Causal Discovery: The secret to more promising data mining leads?

Gabriel Ruiz

University of California, Los Angeles

December 10, 2020

Outline

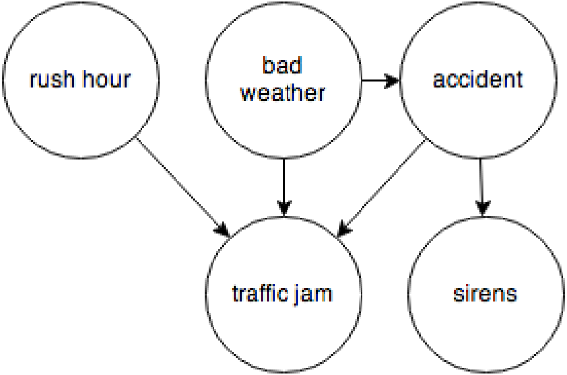
1. Background

2. Causal Discovery Algorithms
   - Constraint-Based Causal Discovery
   - Score-based methods

3. Uniquely Identifiable DAGs

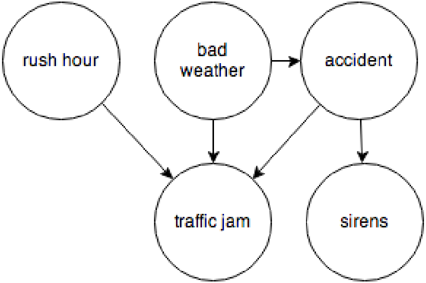
4. Conclusion

Bayesian Network/Causal Diagram



Bayesian Network/Causal Diagram

Motivation for Studying these Structures: With observational data alone, causal inference using an accurate DAG has been shown to provide results that are up to par with the quintessential randomized controlled experiment (**do-calculus**)[1].



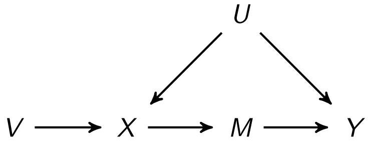
[1] Pearl, J., Glymour, M., and Jewell, N. P. *Causal Inference in Statistics, A Primer*. pgs. 118-124. 2016.

Intervention Distribution From Observational Distribution

Definition (Back-door Criterion)

A set of variabiles Z satisfies the back-door criterion relative to an ordered pair of variables (X, Y) in a DAG $\mathcal{G}$ if:

- no nodes in Z is a descendant of X, the intervention node.
- Z blocks every path between X and Y that contains an arrow in X (backdoor path).



- U satisfies the backdoor-criterion.

Identifying Causal Effects Using a DAG and Observational Distribution

Theorem (Back-door Adjustment)

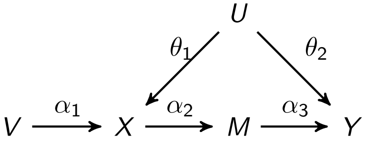
If Z satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is given by:

$$P(y|do(x)) = \sum_z P(y|x, z)P(z).$$

Do-Operator: Signifies an intervention on the node X, exogenous of the other nodes. $P(y|do(x))$ is the distribution of Y given this intervention.
Note: The right-hand side is in terms of the observational distribution.

Other Ways DAGs Help to Augment Causal Inference

Under a linear structural equation model assumption for our system of variables (e.g. $Y = \alpha_3 M + \theta_2 U + \epsilon_Y$), $\gamma_{X \to Y} = \frac{\partial}{\partial x}\mathbb{E}[y|do(x)] = \alpha_2\alpha_3$ is a natural estimand of interest.

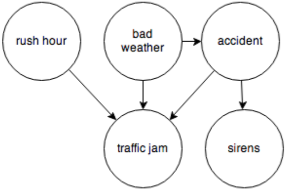


Below, let $\beta_C(A \sim B + C)$ denote the coefficient of variable C in the population-level linear least squares regression of A on B and C.

