Etude de l'hypothèse de non cyclicité dans les modèles graphiques orientés causaux

Gilles Monneret[1,2],Florence Jaffrézic[1],Andréa Rau[1],Grégory Nuel[2]

[1]UMR GABI, AgroParisTech, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[2]LPMA, UMR CNRS 7599, UPMC, Sorbonne Universités, 4 place Jussieu, 75005, Paris.
gilles.monneret@upmc.fr

Résumé.
L'étude des relations entre différentes variables est un enjeu majeur dans de nombreuses disciplines, comme l'économie ou la biologie. Les relations de causalité sont particulièrement recherchées, puisqu'elles éclairent le processus sous-jacent avec une particulière acuité. L'immense majorité des techniques causales utilise une hypothèse d'acyclicité : celle-ci peut-être perçue comme extrêmement limitante, en particulier dans les systèmes où les boucles de rétro-actions sont courantes. Nous allons étudier ici les implications d'un relâchement de l'hypothèse d'acyclicité, en terme de loi jointe et d'identifiabilité. Nous éclairerons ces résultats avec quelques exemples.
Mots clefs Causalité, rétroaction, réseaux bayésiens, modèle graphique, inférence de réseaux
Abstract.
The study of relationships between various variables is a major goal in numerous domain, like economy or biology. Causality relationships are particularly searched, because they describe the underlying mechanism with a particular insight. The vast majority of causation techniques use a hypothesis of acyclicity. We can find this one restrictive, in particular in systems where feedbacks loops are commons. We will study here the implications of a release of this hypothesis of acyclicity, both in term of joint law and identifiability. We will discuss of these results with some examples.
Keywords. Causality, feedback loops, bayesian network, graphical model, network inference

1 Introduction

Lorsque nous essayons de modéliser statistiquement des modèles causaux, l'une des approches les plus parlantes est la modélisation graphique. Celle-ci consiste à représenter par un graphe $G = (E, V)$ un ensemble de nœuds V et de liens orientés E entre eux. Nous pouvons associer une loi de probabilité avec ce graphe, nous donnant alors un modèle graphique probabiliste. L'une des propriétés fondamentale de telles lois est qu'elles sont factorisables de la façon suivante.

Définition 1 *Soit G un graphe orienté. Une loi de probabilité $\mathbb{P}$ se factorise sur G si et seulement si elle s'écrit :*

$$\mathbb{P}(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{i=1,n} \mathbb{P}(X_i | \, \mathrm{pa}(X_i))$$

où $\mathrm{pa}(X_i)$ *désigne les parents du nœud X_i.*

En terme graphique, une arête ira du nœud X_i au nœud X_j si et seulement si X_i est un parent de X_j. Une hypothèse couramment utilisée est l'hypothèse d'acyclicité. Dans celle-ci, la constante de normalisation Z est égale à 1. Dans le cas cyclique, cette constante est à déterminer, ce qui peut potentiellement donner lieu à une totale indétermination ($Z = \infty$). Ce dernier cas peut être modélisé à l'aide de données temporelles ou simplement vu comme un point stationnaire du système considéré.

2 Factorisation d'un réseau bayésien gaussien généralisé

2.1 Moralisation

L'objectif de cette section est de déterminer la loi d'un graphe orienté cyclique gaussien. On cherche à caractériser la loi issue de la relation récursive suivante :

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{X}\boldsymbol{W} + \boldsymbol{\varepsilon} \qquad (1)$$

où $\boldsymbol{X}$ est un vecteur ligne décrivant notre vecteur aléatoire, $\boldsymbol{\mu}$ un vecteur moyenne, $\boldsymbol{W}$ la matrice d'adjacence et $\boldsymbol{\varepsilon}$ un vecteur gaussien de variables décorrélées, de variance $\boldsymbol{\sigma^2}$ et de moyenne nulle. On peut de plus écrire la densité de la loi $\mathbb{P}(X_i|\mathrm{pa}(X_i))$:

$$f(X_i = x| \, \mathrm{pa}(X_i)) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x - \mu_i - \boldsymbol{X}\boldsymbol{W}\boldsymbol{e}_i)^2}{2\sigma_i^2}\right),$$

et la loi jointe s'écrit :

$$f(X_1, \ldots, X_n) = \frac{1}{e^{-Z}} \prod_{i=1}^{n} f(X_i = x| \, \mathrm{pa}(X_i)).$$

