

# Identification d'effets causaux marginaux à partir de données observationnelles et interventionnelles

Gilles Monneret<sup>1,2</sup>  
Florence Jaffrézic<sup>2</sup>

Andrea Rau<sup>2</sup>  
Grégory Nuel<sup>1</sup>

<sup>1</sup> Université Pierre et Marie Curie, 5 place Jussieu, Paris

<sup>2</sup> INRA, Domaine de Vilvert, Jouy-en-Josas

gilles.monneret@upmc.fr

## Résumé

Nous proposons une méthode utilisant le formalisme de la causalité dans le but d'inférer les interactions causales en aval d'un nœud d'intervention. Pour ce faire, nous détaillons comment nous écrivons les vraisemblances des différents modèles, à l'aide d'un jeu d'équivalence paramétriques. Cela permet de comparer les différents modèles à l'aide d'un facteur de Bayes. En guise d'illustration, nous comparons les résultats de notre méthode avec ceux d'une analyse différentielle dans le cadre de données réelles.

## Mots Clef

Causalité, modèles graphiques, méthode bayésienne.

## Abstract

We propose a method to infer downstream causality between an intervention node and other nodes in an unknown graph. In particular, we describe the relevant models used and their respective likelihoods using a parametric equivalence. Models are subsequently compared using a Bayes factor. As an illustration, we compare the results of our method with those of a standard *t*-test in the context of a differential analysis of gene expression data.

## Keywords

Causality, graphical models, bayesian method

## 1 Introduction

Dans le cadre des modèles graphiques probabilistes orientés, des auteurs ont déjà proposé des algorithmes capables d'inférer des squelettes de graphes orientés, c'est-à-dire des graphes non orientés, ou des graphes partiellement orientés via des tests d'indépendance conditionnelle (Kalisch and Bühlmann, 2007) ou des procédures de scores (Chickering et al., 1997). Ceux-ci correspondent à une classe d'équivalence, c'est-à-dire à un sous-ensemble indistinguable de graphes. Nous pouvons aussi ajouter les liens avec les modèles graphiques gaussiens, et notamment l'estimation parcimonieuse de la matrice de précision associée (Fu and Zhou, 2013). Malheureusement ces

techniques sont coûteuses algorithmiquement, et cherchent à inférer l'ensemble du graphe. Eventuellement, on peut chercher non pas à inférer le graphe directement, mais plutôt chercher à obtenir un ordre topologique sur les nœuds. L'utilisation de nombreuses données interventionnelles conjointement aux observations permet de retrouver l'ordre partiel associé au graphe orienté acyclique (Rau et al., 2013). Cependant dans de nombreux cas, seules quelques interventions sont possibles. Dans ce cas, seule une classe d'équivalence restreinte est identifiable (Spirtes et al., 2000). Lorsqu'on cherche à connaître l'effet causal d'un nœud, il peut être alors plus intéressant de se tourner vers une approche marginale, c'est-à-dire se concentrant exclusivement sur les interactions d'un nœud précis avec les autres. La formalisation d'une intervention sous la forme d'un opérateur mathématique (Pearl, 2000) nous permet d'établir une telle méthode. Nous allons ici la développer, puis l'appliquer à des données simulées et des données réelles issues de la génomique. Dans celles-ci, des réplicats interventionnels peuvent être obtenu par invalidation génique (*knock-out*), c'est-à-dire en supprimant l'expression d'un gène.

## 2 Modèles graphiques

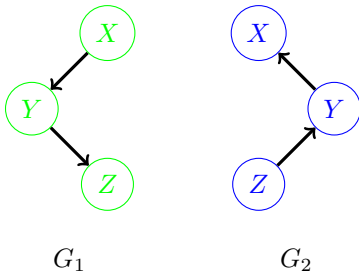
Un graphe orienté  $\mathcal{G}$  est un ensemble de nœuds  $V$  et un ensemble d'arêtes orientées  $\mathcal{E}$ . Pour  $(X, Y) \in \mathcal{E}$ , on dit que  $X$  est un parent de  $Y$  ( $X \in \text{pa}(Y)$ ) ou que  $Y$  est un enfant de  $X$  lorsqu'une arête a pour origine  $X$  et pointe vers  $Y$ . Un chemin orienté est une succession de nœuds telle que chaque élément est un parent du suivant. Un graphe est dit acyclique s'il n'existe pas de chemin orienté d'un nœud vers lui-même. On dit alors que ce graphe est un DAG (*directed acyclic graph*).

