Explaining data using causal Bayesian networks

Jaime Sevilla
Aberdeen University
j.sevilla.20@abdn.ac.uk

Abstract

We introduce Causal Bayesian Networks as a formalism for representing and explaining probabilistic causal relations, review the state of the art on learning Causal Bayesian Networks and suggest and illustrate a research avenue for studying pairwise identification of causal relations inspired by graphical causality criteria.

1 From Bayesian networks to Causal Graphical Models

Bayesian networks (BNs) are a class of probabilistic graphical models, originally conceived as efficient representations of joint probability distributions.

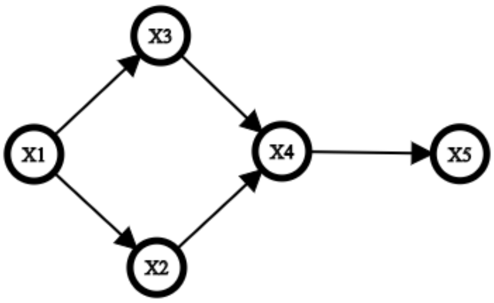

Figure 1: Bayesian network representing the probability distribution $P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1), P(x_4|x_2, x_3)P(x_5|x_4)$

A great deal of work has been dedicated in the last decades to understanding how to represent knowledge as BNs, how to perform efficient inference with BNs and how to learn BNs from data (Koller and Friedman, 2009).

Despite having been overshadowed by subsymbolic approaches, BNs are attractive because of their flexibility, modularity and straightforward statistical interpretation.

On top of that, BNs have a natural interpretation in terms of causal relations. Human-constructed BNs tend to have arrows whose directionality respects the causal intuitions of their architects.

Furthermore, recent work has extended Bayesian Networks with causal meaning (Pearl, 2009; Spirtes et al., 2001). The result are Causal Bayesian Networks and Causal Structural Models, that ascribe new meaning to BNs and extend classical inference with new causal inference tasks such as interventions (eg will the floor get wet if we turn the sprinkler on?) and counterfactuals (eg would this person have received a good credit rating if they had a stable job?).

In this paper we will review work on the area of using Bayesian networks to model causal relationship, and consider one future research direction to explore, concerning the identification of the causal link between pairs of variables.

2 Learning Causal Bayesian Networks

Considerations of causality also affect how Bayesian Networks should be learnt. Manually built Bayesian networks usually respect our causal intuitions. But Bayesian networks learnt from data may not respect the underlying causal structure that generated the data.

Indeed, each probability distribution can be represented by several different Bayesian Networks - and we can group Bayesian Networks graphs in classes capable of representing the same probability distributions, their Markov equivalence class.

Traditional BN learning methods such as score maximization (Cussens et al., 2017) cannot distinguish between members of the same Markov equivalence class, and will be biased towards outputting a Bayesian structure that fits the data well but does not necessarily match the underlying causal mechanisms.

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2020), pages 34–38,
Dublin, Ireland, 18 December 2020. ©2020 Association for Computational Linguistics

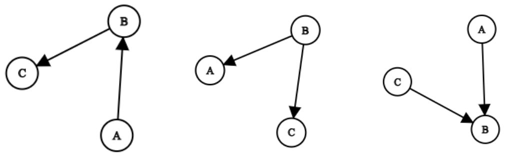

Figure 2: Three Bayesian Networks. The left and the middle one are Markov-equivalent, but the third one isn't equivalent to the other two - in fact, the left one is the only member of its reference class. Hence, if the data is compatible with the right BN and there are no latent variables BN we will be able to conclude that A causes B. But if the data is compatible with the left (and therefore the middle) BN then the orientation of the edge $A - B$ is arbitrary, and we cannot infer just from the data the causal relations between the variables.

This is a key problem for explaining the outcome of Bayesian Network learning algorithms. Experts usually avoid altogether a causal language - instead framing their explanations in terms of association. But we would like to be able to actually explain when a relation is causal and when we do not have enough information to tell one way or another.

