CAUSAL DISCOVERY REPORT ON BASE_DATA

TECHNICAL REPORT

Causal Copilot

October 28, 2024

ABSTRACT

This report focuses on the causal relationships within the Base_data dataset, which was subjected to a rigorous analysis using advanced causal discovery techniques. Our methodology involved an initial data preprocessing phase that included statistical examination and exploratory data analysis to address missing values and understand variable distributions. Subsequently, we employed a large language model (LLM) to assist in selecting suitable causal discovery algorithmsnamely PC, GES, and DirectLiNGAMbased on the dataset's characteristics. We also fine-tuned hyperparameters through LLM guidance to enhance algorithm performance. Our analysis revealed a complex network of causal influences, particularly identifying significant relationships among variables X0, X1, X2, X5, X6, X4, and X7. Notably, although moderate confidence was found in certain causal links, others exhibited lower reliability, indicating a need for further validation. Our contribution lies in providing a structured approach for causal discovery in unobserved contexts, illustrating both the power and limitations of current methodologies in inferential analyses, and underscoring the necessity for cautious interpretation of causal conclusions in the absence of domain knowledge.

Keywords Causal Discovery, Large Language Model, PC

1 Introduction

Causal discovery is a critical area of research that aims to identify the underlying causal relationships among variables in a given dataset. In the absence of prior knowledge about the dataset, the analysis becomes particularly challenging yet intriguing, as it requires the application of various statistical and computational techniques to infer potential causal structures. This report presents an in-depth investigation into the dataset, employing state-of-the-art causal discovery algorithms to uncover meaningful relationships and elucidate the complexities within the data. The findings from this analysis will not only contribute to a better understanding of the interdependencies present but also provide valuable insights for future research and practical applications.

2 Dataset Descriptions and EDA

The following is a preview of our original dataset.

Table 1: Dataset Preview

X0	X1	X2	X3	X4	X5	X6	X7
1	1	0	1	1	0	0	1
1	0	0	0	1	1	1	0
1	0	1	1	1	0	0	1
0	1	0	0	1	1	1	0
0	0	0	0	1	1	1	1

2.1 Data Properties

We employ several statistical methods to identify data properties.

The shape of the data, data types, and missing values are assessed directly from the dataframe. Linearity is evaluated using Ramseys RESET test, followed by the Benjamini & Yekutieli procedure for multiple test correction. Gaussian noise is assessed through the Shapiro-Wilk test, also applying the Benjamini & Yekutieli procedure for multiple test correction. Time-Series and Heterogeneity are derived from user queries.

Properties of the dataset we analyzed are listed below.

Table 2: Data Properties

Shape (n x d)	Data Type	Missing Value	Linearity	Gaussian Errors	Time-Series	Heterogeneity
(1500, 8)	Continuous	False	False	False	False	False

2.2 Distribution Analysis

The following figure shows distributions of different variables. The orange dash line represents the mean, and the black line represents the median. Variables are categorized into three types according to their distribution characteristics.

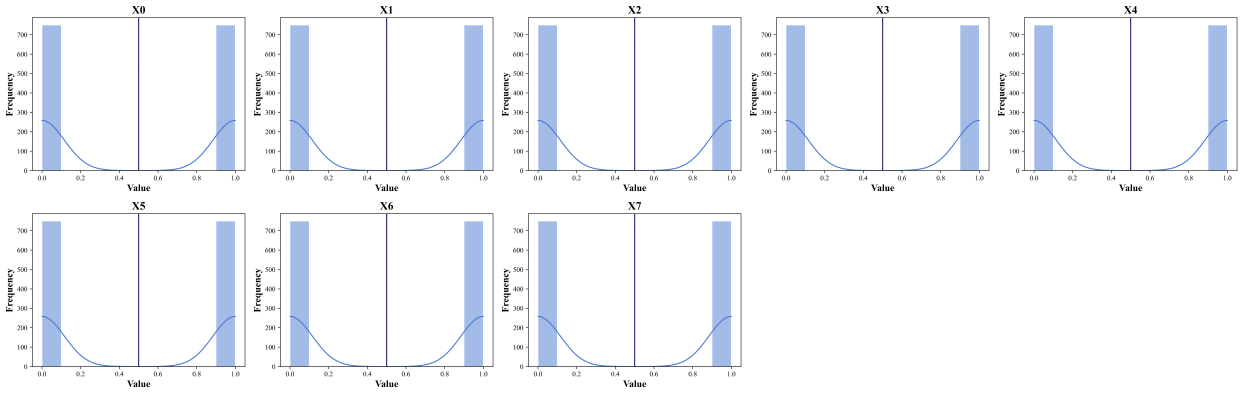


Figure 1: Distribution Plots of Variables

- Slight left skew distributed variables: None
- Slight right skew distributed variables: None
- Symmetric distributed variables: X0, X1, X2, X3, X4, X5, X6, X7

2.3 Correlation Analysis

In this analysis, we will categorize the correlation statistics of features in the dataset into three distinct categories: Strong correlations ($r > 0.8$), Moderate correlations ($0.5 < r < 0.8$), and Weak correlations ($r < 0.5$).

