) &= I(x_{1:t}; y_{1:t}) \\ &+ I(x_{1:t} \rightarrow y_{1:t} | Dx_{1:t}), \end{aligned} \quad (3)$$

où $I(x_{1:t} \rightarrow y_{1:t} | Dx_{1:t})$ est appelée *échange d'information instantané*, montre comment l'information mutuelle se découpe en deux informations dirigées ($Dx_{1:t} = (0, x_{1:t-1})$ est l'opérateur de retard). Notons que l'échange d'information instantané est symétrique. En effet, en utilisant l'équation (2) on obtient

$$I(x_{1:t} \rightarrow y_{1:t} | Dx_{1:t}) = \sum_{i=1}^t I(x_i; y_i | y_{1:i-1}, x_{1:i-1}) \quad (4)$$

qui démontre la symétrie. Ce résultat prouve qu'en général, la somme des informations dirigées de x vers y et de y vers x n'est pas égale à l'information mutuelle, puisqu'il faut lui ajouter un échange d'information instantané. Il y a toutefois égalité si et seulement si $I(x_i; y_i | y_{1:i-1}, x_{1:i-1}) = 0, \forall i$, c'est-à-dire lorsque x et y sont indépendants conditionnellement à leur passé. Cette situation se réalise dans des modèles markoviens multivariés du type $X_t = f(X_{t-1}) + \epsilon_t$ où la suite des ϵ_t est i.i.d. et à composantes indépendantes. Dans l'exemple (1) considéré, la matrice Γ_ε étant non diagonale, il y a un échange instantané d'information entre certaines composantes.

Le second résultat fait le lien avec les graphes de causalité. Face à de multiples mesures, deux approches sont possibles, une analyse bivariée dans laquelle on s'intéresse à l'information dirigée entre paire de signaux sans se soucier des mesures supplémentaires, ou l'analyse multivariée dans laquelle les autres composantes sont prises en compte sous forme de conditionnement.

Analyse bivariée. L'information dirigée peut s'écrire

$$I(x_{1:t} \rightarrow y_{1:t}) = I(x_{1:t} \rightarrow y_{1:t} | Dx_{1:t}) + I(Dx_{1:t} \rightarrow y_{1:t}),$$

somme de l'échange d'information instantané défini en (4) et le terme $I(Dx_{1:t} \rightarrow y_{1:t})$ que nous appellerons *transfert d'entropie* en référence à une définition proposée initialement par T. Schreiber [6]. En l'absence de données supplémentaires, ces deux termes caractérisent les définitions 1 et 2 de la causalité de Granger. En effet, comme développé plus haut, l'échange instantané d'information est nul si et seulement si

x ne cause pas instantanément y . De plus, le transfert d'entropie est nul si et seulement si $I(x_{1:i-1}; y_i | y_{1:i-1}) = 0$, $\forall i$, et donc si et seulement si x ne cause pas y . Ce résultat généralise la décomposition obtenue dans le cas Gaussien dans une précédente contribution [7]. Insistons encore sur le fait que ce résultat n'est valide qu'en l'absence de la prise en compte de données supplémentaires.

Analyse multivariée. La connaissance de données supplémentaires doit être prise en compte dans l'analyse sous forme de conditionnement. Les définitions 1 et 2 montrent que le conditionnement doit se faire de deux manières, suivant que l'on intègre le présent et/ou le passé de m . Nous pouvons donc évaluer deux informations dirigées causalement conditionnelles, $I(x_{1:t} \rightarrow y_{1:t} || m_{1:t})$ et $I(x_{1:t} \rightarrow y_{1:t} || Dm_{1:t})$. Ces deux informations se décomposent de manière différente.

Si l'on considère le passé de m on obtient la décomposition

$$I(x_{1:t} \rightarrow y_{1:t} || Dm_{1:t}) = I(Dx_{1:t} \rightarrow y_{1:t} || Dm_{1:t}) + I(x_{1:t} \rightarrow y_{1:t} || Dx_{1:t}, Dm_{1:t}).$$