1. **Backdoor (Confounder) Adjustment:** $\gamma_{X \to Y} = \beta_X(Y \sim U + X)$.
2. **Instrumental Variable Analysis**: $\gamma_{X \to Y} = \frac{\beta_V(Y \sim V)}{\beta_V(X \sim V)} = \frac{\alpha_1\alpha_2\alpha_3}{\alpha_1}$.
3. **Mediation Analysis**: $\gamma_{X \to Y} = \beta_X(M \sim X) \times \beta_M(Y \sim X + M)$.

These all follow from d-separation queries under the assumption that the DAG is Markov to the underlying joint distribution $\mathbb{P}_{MUVXY}$.

Causal Discovery



Challenge: We do not always have the complete oracle-like knowledge about the graph structure, especially if many variables are involved.

Research Goal: Reconstructing a causal diagram with little to no prior knowledge for how things are related.

Observation	Rush Hour	Bad Weather	Accident	Traffic Jam	Sirens
1	Yes	No	No	Yes	No
2	No	Yes	Yes	Yes	Yes
⋮	⋮	⋮	⋮	⋮	⋮
n	No	No	No	No	No

A Data Mining Use Case

1. Non-experimental data collected on p variables $\mathrm{X} \in \mathbb{R}^p$.
2. We believe there is an underlying DAG $\mathcal{G}$ whose structure is fully or partially identifiable.
3. Before doing a potentially costly experiment, we want to estimate the causal effect of intervening on a set of variables $\mathcal{I} \subset \{1, 2, \ldots, p\}$.
4. We are interstested on its effect on a Response Set $\mathcal{R} \subset \{1, \ldots, p\} \backslash \mathcal{I}$, i.e.

$$P\left(X_{\mathcal{R}} | do(X_{\mathcal{I}})\right)$$

5. We may even wish to iterate our Inference Across different $\mathcal{I}$ or $\mathcal{R}$ using our estimated graphical model to see what experiments are “most promising.”

Related Other Work in this Setting

- Nandy et. al (2017)[2] extends the theory for the earlier method applied by Stekhoven et. al (2012)[3] to validate causal leads from an Arabidopsis Thaliana gene expression dataset.

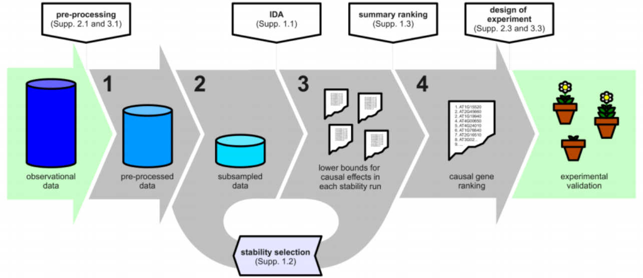


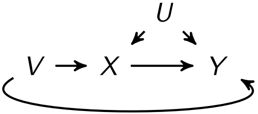
Figure: The causal discovery schema of Stekhoven et. al (2012).

[2] Preetam Nandy, Marloes H. Maathuis, and Thomas S. Richardson. “Estimating the effect of joint interventions from observational data in sparse high-dimensional settings.” *The Annals of Statistics*. 2017.

[3] Stekhoven, Daniel & Moraes, Izabel & Sveinbjörnsson, Gardar & Hennig, Lars & Maathuis, Marloes & Bühlmann, Peter. “Causal Stability Ranking”. *Bioinformatics*. 28. 2819-2823. 2012.

Structure Identifiability

Without extra assumptions or prior knowledge (e.g. temporal order, past experiments), a DAG is generally only identifiable up to its **Markov equivalence class**: all DAGs which have the same skeleton and the same v-structures.



