SYSTEMATIC REVIEW OF CAUSAL DISCOVERY USING CONTINUOUS OPTIMIZATION. REFORMULATING CAUSAL DISCOVERY AS A GRAPH MATCHING PROBLEM

Ali Izadi
University of Tehran
Tehran, Iran
aliizadi.ai@ut.ac.ir

ABSTRACT

Causal inference has an important role in different areas including machine learning, providing new modelings that can answer new prediction tasks. In order to do causal inference, we need to discover causal relationships between variables. There are different approaches for causal discovery from observational data including score-based methods. The primary objective of this systematic review is to identify different score-based methods that formulate the problem as a continuous optimization one. The results show that there are different challenges in causal discovery with continuous optimization including modifications applied to the objective function of causal discovery, the causality constraint of the optimization problem, and the optimization algorithm. Also, we reformulate the problem of causal discovery as a graph matching problem which is a concave quadratic program, introducing a new direction for researching on causal discovery by continuous optimization.

***K*eywords** Causal Discovery · Structure learning · Bayesian networks · Directed acyclic graph · Continuous optimization

1 Introduction

Causal discovery has practical applications in many areas such as genetics [1], biology [2], and economics[3]. The gold standard method for causal discovery is to conduct controlled experiments, which is hard or non-ethical in some cases. Therefore causal discovery from observational data has attracted much research attention in the past decades. Bayesian networks (BN) are interpretable models for finding conditional independencies but they might not recover the true unique causal graphical models. Causal graphical models (CGM) are BNs that have the true causal direction between variables, not just conditional independencies. CGMs can be used to answer interventional queries like What will happen if external one force or intervene on the variable X? An interventional example could be "how does the probability of heart failure change if we convince a patient to exercise regularly?". CGMs also answer more complex queries that are named counterfactual queries. A counterfactual one would be "would a given patient have suffered heart failure if they had started exercising a year earlier?"[4].

In general, recovering the true causal graph $\mathcal{G}$ from purely observational samples from P_X is just possible when relying on a set of assumptions. This is called identifiability of graph $\mathcal{G}$. Also relying on a set of assumptions such as faithfulness, One can identify the Markov equivalence class of causal graphs. A Markov equivalence class is a set of DAGs which encode the same set of conditional independencies. The resulting graph is completed partially directed acyclic graph (CPDAG).

There are two approaches for causal discovery including constraint and score-based. Constraint-based methods use conditional independence tests in the joint distribution and so these methods output CPDAGs or a graph which belongs to a Markov equivalence class. The problem with these methods is that independence tests need large sample sizes to be reliable. The most well-known methods are Spirtes-Glymour-Scheines (SGS), Peter-Clark (PC) and Fast Causal Inference (FCI)[5].

In contrast, score-based approaches test the validity of a candidate graph G according to some scoring function $\mathcal{S}$

$$\hat{\mathcal{G}} = argmax_{\mathcal{G}\ over\ X} S(\mathcal{D}, \mathcal{G})$$

where $\mathcal{D}$ represents the observational samples for variables. The most well-known function $\mathcal{S}$ is bayesian information criteria (BIC).

Score-based methods search in the space of dags heuristically and graphs candidates compared with the best graph obtained so far using scoring function. The most well known method is Greedy Equivalence Search (GES)[5]. The problem with these methods is that the number of possible DAGs increases super-exponentially with the number of variables and the problem is NP-hard. Recently, Zheng proposed DAGs with NO TEARS: Continuous Optimization for Structure Learning[6] and converted the combinatoric graph-search problem into a continuous optimization problem and proposed a smooth and differentiable acyclicity constraint that further can be trained using different optimization algorithms including gradient based optimization approaches.

The primary objective of this systematic review is to identify different continuous optimization methods for causal discovery and improvements methods of NOTEARS article. We also reformulate the problem of causal discovery as a graph matching problem.

Contributions:

- Systematic review of causal discovery methods based on continuous optimization.
- Reformulating the problem of causal discovery as a graph matching problem which is a concave quadratic program.

