Causal Analysis and Gene Expression Data

Silvan de Boer
silvandeboer@gmail.com

Related Work

Constraint-Based Causal Discovery

Constraint-based approaches derive constraints from the data, and use it to restrict the class of possible underlying causal graphs. Causal inference is treated as a constraint-satisfaction problem. Generally, conditional independence relations (CIR) are chosen as constraints, which can be used to derive the separation of variables and the orientation of edges in the graph. Some prominent methods are described below.

Inductive Causation (IC) was introduced by Verma and Pearl [1991] and generally describes how we can induce a PDAG from conditional independences in the data. The algorithm consists of two steps: inducing the skeleton and orienting the edges. For every pair of variables a and b, we check whether there is an edge in the PDAG. Using the faithfulness assumption, we add an edge if there is no separating set of variables S_{ab} that makes a and b conditionally independent ($\nexists S : a \perp\!\!\!\perp b \mid S_{ab}$). One easy step of edge orientation uses the separation criterion of colliders. If two non-adjacent variables a and b share a neighbour c that is not in S_{ab}, it must be a collider on the path and we induce edge orientation $a \rightarrow c \leftarrow b$. Application of additional orientation rules lead us to a maximally oriented PDAG, which describes the Markov equivalence class of all causal graphs that induce the joint data distribution. One early set of such rules was described by Spirtes et al. [2000] in the SGS algorithm, which was named after the authors.

IC is quite limited, because it relies on several assumptions about the underlying SCM (e.g. causal sufficiency), and its naive implementation is costly due to the search over all separating sets S_{ab}.

PC, named after its inventers Peter Spirtes and Clark Glymour [Spirtes and Glymour, 1991], reduces the cost of naive IC. A systematic algorithm finds the separating sets S_{ab} in polynomial time. Starting from a fully-connected graph, edges are systematically removed by considering separating sets of increasing cardinality, and only taking into account the variables that neighbour a and b. For example, first edges are removed between variable pairs that are independent given the empty set (cardinality 0). Then, edges are removed between remaining adjacent variable pairs that are independent given one of their neighbours. Already some possible separating sets can be skipped here, because edges were removed in the previous step. Therefore, as we consider larger possible separating sets, the number of neighbours to choose from decreases.

Fast Causal Inference (FCI) extends PC to allow for selection bias and latent confounding. Dropping the causal sufficiency assumption, means that it should consider some possible separating sets S_{ab} that contain variables not adjacent to a or b (specifically, their ancestors). FCI is a feasible algorithm for datasets with many variables when the underlying graph is sparse and bidirected edges are not too much chained together. It was first introduced by Spirtes et al. [1999], and gradually developed since then. A modern version named FCI+ by Claassen et al. [2013] is relatively fast. It finds the complete PAG in polynomial time.

Invariant Causal Prediction (ICP) does not infer a complete equivalence class of causal graphs from the data, but is concerned with local features of that graph. In its original formulation by Peters et al. [2016] it outputs a subset of parents of a target variable, and in a later adaptation by Mooij et al. [2016] the output is a subset of ancestors. In comparison to the previous methods, this approach uses data from different contexts.

When we model the context (e.g. intervention) as a variable that is not a cause of the target variable, then the conditional distribution of the target variable given its direct causes is invariant to the value of the context variable. This property is used by ICP to find parent or ancestor relations in data from different contexts.

In a naive implementation, we would have to investigate every possible parent set for every target variable, making the complexity exponential in the number of variables. Since (direct) causes tend to correlate with their effect, it is reasonable to only consider variables that have a relatively high correlation with the target variable. Such an approximation was used by Peters et al. [2016] and Meinshausen et al. [2016] to apply ICP to the dataset of Kemmeren et al. [2014] with over six thousand variables.

Local Causal Discovery (LCD) is a simple local method that looks at ancestral causal relations in the context of a changing environment. In the data, we marginalize over three variables (C, X, Y), one of which is exogenous (C). This means that we have the domain knowledge that this variable is not a direct or indirect effect of any endogenous variables. Next, we perform three statistical tests: $C \perp\!\!\!\perp Y \mid X$, $C \not\perp\!\!\!\perp X$ and $X \not\perp\!\!\!\perp Y$. If these tests are positive, [Cooper, 1997] proves that there is a relationship $X \to Y$.

