UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR
THEORETISCHE INFORMATIK

Algorithmic Causality AG

Maciej Liśkiewicz
Katharina Dannenberg
Benito van der Zander
Marcel Wienöbst

List of Bachelor and Master Thesis Proposals

Algorithmic Challenges in Causal Inference

9.04.2020

Background and Projects Goals

Discovering and understanding causal relationships is an important topic of empirical sciences. Analyzing causes of diseases, economic crises and other complex phenomena is of great social and economic importance. However, for ethical or economic reasons questions as "Does smoking cause lung cancer?" or "What are the major causes of economic crises?" can be difficult, and in many cases impossible, to examine through direct experimentation. On the other hand, there are often large amounts of observed data available that can provide relevant information about these issues.

Causal inference has recently become a quickly growing field of Artificial Intelligence and Machine Learning, with the goal to explore from observed data and phenomena, the causal relationships between different objects and actions like e.g. between medical treatment and recovery. A key turning point in causal theory was marked by development of *graphical causal models* which allow intuitive, mathematically sound modelling of causal relationships. A most basic model is represented by a directed acyclic graph (DAG) whose vertices represent random variables of interest and whose edges express direct causal effects of one variable on another. The next very important milestone in the theory was Judea Pearl's[1] invention of the concept of *do-operator* [4] to analyze the effect of interventions on a system of causally related variables.

Consequently, the causal theory allows modelling of direct experimentation and inference of causal effects, however, only on observed and/or interventional data with existing knowledge. This approach is gaining increasing attention in Epidemiology, Sociology, and other empirical disciplines. In Artificial Intelligence and Machine Learning, causal inference leads, e.g., to an improved performance of the learning procedures and selecting optimal strategies.

The main goal of these projects is to cope with algorithmic issues in causality which are essential for the development of scalable methods. The proposed specific topics concern two general problems of causality:

- *Learning Graphical Causal Models*
  The basic task here is, for given observational data, to find the DAG representing the causal structure or, if this is not possible, to give a class of DAGs to which the true DAG belongs.
- *Inferring Causal Effects from Causal Structures (and Data)*
  Given causal structure (e.g., as a DAG) and observational data, compute the causal effect of the exposure variable, e.g., medical treatment, on the outcome variables, e.g., recovery of a patient.

[1] In 2011 Judea Pearl won the ACM A.M. Turing Award – regarded as the "Nobel Prize in Computing" – for his groundbreaking work in the field of Bayesian networks, which greatly advanced both Artificial Intelligence and Causality.

Learning Graphical Causal Models

1. **Improving the Learning from Low-Order Conditional Independencies.**
   One of the common obstacles for learning graphical causal models from data is that high-order conditional independence (CI) relationships between random variables are difficult to estimate. In [11] we present an algorithm, named LOCI, which computes a causal structure from a given set of CIs of order less or equal to k, where k is a small fixed number. A drawback of this algorithm is that it is not scalable: while it is quite effective for $k = 0, 1$, it requires a huge amount of input data in cases $k \geq 2$.

   The goal of this project is to analyse and to evaluate more efficient implementations of the LOCI algorithm which approximate the exact representation. A natural approach will be to use the well-known PC algorithm in the first step of LOCI. Interesting issues, which could be evaluated experimentally (and theoretically), include:

   - The properties of the resulting structure?
   - Is the resulting graph a CPDAG – a graphical representation of the equivalence class of DAGs? If not, can the algorithm be adapted to produce a CPDAG?
   - What is the Hamming distance to the true representation (true CPDAG).
   - Is the resulting graph an extendable graph? Can the algorithm be adapted to produce an extendable graph?

   Supervisor: Prof. M. Liśkiewicz, M.Sc. M. Wienöbst

2. **Learning from Low-Order Conditional Independencies in the Sample Setting.**
   While Project 1 concerns learning in the "oracle setting", i.e., such that the conditional independence (CI) queries are assumed to provide correct answers, the task of this project is to analyse the LOCI algorithm [11] and its potential improvements using statistical tests to estimate CIs.

   The study should provide the evidence to what extent the LOCI algorithm can be useful in practice. To this aim, different implementations of the algorithm providing approximate structures should be considered. One can start with a modification which uses the Conservative PC algorithm in the initial step. The next step could be to use specific procedures for independence tests for non-parametric models (e.g. Kernel-based CI test). Then the task is to compare these modifications with variants of the classical PC algorithm on different families of random DAGs.