In order to do this, we need our learning methods to distinguish and explain when their edge orientation decisions are arbitrary (ie there is another BN compatible with the data [1]where the edge is oriented in a different way) or necessary (ie the edge has this orientation in every diagram compatible with the data we have) - since only in the latter situation can we guarantee that the orientation will respect causality.

3 Previous work

This problem of causal discovery based on graphical models is reviewed in depth in (Glymour et al., 2019). In this article the authors introduce three families of causal discovery algorithms:

- Constraint based algorithms that rely on conditional independence tests to orient the edges in a graph. See for example the PC algorithm (Spirtes et al., 2001).

- Score based algorithms that greedily optimize a score function to orient the edges in a graph. See for example the Greedy Equivalence Search (Chickering, 2002).

- Functional algorithms that use stronger assumptions about the relation between two directly related variables to distinguish cause and effect. See for example the post-nonlinear causal model (Zhang and Hyvarinen, 2009).

The problem is considerably more difficult when we allow the possibility of unmeasured ('latent') common causes of the variables in our dataset.

This situation is arguably more representative of usual datasets, and requires specialized methods to be addressed. (Zhang, 2008) proposed a constraint-based learning algorithm that is provably sound and complete, assuming correct conditional independence decisions. The algorithm was later refined in (Claassen and Heskes, 2011).

4 A graphical test of causality and missing confounders

However, J. Zhang's and similar methods rely on frequentist and high order conditional independence tests to learn the causal structure, which are prone to error. The serial nature of the algorithm means that early errors in the conditional independence decisions lead to more errors later.

Ideally, we would like to have our methods of learning causality from observational data be more robust to statistical noise, and do not let errors propagate through the graph.

This is especially important when we are not interested in learning the complete structure of the graph, but rather we want to study the particular relation between a variable we could manipulate (the 'exposure') and a variable we care about (the 'outcome').

This problem has been discussed in depth in the context of econometrics, where structural equation modelling (Kaplan, 2020) and instrumental variable estimation methods (Reiersöl, 1945) are widely used tools for causal inference.

While structural equation modelling provides satisfactory answers to many questions of causal estimation, they are hard to interpret and use. Graphical models could lead us to better explanations of

[1] We have glossed over what compatible exactly means. A necessary condition is that all the independence conditions represented via d-separation in the graph are present in the joint probability distribution of the data (Spirtes, 1996). We usually also require the reverse, that all conditional independencies in the joint probability distribution are represented via d-separation in the graph - this is called the faithfulness assumption. The faithfulness assumption renders conditional independence and d-separation effectively equivalent, and restricts the output of the algorithm to a single Markov equivalence class. A justification of why we should expect our data to be faithful to the underlying model can be found in (Pearl, 2009, Chapter 2).

the models of causality used in econometrics and other contexts. For example, instead of providing models as mathematical equations, the causality we can infer from the data could be represented graphically, and described via text using similar techniques to those that apply to explaining Bayesian Networks with natural language generation techniques (see (Reiter, 2019) for discussion).

In particular, under certain conditions we can use insights derived from causal discovery in graphical models to test conditions usually taken on faith.

For example, if we identify two additional variables Z, W and a context $S = s$ such that:

- A and B are conditionally dependent given $S = s$
- Z and W are conditionally independent given $S = s$, but are conditionally dependent given $S = s$ and $A = a$ for some value a
- Z and B are conditionally dependent given $S = s$, but conditionally independent given $S = s$ and $A = a$ for every value a

then lemma 1 from (Claassen and Heskes, 2011) implies under mild assumptions that there is a directed path from A to B in every causal bayesian network compatible with the data we have observed.

To ground this example, let's suppose that we are interested in studying the effect of a drug (A) on the health of a patient (B). We furthermore have access to information about the patient's income (Z) and whether they have health insurance (W). We also have access to a set of background information variables (O) like for example age and gender.