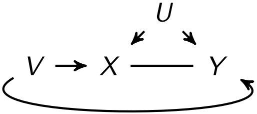
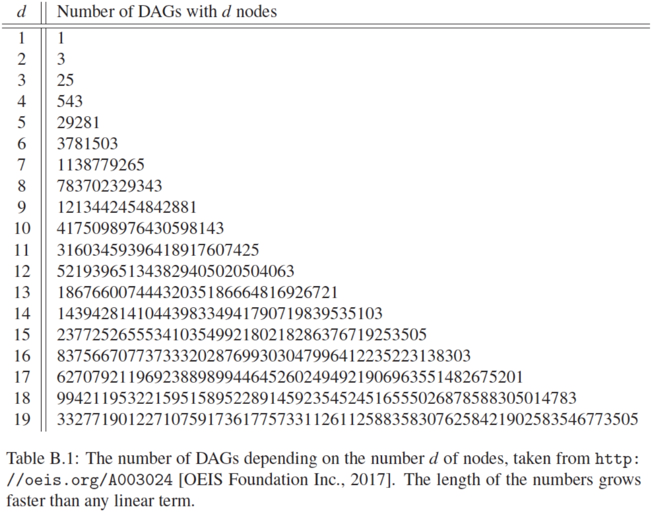


Figure: The original DAG (left) and its corresponding CPDAG (right), obtained by keeping the orientation of edges corresponding to the v-structures $V \to X \leftarrow U$ and $V \to Y \leftarrow U$, and removing the orientation from all other edges. Note the ambiguity about the causal direction $X \to Y$ vs. $X \leftarrow Y$ in the CPDAG.

For this system of variables, we may resolve the ambiguity on the causal direction $X \to Y$ vs. $X \leftarrow Y$ by conducting an experiment: if Y is associated with X when we intervene on X, then it must be that $X \to Y$. Or visa versa if we intervene on Y.

Structure Identifiability



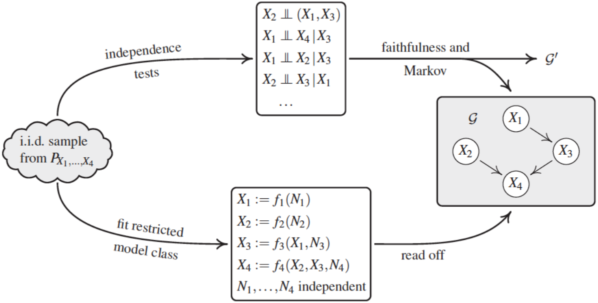
d	Number of DAGs with *d* nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103
15	237725265553410354992180218286376719253505
16	83756670773733320287699303047996412235223138303
17	62707921196923889899446452602494921906963551482675201
18	99421195322159515895228914592354524516555026878588305014783
19	332771901227107591736177573311261125883583076258421902583546773505

Table B.1: The number of DAGs depending on the number *d* of nodes, taken from `http://oeis.org/A003024` [OEIS Foundation Inc., 2017]. The length of the numbers grows faster than any linear term.

Structure Identification

Idenitifying DAGs (or CPDAGs) is challenging and an open area of inquiry.

- Main methods (3+ variables)[4]


i.i.d. sample from $P_{X_1,\ldots,X_4}$
independence tests
$X_2 \perp\!\!\!\perp (X_1,X_3)$
$X_1 \perp\!\!\!\perp X_4 \mid X_3$
$X_1 \perp\!\!\!\perp X_2 \mid X_3$
$X_2 \perp\!\!\!\perp X_3 \mid X_1$
…
faithfulness and Markov
$\mathcal{G}'$
$\mathcal{G}$
X_1
X_2
X_3
X_4
fit restricted model class
$X_1 := f_1(N_1)$
$X_2 := f_2(N_2)$
$X_3 := f_3(X_1,N_3)$
$X_4 := f_4(X_2,X_3,N_4)$
$N_1,\ldots,N_4$ independent
read off

- Special methods (2 variables)
  - Direction-learning methods

[4] Peters, Jonas, et al. Elements of Causal Inference Foundations and Learning Algorithms. MIT Press. 2017

Structure Learning By Conditional Independence Tests

Constraint based methods, e.g. the PC Algorithm[5,6]:

1. Find the Skeleton of $\mathcal{G}$ by CI tests: Check independence between every X and Y conditional on all $S \subseteq V \backslash \{X, Y\}$ of size at most k;
2. Identify v-structures: relations between triplets (A, B, C) such that $A \rightarrow B \leftarrow C$ and A, C not adjacent;
3. Orient other edges, for example by noting that the opposite orientation introduces additional v-structures which were not found in the previous step.
4. **Output:** CPDAG (or PDAG).