This method allows for latent confounders, and assumes that there is no selection bias. Note that this method can find only a limited subset of ancestral relations. [Cooper, 1997] prove the theory by listing all possible causal graphs of three variables in which the context variable is not caused by endogenous variables. They note that of the 32 networks in which X causes Y, LCD is only able to identify this relation in 3 cases. The others are not identified, because there are other configurations that induce the same independence relations. For example, if C causes Y not only through X, but via another path as well, we will not find that X causes Y.

Y-structures are a pattern in a Partial Ancestral Graph (PAG), which has information about ancestral relations between four random variables. Mooij et al. [2015] showed that a set of independence tests can be used to infer some ancestral relations from observational data. This local method can be used to find an incomplete set of ancestral relations in the underlying SCM.

Specifically, we marginalize over four random variables X, Y, Z and U. Then we perform four statistical tests: $Z \perp\!\!\!\perp Y \mid X$, $Z \not\perp\!\!\!\perp Y$, $Z \not\perp\!\!\!\perp U \mid X$ and $Z \perp\!\!\!\perp U$. If all tests are positive, there are only two possible PAGs representing the marginalization over the four variables, which are called Y-structure (a term coined for a score-based method by Mani [2006]) and Extended Y-structure. In both PAGs, we can use the backdoor criterion to infer that $p(Y \mid \mathrm{do}(X = x)) = p(Y \mid X)$.

Mooij et al. [2015] investigate the performance effect of adding more independence tests. By adding two more tests, the Y-structure remains the only possible PAG. After that, more redundant independence tests can be added. Adding more tests necessarily reduces recall, but precision can improve if the extra tests eliminate false positives. Interestingly, in their experiments on synthetic data, maximum precision is not always achieved with the minimum or maximum number of tests.

An interesting insight from Mooij et al. [2015] is that the faithfulness assumption becomes more problematic as the number of random variables grows. The data is sampled from a SCM, which makes the faithfulness assumption very reasonable. However, it appears that the marginalized data can be almost faithless to the supposed PAG.

Score-Based Causal Discovery

Score-based approaches search for the causal graph that optimizes some loss function, often based on independences in the data. Some methods are discussed below.

Accounting for Strong Dependences (ASD) is a method that focusses on graphs satisfying dependences in the data, whereas most methods focus on the independences. The groundwork of the method was layed by Hyttinen et al. [2014]. Dependence and independence relations in the data are encoded as soft constraints in ASP (Answer Set Programming), together with rules that determine that the solution should be a valid graph. The solution is then found by an ASP solver, that minimizes the loss computed as the sum of the weights of the constraints that are not satisfied. This method can only be applied to problems with a small number of variables, because the number of dependence and independence relations quickly becomes very large.

Magliacane et al. [2016] compute weights based on the p-value of the conditional independence test and the significance level, such that strong dependences obtain a larger weight than independences. They also provide a method to compute a confidence score for single features of the graph (like an ancestral relation) by solving two optimization problems and subtracting the losses. In one problem, the presence of the feature is added as a hard constraint, in the other the absence is added as a hard constraint.

As part of their Joint Causal Inference (JCI) framework, Mooij et al. [2016] adapt the ASD method to include interventional data. They simply add some constraints to the ASP problem that encode their JCI assumptions for interventional data.

Greedy Equivalence Search (GES) was first introduced by Meek [1997] and further detailed by Chickering [2002]. A search for the Markov equivalence class (MEC) is performed in two phases. Our graph is initialized without edges. In the first phase, we move between MECs by adding edges. In each step, we consider the set of MECs that are one edge addition away from the current MEC. Reversing a covered edge does not change the MEC, so we need to consider every combination of covered edge reversals, followed by a single edge addition, followed by covered edge reversals. We score the MECs and move to the one with the highest score. Meek [1997] proposed the Bayesian scoring criterion to score DAGs in a MEC. Under some conditions this can be approximated by the Bayesian information criterion (BIC).