We assume that the causal relationships between the variables can be represented as an acyclic graphical model.

We check that the income (Z) and the drug (A) are independent conditional on some of the background variables ($S \subset O$), but dependent when we condition on $S \cup \{A\}$.

Then we check that the income (Z) and the patient's health outcome (B) are conditionally dependent given the same subset of background variables S, but independent when we condition on the drug A.

Then we will be able to assert that no matter what the true acyclic causal diagram is, there will always be a causal path that starts in the treatment (A) and ends in the patient's health outcome (B). This guarantee holds as long as we can guarantee acyclity - even if there are unmeasured latent variables in the true causal diagram.

Hence it would be appropriate to describe the data as providing evidence for the natural language explanation ”the drug has an effect on the health of the patient”. Note that we can only provide this explanations because of our explicit causal analysis. A traditional Bayesian analysis would only be able to conclude that the drug and the health outcome are somehow related - but it would not have been able to distinguish the direction of causation (perhaps sicker patients are more likely to be treated with the new drug!) or rule out confounding common causes (perhaps richer patients are both more likely to receive the treatment and have better health outcomes for reasons unrelated to the drug!).

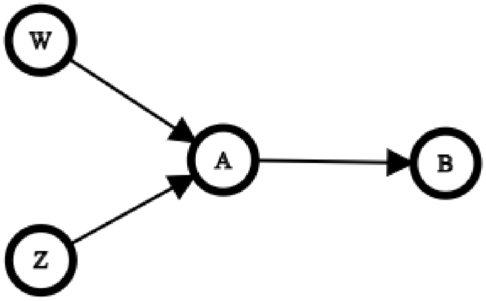

Figure 3: If the underlying causal structure follows this diagram, then because of d-separation properties we will be able to conclude that A, B, Z, W and $S = \emptyset$ satisfy the conditions we listed. Hence, every Bayesian network in the Markov-equivalence class of the diagram (including diagrams with latent variables) includes a directed path from A to B. So we will be able to unequivocally conclude that A causes B.

Like J. Zhang's causal discovery algorithm, this criteria allows the possibility of latent common causes. Unlike J. Zhang's, this criteria only depends on locally testing the conditional independence relations between A, B, Z, W, S to conclude that A is a cause of B. A similar approach is considered in (Mani et al., 2012), though in the context of global structure learning.

From an econometric perspective, the interest of the criteria above is that this condition provides a graphical test for causality and missing confounders, under the assumption of no cyclical causal relations. In particular, the conditions outlined above imply that $S = s$ blocks all confounding paths that spuriously relate A and B but blocks no causal path between the variables. Thus if we found that the criteria holds we would be able to use standard tools such as ordinary least squares regression to quantify the strength of the causal relation $A \to B$.

Related tests already exist in the literature - for example the overidentification J-test for testing the exogeneity of instrumental variables (Stock and Watson, 2011, Section 12.3), and selection of observables for testing the sensibility of an estimate to hidden confounders (Altonji et al., 2000).

Understanding the relationship between these traditional tests and the tests derived from the graphical criteria will be an interesting multidisciplinary exercise.

5 Conclusion and next steps

In summary, the development of a local causal criteria will give us a powerful tool to build causal explanations of data, that under certain conditions can distinguish the direction of causality and quantify the strength of the underlying causal relation.

The development of this criteria will be of great help to fields eager to extract causal conclusions from historical data. For example, this could help medics and patients gain an understanding of how much of a difference a treatment would make based on the history of past patients, so they can make an informed decision about it.

It is unclear how to generalize these conditions to cover more cases and possible causal relations, how often these conditions are met, how reliable procedures of proving causality based on this type of criteria would be and how to deal with possibly contradictory evidence of causality.

Our intention is to explore these questions through our work. This will involve three avenues of research:

- Formulating and formally studying criteria for proving causal relations through mathematical definitions and proofs
- Developing my own algorithms of causal discovery based on such criteria and refining them by evaluating them on synthetic data
- Testing the performance the resulting algorithms on real datasets

We do not expect this work to be easy.