R Package: `pcalg`.

[5] Peter Spirtes and Clark Glymour. "An algorithm for fast recovery of sparse causal graphs." *Social Science Computer Review*, 9(1):62–72, 1991.

[6] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* Springer, 1993.

Can we optimize a score function?

Score-based methods:

$$\hat{\mathcal{G}} = \arg \max_{G \in \{acyclic\}} S(G, \mathrm{D}_n).$$

1. $\mathrm{D}_n = (x_{ij})_{n \times p}$ iid data from $(\mathcal{G}, \mathbb{P})$.
2. $S(G, \mathrm{D}_n)$ is a scoring function, e.g.:

$$S_{BIC}(G, \mathrm{D}_n) = \log \; p(\mathrm{D}_n | \hat{\theta}, G) - \frac{d}{2} \log n,$$

$\hat{\theta}$: MLE of parameters under G, $d =$ dimension of $\hat{\theta}$.

Can we optimize a score function?

Theorem (Chickering (2002))

Assume that $(\mathcal{G}, \mathbb{P})$ satisfies faithfulness. If the score function $S(G, \cdot)$ is consistent and score-equivalent, then:

$$\lim_{n \to \infty} \mathbb{P} \left\{ \arg\max_{G} S(G, \mathrm{D}_n) \in [\mathcal{G}] \right\} = 1,$$

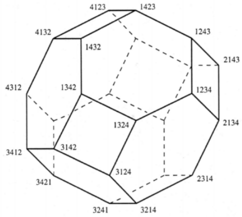
where $[\mathcal{G}] := \{G : G \sim \mathcal{G}\}$ is the MEC of the true $\mathcal{G}$.

Consistency roughly says that a DAG G gives the the optimal score with probability → 1 only if it is faithful to the underlying distribution and is sparsest among all faithful DAGs[7].

[7] David Chickering. "Optimal structure identification with greedy search." *Journal of Machine Learning Research*, 3:507–554, 01, 2002.

Notable Score-based methods

edges in polytope of permutations (i.e., permutohedron) connect neighboring transpositions, e.g. $(3,1,4,2) - (3,4,1,2)$



- Solus et. al (2018) search across a permutohedron and simply use the number of edges in the DAG as the score $S(G, \mathrm{D}_n)$[8].
- Ye et. al (2020) use a Gaussian regularized likelihood score and Simulated Annealing to search permutations of nodes[9].

[8] Solus, L., Wange, Y., Uhler, C. “Consistency Guarantees for Greedy Permutation-Based Causal Inference Algorithms.” arXiv:1702.03530, 2018.

[9] Ye, Q., Amini, A.A., and Zhou, Q. “Optimizing regularized Cholesky score for order-based learning of Bayesian networks.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

What if a CPDAG is not informative enough?

Identifiability Problem: Nandy et. al (2017) use the PC Algorithm (or similar) which outputs an equivalence class, so a given total causal effect query can be trivially lower bounded by zero.
Linear Non-Gaussian Acyclic Model (LiNGAM): Shimizu et. al (2011) propose an unconfounded linear Bayesian network with non-gaussian errors whose DAG structure (and causal ordering) is identifiable[10].

The LiNGAM Structural Causal Model:

$$X = B_{p\times p}^{\top} X_{p\times 1} + \epsilon \in \mathbb{R}^p;$$

B acyclic, $\epsilon_j \sim \mathbb{P}(\epsilon_j; \theta_j)$ independent non-Gauassian entries.
In terms of a Mixing Matrix: Let $M := (\mathbb{I}_p - B)^{-T}$

$$X = M\epsilon \implies X_k = \epsilon_k + \sum_{j \in AN(k)} M_{kj}\epsilon_j.$$

[10] Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K., "DirectLiNGAM: a direct method for learning a linear non-Gaussian structural equation model," *Journal of Machine Learning Research*, 12, 1225-1248. 2011.

Topological Ordering of a DAG

Goal: Do inference using the topological ordering.

