DISCUSSION ON "BAYESIAN INTERVENTION OPTIMIZATION FOR CAUSAL DISCOVERY" BY WANG ET AL.

Shuo Hao
The Hong Kong Polytechnic University
cee-shuo.hao@connect.polyu.hk

February 6, 2025

ABSTRACT

This report presents a critical analysis of the study "Bayesian Intervention Optimization for Causal Discovery" by Wang et al. [1], accompanied by the replication of their synthetic experiments utilizing the TensorFlow Probability framework. Prepared in fulfillment of the requirements for the AI & Data Science Associate Internship Program interview, the answers to the given questions are also covered in this report.

1 Overview

Overall, the paper does not contain errors that affect its main arguments, but there are several minor issues. Here are the specific details:

- **Page 5:** To estimate the interventional distribution using the Maximum Likelihood Estimation (MLE) principle, the optimization should be $\max P(Y)$, instead of the joint distribution $P(X, Y)$.
- **Page 6, Algorithm 1:** For step 5, the hypothesis should be $\mathbb{H}_0$ and $\mathbb{H}_1$, not two models. For steps 7 and 8, The authors seem to want to express that P_{DC} can be evaluated and thus the optimal value of x can be computed. In this context, step 7 is is redundant.
- **Page 8:** In equation for computing π, the subscript on '**1**' is omitted.
- **Captions for Figures 3 to 6:** The values for k_0 and k_1 are incorrect.
- **Comparison of three causal discovery schemes:** In all figures showing the comparison of results obtained from different strategies, the values of P_{DC}, $P(\mathbb{H}_0|\mathrm{D_{int}})$ and $P(\mathbb{H}_1|\mathrm{D_{int}})$ should not exceed 1. However, the figures display colorful regions where these probabilities are greater than 1. This discrepancy may arise because the regions represent wide intervals, such as three standard deviations, resulting from unstable performance across the ten replications. Nevertheless, the authors did not provide any explanation of what these regions signify.

The central problem this paper investigates is how to design an optimal intervention scheme to expedite causal discovery. The paper's key contribution is the introduction of the metric P_{DC}, formulated to determine the probability of achieving decisive and correct evidence by combining new data obtained from the intervention $\mathrm{do}(X = x)$ with existing interventional data $\mathrm{D_{int}}$. In essence, this metric is a function of x, and by maximizing it, one can seek the next optimal intervention based on the existing intervention dataset.

As a serious reviewer, I would not consider this paper to make a substantial theoretical contribution. The only notable aspect is the proposal of P_{DC} as a function of $\mathrm{do}(x)$, which effectively serves as an objective function to guide the next intervention based on known information. However, the concept of P_{DC}, i.e., "probability of achieving decisive and

correct evidence", is not introduced here for the first time. De Santis [2] already provided a comprehensive definition of decisive and correct evidence, in addition to two more matrices, the misleading and weak evidence. While De Santis did not employ any of these evidence types as an objective function for causal discovery optimization, the paper under review does. Essentially, the approach can be interpreted as MLE over the two hypothesis $\mathbb{H}_0$ and $\mathbb{H}_1$. Consequently, this work does not represent a significant theoretical advancement.

In addition, there are two aspects of the paper that the authors present as advanced, but upon careful examination, I do not agree with their claims. First, the author claims that the optimization process requires a Bayesian optimization strategy. However, in practical applications, computing P_{DC} can be fully executed by a computer without relying on experimental evaluation. In our reproduction using TensorFlow Probability, we followed the author's requirement of using 4096 Monte Carlo samplings to evaluate P_{DC} at every do(x_i), and each evaluation took approximately 0.0557 s. This level of computational cost is minimal and does not justify the need for Bayesian optimization; instead, one can directly identify the optimum by performing a Monte Carlo estimation of P_{DC} across the full domain of x.