Le premier terme, appelé *transfert d'entropie conditionnel*, est nul si et seulement si $I(x_{1:i-1}; y_i | y_{1:i-1}, m_{1:i-1}) = 0, \forall i > 1$, où de manière équivalente (définition 1) si x ne cause pas y relativement à m . Le deuxième terme $I(x_{1:t} \rightarrow y_{1:t} || Dx_{1:t}, Dm_{1:t})$ est symétrique en x et y et s'annule si et seulement si $I(x_i; y_i | y_{1:i-1}, x_{1:i-1}, m_{1:t-1}) = 0$ pour tout i . Il n'y a donc pas d'équivalence avec les assertions de la définition 2. La grandeur $I(x_i; y_i | y_{1:i-1}, x_{1:i-1}, m_{1:t-1})$ est appelée *échange d'information instantané INconditionnel*. Pour justifier le terme inconditionnel examinons l'exemple (1). Savoir si le terme est nul est équivalent à savoir si $P(y_i | x_{1:i}, y_{1:i-1}, (wz)_{1:i-1}) \stackrel{?}{=} P(y_i | x_{1:i-1}, y_{1:i-1}, (wz)_{1:i-1})$. Dans le cas de l'exemple, cela revient à savoir si les bruits de dynamiques $\varepsilon_{x,i}$ et $\varepsilon_{y,i}$ sont indépendants, et ce sans être conditionnés par les deux autres. Pour obtenir un conditionnement par rapport à $\varepsilon_{w,i}$ et $\varepsilon_{z,i}$, il faut considérer m_t et non Dm_t pour le conditionnement causal.

Si l'on considère le passé de m et son présent on obtient la décomposition

$$I(x_{1:t} \rightarrow y_{1:t} || m_{1:t}) = I(Dx_{1:t} \rightarrow y_{1:t} || Dm_{1:t}) + I(x_{1:t} \rightarrow y_{1:t} || Dx_{1:t}, m_{1:t}) \quad (5) + \Delta I(m \rightarrow y || Dx, Dy, Dz)$$

Dans ce cas, l'information dirigée causalement conditionnelle se découpe en la somme du transfert d'entropie conditionnelle, de l'échange instantané d'information conditionnelle, et d'un troisième terme qui s'écrit

$$\Delta I(m \rightarrow y || Dx, Dy, Dz) = I(m_t \rightarrow y_t || Dx_t, Dm_t) - I(m_t \rightarrow y_t || Dm_t)$$

qui est également une grandeur instantanée.

Nous n'avons donc pas dans le cas des réseaux le même résultat que dans le cas bivarié. Toutefois, l'analyse précédente met en lumière les deux grandeurs fondamentales pour inférer les graphes de causalité et leur lien avec l'information dirigée.

3 Exemples et considérations pratiques

Pour estimer l'information dirigée, l'hypothèse de stationnarité est nécessaire. Dans ce cas et dans la limite des temps infinis, on montre que l'information dirigée est égale à la limite d'un seul terme d'information conditionnelle de la somme (2). Ensuite, ne pouvant réaliser pratiquement cette limite, l'estimation des mesures d'information dirigée se fait sur une fenêtre de taille finie. Si l'on se restreint au transfert d'entropie conditionnelle, on estime alors $I(x_{t-p:t-1}; y_t | y_{t-p:t-1}, m_{t-p:t-1})$. L'information conditionnelle est estimée à l'aide de l'estimation de l'entropie H . Différents estimateurs d'entropie peuvent être utilisés. Nous utilisons des estimateurs fondés sur les distances au plus proches voisins développés par Leonenko et ses collaborateurs [8, 9]. L'estimateur de l'entropie de Shannon d'un vecteur aléatoire x s'écrit

$$\hat{H}(x) = \frac{1}{N} \sum_{i=1}^N \log \left((N-1) C_k V_n d_{i,(k)}^m \right)$$

où $d_{i,(k)}$ est la distance entre l'observation i et son k ème plus proche voisin, N est le nombre d'observation, n la dimension de l'espace, $\log C_k = \psi(k)$ est la fonction digamma, V_n le volume de la boule unité. Cet estimateur converge en moyenne quadratique pour tout $k \in \{1, \dots, N-1\}$. Toutefois, sa vitesse de convergence est inconnue, et n'est approchée que par des simulations. Ces simulations montrent que l'estimateur est supérieur à des estimateurs de type histogramme ou noyau, surtout lorsque la dimension augmente. Les simulations montrent que l'estimateur de Leonenko possède un compromis biais-variance dépendant de k , et que le choix optimal de k en ce sens dépend de la dimension. Pour des dimensions élevées, k doit être choisi petit, typiquement de l'ordre de l'unité. D'autre part, l'utilisation de cet estimateur dans $I(a, b|c) = H(b, c) + H(a, c) - H(a, b, c) - H(c)$ conduit à un biais assez fort pour l'information mutuelle conditionnelle. On utilise alors une astuce proposée par Kraskov et étendue par Frenzel&Pompe [10, 11] pour limiter ces erreurs.