En notant $\boldsymbol{P}$ la matrice de précision associée à notre loi jointe, la constante de normalisation devient :

$$Z = \frac{1}{2}\log(|\boldsymbol{P}|) + \frac{1}{2}\sum_{i=1}^{n} \log(\sigma_i^2).$$

Proposition 1 *Soit X un vecteur aléatoire suivant* (1), *et si $(\boldsymbol{I} - \boldsymbol{W})$ est inversible, alors :*

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{I} - \boldsymbol{W})^{-1}, ((\boldsymbol{I} - \boldsymbol{W})^{-1})^T \mathrm{diag}(\boldsymbol{\sigma^2})(\boldsymbol{I} - \boldsymbol{W})^{-1}).$$

Cette proposition est très utile puisqu'elle nous permet en pratique de simuler un réseau gaussien cyclique.

2.2 Effets causaux

On introduit l'opérateur do : la variable aléatoire $\boldsymbol{X}|\,\mathrm{do}\,(X_i = x_i)$ est définie par le jeu d'équation (1) modifié de telle sorte que la i-eme composante est supprimée, et où X_i est remplacé dans toutes les autres lignes par x_i. Ceci correspond a fixer la valeur de X_i et de supprimer tout lien avec ses ascendants. Nous proposons pour formaliser la notion de causalité d'utiliser les définitions des effets causaux selon Pearl [2000] :

Définition 2 *L'effet causal direct de X_i sur X_j est défini par :*

$$\frac{\partial}{\partial x}\mathbb{E}(X_j|\,\mathrm{do}(X_i = x_i), X_{-j}).$$

L'effet causal total de X_i sur X_j est défini par :

$$\frac{\partial}{\partial x}\mathbb{E}(X_j|\,\mathrm{do}(X_i = x_i)).$$

Nous pouvons alors entre autre retrouver la loi de $X_j|\boldsymbol{X}_{-j} = \boldsymbol{y}$:

$$\begin{array}{rcl}\mathbb{E}(X_j|X_{-j}) & = & \mu_j + \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{-j}) \\ \mathrm{Var}(X_j|X_{-j}) & = & \Sigma_{j,j} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j},\end{array}$$

où $\Sigma_{j,-j}$ correspond à la j-ieme ligne de la matrice de covariance, sans la j-ieme colonne. Vouloir calculer la loi de $(X_j|\,\mathrm{do}(X_i = x), \boldsymbol{X}_{-j} = \boldsymbol{y})$ revient à calculer $\left(\tilde{X}_j|\tilde{\boldsymbol{X}}_{-j} = \boldsymbol{y}\right)$, respectant

$$\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{X}}\tilde{\boldsymbol{W}} + \tilde{\varepsilon},$$

et où les différentes quantités respectent ces caractéristiques : $\tilde{\boldsymbol{X}}$ est de dimension $N-1$, $\tilde{\boldsymbol{W}}$ correspond à la matrice $\boldsymbol{W}$ mais avec la ligne et la colonne i supprimée, $\tilde{\boldsymbol{\mu}}$ correspond à $\boldsymbol{\mu} + x\boldsymbol{e}_i^T\boldsymbol{W}$, et $\tilde{\boldsymbol{\varepsilon}}$ est un vecteur gaussien centré de dimension j.

Avec cette formulation, on obtient :

$$\begin{array}{rcl}\mathbb{E}(X_j|\,\mathrm{do}(X_i = x), X_{-j}) & = & \mu_j + x\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{e}_j + \tilde{\Sigma}_{j,-j}\tilde{\Sigma}_{-j,-j}^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}_{-j}) \\ \mathrm{Var}(X_j|\,\mathrm{do}(X_i = x), X_{-j}) & = & \tilde{\Sigma}_{j,j} - \tilde{\Sigma}_{j,-j}\tilde{\Sigma}_{-j,-j}^{-1}\tilde{\Sigma}_{-j,j}\end{array}$$

On voit que l'espérance est linéaire en x, l'effet causal direct vaut donc :

$$\frac{\partial}{\partial x}\mathbb{E}(X_j|\,\mathrm{do}(X_i = x), \boldsymbol{X}_{-j}) = W_{i,j}$$

De la même façon, en partant sur l'expression du vecteur gaussien $\tilde{\boldsymbol{X}}$:

$$\tilde{\boldsymbol{X}} \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}(\boldsymbol{I} - \tilde{\boldsymbol{W}})^{-1}, ((\boldsymbol{I} - \tilde{\boldsymbol{W}})^{-1})^T\mathrm{diag}(\tilde{\boldsymbol{\sigma}}^{\boldsymbol{2}})(\boldsymbol{I} - \tilde{\boldsymbol{W}})^{-1}).$$