2 Background

2.1 Structure Equation Model

The following articles use Structure Equation Model (SEM) as a model for the data generating procedure. Each variable x_i is associated with its parent in a DAG $\mathcal{G}$ with a function of f_i parameterized by $x_{pa(i)}$ and noise variable n_i as

$$x_i := f_i(x_{pa(i)}, n_i) \quad \text{where } x_{pa(i)} \text{ independent of } n_i$$

In general, SEM is not identifiable except in some cases including linear additive noise model with non-gaussian noise, non-linear additive noise model, and post non-linear models.

2.2 DAGs with NO TEARS: Continuous Optimization for Structure Learning

The NOTEARS method changed previous heuristic search-based methods of equation 1 to continuous optimization one of equation 2,

$$\begin{aligned} \min_{A \in \mathbb{R}^{d \times d}} \quad & S(A) \\ \text{s.t.} \quad & G(A) \in DAGs \end{aligned} \tag{1}$$

$$\begin{aligned} \min_{A \in \mathbb{R}^{d \times d}} \quad & S(A) \\ \text{s.t.} \quad & h(A) = 0 \end{aligned} \tag{2}$$

where the constraint $h(A) = 0$ enforces acyclicity and A is the learnable adjacency matrix of the output DAG. In NOTEARS paper $h(A)$ defined as equation 3

$$h(A) \ = \ tr(e^{A \odot A}) - d = 0 \tag{3}$$

Also, NOTEARS used augmented lagrangian optimization algorithm to solve the optimization problem.

2.3 Challenges

The are multiple challenges with the proposed method including linearity of SEM, difficulties of optimization due to the hard acyclicity constraint, and use of least square objective which does not directly maximize the data likelihood.

Following the NOTEARS method, The primary objective of this review is to identify what improvements and modifications applied to the objective function S, the acyclicity constraint and the optimization algorithm.

2.4 Graph matching

Graph matching is a problem to find similarity between two graphs. For two graphs G, H with the same number of vertices, the graph matching problem tries to find an alignment between two graphs that shows a correspondence between vertices of G and H in some optimal way [7]. To optimize graph matching problem we minimize the loss 4:

$$\begin{aligned} ||A_G - A_{P(H)}|| &= ||A_G - PA_HP^T|| \\ &= tr((A_G - PA_HP^T)^T(A_G - PA_HP^T)) \\ &= tr(A_G^TA_G) + tr(A_H^TA_H) - tr(A_G^TPA_HP^T) - tr(PA_H^TP^TA_G) \\ &= tr(A_G^TA_G) + tr(A_H^TA_H) - 2 \times tr(A_G^TPA_HP^T) \\ &= const - 2 \times tr(A_G^TPA_HP^T) \\ &= const - 2 \times tr(P^TA_G^TPA_H) \end{aligned} \tag{4}$$

3 Causal discovery as a Graph matching problem

In this section by reformulating the constrained loss function introduced in NOTEARS, we will see that the problem of linear causal discovery is equal to a graph matching problem. The linear causal discovery problem introduced in NOTEARS is formulated as problem 5:

$$\begin{aligned} \min_{A} \quad & \frac{1}{n}\sum_{i=1}^{n}||A^Tx_i - x_i||_2^2 \\ \text{s.t.} \quad & A \in DAGs \end{aligned} \tag{5}$$

where A is the adjacency matrix between variables that must be a directed acyclic graph. The adjacency matrix A can be reformulated as P^TBP where P is a permutation matrix and B is a strictly upper triangular matrix. Actually, we can find an upper triangular matrix B and permute the rows and columns of B by permutation matrix P to get the final adjacency matrix A. See equation 6:

$$A = P^TBP \tag{6}$$

Thus the problem 5 is formulated as equation 7 :

$$\begin{aligned} \min_{B,P} \quad & \frac{1}{n}\sum_{i=1}^{n}||(P^TBP)^Tx_i - x_i||_2^2 \\ \text{s.t.} \quad & B \in \text{strictly upper triangular matrices} \\ & P \in \text{permutation matrices} \end{aligned} \tag{7}$$