Nous pouvons associer à un graphe orienté acyclique une densité de probabilité  $P$ . Une telle densité doit nécessairement respecter le critère de factorisation : la densité jointe du graphe se factorise comme suit :

$$P(\mathbf{V}) = \prod_{V \in V} f(V | \text{pa}(V)).$$

Un inconvénient d'une telle représentation est qu'elle n'est pas stable par marginalisation. Nous cherchons ici à trouver des liens de causalité, c'est-à-dire à s'assurer de trouver des effets à des causes, ou des causes à des effets. Pour la suite, nous allons toujours admettre une interprétation causale des DAGs. Deux graphes sont dits équivalents au sens de Markov lorsque la probabilité jointe à chacun de ces deux graphes peut se factoriser indépendamment dans l'un ou l'autre graphe. Par exemple les deux graphes suivant appartiennent à la même classe d'équivalence.

FIGURE 1 – Graphes équivalents



$$\begin{aligned} \mathbb{P}(G_1) &= \mathbb{P}(X) \mathbb{P}(Y|X) \mathbb{P}(Z|Y) \\ \mathbb{P}(G_2) &= \mathbb{P}(Z) \mathbb{P}(Y|Z) \mathbb{P}(X|Y) \end{aligned}$$

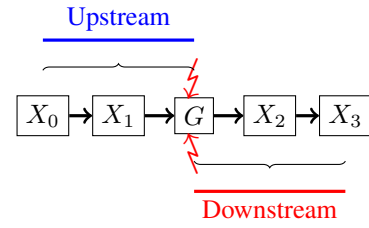
Les lois de probabilité sous-jacentes se factorisent toutes deux dans l'un ou l'autre graphe. On définit l'opérateur  $\text{do}(X = x)$  graphiquement par l'opération consistant à supprimer toutes les arêtes provenant des parents de  $X$  vers  $X$  (Pearl, 2000). Pour la loi de probabilité associée, cela consiste à remplacer la loi conditionnelle  $f(X|\text{pa}(X))$  par  $\mathbf{1}_{X=x}$ .

## Sélection de modèle

Plaçons nous dans un graphe  $\mathcal{G}$  dans lequel un nœud d'intérêt, noté  $G$  peut subir des interventions. Nous proposons de distinguer 2 types de causalité : la causalité amont (*upstream*) et la causalité aval (*downstream*). La première fait référence aux arêtes pointant vers  $G$  (les *causes* de  $G$ ), tandis que la seconde fait référence aux arêtes provenant de  $G$  (les *conséquences* de  $G$ ). Nous choisissons cette formulation pour la raison suivante. La grande majorité des plans d'expérience consiste à perturber un système en fixant une variable, toutes choses égales par ailleurs, ce qui revient à appliquer l'opérateur  $\text{do}$ . Or l'opérateur  $\text{do}$  permet de détecter la causalité aval, par construction.

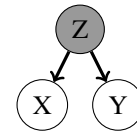
Notre problématique est la suivante : comment détecter ce type de causalité en particulier ? Nous savons qu'à partir de données observationnelles, seule une classe d'équivalence pour les DAG est identifiable. Avec des interventions, nous allons réduire cette classe d'équivalence, mais nous ne pourrons toujours pas identifier un DAG précis. De plus, les modèles que l'on cherche à décrire peuvent admettre un nombre conséquent de variables latentes, notamment en génétique. En effet, considérons le graphe suivant

FIGURE 2 – Causalité amont/avale



muni d'une loi qu'on ne précise pas. Si on marginalise sur

FIGURE 3 – Corrélation par variable latente



$Z$ , nous obtenons un graphe où les variables  $X$  et  $Y$  sont corrélées, et nous perdons l'identifiabilité du modèle. En effet, par l'équivalence de Markov pour les graphes orientés, les deux orientations sont possibles. Cela pose particulièrement problème lorsqu'on cherche à effectuer une interprétation causale de ce DAG : il n'existe plus forcément de relation causale entre  $X$  et  $Y$ .