   Supervisor: Prof. M. Liśkiewicz, M.Sc. M. Wienöbst

3. **Counting Problems Concerning k-Markov Equivalence Classes.**
   A CPDAG is a compact and elegant representation of all DAGs over the same set of nodes which are Markov equivalent (two equivalent DAGs encode the same causal information). One of the crucial problems concerning CPDAGs is to count the number of Markov equivalent DAGs encoded by a given CPDAG. An important application of this problem is to sample efficiently a DAG from the class of Markov equivalent structures. Recently [5, 6, 1] have provided some counting and sampling algorithms for this general setting.

   The goal of this project is to cope with the counting problems for the CPDAGs which represent all k-Markov Equivalence Classes (k-MEC) – a k-MEC contains all DAGs which entail the same independencies up to order k. The problems of interest here are the following ones:

   - How efficiently can such a k-MEC be enumerated given its representation?
   - Counting the size of a k-MEC given its representation. Starting point could be a reduction to the problem of counting the size of a single MEC which has attracted attention recently [6, 1].

- Counting the number of k-MECs with n nodes. Starting point could be experimental analysis for small k and n. Theoretical analysis could follow (note that in case of $k \geq n-2$ we have the likewise open question of counting the number of CPDAGs with n nodes (see Radhakrishnan et al., [5])).

Supervisor: Prof. M. Liśkiewicz, M.Sc. M. Wienöbst

4. **Evaluation of the Sample Variant of the flowPC Algorithm.**
The PC algorithm belongs to the most prominent and popular methods for learning causal structures from observational data. It is well-known that in the "oracle setting" – i.e., assuming that the conditional independences (CIs) are estimated perfectly – PC computes causal structures correctly. However in practice, i.e. when the CIs are estimated from sample data by statistical tests, the resulting structure may deviate from the correct one. The main obstacle here is that higher-order CI testing, i.e. in case of large conditioning sets, remain still a difficult and challenging task. In [10] we propose a modification of the PC algorithm, called flowPC, that reduces the order of CI tests. The experimental evaluation of the algorithm has been done in the "oracle setting".

   The goal of this project is to evaluate flowPC using statistical tests for estimating CIs on real data (particularly in Gaussian and non-parametric setting). An interesting issue is to compare the flowPC with the classical variants of the PC-algorithm on different families of random DAGs. If the experiments confirm the outperformance of flowPC compared to the classical variants of PC, the results of this project can be presented at a high level conference in Machine Learning.

   Motivating questions here are the following ones: are there settings in which flowPC is (significantly) better than PC? Is there a way to implement flowPC in a more robust way?

   Supervisor: Prof. M. Liśkiewicz, M.Sc. M. Wienöbst

5. **Properties of Incompatible Nodes in Directed Acyclic Graphs.**
The notion of *incompatible nodes* in a causal graphical structure plays a crucial role in the flowPC algorithm proposed in [10]. Such kind of nodes allow to reduce the order of conditional independence tests in the popular PC algorithm significantly and consequently to improve the performance of the algorithm.

   The aim of this project is to analyse the properties of incompatible nodes in directed acyclic graphs (DAGs) both experimentally and theoretically. In particular, the following tasks should be investigated:

   - What is the expected number of incompatible nodes for $l = 0, 1, \ldots$ (i.e. for conditioning sets of size $0, 1, \ldots$) in the Erdös-Renyi model. In particular, the simplest case of $l = 0$ is of high interest. The study (both theoretical and experimental) of asymptotic behaviour for $n \to \infty$ is expected.
   - Similar investigations for other graph classes and, in particular, real world graphs (alarm network, etc.).

   Supervisor: Prof. M. Liśkiewicz, M.Sc. M. Wienöbst

6. **Dynamic or Soft Thresholding in the PC Algorithm.**
Like the previous projects, this one also concerns the PC algorithm – one of the most popular methods for learning causal structures from observational data. When working with data, the conditional independences, on the basis of which the PC algorithm constructs causal structures, are assessed via statistical tests on the data. These statistical tests use a level of significance α. Since many co-dependent statistical tests are executed, the parameter α can not be interpreted as an overall level of significance, but rather as some sort of tuning parameter.

