Causal Analysis and Gene Expression Data

Silvan de Boer
silvandeboer@gmail.com

Abstract

Abstract

Introduction

Background

SCMs Cycles, latent confounding, selection bias, interventions, constraint VS score-based, faithfulness, causal sufficiency (Markov properties), graph types,

Related Work

Data types: observational, interventional, fixed variables

Factors

- Confounding
- Mechanism function
- Cycles
- Intervention: perfect, stochastic, mechanism change, activity intervention, other
- Known intervention targets
- Taks/Output: global causal discovery (MAG/MG?, strength of links?), ancestral relations, predict expression after intervention, ...
- Fixed variables
- Computation time

Papers

Yang et al. [2018] Generalize Hauser and Bühlmann [2012] from perfect to general interventions. Define interventional Markov equivalence class, which is identified from general intervention experiments. First provably consistent algotithm for learning DAGs. Simulated and biological datasets, with observational and intervention data.

Mooij and Heskes [2013] Observational and interventional equilibrium data (cytometry Sachs et al. [2005]). Deals with feedback loops and continuous data. Model activity (how the equilibrium distributions of direct effects is influenced) instead of abundance of compounds, so standard intervention formalism of Pearl [2009] is not applicable. Nonlinear mechanisms are approximated by coupled local linearizations.

Hauser and Bühlmann [2012] Every interventional Markov equivalence class can be represented by an interventional essential graph (like CPDAG in obs. case). Generalize GES to the Greedy Interventional Equivalence Search algorithm (GIES). Tested in simulation study.

Tian and Pearl [2001] Possibly infer some environmental change?

Forré and Mooij [2018] Introduces modelling framework of modular SCMs (mSCM) and σ-connection graphs, extending σ-separation to work on them. ASD (Accounting for Strong Dependences; score-based) method tested on synthetic data and solved with an ASP (Answer Set Programming) solver. Very computationally expensive.

Versteeg and Mooij [2019] Estimates of LCD and ICP following JCI framework.

Datasets

Kemmeren et al. [2014] Gene expression dataset.

Sachs et al. [2005] Flow cytometry in human immune system cells; signalling network; compare to consensus network

Methods General description and details about different implementations in papers (OG and SotA and more).

- Constraint-based: using the faithfulness assumption, restrict the class of possible causal graphs with independence tests.
  - 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 ($\nexists S : a \perp\!\!\!\perp b \,|\, S_{ab}$). 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, thus dropping the causal sufficiency assumption. It 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 complete and relatively fast.
- LCD: e.g. Trigger[Chen et al., 2007]. Local strategy.
- Y-Structures: local strategy. Mooij et al. [2015] are OG?
- ICP, look at Meinshausen et al. [2016]

- Score-based: search for the causal graph that optimizes some loss function based on independences. (e.g., Cooper and Herskovits, 1992;

Heckerman et al., 1995; Chickering, 2002; Koivisto and Sood, 2004) (see: Peters 7.2)

  - ASD

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

Methods

Experiments

Data

A biological dataset is used to evaluate the proposed methods. The DNA of a cell contains genes that are involved in many of the cell's functions. They are often responsible for the production of a protein. The first step in this process is to copy its information to a Messenger RNA (mRNA) strand. To measure how active a specific gene is, we can measure how much of its mRNA we find in the cell.

Genes interact to fulfill a plethora of cell functions. For example, the expression of one gene might up- or down-regulate the expression of some other gene. This interaction is regulated by some biochemical process.

For a variety of reasons, it is interesting to know how genes interact precisely, that is: what the regulatory network looks like. By jointly measuring the expression of a large set of mRNA strands we obtain an mRNA profile. Collecting a set of these profiles allows us to model the joint distribution of mRNA expression and the causal relations.

Specifically, we use mRNA profiles from Kemmeren et al.. They measured a profile of 6.182 genes in cells of the yeast species *Saccharomyces cerevisiae* (baker's yeast). The dataset consists of 262 observational samples obtained from unaltered wild type cells, and 1.484 interventional samples obtained from mutant cells where one gene was deactivated.

Both the observational and interventional profiles are reported relatively to some average wild type profile. Kemmeren et al. report that the interventional profiles are compared to a set of 428 wild type profiles. [It is unclear if the observational profiles are compared to the same set.]

There are some details of the experiments that might be of relevance in a discussion of underlying assumptions. First of all, the researchers chose to measure only a subset of about 25% of all genes. Selection criteria included whether genes were expected to be involved in regulating other genes, and only genes were selected that do not play a vital role in keeping the cell alive (viability).

Furthermore, the profile resulting from an experiment had to pass a quality control before being admitted to the dataset. Failing this test resulted

either in repeating the experiment, or excluding the mutant. Although these checks improve the quality of the data by removing some failed experiments, they might also admit some selection bias.

A final factor to consider is that data from previous work of the same institute is included in the dataset, specifically from Lenstra et al. and Van Wageningen et al.. The authors note that they could not find any significant differences in the data. Nevertheless, this information can be seen as a context variable, and ignoring it can be an explicit modelling assumption.

Ground truth

Value and slope, absolute and 'normalized'

Train-test splits 5-fold cross-validation; separate test set?

Simple analysis

Visualisation of gene VS gene Cycles and latent confounders

Results

Analysis

Conclusion

References

L. S. Chen, F. Emmert-Streib, and J. D. Storey. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome biology*, 8(10):R219, 2007.

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

P. Forré and J. M. Mooij. Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. *arXiv preprint arXiv:1807.03024*, 2018.

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.

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.

T. L. Lenstra, J. J. Benschop, T. Kim, J. M. Schulze, N. A. Brabers, T. Margaritis, L. A. van de Pasch, S. A. van Heesch, M. O. Brok, M. J. G. Koerkamp, et al. The specificity and topology of chromatin interaction pathways in yeast. *Molecular cell*, 42(4):536–549, 2011.

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. Mooij and T. Heskes. Cyclic causal discovery from continuous equilibrium data. *arXiv preprint arXiv:1309.6849*, 2013.

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. Pearl. *Causality*. Cambridge university press, 2009.

K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

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.

J. Tian and J. Pearl. Causal discovery from changes. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 512–521. Morgan Kaufmann Publishers Inc., 2001.

S. Van Wageningen, P. Kemmeren, P. Lijnzaad, T. Margaritis, J. J. Benschop, I. J. de Castro, D. Van Leenen, M. J. G. Koerkamp, C. W. Ko, A. J. Miles, et al. Functional overlap and regulatory links shape genetic interactions between signaling pathways. *Cell*, 143(6):991–1004, 2010.

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

P. Versteeg and J. M. Mooij. Boosting local causal discovery in high-dimensional expression data. *arXiv preprint arXiv:1910.02505*, 2019.

K. D. Yang, A. Katcoff, and C. Uhler. Characterizing and learning equivalence classes of causal dags under interventions. *arXiv preprint arXiv:1802.06310*, 2018.