First, we rewrite the $||.||_2$ part in above loss function as equation 8:

$$\begin{aligned} ||P^TB^TPx_i - x_i||_2^2 &= (P^TBPx_i - x_i)^T(P^TB_TPx_i - x_i) \\ &= x_i^TP^TBPP^TB^TPx_i - x_i^TP^TBPx_i - x_i^TP^TB^TPx_i + x_i^Tx_i] \\ &= tr(P^TBB^TPx_ix_i^T) - tr(P^TBPx_ix_i^T) - tr(P^TB^TPx_ix_i^T) + tr(P^TIPx_ix_i^T) \end{aligned} \tag{8}$$

Then by adding the summation part we can rewrite the loss function as equation 9:

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}||(P^TBP)^Tx_i - x_i||_2^2 = tr(P^TCPS) \\ \text{where:} \quad & S = \frac{1}{n}\sum_{i=1}^{n}x_ix_i^T \text{ : sample covariance} \\ & C = (BB^T - B - B^T + I) \end{aligned} \tag{9}$$

Thus the problem of causal discovery 5is converted to problem 10:

$$\begin{aligned} \min_{B,P} \quad & tr(P^TCPS) \\ \text{s.t.} \quad & S = \frac{1}{n}\sum_{i=1}^{n}x_ix_i^T \\ & C = (BB^T - B - B^T + I) \end{aligned} \tag{10}$$

By rewriting the problem as below using equation 4, we conclude that the problem of causal discovery5 is equivalent to a graph matching problem. 11

$$\begin{aligned} min_{P\in P} tr(P^T CPS) = max_{P\in P} tr(P^T A_G^T P A_H) : \text{graph matching} \\ \text{where:} \quad A_G^T = -C \\ A_H = S \end{aligned} \tag{11}$$

So in the graph matching problem A_G and A_H are computed as equations 12, 13:

$$A_G = -BB^T + B^T + B - I \tag{12}$$

$$A_H = S = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \tag{13}$$

As a result, the problem of causal discovery is a graph matching between A_H which is the covariance matrix and A_G which is the defined transformation of strictly upper triangular matrix A by equation 12.

3.1 Concave Quadratic Graph matching

Here we show that the derived graph matching problem 11 (Finding the best permutation matrix) is also a concave quadratic program. We prove this statement as follows 14, 15:

$$\begin{aligned} max_{P\in P} tr(P^T A_G^T P A_H) = max_{P\in P} vec(P)^T Q vec(P) \\ \text{s.t.} \quad Q = A_H^T \otimes A_G^T \end{aligned} \tag{14}$$

And Q is also negative semi-definite. proof 15:

$$\begin{aligned} A_G = -BB^T + B^T + B - I = -(A_G^T - I)^T (A_G^T - I) \\ \rightarrow -x^T (A_G^T - I)^T (A_G^T - I) x = -||(A_G^T - I)x||_2^2 \leq 0 \end{aligned} \tag{15}$$

And because $A_H = S$ is positive semi-definite the $A_H^T \otimes A_G^T$ is negative semi definite and as a result the quadratic program is concave.

4 Systematic review

In this section we identify different continuous optimization methods for causal discovery.