Ces considérations nous ont fait choisir une approche marginale de notre problème. Nous choisissons de regarder les interactions du nœud d'intérêt avec les autres nœuds pour en déduire les interactions causales avals les plus vraisemblables. On se place dans un modèle d'équations structurelles gaussiennes : pour chaque nœud  $X$  du graphe, si  $X$  est un enfant de  $G$ , alors l'équation suivante est vérifiée :

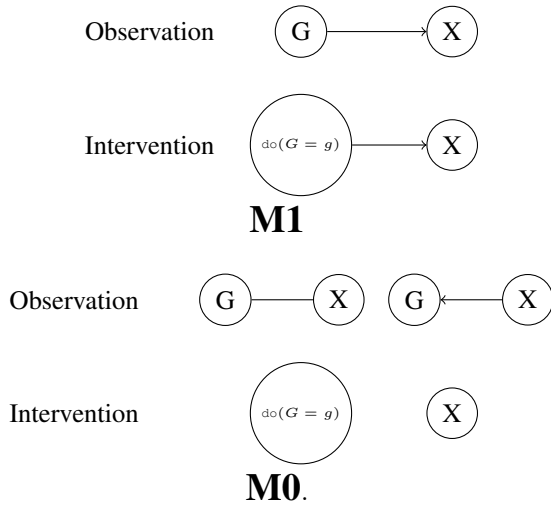
$$X = \mu_X + \alpha G + \epsilon_X, \quad (1)$$

où  $\mu$  est la moyenne résiduelle de  $X$ ,  $\alpha$  est un coefficient d'effet total de  $G$  vers  $X$  et  $\epsilon_X$  est un bruit gaussien centré de variance  $\sigma_X^2$ . Si  $X$  est un parent de  $G$  alors de la même façon on peut écrire l'équation suivante :

$$G = \mu_G + \alpha X + \epsilon_G.$$

Enfin, si  $X$  n'est ni un enfant, ni un parent de  $G$ , alors on sort du modèle d'équations structurelles et on représente le couple  $(X, G)$  comme une variable aléatoire suivant une loi gaussienne multivariée. Ayant écrit ces trois possibilités, nous pouvons écrire les vraisemblances associées, pour les données d'observations et d'interventions. Nous pouvons y associer une représentation graphique, présentée dans la figure 4. Celle ci fait apparaître une différence importante dans l'apport des données interventionnelles : tandis que pour le modèle M1, nous avons besoin de la loi de  $X$  sous l'opérateur  $\text{do}$ , dans le modèle M0 nous avons besoin de la loi marginale de  $X$ . Écrit de cette façon, cela fait ap-

FIGURE 4 – Modèles sous observations/interventions



paraître le fait qu'ici les modèles amont et corrélés sont indistinguables. Ces modèles étant posés, grâce à l'équivalence de Markov pour les données observationnelles, nous pouvons tout reparamétriser sous la forme (1), avec l'identification suivante. Nous avons alors toutes les lois néces-

$$M1 : Z_1 \sim \mathcal{N}(\mu_1, \sigma_1^2) \quad Z_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

$$G = Z_1 \quad X = \alpha Z_1 + Z_2$$

$$M0 : \tilde{Z}_1 \sim \mathcal{N}(\tilde{\mu}_1, \tilde{\sigma}_1^2) \quad \tilde{Z}_2 \sim \mathcal{N}(\tilde{\mu}_2, \tilde{\sigma}_2^2)$$

$$G = \beta \tilde{Z}_1 + \tilde{Z}_2 \quad X = \tilde{Z}_1$$

$$M0 : \begin{pmatrix} G \\ X \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \begin{pmatrix} s_1^2 & \rho s_1 s_2 \\ \rho s_1 s_2 & s_2^2 \end{pmatrix} \right)$$

