Un algorithme efficace fondé sur les scores pour l'apprentissage de structure de réseaux causaux avec variables latentes

Amir-Hosein Valizadeh, Christophe Gonzales
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

L'apprentissage de réseaux causaux [1] à partir données est une problématique importante en Intelligence artificielle car ces réseaux permettent d'effectuer des raisonnements tels que le pratiquent les humains, notamment le raisonnement contrefactuel (si j'avais fait ceci, que se serait-il passé ?). Un tel apprentissage s'appuie souvent sur celui de réseaux bayésiens [2]. On parle alors d'apprentissage à partir de données observationnelles et c'est le cadre dans lequel se place cet article. Les réseaux bayésiens et causaux sont constitués, d'une part, d'un graphe orienté sans circuit (DAG – *Directed Acyclic Graph*) dont les nœuds représentent des variables aléatoires, une absence d'arc entre deux nœuds indiquant une indépendance conditionnelle probabiliste entre ces variables, et d'autre part, de distributions de probabilité conditionnelles des variables sachant leurs parents dans le graphe. Lorsqu'aucune variable n'est latente (non observée), deux grandes classes d'algorithmes ont été développées pour apprendre la structure de ces réseaux : ceux fondés sur les contraintes, qui s'appuient sur des tests d'indépendance conditionnelle statistique pour chercher la structure représentant au mieux ces indépendances, et les algorithmes à base de score, qui recherchent le graphe dont la vraisemblance est la plus élevée ou bien dont le regret est minimal. Les algorithmes sous contraintes, par exemple PC [3], sont assez sensibles aux erreurs dans les tests statistiques, commises notamment sur les petites bases de données, ainsi qu'à l'ordre dans lequel les tests sont réalisés. Les algorithmes à base de score, comme *Greedy Hill Climbing*, sont en ce sens plus robustes et sont souvent préférés. Les algorithmes hybrides combinant les deux approches donnent souvent les meilleurs résultats.

Malheureusement, dans la réalité, il existe souvent des variables latentes et leur prise en compte est importante pour le raisonnement causal. Les algorithmes sous contraintes peuvent être étendus pour en tenir compte, c'est le cas par exemple de l'algorithme *Fast Causal Inference* (FCI) [3]. *Multivariate Information-based Inductive Causation* (MIIC) [4] en est un autre exemple, fondé sur la théorie de l'information et se focalisant sur la détermination de v-structures (des motifs $X \rightarrow Y \leftarrow Z$). En revanche, étant donné que les algorithmes à base de score calculent des vraisemblances, ils nécessitent de compter les occurrences d'apparition des différentes valeurs que peuvent prendre des ensembles de variables, ce qui est impossible dès lors que certaines d'entre elles sont non observées. C'est pourquoi ils ne sont pas exploités pour déterminer la structure de réseaux bayésiens ou causaux en présence de variables latentes. On peut toutefois utiliser des scores pour apprendre d'autres types de graphes comme les PAG (Partial Ancestral Graphs) ou les MAG (Maximal Ancestral Graphs) [5] qui sont des extensions des DAG, mais que l'on ne sait scorer que dans un cadre continu gaussien.

Dans ce papier, nous nous intéressons au cas de variables discrètes. Plus précisément, nous proposons un nouvel algorithme fondé uniquement sur des scores et n'effectuant qu'une recherche dans l'espace des DAG afin de déterminer la structure causale la plus probable, même en présence de variables latentes. Sans perte de généralité [6], nous nous plaçons dans le cadre des réseaux causaux semi-markoviens [1], c'est-à-dire dans le cas où les variables latentes n'ont aucun parent et précisément deux enfants, observés (non latents), dans le graphe. Par ailleurs, nous supposons que la base de données servant pour l'apprentissage a été générée à partir d'une distribution de probabilité P qui admet une *perfect map* G*, c'est-à-dire un réseau bayésien dans lequel toute présence (resp. absence) d'arc représente une dépendance (resp. indépendance) probabiliste. Notre algorithme s'appuie sur l'observation suivante : lorsque A et B sont deux enfants, dans le graphe G*, d'une variable latente L et qu'ils ont au moins un autre parent

