Abstract: To increase access to causal learning without having to manually choose free hyperparameters, we introduced Bayesian Optimization based techniques that treats each causal model as a set of predictive models from each node's Markov Blanket. We demonstrated higher performance than out-of-the-box NOTEARS-MLP algorithm using stimulated datasets. The code is available at https://github.com/findalexli/autobayes

Introduction:

Motivation: Bayesian Network provides a graphical framework for domain experts to embed their understanding of causal effects in the system via Directed Acalytic Graphs. While frameworks like CausalNex provides a nice DIY toolbox to build the graph semi-automatically via NOTEARS (Zheng 2018) and fit the conditional probabilities for do-calculus, it remains challenging for non-experts to use due to the the number of free parameters, especially the weight cutoff for determining the edges. Given that the underlying true graph is unknown, we have an unsupervised task that presents challenges to tune those parameters.

Related Work

A key innovative in causal inference was transforming the discrete DAG constraints into a smotth function proposed by Zheng 2018, namely the NOTEARS algorithm. The authors conver the combinatorial acyclic constraints on the DAG into a constrained continuous optimization problem. (Zheng et al. 2018). NOTEARS states a differentiable object function that can be optimized via common optimization techniques. More recently, Lachapelle et al. proposed to learn the DAG from the weights of a neural network while still considering the DAG constraints. (Lachapelle et al. 2019). Zheng et al. further extended the DAG learning into a nonparametric problem, proposing a generic formula of the DAG learning as E[X_j | X_parentj] = g_j (f_j ()) where f_j(u1, … uj) depends only on the set parents. As a example, they proposed a Multilayer Layer Perception model.

While NOTEARS and its descendants proposed a wide range of algorithms, there are regularization parameters and weight cutoff parameters left to be tuned. There have been several techniques proposed to tune causal discovery networks, including StARs method (Liu et al. 2010) introduced to tune the lambda penalty hyper-parameter for graphical lasso (Friedman et al. 20080. Another framework to select model configuration based on in-sample fitting of the data and model complexity has been introduced to causal disvoery tuning, namely the Bayesian Information Criterion(BIC) (Maathuis et al. 2009). Most recently, a Out-of-Sample Causal Tuning techniques (OCT) was introduced that treats a causal model as a set of predictive model, one for each nodes given its Markov Blanket, then tune the choices using cross-validation (Biza et al. 2011).

Methods:

In this work, our main contribution is

- Extension of the OCT technique to suit a non-parametric optimization presented by NOTEARS-MLP. We are able to tune four free parameters in the NOTEARS-MLP setup: L1 regularization parameter, L2 regularization parameter, Number of Nodes in the Hidden Layer and weight cutoff.

- While the OCT authors admitted limitation on their algorithm by inability to account for false positives, we introduced additional penality term for weakly correlated features in the Markov Blanket

Start Bayesian Optimization with L1, L2, weight cut off and number of nodes.

Split the training and test datasets

Find the estimated Graph using inner fold training data.

For each node,

find the markov blanet of the node, and fit a predictive model using catboost (50 iterations is enough) using the markov blanket features. Obtain the performance (RMSE) from Linear Regression Model with outer fold test set. Find the number of weakly correlated features (<0.15 correlation coefficient). Add the count of weakly correlated features factorized by a constant to the RMSE We also implemented a decision tree left to be explored.

Find average performance over all nodes

Find average performance across all nodes

Figure (1): Algorithm Introduced

We evaluated the results using SHD as well as a new metric we call G-score. In causal structural learning, NOTEARS and its derivations paper typically report on the Structural Hamming distance (SHD), the number of required changes to the graph to match the ground truth, as well as True Positive Rate and False Positive Rate. Structural Hamming distance is intuitive to understand and a proper indicator of structural distance. In an empirical study, de Jongh and Drudzel examined the direction of individual components of the Hamming distance metric shown in Figure (1). As more data is available, while it seems the skeleton of structure improves, the increase of incorrect edges increases. This example shows that the Hamming distance is subject to variations in data size and should not be used as a metric.

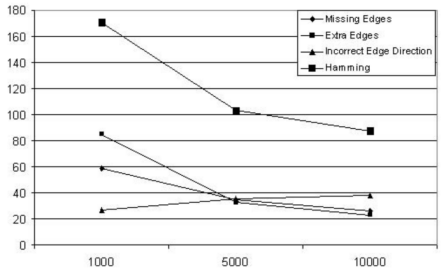

Figure (2): Hamming metrics and its components on the Hepar network, with n = {1000, 5000, 10000}

We combined the True Positives, False Positives and False Negatives to form one aggregated metric, defined as max(0, (True Positives - False Positives) / (True Positives + False Negatives). True positive means correctly estimated edge, False positive means estimated directed edge that is not present in ground truth, and false negative means true directed edge not present in true graph (Trustworthy AI). This is to reflect real-life application requirements in which 1) The denominator and numerator scales together, and normalize the metric invariant of the scale of the network 2) We would like to avoid false positives while seeking to pick up as much true causal relations as possible.

Once a validation metric is chosen, we seek to seek how we can choose the optimization-based approach to causal discovery. We evaluated the ease of generalization and code implementation of the following gradient-based structural learning model. MLP-based models are considered more flexible as multilayer perceptions can capture non-linear relationships.

NOTEARS	A gradient-based algorithm for linear data models (with least-squares loss). Generally considered as first to recast combinatorics graphs search problem as continuous optimization
NOTEARS-MLP	A gradient-based algorithm using Multilayer Perception modeling for non-linear causal relationships, with l1 and l2 regularization on its first layer
GOLEM	Extension of linear NOTEARS from least-square score function to likelihood based score function and soft acyclicity penalty

We implemented the following feature based on NO-TEARS-MLP: There are a number of free hyper parameters pending user-choice. Published NOTEARS-MLP paper notably used fixed parameters across graph types of weight threshold of 0.5, Lamba 1 and Lambda 2 = 0.03 and [d, 10, 1] fully connected layers.