Since the reliability of the conditional independence tests decreases with the progress of the PC algorithm, that is, when they are not of low order anymore, it makes sense to adapt the level of significance accordingly. The aim of this project is to investigate experimentally the benefits of different strategies that use a dynamic value of α.

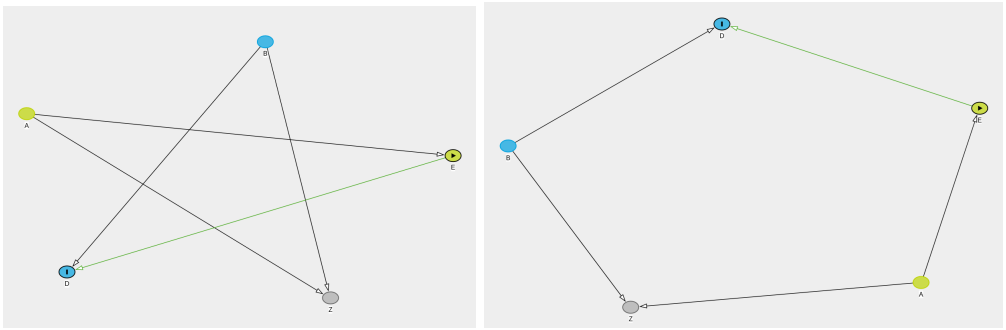
Supervisor: Prof. M. Liśkiewicz, M.Sc. K. Dannenberg

7. **Heuristics for the Evaluation of v-structures found by the PC Algorithm.**
This project, as the previous one, concerns the sample variant of the PC algorithm. After completing the skeleton of the graph, that is, finding out which nodes should be connected in some way, the PC Algorithm starts to direct edges. What is crucial for this, is the detection of so called v-structures, simply said, pairs of edges directed to their common node. These are the basis for the direction of other edges, and it is therefore important to make sure they are identified correctly. However, conflicting v-structures indicate that errors must have occurred in finding the v-structures. The goal of this project is to find heuristic measures to assess the confidence in each v-structure, and based on that, methods to delete weak v-structures and thus also solve conflicts.

   Supervisor: Prof. M. Liśkiewicz, M.Sc. K. Dannenberg

8. **Visualizing causal graphs.**

   After a causal graph has been learned by an algorithm, it is important to present the resulting graph in a format that can be understood intuitively by humans, so experts can interpret and verify the graphical model. DAGitty (`http://dagitty.net`) can automatically arrange the nodes of a graph to display the graph on its web interface, but it chooses the layout almost randomly:



   The goal of this project is to implement various layouting strategies in DAGitty, perhaps similarly to the layouts available in the tikz graph library. Furthermore the layouts should be customizable, for example to reduce the number of overlapping edges, group so-called exposure nodes and/or outcome nodes together, etc.

   Supervisor: Prof. M. Liśkiewicz, M.Sc. B. van der Zander

Inferring Causal Effects from Causal Structures

9. **A Comparative Study of Efficient Methods for Inferring Causal Effects from Causal Structures**
Based on, so-called, instrumental variables (IVs), a bunch of methods has been developed to identify causal effects in linear graphical causal models. However, the most powerful such methods rely on exponential-time procedures. Recently, several efficient identification algorithms have been given (see e.g. [9, 7, 2]). An important issue is that the proposed methods are not complete in the sense, that there exist instances of causal graphs which are identifiable, but the methods fail to identify causal effects for such instances. Also the power of the methods is incomparable: there exist instances which are identifiable by one method but not by the another and vice versa.

The goal of this project is to study systematically (both experimentally and theoretically) the power of the selected identification methods. An experimental analysis can be conducted to investigate the power of the identifiability, particularly of the algorithms proposed in [9] and in [2]. In experiments, different families of random causal linear structures can be used.