observé, autrement dit quand G* contient un motif tel que C → A ← L → B ← D, si la base de données est suffisamment grande, les algorithmes d'apprentissage à base de score produisent une structure G dans laquelle il existe un arc entre A et B (car ces variables sont dépendantes d'un point de vue probabiliste). Supposons qu'il s'agisse, ici, de l'arc A → B. Dans ce cas, C et B sont dépendants conditionnellement à la variable A, ce qui amène les algorithmes à base de score à ajouter un arc entre C et B. Autrement dit, (C,A,B) forme un triangle dans G. Par ailleurs, il n'y a pas de dépendance directe entre B et C. Conjointement, ces deux propriétés permettent de déterminer que ces triangles proviennent de variables latentes. Notre algorithme consiste donc, dans un premier temps, à exploiter un algorithme à base de score afin de trouver un graphe G, qui est une structure de réseau bayésien (mais pas de réseau causal). On recherche ensuite les triangles (C,A,B) présents dans ce graphe. En réalisant des comparaisons de scores impliquant les variables B, C, A (plus, éventuellement, d'autres variables permettant d'assurer qu'une dépendance entre B et C ne passerait pas par un autre « chemin » que C → A ← B), on peut déterminer si le triangle indique ou non la présence d'une variable latente. En effet, lors que le graphe G* contient des triangles n'impliquant pas de variable latente, les triangles (C,A,B) correspondants dans G ne n'impliquent pas non plus la présence d'une variable latente. Lorsque l'on détermine la présence d'une variable latente dans un triangle, on supprime les arcs entre A et B, et entre C et B, et on rajoute un nœud L dans G pour cette variable latente ainsi que deux arcs L → A et L → B, afin de se rapprocher, progressivement, du graphe G*. Lorsqu'il n'existe plus de triangle à examiner, on peut le transformer en CPDAG (*Completed Partially Directed Acyclic Graph*), c'est-à-dire en son représentant dans sa classe d'équivalence de Markov. Ce dernier graphe représente précisément tout ce que l'on inférer d'un point de vue structure causale à partir de la base de données.

Nous avons comparé notre algorithme avec FCI et MIIC sur des nombreuses bases de données générées aléatoirement en utilisant pyAgrum [7] à partir de réseaux bayésiens classiques (alarm, child, insurance, *etc*.), de différentes tailles (1k, 5k, 10k, 20k enregistrements), contenant différents nombres de variables latentes, de domaines de différentes tailles. Les résultats montrent que notre méthode surpasse FCI et MIIC dans la découverte des variables latentes, notamment elle commet beaucoup moins d'erreurs. Cela lui permet d'obtenir de très bons résultats en termes de précision quand on compare la structure de G* et celle apprise, ce qui est crucial si l'on souhaite s'en servir pour réaliser du raisonnement causal.

References

[1] J. Pearl (2009) *Causality*. Cambridge university press.

[2] J. Pearl (1988) Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufman.

[3] P. Spirtes, C. Glymour, R. Scheines (2001) *Causation, Prediction, and Search.* MIT Press.

[4] L. Verny, N. Sella, S. Affeldt, P.P. Singh, H. Isambert (2017) *Learning causal networks with latent variables from multivariate information in genomic data*. PLoS computational biology vol. 13, no. 10.

[5] T. Richardson, P. Spirtes (2002) Ancestral graph Markov models. Technical Report 375, Dpt of statistics, University of Washington.

[6] J. Tian, J. Pearl (2002) On the identification of causal effects. Technical Report R-290-L, UCLA Computer Science Lab.

[7] G. Ducamp, C. Gonzales, P.-H. Wuillemin (2020) aGrUM/pyAgrum: a toolbox to build models and algorithms for probabistic graphical models in Python. Proc. of Probabilistic Graphical Models, pp. 609-612.