Once we reach a local maximum, Chickering [2002] shows that the MEC contains the generative distribution. The second phase is analogous to the first, but this time single edges are removed. Chickering [2002] shows that this final MEC is a perfect map of the generative distribution.

Hauser and Bühlmann [2012] generalize the method to datasets with interventional data (GIES). They introduce the interventional MEC as the object of the search. The algorithm is adapted by introducing the turning phase. Repeated application of the three phases leads to the interventional MEC underlying the generative distribution. They note that the space of interventional MECs is more fine-grained, meaning that this type of data leads to improved identifiability.

Hybrid methods

Sparsest Permutation (SP) is a method proposed by Raskutti and Uhler [2018] that searches for the Markov equivalence class in a space of topological orderings of random variables, using observational data. The topological ordering is a permutation of variables in which ancestors precede descendents. Given an ordering and a method to infer dependence and independence relations, we can infer a DAG that satisfies causal minimality. The SP algorithm searches over these DAGs infered from all possible permutations and selects the DAG with the smallest number of edges. Raskutti and Uhler [2018] show that this algorithm is consistent, which means that it finds a DAG in the correct Markov equivalence class. This consistency relies on an assumption that is weaker than faithfulness.

Solus et al. [2017] propose a greedy algorithm (GSP) that increases the number of variables that can be handled from about 10 to hundreds. The only sacrifice is that a somewhat stronger assumption is required, which is still weaker than faithfulness. They introduce the DAG associahedron, a sub-polytope of the permutohedron which indicates possible directions for the depth-first search. These directions are based on the concept of covered edge.

Wang et al. [2017] adapt the algorithm further to take interventional data into account (IGSP). The score is now summed over a score for each intervention distribution. This score is a weighted sum of the number of edges in the inferred graph, excluding edges into the intervened variable, and some maximum likelihood score of the interventional data given some model assumptions.

Other

other statistical patterns in the joint distribution can be exploited too (e.g Mooij et al., 2016; Peters et al., 2017)

References

D. M. Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

T. Claassen, J. Mooij, and T. Heskes. Learning sparse causal models is not np-hard. *arXiv preprint arXiv:1309.6824*, 2013.

G. F. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1(2):203–224, 1997.

A. Hauser and P. Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(Aug):2409–2464, 2012.

A. Hyttinen, F. Eberhardt, and M. Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, pages 340–349, 2014.

P. Kemmeren, K. Sameith, L. A. van de Pasch, J. J. Benschop, T. L. Lenstra, T. Margaritis, E. O'Duibhir, E. Apweiler, S. van Wageningen, C. W. Ko, et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157(3):740–752, 2014.

S. Magliacane, T. Claassen, and J. M. Mooij. Ancestral causal inference. In *Advances in Neural Information Processing Systems*, pages 4466–4474, 2016.

S. Mani. *A bayesian local causal discovery framework*. PhD thesis, University of Pittsburgh, 2006.

C. Meek. *Graphical Models: Selecting causal and statistical models*. PhD thesis, PhD thesis, Carnegie Mellon University, 1997.

N. Meinshausen, A. Hauser, J. M. Mooij, J. Peters, P. Versteeg, and P. Bühlmann. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27):7361–7368, 2016.

J. M. Mooij, J. Cremers, et al. An empirical study of one of the simplest causal prediction algorithms. In *UAI 2015 Workshop on Advances in Causal Inference*, number 1504, pages 30–39, 2015.

J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *arXiv preprint arXiv:1611.10351*, 2016.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5): 947–1012, 2016.

G. Raskutti and C. Uhler. Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7(1):e183, 2018.

L. Solus, Y. Wang, L. Matejovicova, and C. Uhler. Consistency guarantees for permutation-based causal inference algorithms. *arXiv preprint arXiv:1702.03530*, 2017.

P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

P. Spirtes, C. Meek, and T. Richardson. *An algorithm for causal inference in the presence of latent variables and selection bias*, volume 1. MIT Press, 1999.

P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson. *Causation, prediction, and search*. MIT press, 2000.

T. Verma and J. Pearl. *Equivalence and synthesis of causal models*. UCLA, Computer Science Department, 1991.

Y. Wang, L. Solus, K. Yang, and C. Uhler. Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems*, pages 5822–5831, 2017.