Continuous Causal Structural Learning without Acyclic Constraint

Prince Zizhuang Wang, William Yang Wang
University of California, Santa Barbara
{zizhuang_wang, william}@cs.ucsb.edu

Introduction

Learning the causal structure from samples of a joint distribution is a challenging problem, since the search space of the underlying Directed Acyclic Graphs (DAGs) is combinatorial. Some recent formulations turn this problem into a continuous optimization with a structural constraint enforcing acyclicity. We propose a novel formulation that ensures the learned graphs are acyclic without adding the acyclicity constraint and therefore turn the constrained optimization problem into an unconstrained one. We compare our approach to existing continuous optimization methods on real and synthetic data, and demonstrate that our method learns appropriate DAGs more efficiently thanks to the relaxation of the acyclic constraint.(3) Can generate diverse sequences of texts.

Problem Setup

This section reviews the recently proposed gradient-based method for causal structure learning and its variants.

The *NOTEARS* is the first one to propose a continuous optimization approach for learning the DAG structure of a linear structural equation model (SEM). Let $A \in \mathbb{R}^{d\times d}$ be a weighted adjacency matrix of a directed acyclic graph (DAG) $\mathcal{G}$, the linear SEM entailed by $\mathcal{G}$ reads,

$$X = A^\top X + Z$$

where $Z \in \mathbb{R}^d$ is the noise variable. The adjacency matrix A encodes the causal structure of the graph, in which the indices of non-zero elements in a column A_i correspond to the parents of node x_i in the causal graph.

The scored based method in *NOTEARS* solves the following constrained optimization problem to learn A,

$$\min_{A\in\mathbb{R}^{d\times d}} \frac{1}{2n}\sum_{i=1}^{n} \|X^{(i)} - A^\top X^{(i)}\|_F^2 + \lambda\|A\|_1$$
$$\text{s.t. } h(A) = tr(e^{A\circ A}) - d = 0$$

Approach

In this section, we propose a simple and elegant approach to learn the causal DAG from observational data without using *NOTEARS*' setting.

Given an arbitrary DAG $\mathcal{G}$, it is widely known that there exists an ordering π such that $\pi(i) < \pi(j)$ if and only if x_i is the parent or ancestor of x_j in the graph. Let X_π be the ordered set of variables, then it is obvious that the graph adjacency matrix A of the structural equation $X_\pi = A^\top X_\pi + Z$ is an upper triangular matrix. In this case, any upper triangular matrix A is guaranteed to be acyclic.

Contributions

- **We propose to solve the following optimization problem**

$$\min_{A,\theta} \frac{1}{2n}\sum_{i=1}^{n} \|P_\theta X - A^\top P_\theta X\|_F^2 + \alpha\|A\|_1$$

- **A is an upper-triangular matrix.**
- **A' = P^T A P would automatically be acyclic if the optimal solution is found**
- **However, permutation P is a discrete variable. We need to find a way to parameterize P using a continuous variable**

How to parameterize the permutation P

Definition 1. *A set of permutations $\mathcal{P}$ is a class of bijective mappings of a set onto itself, in which every element is a binary matrix $P \in \{0,1\}^{d\times d}$ where both rows and columns sum to one.*

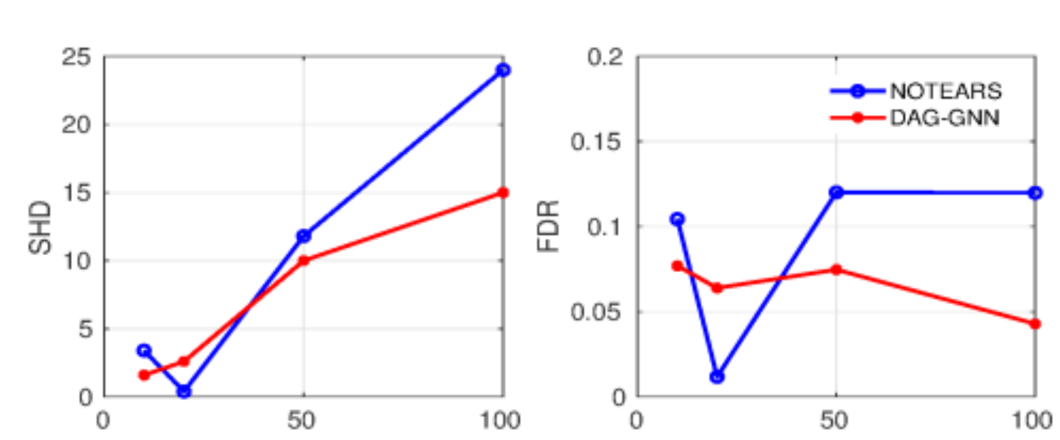
$$\mathcal{P}_d := \{P \in \{0,1\}^{d\times d} : P\mathbf{1}_d = \mathbf{1}_d, \mathbf{1}_d^\top P = \mathbf{1}_d^\top\}$$

Definition 2 (Birkhoff Polytope: Doubly Stochastic Matrix)**.** *Birkhoff Polytope is a set of Doubly Stochastic (DS) matrices defined as a set of non-negative matrices whose rows and clumns sums to one,*

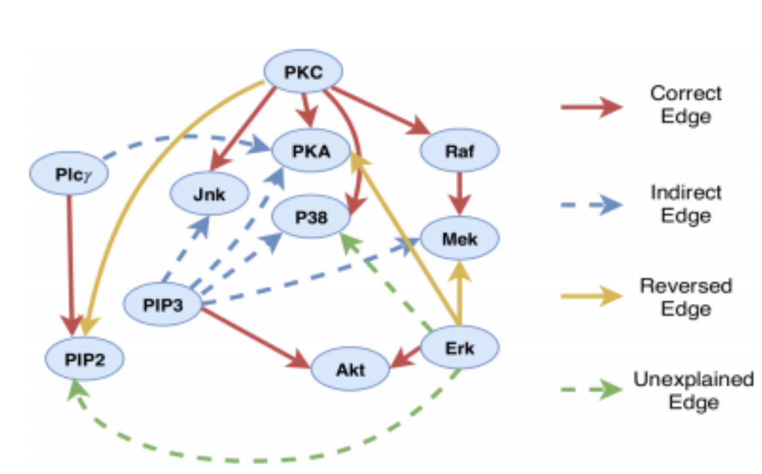
$$\mathcal{DP}_d := \{X \in \mathbb{R}_+^{d\times d} : \sum_{i=1}^{d} X_{ij} = \sum_{j=1}^{d} X_{ij} = 1\}$$

- **Birkhoff Polytope can be viewed as a continuous relaxation of permutation**
- **Permutation is the intersection of Birkhoff Polytope and Orthogonal Matrices**
- **We can parameterize P as Birkhoff Polytope, and then enforce orthogonality during training**

Results



SHD (structural Hamming distance), lower the better



Example of the learned protein signaling structure

Acknowledgement

This research is funded by CCS SURF. We thank Mrs. Carol and Mr. Paul George for their generous support!

References

[1] Zheng X, Aragam B, Ravikumar P K, et al. DAGs with NO TEARS: Continuous optimization for structure learning[C]//Advances in Neural Information Processing Systems. 2018: 9472-9483.