Second, the approximate expression for P_{DC} is derived by replacing a step function with a differentiable variant. Although this transformation renders the function differentiable, it is debatable whether differentiability is essential in this context. First, Bayesian Optimization does not inherently require the objective function to be differentiable. Second, our Monte Carlo method does not depend on gradient information. The paper does not provide sufficient discussion on the benefits of this modification. Consequently, despite being a central element of the paper, this approach appears to offer limited practical advantages compared to directly evaluating P_{DC} using the original formulation for improved accuracy.

2 Accept or Reject?

For a top-tier machine learning conference, I lean toward rejecting the paper. Although the paper introduces an innovative use of P_{DC} as an objective function to guide optimal interventions in causal discovery, a concept that could inspire further research on alternative metrics (such as minimizing misleading evidence or weak evidence in [2]), it fails to address several core issues. For instance, the claim that the optimization process requires Bayesian optimization is not substantiated by practice. Additionally, the paper derives an approximate, differentiable expression for P_{DC} by replacing a step function with a differentiable one; however, neither Bayesian optimization nor the Monte Carlo method used relies on differentiability and the authors do not adequately discuss the benefits of this modification.

Moreover, the framework considers only two hypotheses $\mathbb{H}_0$ and $\mathbb{H}_1$, which is overly simplistic for real-world causal inference problems. In scenarios where undetected confounders exist (e.g., $X \to Z \to Y$), the relationship $p(Y|\text{do}(X)) = \sum_Z p(Y|Z=z)p(Z|X=x)$ is not fully captured by a binary hypothesis framework, potentially leading to biased estimations.

3 Replication Using Tensorflow Probability

The source code for our replication is located in the `code_for_part_2` folder, with a demonstration on applying the code for causal discovery available in the `demo` sub-folder. In this report, we present results that closely match those shown in Figure 1 of the original paper [1].

All critical experimental parameters, such as observation data sample sizes, k_0, k_1, β, and others, are maintained exactly as described by the authors. Three distinct causal scenarios are evaluated by performing ten independent trials for each. In every trial, the algorithm incrementally increases the number of intervention samples from 0 to 9, recording four metrics: P_{DC}, $\log \text{BF}_{01}$, $P(\mathbb{H}_0|\text{D}_{\text{int}})$, and $P(\mathbb{H}_1|\text{D}_{\text{int}})$. To clearly illustrate the evolution of these metrics, the mean and variance across the ten trials for each intervention sample size are calculated. In the accompanying Figure 1, the blue-shaded area represents $\pm$ one standard deviation. The results indicate that causal discovery was most stable in the third scenario, followed by the first scenario, while the $X \leftarrow U \to Y$ configuration proved to be the most challenging even using the optimal causal discovery.