4.1 Search Strategy

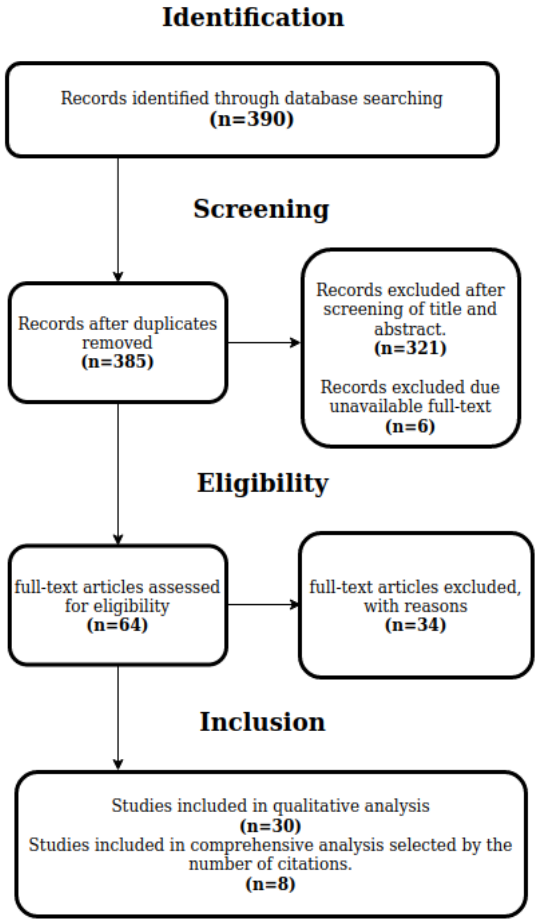
The search was carried out in the **Scopus** database. Searching terms were developed based on the research question that is finding causal discovery methods using continuous optimizations. We have developed different Boolean operators to have comprehensive database searches on Causality and discovery. Thus, the Scopus must include; **causal discovery OR causal learning OR (structure learning AND ("bayesian network" OR "directed acyclic graph")) AND optimization**. Some articles used the term Bayesian network or directed acyclic graph along with structure learning for continuous optimization methods but they don't mention causality but their approaches are important for this review. There the third OR term has been designed to generalize the search term. Finally, these papers must include the optimization term to just include causal discovery score-based methods that used optimization algorithms. The results are also limited to the **computer science** (CS) area because The objective is to find causal discovery and optimization methods in CS literature and doesn't consider applications in other areas. Also, Theere was a limit on the publication date from **2012**, and the search is updated until **March 19, 2022**

Google Scholar has also been consulted for unpublished literature and finding articles that have references to NOTEARS, the main article in continuous optimization for causal discovery.

4.2 Screening and Selection

The initial search from 2 databases results in 390 articles. on. By removing duplicates number of searched results decreased to 385 articles. Then articles were selected based on inclusion and exclusion criteria by title and abstract screening which results in 64 articles. Then selected papers were reviewed based on full-text and 30 articles were selected also based on criteria. Finally, 8 articles with the most citations were selected comprehensively compared in section 4.4.

Figure 1: Flow diagram of study selection and identification process



4.3 Eligibility criteria

Inclusion criteria

- Articles that proposed a new methods for causal discovery and not just used previous methods for causal inference tasks were included.
- Articles that proposed a continuous optimization method were included.
- The results that just proposed score-based method for causal discovery were selected.
- The results are also limited to the computer science.

Exclusion criteria

- There are optimization methods for causal discovery that used heuristic and greedy methods like genetics and particle swarm optimization that were excluded from results.
- Systematic or literature reviews were excluded.
- Results that are before the year 2012 were excluded.
- Few results are just in the list of all articles of a journal and actually are duplicated articles.
- Records that doesn't have full-text available were removed.

The flow diagram of selection process is presented in figure 1.

4.4 Results

Main articles comparison The main articles compared based on following features.

- Linearity or non-linearity of structure equation model?
- Does the article provide an application of causal inference for the tasks of intervention and counterfactual or not?
- Is there any proof of identifiability for new proposed model? Does the article use an identifiable model?
- Is there any change to the DAG constraint or not?
- What are the metrics used for evaluations? Most of them used Structural Hamming distance (SHD) value which is the most well-known metric for causal discovery. SHD denotes number of edge insertions, deletions or flips in order to transform one DAG to the true causal DAG. Other metrics are Least squares (LS), False Discovery Rate (FDR), True Positive Rate (TPR), Structural Intervention Distance (SID)
- What are the datasets used in the article for causal discovery?
- What is the main contribution of the article?