- Strong Correlated Variables: None
- Moderate Correlated Variables: None
- Weak Correlated Variables: None

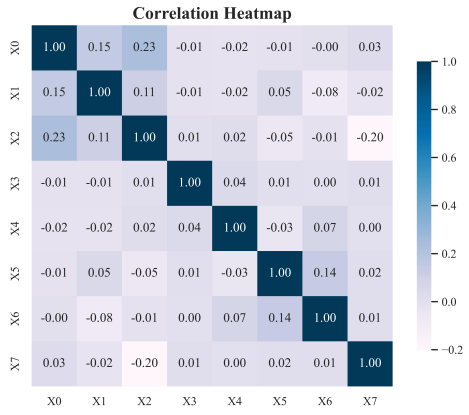


Figure 2: Correlation Heatmap of Variables

3 Discovery Procedure

In this section, we provide a detailed description of the causal discovery process implemented by Causal Copilot. We also provide the chosen algorithms and hyperparameters, along with the justifications for these selections.

3.1 Data Preprocessing

In this initial step, we preprocessed the data and examined its statistical characteristics. This involved cleaning the data, handling missing values, and performing exploratory data analysis to understand distributions and relationships between variables.

3.2 Algorithm Selection assisted with LLM

Following data preprocessing, we employed a large language model (LLM) to assist in selecting appropriate algorithms for causal discovery based on the statistical characteristics of the dataset and relevant background knowledge. The top three chosen algorithms, listed in order of suitability, are as follows:

- **PC**:
  - **Description**: The PC algorithm is a constraint-based method that learns the structure of a causal graph from data by testing conditional independencies between variables. It constructs a directed acyclic graph (DAG) representing the causal relationships.
  - **Justification**: Given the dataset's large sample size, PC is efficient for discovering causal structures when all relevant variables are observed. The assumption of sufficient data allows it to operate effectively despite non-linearities.
- **GES**:
  - **Description**: Greedy Equivalence Search (GES) is a score-based causal discovery algorithm that identifies the optimal causal structure by navigating the space of equivalence classes of Directed Acyclic Graphs (DAGs).
  - **Justification**: GES is suitable due to the dataset's characteristics: with no missing values and a large sample size, it can effectively handle the non-linear relationships present in a more complex search space efficiently.
- **DirectLiNGAM**:
  - **Description**: DirectLiNGAM improves the original LiNGAM framework by introducing an efficient stepwise linear regression approach to directly estimate the causal order.
  - **Justification**: Although DirectLiNGAM assumes linear relationships, its effectiveness in handling non-Gaussian errors can still be leveraged in this dataset, making it a valuable option for causal discovery.

3.3 Hyperparameter Values Proposal assisted with LLM

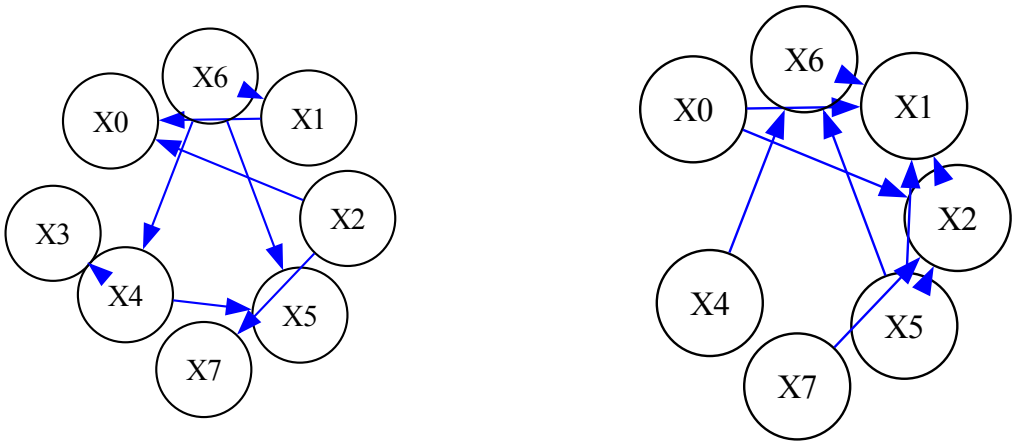
Once the algorithms were selected, the LLM aided in proposing hyperparameters for the [ALGO] algorithm, which are specified below:

- **alpha**:
  - **Value**: 0.1
  - **Explanation**: Given that the dataset is relatively large (1500 samples) and does not exhibit predominantly linear relationships, a slightly higher alpha value of 0.1 is suggested to avoid missing true causal relationships. This balances the need for detecting edges without being overly conservative.
- **indep_test**:
  - **Value**: fisherz
  - **Explanation**: Although Fisher's Z test assumes normality and linear relationships, it is still appropriate given that the data is continuous even if it does not meet all assumptions. As this dataset features continuous data types, Fisher's Z is the best method among the alternatives since chi-squared tests are not suitable for continuous data without discretization.
- **depth**:
  - **Value**: -1

– **Explanation**: Setting the depth to -1 allows the algorithm to explore all possible relationships between the variables without restricting the depth of the search. This is justified given the dataset's characteristics, as there are no known complexities that would necessitate limiting the search.

This structured approach ensures a comprehensive and methodical analysis of the causal relationships within the dataset.

4 Results Summary



(a) True Graph

(b) Initial Graph

Figure 3: Graphs Comparison of PC

The above are the true graph and the result graph produced by our algorithm.

The causal relationships among the variables illustrate a complex network of influence where X0 serves as a precursor to both X1 and X2, indicating that changes in X0 can directly affect these subsequent variables. Additionally, X2 has a causative effect on X1, suggesting that the dynamics involving X0 and X2 play a significant role in shaping the behavior of X1. Furthermore, X5 emerges as a pivotal variable, exerting influence over X1, X2, and X6, thereby linking these variables in a chain of causation that underscores the interconnectedness of their relationships. The causation extends to X6 affecting X1, establishing an additional layer of influence from X5 through X6 to X1. Meanwhile, X4 impacts X6, positioning it as a potential regulator within the system. Finally, X7 influences X2, indicating that X2's response may also be shaped by external factors represented by X7, which adds complexity to the causal web. Overall, each variable plays a distinct role, contributing to the network of cause and effect that characterizes this system.

4.1 Graph Reliability Analysis

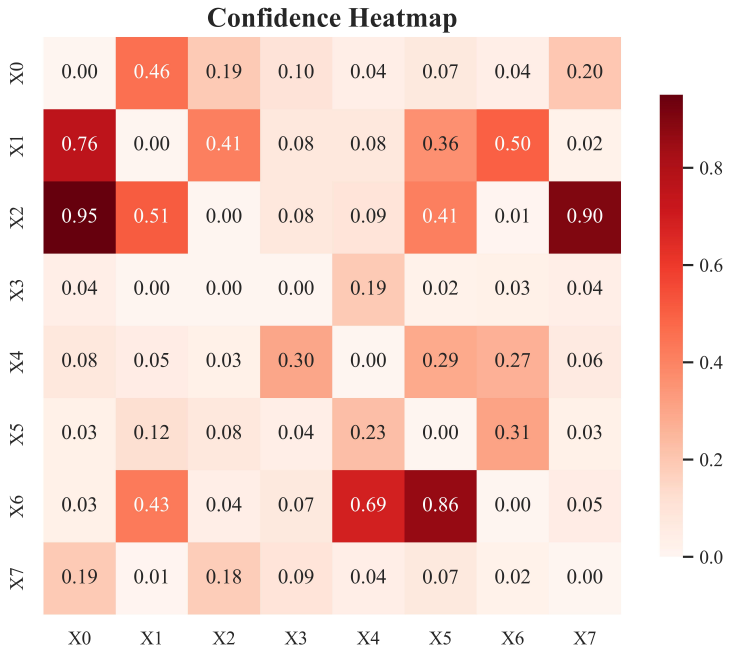


Figure 4: Reliability Graph

Based on the confidence probability heatmap and background knowledge, we can analyze the reliability of our graph.

From the statistics perspective, we have moderate confidence to believe that these edges exist: X2 $\rightarrow$ X1 (bootstrap probability of 0.51) and X6 $\rightarrow$ X1 (bootstrap probability of 0.43), indicating a significant likelihood of these causal relationships. However, we have low confidence in the existence of the following edges due to their lower bootstrap probabilities: X0 $\rightarrow$ X1 (0.46), X4 $\rightarrow$ X6 (0.27), X5 $\rightarrow$ X1 (0.12), X5 $\rightarrow$ X2 (0.08), X5 $\rightarrow$ X6 (0.31), and X7 $\rightarrow$ X2 (0.18), suggesting that these relationships may not be robust.

However, based on the expert knowledge, since we have no specific background information regarding the variables in question, we cannot definitively assert the existence or non-existence of any causal relationships. The absence of background knowledge implies that we cannot corroborate or dispute any of the statistical findings with domain expertise.

Therefore, the result of this causal graph is somewhat unreliable. The moderate to low confidence levels indicated by the bootstrap probabilities suggest that many proposed edges may not represent true causal relationships, highlighting the need for further investigation and validation before drawing concrete conclusions regarding these variables.