Definition: Topological Ordering via a Permutation

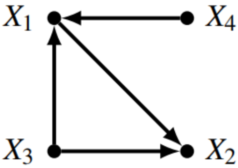
1. A bijective function (permutation)

$$\pi : \{1, 2, \ldots, p\} \mapsto \{\pi(1), \pi(2), \ldots, \pi(j), \ldots, \pi(p)\}$$

2. $\pi(1), \pi(2), \ldots, \pi(p)$ are the observed node labels.
3. Every parent node precedes its child in the ordering via π:

$$j \in PA_k \implies \pi^{-1}(j) < \pi^{-1}(k).$$

Topological Ordering of a DAG: An Example



$$X_3 = f_3(U_3)$$
$$X_4 = f_4(U_4)$$
$$X_1 = f_1(X_3, X_4, U_1)$$
$$X_2 = f_2(X_1, X_3, U_2)$$

A Possible Permutation:

- $\pi(1) = X_3$,
- $\pi(2) = X_4$,
- $\pi(3) = X_1$, and
- $\pi(4) = X_2$.

A Generic Algorithm

Let

- $\hat{\pi}$ be our estimate of a topological ordering for DAG $\mathcal{G}$.
- $\mathcal{A}_t = \{\hat{\pi}(1), \ldots, \hat{\pi}(t-1)\}$

Algorithm: Continuing an Estimate Ordering

1. $$\hat{\pi}(t) \leftarrow \arg\min_{k \notin \mathcal{A}_t} \mathcal{S}\left(k, \mathcal{A}_t; \mathrm{X}\right)$$

   for summary statistic $\mathcal{S}$ comparable between different nodes.
2. The continued Partial Ordering: $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_t \cup \{\hat{\pi}(t)\}$.

We apply this algorithm sequentially for $t = 1, 2, \ldots$ until completion.

Linear Non-Gaussian Acylcic Model

We have:

$$X = B_{p\times p}^{\top} X_{p\times 1} + \epsilon = M\epsilon \implies X_k = \epsilon_k + \sum_{j\in AN(k)} M_{kj}\epsilon_j.$$

Shimizu et. al (2011) show that node $j \notin \mathcal{A}_t$ is a valid node to append to $\mathcal{A}_t$ if and only if

$$X_k - \mathbb{E}[X_k | X_j, X_{\mathcal{A}_t}]$$

is independent of $X_j - \mathbb{E}[X_j | X_{\mathcal{A}_t}]$ for each $k \notin \mathcal{A}_{tj} \triangleq \mathcal{A}_t \cup \{j\}$.

Intuition for sufficiency: If $PA(j) \subseteq \mathcal{A}_t$, then $X_j - \mathbb{E}[X_j | X_{\mathcal{A}_t}] = \epsilon_j$. Also,

$$X_{\mathcal{A}_{tj}} = M_{\mathcal{A}_{tj},\mathcal{A}_{tj}} \epsilon_{\mathcal{A}_{tj}}.$$

So

$$X_k - \mathbb{E}[X_k | X_j, X_{\mathcal{A}_t}] = X_k - \mathbb{E}[X_k | \epsilon_{\mathcal{A}_{tj}}] = \sum_{l\notin \mathcal{A}_{tj}} M_{kl}\epsilon_l.$$

Conclusion

Takeaway: Causal discovery, a small initial step in the scientific pipeline, should be used with caution.

- The idea behind Causal Discovery is appealing, but it is difficult. Nonetheless, there has been progress.
- We have a trade-off between generality (e.g. equivalence class estimation which gives ambiguity of causal effects) and stronger assumptions which may or may not be realistic.
- What about unobserved confounders? What about non-iid data, such as data from a system of variables that varies in time?[11].

[11] Glymour, C., Zhang, K., & Spirtes, P. (2019). "Review of Causal Discovery Methods Based on Graphical Models." *Frontiers in genetics*, 10, 524. https://doi.org/10.3389/fgene.2019.00524

Acknowledgements

Qing, Oscar

NSF-GRFP DGE-1650604