Description of Articles The Recent (2018) method DAGs with NO TEARS [6] is considered as the first method that converted the combinatoric graph-search problem into a continuous optimization problem and proposed a smooth and differentiable acyclicity constraint that further can be trained using different optimization algorithms including gradient-based optimization approaches. DAG-GNN [8] extended NO TEARS by adding non-linearity to SEM by using neural network functions and variational inference. DAG-GNN also introduced a modified DAG constraint that is more computationally effective. GAE[9] extended NO TEARS and DAG-GNN formulations for structure learning into a graph autoencoder model non-linear structural relationships. They demonstrated that GAE performs significantly better than NO TEARS and DAG-GNN. GranDAG [10] proposed a score function that directly maximizes the data likelihood. RL-based Causal discovery method [11] used the reinforcement learning method to train an encoder-decoder model to generate DAG. MaskedNN [12] improved NOTEARS by adding non-linearity to SEM with neural network and comprehensive discussion of identifiability. Sparse non-parameteric dags [13] is from the authors of NOTEARS that extended linearity of SEM to a non-linear one and introduced partial derivatives as a measure of acyclicity. The soft constraint dag method [14] studied the role of hard DAG constraint and showed that it is just necessary to apply soft sparsity so the problem converted to an unconstrained one.

Details of included articles are provided in Table 1

5 Conclusion

In this systematic review, 8 main articles on continuous optimization for causal discovery reviewed and compared with each other based on the challenges and information on causal discovery. The results show that there are different challenges in causal discovery with continuous optimization including modifications applied to the objective function of causal discovery, the causality constraint of the optimization problem, and the optimization algorithm. We also reformulate the problem of linear causal discovery as a graph matching problem and proved it as concave quadratic program.

Table 1: Details of included articles

References	Modeling	Causal Inference	Identifiability	DAG constraint	Metric	Datasets	Main contribution
[6]NOTEARS	Linear	No	No	Yes	SHD, LS	Synthetic, Protein[2]	continuous optimization for causal discovery
[8]DAG-GNN	Non-linear	No	No	Yes	SHD, FDR	Synthetic, Protein, Knowledge-Base[15]	First optimization method for non-linear SEM
[10]GranDAG	Non-linear	No	Yes	No	SHD, SID	Synthetic, Protein, SynTReN[16]	Likelihood optimization
[11]RL-Causal	Non-linear	No	Yes	No	SHD, FDR, TPR	Synthetic, Protein	Causal discovery with RL
[12]Masked NN	Non-linear	No	Yes	No	SHD, TPR	Synthetic, Protein,	Gradient-based optimization and comprehensive discussion of identifiability
[9]GAE	Non-linear	No	No	No	SHD, TPR	Synthetic	Causal discovery with Graph-Auto-Encoder
[13]Sparse-non-parameteric DAG	Non-linear	No	Yes	Yes	SHD	Synthetic, Protein	Causal discovery for more general SEMs
[14]Soft-Constraint	Linear	No	Yes	Yes	SHD, SID	Synthetic, Protein	Soft DAG constraint instead of hard one

References

[1] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[2] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[3] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.

[4] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. *arXiv preprint arXiv:2102.11107*, 2021.

[5] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[6] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[7] Mikhail Zaslavskiy, Francis Bach, and Jean-Philippe Vert. A path following algorithm for the graph matching problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2227–2242, 2008.

[8] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR, 2019.

[9] Ignavier Ng, Shengyu Zhu, Zhitang Chen, and Zhuangyan Fang. A graph autoencoder approach to causal structure learning. *arXiv preprint arXiv:1911.07420*, 2019.

[10] Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*, 2019.

[11] Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. *arXiv preprint arXiv:1906.04477*, 2019.

[12] Ignavier Ng, Zhuangyan Fang, Shengyu Zhu, Zhitang Chen, and Jun Wang. Masked gradient-based causal structure learning. *arXiv preprint arXiv:1910.08527*, 2019.

[13] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. PMLR, 2020.

[14] Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

[15] Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. Representing text for joint embedding of text and knowledge bases. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1499–1509, 2015.

[16] Tim Van den Bulcke, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchal. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics*, 7(1):1–12, 2006.