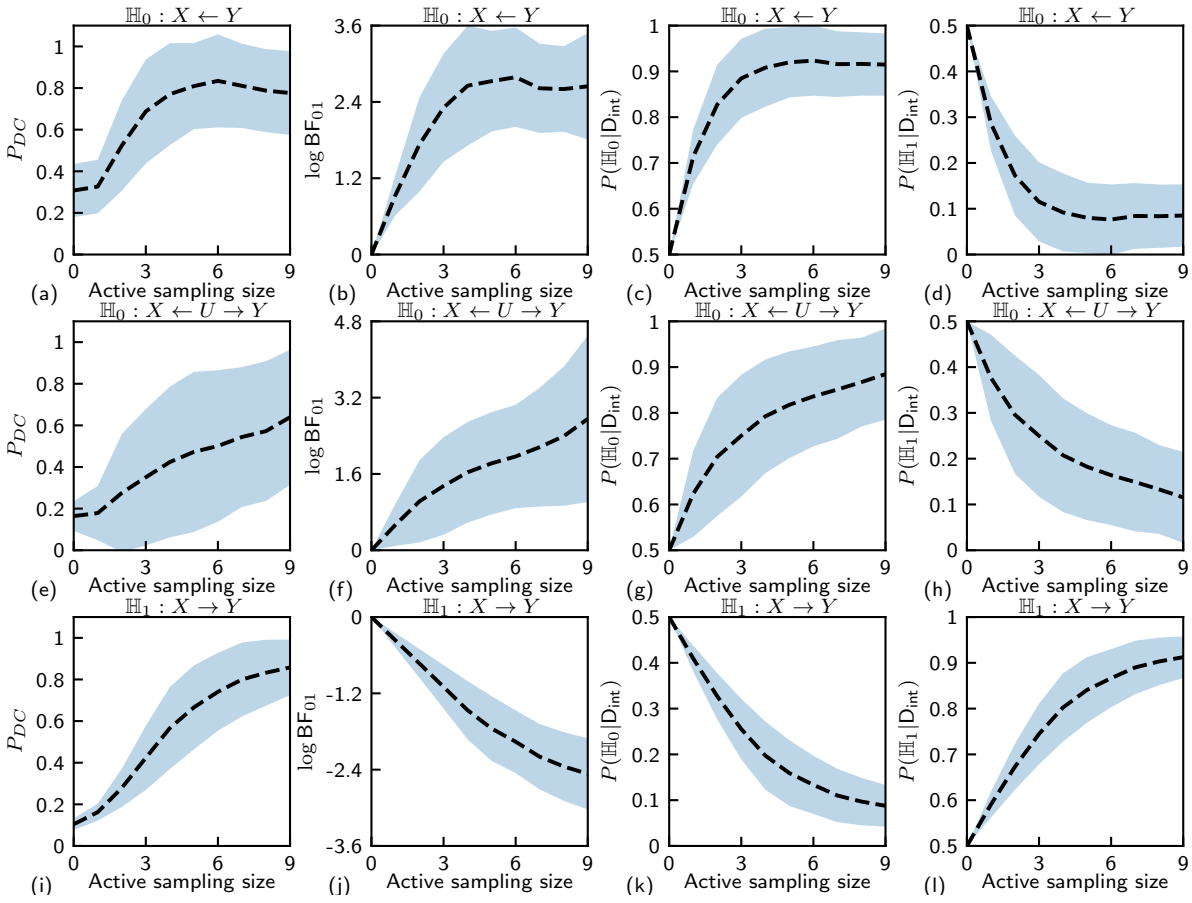

Figure 1: Results under different ground truths. The first row corresponds to $\mathbb{H}_0(X \leftarrow Y)$. The second row corresponds to $\mathbb{H}_0(X \leftarrow U \rightarrow Y)$. The third row corresponds to $\mathbb{H}_1(X \rightarrow Y)$.

4 Optimal Causal Discovery Considering Multiple Variables

Extending the approach to handle multiple variables is possible. When additional variables are involved, the problem shifts to identifying, among several competing candidate causal schemes, the most plausible causal structure based on data collected through a series of intervention experiments. In effect, the objective is to determine which candidate graph best captures the true causal relationships so that it may serve as a reliable foundation for subsequent analyses.

A practical approach in the multivariable setting is to draw on the idea of maximizing the probability of achieving decisive and correct evidence, P_{DC}, to guide intervention selection. At each stage of experimentation, the intervention that maximizes P_{DC} is chosen, thus strategically directing the experimental process toward distinguishing among competing causal hypotheses and enhancing the overall precision of the causal discovery effort.

More specifically, consider a collection of candidate directed acyclic graphs (DAGs) denoted as $\mathcal{G}_i = \langle \mathbf{U}, \mathbf{V}, F, P(\mathbf{U}) \rangle$, with $i = 1, \cdots n$. In each DAG, $\mathbf{U}$ consists of independent exogenous background variables distributed according to $P(\mathbf{U})$, $\mathbf{V}$ represents observed endogenous variables, which we further divide into three disjoint subsets: non-manipulative variables $\mathbf{C}$, manipulative variables $\mathbf{X}$ that can be intervened in, and output variables $\mathbf{Y}$. The function $F = \{f_1, \cdots, f_{|\mathbf{V}|}\}$ comprises mappings that capture the causal relationships between these variables. Although each candidate DAG shares the same $\mathbf{U}$ and $\mathbf{V}$, they differ in terms of their underlying topology.