Dans l'estimation de l'information dirigée, deux problèmes, indépendants du type d'estimateur, surgissent. Le premier est celui de la dimension. Pour une fenêtre d'horizon de taille p et une dimension n du signal multivarié, le transfert d'entropie conditionnel requiert l'estimation d'entropie de variables de dimension $O(n \times p)$. Quel que soit l'estimateur d'entropie, la taille des données requises pour une qualité raisonnable croît vite avec la dimension. Le deuxième problème concerne le comportement de l'estimateur qui est inconnu en général dans les cas non gaussiens. Pour utiliser pratiquement les estimations, il faut avoir recours à des techniques statistiques (type bootstrap, permutations) pour évaluer biais et variance, et pour créer l'hypothèse nulle.

Un premier exemple montrant l'importance du conditionnement est celui de la chaîne $x_t = bx_{t-1} + \varepsilon_{x,t}; y_t = cy_{t-1} + d_{xy}x_{t-1}^2 + \varepsilon_{y,t}; z_t = dz_{t-1} + c_{yz}y_{t-1} + \varepsilon_{z,t}$ où pour simplifier on considère des bruits de dynamiques indépendants et normalisés, et où les coefficients en jeu ont été choisis de sorte à assurer la stabilité du système.

Les tests de Geweke développés dans le cas linéaire-gaussien [12, 7] sont non satisfaisants dans cet exemple à cause de la non linéarité introduite. Ils ne prévoient pas le lien de x vers y par exemple. On évalue alors les transferts d'entropie conditionnels $I(x_{t-2:t-1}; z_t | z_{t-1}, y_{t-1})$ ou non $I(x_{t-2:t-1}; z_t | z_{t-1})$ (uniquement dans un sens ici à titre d'illustration). On représente les deux valeurs par les histogrammes de la figure (1). Les histogrammes blancs représentent les grandeurs conditionnelles, et les noirs les grandeurs inconditionnelles. Les histogrammes sont calculés sur 100 réalisations de 2000 échantillons de la chaîne. Nous utilisons l'estimateur présenté précédemment avec $k = 5$ (une implantation naïve matlab est disponible sur demande aux auteurs). On utilise 100 permutations pour simuler l'hypothèse nulle et calculer un seuil correspondant à un test de niveau 0.05 (probabilité de fausse alarme). Ce seuil est représenté par le trait vertical plein. L'histogramme des transferts inconditionnel montre clairement un lien de x vers z . Toutefois, la prise en compte du conditionnement par rapport à y démontre que ce lien passe par y , puisque le transfert d'entropie conditionnel est jugé nul avec confiance. Le transfert d'entropie conditionnel, à l'opposé du transfert inconditionnel, permet de conclure que x ne cause pas y (relativement à y).

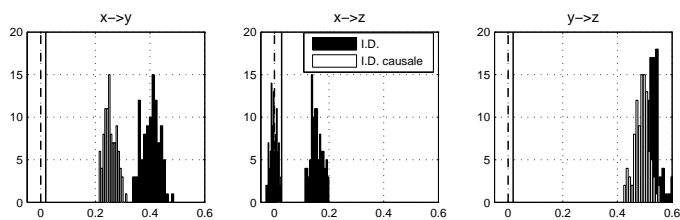


FIG. 1 – Histogramme de transferts d'entropie entre paires de signaux dans la chaîne considérées. Les barres blanches concernent le transfert entre deux signaux conditionnellement au troisième, alors que les barres noires concernent le transfert inconditionnel.