$$\mu_1 = m_1 \quad \mu_2 = m_2 - \alpha m_1 \quad \sigma_1 = s_1$$

$$\alpha = \rho s_2 / s_1 \quad \sigma_2 = \sqrt{s_2^2 - \alpha^2 s_1^2}$$

$$\tilde{\mu}_1 = m_2 \quad \tilde{\mu}_2 = m_1 - \beta m_2 \quad \tilde{\sigma}_1 = s_2$$

$$\beta = \rho s_1 / s_2 \quad \tilde{\sigma}_2 = \sqrt{s_1^2 - \beta^2 s_2^2}$$

saires : la loi jointe, la loi conditionnelle de  $X$  sachant  $G$  ainsi que la loi marginale de  $X$ , le tout écrit avec le même jeu de paramètres et pour tous les modèles. Par simplification, nous noterons par la suite  $\mu_X, \sigma_X, \mu_G, \sigma_G$  le jeu de paramètres résiduels associés respectivement à  $X$  et  $G$ . Nous obtenons alors deux vraisemblances, où  $WT$  désigne le sous-ensemble de données d'observations et  $KO$  le sous-ensemble de données d'interventions :

$$\ell_0(\theta) = \sum_{k=1;WT} -\frac{1}{2} \log(2\pi\sigma_X^2) - \frac{1}{2\sigma_X^2} (X_k - \mu_X - \alpha G_k)^2$$

$$- \frac{1}{2} \log(2\pi\sigma_G^2) - \frac{1}{2\sigma_G^2} (G_k - \mu_G)^2$$

$$+ \sum_{k=1;KO} -\frac{1}{2} \log(2\pi(\alpha\sigma_G^2 + \sigma_X^2))$$

$$- \frac{1}{2(\alpha\sigma_G^2 + \sigma_X^2)} (X_k - \mu_X - \alpha\mu_G)^2$$

$$\ell_1(\theta) = \sum_{k=1;WT} -\frac{1}{2} \log(2\pi\sigma_X^2) - \frac{1}{2\sigma_X^2} (X_k - \mu_X - \alpha G_k)^2$$

$$- \frac{1}{2} \log(2\pi\sigma_G^2) - \frac{1}{2\sigma_G^2} (G_k - \mu_G)^2$$

$$+ \sum_{k=1;KO} -\frac{1}{2} \log(2\pi\sigma_X^2) - \frac{1}{2\sigma_X^2} (X_k - \mu_X - \alpha G_k)^2.$$

Une fois ces quantités maximisées, on peut considérer un facteur de Bayes :

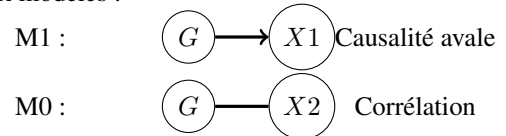
$$B = \mathbb{P}(\text{data}|M_0) / \mathbb{P}(\text{data}|M_1).$$

On utilise ce facteur pour classer les différents nœuds en terme de probabilité d'être en aval de  $G$  : s'il est supérieur à 1, on choisira le modèle  $M0$ , tandis que s'il est inférieur à 1 le modèle  $M1$  s'imposera.

## Application

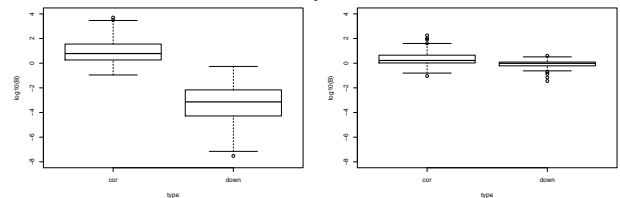
### Simulation

Pour vérifier la pertinence de notre méthode, nous commençons par une étude de simulations. Dans un premier temps, nous nous posons la question de savoir si notre modèle distingue bien la causalité avale de la corrélation. Pour cela, nous simulons 24 observations et 24 interventions selon deux modèles :



Nous nous plaçons toujours sous le modèle gaussien. Nous calculons notre facteur de Bayes et répétons cette expérience une centaine de fois. Les résultats sont les suivants (FIGURE 1). En gardant les mêmes paramètres de moyenne