Specially challenging in the context of econometrics will be the validation of the methods used. Only seldom do we have direct experimental evidence of the causal relations in a economic domain. Because of this, initial experimentation should focus on explaining observational data in domains where there is a strong and well-established theory of causation, such as price-demand modelling.

Another key difficulty is the requirement of conditional independencies - it will often be impossible in econometric contexts to render variables conditionally independent. Thus part of our work will require us to relax the conditions of Y-structure based causal discovery to exploit weaker forms of conditional independence. For example, we could look into interaction information (McGill, 1954) or related concepts from information theory.

Finally, there is a problem on explaining this graphical reasoning to users. It is not obvious why Y-structures imply a causal relationship. It may be fruitful to draw an analogue between this method and how humans infer causation, to make them more intuitive.

We believe that this work will help us better understand how to study causal relationships from observational data, which will have long reaching applications in econometrics, medicine and other fields of practice that routinely need to rely on observational data for their analyses.

Furthermore, causal graphical models have an advantage compared to black box learning and reasoning models due to their ability to address causal queries. This could be leveraged to marginally push the field of AI towards methods inspired by probabilistic graphical models, which are arguably more transparent and will facilitate goal alignment.

Acknowledgements

I thank my supervisors Ehud Reiter and Nava Tintarev for thorough discussion and support.

I also thank the anonymous reviewers for the NL4XAI for kindly providing constructive feedback to improve the paper.

This research has been supported by the NL4XAI project, which is funded under the European Union's Horizon 2020 programme, grant agreement 860621.

References

Joseph G Altonji, Todd E Elder, and Christopher R Taber. 2000. Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic

Schools. Working Paper 7831, National Bureau of Economic Research. Series: Working Paper Series.

David Maxwell Chickering. 2002. Optimal Structure Identification With Greedy Search. *Journal of Machine Learning Research*, 3(Nov):507–554.

Tom Claassen and Tom Heskes. 2011. A structure independent algorithm for causal discovery. In *In ESANN'11*, pages 309–314.

James Cussens, Matti Järvisalo, Janne H. Korhonen, and Mark Bartlett. 2017. Bayesian Network Structure Learning with Integer Programming: Polytopes, Facets and Complexity. *Journal of Artificial Intelligence Research*, 58:185–229.

Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, 10. Publisher: Frontiers.

David Kaplan. 2020. *Structural Equation Modeling (2nd ed.): Foundations and Extensions*, 2nd edition. Thousand Oaks, California.

Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press.

Subramani Mani, Peter L. Spirtes, and Gregory F. Cooper. 2012. A theoretical study of Y structures for causal discovery. *arXiv:1206.6853 [cs, stat]*. ArXiv: 1206.6853.

William J. McGill. 1954. Multivariate information transmission. *Psychometrika*, 19(2):97–116.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*, 2nd edition edition. Cambridge University Press, Cambridge, U.K. ; New York.

Olav Reiersöl. 1945. Confluence analysis by means of instrumental sets of variables.

Ehud Reiter. 2019. Natural Language Generation Challenges for Explainable AI. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, pages 3–7. Association for Computational Linguistics.

Peter Spirtes. 1996. Using d-separation to calculate zero partial correlations in linear models with correlated errors. Publisher: Carnegie Mellon University.

Peter Spirtes, Clark Glymour, and Richard Scheines. 2001. *Causation, Prediction, and Search, 2nd Edition*, volume 1. The MIT Press. Publication Title: MIT Press Books.

James Stock and Mark Watson. 2011. *Introduction to Econometrics (3rd edition)*. Addison Wesley Longman.

Jiji Zhang. 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896.

Kun Zhang and Aapo Hyvarinen. 2009. On the Identifiability of the Post-Nonlinear Causal Model. page 9.