1. Weight Threshold: threshold to transform learned adjacency matrix to estimated edges
2. Lamba 1: L1 regularization constant
3. Lambda 2: L2 regularization
4. Multilayer-perceptron neural structure.

We setup the tuning experiments using Bayesian Optimization as experiment configuration generalization techniques. A prior probability capture observed behavior of the objective function, and is being updated to estimate the posterior distribution over the objective function. We update the prior when we pick a combination of hyperparameter values at each iteration, i.e. P(metric | hyperparameter combination), and stop when we meet preset stopping criteria. An acquisition function is made from the estimated posterior distribution that would infer the next configuration setup. This choice is motivated by extensive computer time.

Experiments and Results

We generate simulated ground-truthDAG with Erdos-Renyi (ER), and sample DAG with 2d edges denoted by ER2 with 200 samples. With ground-truth DAG, we stimulated X_j = f_j (X) + zj where f_j are Additive Noise Models with Gaussian processes (ANM with GP) and z_j ~ N(0, 1).

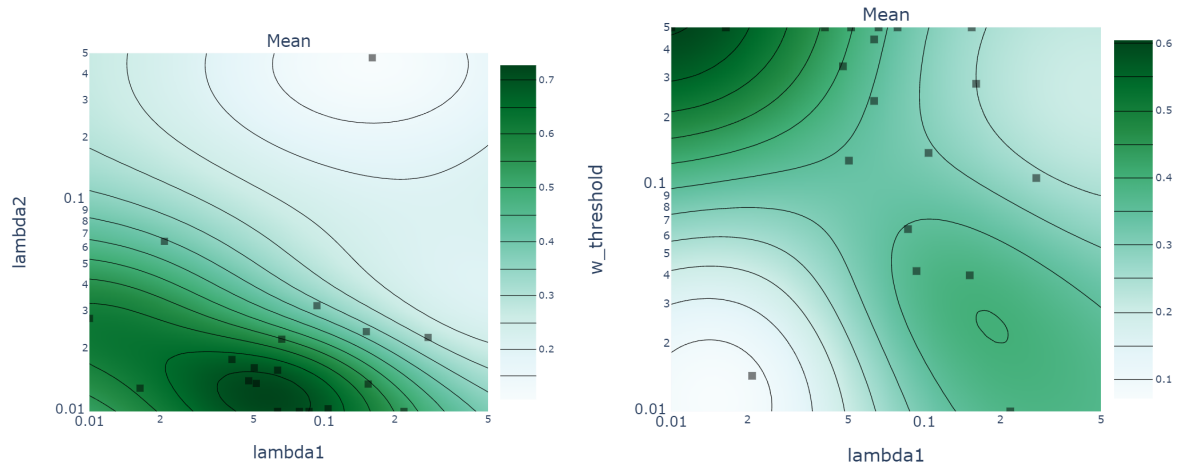

Figure (2): Inferred contour map based on 20-iteration of Bayesian Optimization. Based on (d=10) ER2 stimulation with additive noise model + gaussian process

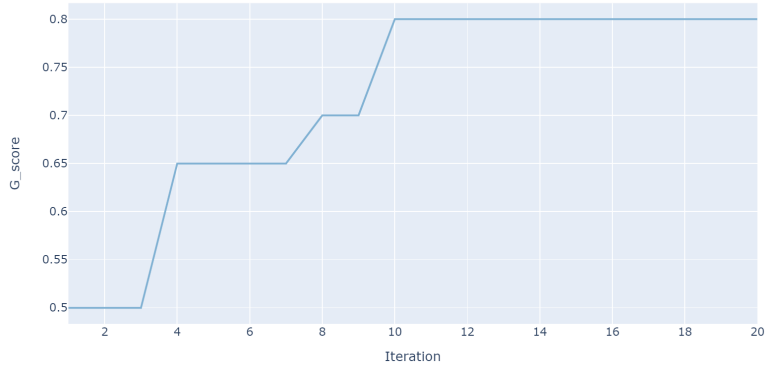

Figure (3) G score over iterations for the experiment described in Figure (2)

shd	shd_paper	d (number of edges)
5	10	10
15	19	20
16	26	30

Figure (4): Comparison of SHD with published SHD (Zheng et, al. 2020). The Bayesian optimization found more optimal set of (weight threshold, lambda 1, lambda 2) compared to the original NOTEARS-MLP result

Future directions

We have yet to fully go through the full search space that the NOTEARS-MLP has covered in terms of stimulation data. Additionally, we will validate the approach on real-world data.

Contribution

This is a solo project.

References:

Biza, K., Tsamardinos, I., & Triantafillou, S. (2020, February). Tuning causal discovery algorithms. In *International Conference on Probabilistic Graphical Models* (pp. 17-28). PMLR

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., & Xing, E. P. (2020). *Learning Sparse Nonparametric DAGs*. https://github.com/xunzheng/notears.

https://github.com/huawei-noah/trustworthyAI

Vowels, M. J., Cihan Camgoz, N., & Bowden, R. (2021). *D'ya like DAGs? A Survey on Structure Learning and Causal Discovery; D'ya like DAGs? A Survey on Structure Learning and Causal Discovery*. https://arxiv.org/pdf/2103.02582.pdf

Ng, I., Ghassami, A., & Zhang, K. (n.d.). *On the Role of Sparsity and DAG Constraints for Learning Linear DAGs*.