Le deuxième exemple reprend (1) sous la forme

$$\begin{cases} w_t = aw_{t-1} + d_{xw}x_{t-1}^2 + c_{zw}z_{t-1} + \varepsilon_{w,t} \\ x_t = bx_{t-1} + c_{zx}z_{t-1} + \varepsilon_{x,t} \\ y_t = cy_{t-1} + c_{xy}x_{t-1} + d_{xy}x_{t-1}^2 + \varepsilon_{y,t} \\ z_t = dz_{t-1} + c_{wz}w_{t-1} + \varepsilon_{z,t} \end{cases} \quad (6)$$

où pour simplifier on considère des bruits de dynamiques indépendants et normalisés, et où les coefficients en jeu ont été choisis de sorte à assurer la stabilité du système. On représente figure (2) les histogrammes des transferts d'entropie conditionnels, calculés sur 100 réalisations du système, l'estimation des entropie étant réalisés sur 2000 échantillons. Le seuil correspondant à un test de niveau 0.05 est représenté sur chaque graphe. La structure du graphe est correctement inférée lorsque l'on conditionne.

Remerciements. P.-O. Amblard est supporté par une bourse ExploraPro de la Région Rhône-Alpes (France), et par une bourse Marie Curie (International Outgoing Fellowship) de la Communauté Européenne.

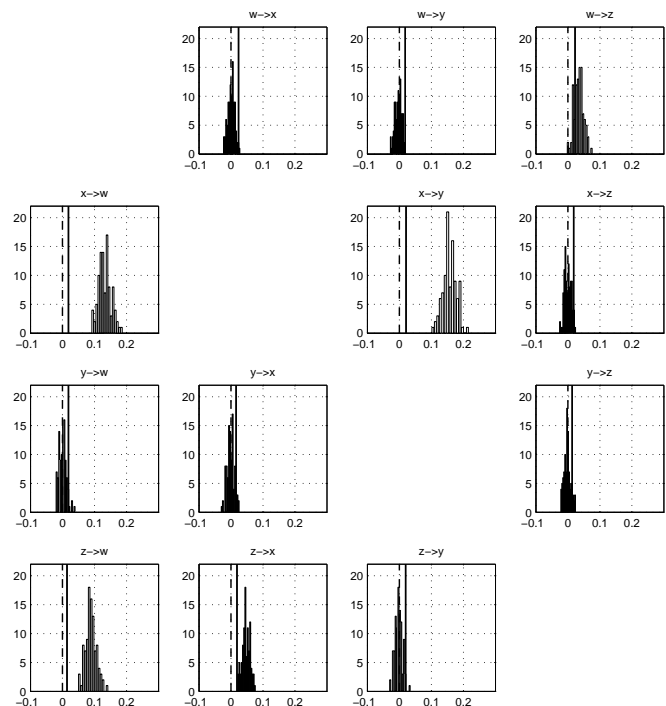


FIG. 2 – Histogramme de transferts d'entropie conditionnels entre paires de signaux dans l'exemple (6). Le trait pointillé représente 0 et le trait plein le seuil correspondant à un test de niveau 0.05.

Références

- [1] R. Dahlaus and M. Eichler. *Highly structured stochastic systems*, chapter Causality and graphical models in time series analysis. P. Green, N. Hjort, and S. Richardson (eds.), University Press, Oxford., 2003.
- [2] P. O. Amblard and O. J. J. Michel. On directed information theory and granger causality graphs. *Journal of Computational Neuroscience*, 30(1) :7–16, january 2011.
- [3] J.L. Massey. Causality, feedback and directed information. In *Proc. Intl. Symp. on Info. th. and its Applications, Waikiki, Hawaii, USA*, nov 1990.
- [4] T. Cover and J. Thomas. *Elements of information theory*. Wiley, 1993.
- [5] G. Kramers. *Directed information for channels with feedback*. PhD thesis, Swiss Federal Institute of Technology Zurich, 1998.
- [6] T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85(2) :461–465, 2000.
- [7] P. O. Amblard and O. J. J. Michel. Sur différentes mesures de dépendances causales entre signaux aléatoires (On different measures of causal dependencies between random signals). In *proc. Grets, Dijon, France, Sept.*, 2009.
- [8] L. F. Kozachenko and N. N. Leonenko. Sample estimate of the entropy of a random vector. *Problems Infor. Transmiss.*, 23 :95–101, 1987.
- [9] M. N. Goria, N. N. Leonenko, V. V. Mergel, and P. L. Novi Invardi. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *Jour. Nonparam. Stat.*, 17(3) :277–297, 2005.
- [10] A. Kraskov, H. Stogbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69 :066138, 2004.
- [11] S. Frenzel and B. Pompe. Partial mutual information for coupling analysis of multivariate time series. *Phys. Rev. Lett.*, 99 :204101, 2007.
- [12] J. Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77 :304–313, 1982.