On obtient en particulier la moyenne de la variable X_j, valant $\left[\mu_j + x\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{e}_j\right](\boldsymbol{I} - \tilde{\boldsymbol{W}})^{-1}$. Encore une fois, elle est linéaire en x. L'effet causal total est donc :

$$\frac{\partial}{\partial x}\mathbb{E}(X_j|\,\mathrm{do}(X_i = x)) = W_{i,j}(\boldsymbol{I} - \tilde{\boldsymbol{W}})^{-1}$$

3 De la corrélation à la causalité

Identifibialité cas observationnel

Nous proposons ici une simulation numérique décrivant le problème de l'interprétation d'un graphe cyclique issue de simples observations. Commençons par définir une immoralité :

Proposition 2 *Soit (A, B, C) trois nœuds d'un graphe orienté acyclique G. Une immoralité est une structure du type $A \to B \leftarrow C$, et où il n'existe pas de flèche entre A et C.*

Dans le cas purement observationnel, c'est-à-dire sans données d'interventions, il existe déjà un problème de classe d'équivalence pour les graphes acycliques.

Proposition 3 *Deux graphes orientés acycliques sont Markov-équivalent si et seulement si ils ont le même squelette et les mêmes immoralités.*

Nous nous attendons a minima à retrouver ce type de classe d'équivalence. Nous obtenons un espace bien plus grand pour cette équivalence :

Exemple 1 *Considérons le modèle causal gaussien défini par les paramètres suivants :*

$$\boldsymbol{\mu_A} = (1 \quad 1 \quad 1 \quad -1)$$

$$\boldsymbol{W_A} = \begin{pmatrix} 0 & 0 & 0 & 0.3 \\ 2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$\boldsymbol{\sigma_A^2} = \begin{pmatrix} 10^{-10} & 10^{-10} & 10^{-3} & 10^{-6} & 10^{-8} \end{pmatrix}$$

Alors dans le cadre d'observations uniquement, celui ci est équivalent au modèle suivant :

$$\boldsymbol{\mu_B} = (1.875 \quad -0.5 \quad 0.001008981 \quad -1)$$

$$\boldsymbol{W_B} = \begin{pmatrix} 0 & 0.5 & 0.299697306 & 2.970297 \times 10^{-1} \\ 0 & 0 & -0.001008981 & -1.009059 \times 10^{-9} \\ 0 & 0 & 0 & 9.900990 \times 10^{-3} \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\boldsymbol{\sigma_B^2} = \begin{pmatrix} 1.564078 \times 10^{-3} & 2.500000 \times 10^{-11} & 1.008981 \times 10^{-6} & 9.900990e \times 10^{-9} \end{pmatrix}$$

*On peut représenter les graphes G_A et G_B associés à ces paramètres (*Figure 1*).*

Lorsque l'on relâche l'hypothèse d'acyclicité, la classe d'équivalence associée aux observations :

— ne conserve plus le même squelette

— ne conserve plus les immoralités.

En résumé, si X est un vecteur aléatoire suivant un processus causal linéaire à bruit gaussien non récursif, alors X est identifiable dans une classe d'équivalence plus large que la classe d'équivalence markovienne.

Figure 1 – A gauche, le modèle A, à droite le modèle B

4 Discussion

Cette étude sur l'hypothèse d'acyclicité montre qu'elle est nécessaire dans le cas de données purement observationnelles. Dans ce cas, on arrive toujours à se ramener à une classe d'équivalence facilement compréhensible. Le modèle cyclique, même s'il dispose de l'avantage d'être stable par marginalisation, ne permet pas de conclusion sur de simples observations. Une question qui reste ouverte est cette même étude dans le cas d'interventions. Ces données sont de plus en plus d'actualité et elles permettent dans le cas acyclique de réduire la classe d'équivalence. Nous pouvons espérer atteindre le même type de résultat ici.

Références

Antti Hyttinen, Frederick Eberhardt, and Patrik O Hoyer. Learning linear cyclic causal models with latent variables. *The Journal of Machine Learning Research*, 13(1) :3387–3439, 2012.

Judea Pearl. *Causality : models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

Andrea Rau, Florence Jaffrézic, and Grégory Nuel. Joint estimation of causal effects from observational and intervention gene expression data. *BMC Systems Biology*, 7(1) :111, 2013.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*, volume 81. MIT press, 2000.