FIGURE 5 – Facteur de bayes : corrélation et causalité

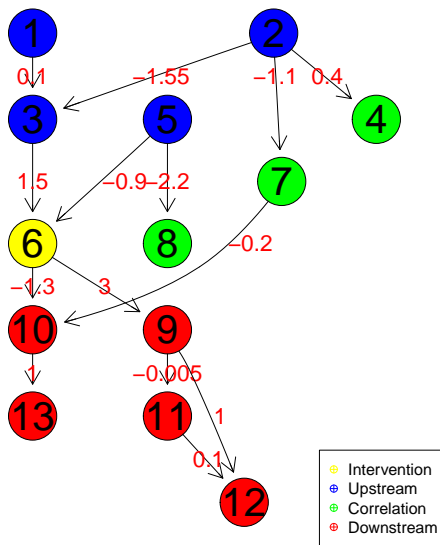


et d'effet causal, nous avons représenté deux résultats en fonction des variances résiduelles. A gauche, le modèle est simulé avec  $\sigma_G = 0.3, \sigma_X = 0.5$ . A droite, avec  $\sigma_G = 3, \sigma_X = 5$ . Nous voyons qu'à gauche, le logarithme du facteur de Bayes reste négatif pour le modèle causalité avale, tandis que pour le modèle corrélaté ce facteur est plutôt positif ou nul. A droite, nous montrons que

nous sommes bien sensibles au bruit et que si ce dernier est trop fort comparé à l'effet causal, nous n'arriverons pas à capturer l'éventuelle causalité. Nous choisirons toujours dans ce cas le modèle corrélé.

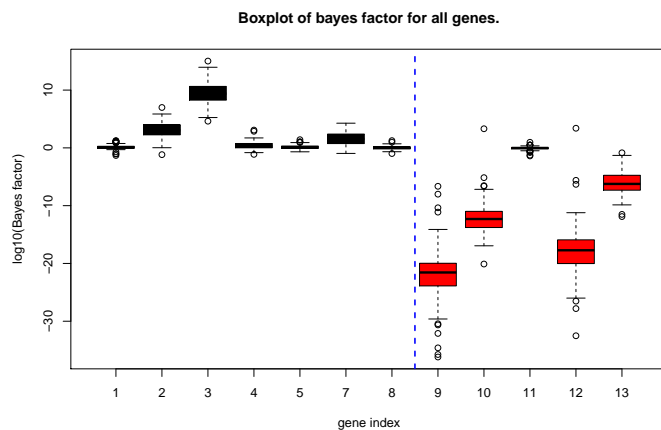
Pour une seconde approche, nous tentons de voir si nous sommes capables de retrouver les partenaires causaux aval y compris dans un graphe simulé. Nous simulons une centaine de réplicats autant observationnels qu'interventionnels selon le graphe orienté acyclique suivant, sous un modèle d'équations structurelles gaussiennes. Notre nœud

FIGURE 6 – Graphe de simulation



d'intervention, en jaune, a plusieurs type d'interactions avec les autres nœuds. En bleu, ce sont des ascendants causaux, ceux admettant une causalité amont, en rouge la causalité avale et en vert les simples corrélations. Nous indiquons d'autre part la force des effets causaux sur les différentes arêtes. Plus le logarithme du facteur de Bayes est