For each candidate DAG $\mathcal{G}_i$, we assume the existence of a minimal intervention set denoted by MIS_i, which is the smallest subset of manipulable variables necessary to unearth the causal structure inherent in $\mathcal{G}_i$. Consolidating the

intervention requirements across all candidate models, we define the overall minimal intervention set as:

$$\text{MIS} = \bigcup_{i=1}^{n} \text{MIS}_i. \tag{1}$$

Thus, any candidate intervention X_s is selected from this union.

In order to evaluate the effectiveness of a proposed intervention $X_s \in \text{MIS}$, the metric $P_{DC}(\text{do}(X_s))$ is adopted to quantify the likelihood that the intervention will produce evidence sufficient to identify the true causal structure. Given the intervention dataset D_{int} and corresponding hypotheses $\mathbb{H}_1 \cdots \mathbb{H}_n$(each associated with a candidate DAG), the metric is formulated as a weighted sum, given by: the optimal intervention selection problem for causal discovery is given by:

$$P_{DC}(\text{do}(X_s)) = \sum_{i=1}^{n} P^i_{DC}(\text{D}_{\text{int}}, \text{do}(X_s)) P(\mathbb{H}_i|\text{D}_{\text{int}}), \tag{2}$$

Here, $P(\mathbb{H}_i|\text{D}_{\text{int}})$ is the probability of hypothesis $\mathbb{H}_i$ given the intervention data, and $P^i_{DC}(\text{D}_{\text{int}}, \text{do}(X_s))$ denotes the probability that, under hypothesis $\mathbb{H}_i$, the combined observational and intervention data will yield decisive and correct evidence about the causal structure.

More concretely, for a given hypothesis $\mathbb{H}_i$, $P^i_{DC}(\text{D}_{\text{int}}, \text{do}(X_s))$ is defined via the pairwise Bayes factor criterion:

$$P^i_{DC}(\text{D}_{\text{int}}, \text{do}(X_s)) := P(\sum_{\substack{j=1 \\ j\neq i}}^{n} \text{BF}_{ij}(\text{D}_{\text{int}} \cup \text{D}_{\text{new}}) > k \mid \text{D}_{\text{int}}, \text{do}(X_s), \mathbb{H}_i), \tag{3}$$

where the pairwise bayes factor $\text{BF}_{ij}(\text{D})$ is given by the ratio:

$$\text{BF}_{ij} = \frac{P(D|\mathbb{H}_i)}{P(D|\mathbb{H}_j)} \tag{4}$$

with D_{new} representing the new data obtained from $\text{do}(X_s)$ subjected to $\mathbb{H}_i$, the threshold k is set to decide when the combined evidence against all competing hypotheses is sufficiently strong. In this context, the overall optimal intervention $X^\star_s$ is chosen as the one maximizing the P_{DC}.

From a computational perspective, the assessment of $P^i_{DC}(\text{D}_{\text{int}}, \text{do}(X_s))$ can be challenging to the necessity of handling multiple competing hypotheses and the potential high-dimensionality of the intervention space. In such cases, approximate strategies, such as the use of surrogate models like causal Gaussian processes, can be employed to facilitate efficient maximization of $P_{DC}(X_s)$.

5 Potential Application

Assuming the findings of this paper are accurate, banks can significantly enhance their risk management strategies by integrating causal insights into their predictive models for loan defaults. Utilizing the optimal causal discovery method proposed in this study, it is anticipated that numerous indicators, those not only statistically correlated with loan repayment rates but causally impact the loan repayment, can be identified at a minimal cost. By discerning which factors genuinely influence loan repayment, banks can effectively mitigate bad debt rates and improve the accuracy of their creditworthiness assessments.

References

[1] Yuxuan Wang, Mingzhou Liu, Xinwei Sun, Wei Wang, and Yizhou Wang. Bayesian intervention optimization for causal discovery, 2024.

[2] Fulvio De Santis. Statistical evidence and sample size determination for bayesian hypothesis testing. *Journal of Statistical Planning and Inference*, 124(1):121–144, 2004.