Supervisor: Prof. M. Liśkiewicz, M.Sc. B. van der Zander

10. **Improving the runtime of finding minimum d-separating sets and adjustment sets.**

    One of the most common methods of identifying causal effects is to adjust a set of confounding variables to block biasing paths. The relevant set of confounders can be found by searching for a separating set, a set that blocks all paths in the proper back-door graph. A set of confounders that has minimal or minimum size can be found by transforming the back-door graph to an undirected graph and using standard methods on the undirected graph [8]. Unfortunately the undirected graph created by the transformation is larger than the original causal graph, inflating a graph of n nodes and m edges to a graph of n nodes and n^2 edges, so algorithms on the undirected graph are slower than usual. We have shown that the transformation of the causal graph to the undirected graph is not necessary, when searching for a minimal set, because it can be searched directly in the causal graph; and thereby we have improved the runtime of finding a minimal set from $O(n^2)$ to $O(m)$.

    Similarly, finding a minimum set should also be possible in the causal graph itself, without the transformation to an undirected graph. A minimum set can be found in a graph of n nodes and m edges in time $O(nm)$ [3] using a max-flow algorithm, while applying that algorithm to the transformed, undirected graph takes of course $O(n^3)$ time. In this project it should be investigated if the max-flow algorithm can be applied directly to the causal graph in time $O(nm)$. This probably requires a straightforward modification of the max-flow algorithm to use paths as paths are defined in causal graphs rather than paths as they are defined in standard graphs, but first you need to understand the max-flow algorithm perfectly.

    The algorithm should be implemented in DAGitty (`http://dagitty.net`), preferably for different classes of causal graphs, DAGs, MAGs, CPDAGs, and RCGs. Finally, the performance of the implementation can be benchmarked.

    Supervisor: Prof. M. Liśkiewicz, M.Sc. B. van der Zander

11. **Equivalence of causal, probabilistic formulas.**

    Causal effects and other probabilities can be combined in probabilistic formulas to calculate arbitrary functions. For example, $\sum_x P(x) + \sum_y P(y) = 2$. Although these formulas can be evaluated numerically rather straightforwardly, when a certain graph and probability distribution are given, it is difficult to calculate them for unknown probability distributions.

    The goal of this project is to implement an algorithm to test if two formulas are equivalent. First the formulas need to be normalized by expanding sums, by removing subtractions and divisions, and by choosing a canonical form for probabilistic terms. Once the formulas contain only additions, multiplications and canonical probabilistic terms, the terms can be replaced by variables, which makes the equation equivalent a polynomial equation. This equation can be verified using the standard method of a randomized polynomial identity test (PIT).

    The algorithm should be integrated in DAGitty (`http://dagitty.net`), such that the formulas can be entered in a user-friendly format. Finally the performance of the algorithm should be evaluated on different causal graphs and formulas, in order to see if it is efficient enough to have any real-world use.

    Supervisor: Prof. M. Liśkiewicz, M.Sc. B. van der Zander

References

[1] Robert Ganian, Thekla Hamm, and Topi Talvitie. An efficient algorithm for counting markov equivalent DAGs. In *34th AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[2] Daniel Kumor, Bryant Chen, and Elias Bareinboim. Efficient identification in linear structural causal models with instrumental cutsets. In *Advances in Neural Information Processing Systems*, pages 12477–12486, 2019.

[3] James B Orlin. Max flows in o (nm) time, or better. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 765–774, 2013.

[4] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

[5] Adityanarayanan Radhakrishnan, Liam Solus, and Caroline Uhler. Counting markov equivalence classes for DAG models on trees. *Discrete Applied Mathematics*, 244:170–185, 2018.

[6] Topi Talvitie and Mikko Koivisto. Counting and sampling markov equivalent directed acyclic graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7984–7991, 2019.

[7] Benito van der Zander and Maciej Liśkiewicz. On searching for generalized instrumental variables. In *AISTATS*, pages 1214–1222, 2016.

[8] Benito van der Zander, Maciej Liskiewicz, and Johannes Textor. Constructing separators and adjustment sets in ancestral graphs. In *CI@ UAI*, pages 11–24, 2014.

[9] Benito van der Zander, Johannes Textor, and Maciej Liśkiewicz. Efficiently finding conditional instruments for causal inference. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[10] Marcel Wienöbst and Maciej Liśkiewicz. An approach to reduce the order of conditional independence tests in the PC algorithm. In *Submitted*, 2020.

[11] Marcel Wienöbst and Maciej Liśkiewicz. Recovering causal structures from low-order conditional independencies. In *34th AAAI Conference on Artificial Intelligence (AAAI)*, 2020.