FIGURE 7 – Facteur de Bayes obtenus sur les données de simulations

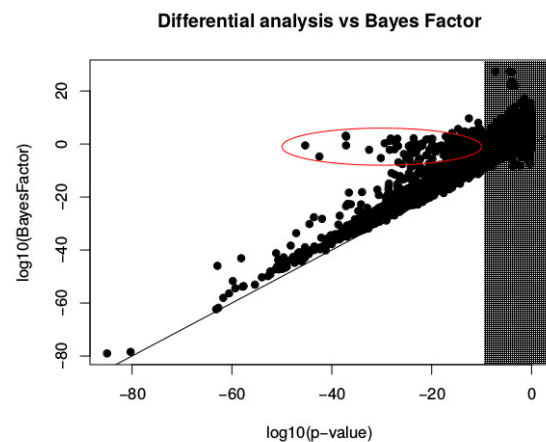


negatif, plus nous sommes certains d'être dans le modèle causalité avale. Nous voyons ici qu'effectivement, pour les nœuds en rouge nous obtenons un facteur de Bayes cohérent. Seul le nœud 11 n'est pas détecté en aval : cela s'explique par le faible effet causal total, égal à  $-0.015$ , du nœud 6 sur le nœud 11. Nos simulations permettent de valider l'efficacité de notre méthode.

## Données réelles

Nous proposons de tester cette méthode dans le contexte des réseaux de régulations géniques. Bien que ces derniers peuvent admettre des interactions non linéaires et des cycles, nous faisons le choix de continuer sous des hypothèses d'acyclicité et de linéarité. Nous travaillons sur des données transcriptomiques chez la poule. Celles-ci sont généralement traitées, après normalisation, comme des réalisations d'une loi gaussienne multivariée : nous nous plaçons dans ce cas. Nous avons deux lignées, une témoin et une présentant une mutation sur le gène du récepteur de l'hormone de croissance, ce qui induit un phénotype de nanisme. Nous prenons donc cette mutation comme une intervention, et nous cherchons à savoir quels vont être parmi les 43077 autres gènes détectés ceux qui sont le plus vraisemblablement en aval de ce gène. Nous avons 24 réplicats observationnels et 24 réplicats interventionnels. L'approche standard utilisée pour traiter ce type de modèle est l'analyse différentielle. Il s'agit d'une démarche de test bilatéral portant sur le paramètre de moyenne : c'est, dans le cas gaussien, un test de Student appliqué sur les deux groupes. Nous comparons dans la figure 8 le facteur de Bayes obtenu par notre méthode et la  $p$ -value obtenue par l'analyse différentielle. Nous voyons sur cette figure que

FIGURE 8 – Comparaison entre l'analyse différentielle et notre méthode causale



nos résultats sont extrêmement proches de ceux de l'analyse différentielle. Seuls une vingtaine de gènes différentiellement exprimés semblent ne pas être causalement en aval du gène  $G$ , à mettre en comparaison avec les 403077 gènes. Nous voyons donc qu'il semble dans ce cas y avoir

une relation très proche entre l'analyse différentielle et notre méthode causale. Ceci justifie donc a priori l'utilisation de l'analyse différentielle comme outil de détection de la causalité : nous obtenons les mêmes résultats lorsque l'on formalise explicitement la causalité.

## Discussion

Nous avons proposé une nouvelle méthode causale, qui bien qu'elle reste extrêmement proche de l'analyse différentielle en terme de résultats, propose une formulation causale plutôt qu'en terme de différence de moyenne. Nous l'avons confronté à des simulations, et celles-ci valident notre démarche. Même si nous nous plaçons dans un cadre plus restrictif, du fait des équations structurelles linéaires, elle possède le mérite d'être facilement interprétable. En terme de plus-value, en plus d'insérer ce formalisme de la causalité, cette approche permet d'entrevoir toute une série d'approche bayésienne autour de modèles à multiples invalidations géniques, difficiles à traiter avec une analyse différentielle.

## Références

- Chickering, D. M., Heckerman, D., and Meek, C. (1997). A bayesian approach to learning bayesian networks with local structure. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pages 80–89. Morgan Kaufmann Publishers Inc.
- Fu, F. and Zhou, Q. (2013). Learning sparse causal gaussian networks with experimental intervention : regularization and coordinate descent. *Journal of the American Statistical Association*, 108(501) :288–300.
- Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *The Journal of Machine Learning Research*, 8 :613–636.
- Pearl, J. (2000). *Causality : models, reasoning and inference*, volume 29. Cambridge Univ Press.
- Rau, A., Jaffrézic, F., and Nuel, G. (2013). Joint estimation of causal effects from observational and intervention gene expression data. *BMC Systems Biology*, 7(1) :111.
- Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*, volume 81. MIT press.