CAUSAL DISCOVERY REPORT ON 2021ONLINE SHOP

TECHNICAL REPORT

April 9, 2025

ABSTRACT

This report presents a comprehensive causal discovery analysis of a dataset from an online shopping context, examining key variables such as Shopping_Event, Ad_Spend, Page_Views, Unit_Price, Sold_Units, Revenue, Operational_Cost, and Profit. Applying the PCMCI algorithm, we leveraged advanced causal discovery techniques to uncover the intricate relationships among these variables, particularly the dynamic interplay between Ad_Spend and other factors like Revenue and Operational_Cost. Our findings reveal a complex feedback loop where Ad_Spend not only drives Shopping Events and Page Views but is also influenced by Sold Units and Revenue, indicating a cyclical relationship vital for decision-making. Additionally, we compared PCMCI with VARLiNGAM, highlighting that while PCMCI captures a broader set of causal relationships, VARLiNGAM offers more stable and interpretable results with less uncertainty. This study contributes to the understanding of online shopping dynamics, emphasizing the importance of considering reciprocal relationships in marketing strategies to enhance profitability and operational efficiency.

Keywords Causal Discovery, Large Language Model, PCMCI, 2021online shop

1 Introduction

In this report, we aim to conduct a comprehensive causal discovery and inference analysis of a dataset derived from an online shopping context. The dataset encapsulates various variables, including Shopping_Event, Ad_Spend, Page_Views, Unit_Price, Sold_Units, Revenue, Operational_Cost, and Profit, each providing valuable insights into consumer behavior and business performance. By examining the intricate relationships between these variables, such as the impact of Ad_Spend on Page_Views and the direct link between Sold_Units and Revenue, we intend to unveil the underlying causal structures that govern online shopping dynamics. Additionally, we will account for contextual factors, such as seasonality effects and competitive influences, which may further complicate these relationships. Our investigation will leverage causal discovery algorithms to enhance our understanding of how these variables interact over time, ultimately guiding strategic decision-making for improved business outcomes.

2 Background Knowledge

2.1 Detailed Explanation about the Variables

The dataset features several meaningful variables associated with online shopping, including a **Date** timestamp for tracking shopping events, a **Shopping Event** indicator to categorize customer activities, and **Ad Spend** representing the advertising budget for the period. Additionally, **Page Views** provides insights into customer interest, while **Unit Price** and **Sold Units** directly affect revenue and sales performance. Finally, **Revenue**, **Operational Cost**, and **Profit** are essential for assessing the financial success of the online storefront.

Crucial background knowledge includes understanding seasonal traffic variations, which can significantly influence shopping behavior, and customer behavior analysis, where product visibility and pricing strategies are key factors. Moreover, it is important to consider attribution models for revenue credit allocation, as well as the macroeconomic

context and competitive landscape that can sway both pricing and sales figures. Together, these insights can refine the causal discovery process, enhancing the analysis of variable interactions within the dataset.

2.2 Possible Causal Relations found by LLM

The following are potential causal relationships suggested by the language model, which are visualized in Figure 1. Please note that only variables present in our dataset are included in the figure.

- **Revenue** $\rightarrow$ **Profit**: Profit is directly influenced by revenue, calculated as Revenue - Operational Cost - Ad Spend.

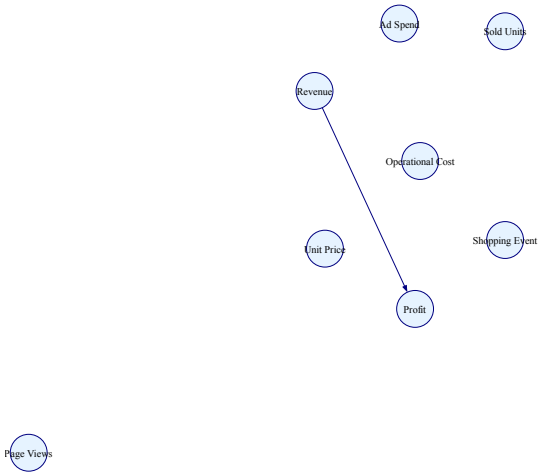


Figure 1: A Causal Graph Suggested by LLM.

3 Dataset Descriptions and EDA

The following provides a preview of our original dataset. If the dataset contains more than 10 columns, a random subset of 10 columns is displayed for illustrative purposes.

Table 1: Dataset Preview

Shopping Event	Ad Spend	Page Views	Unit Price	Sold Units	Revenue	Operational Cost	Profit
0	1490.490265	11861	999.000000	2317	2314683.000000	1659999.894000	654683.105600
0	1455.917463	11776	999.000000	2355	2352645.000000	1678959.080000	673685.919900
0	1405.824790	11861	999.000000	2391	2388609.000000	1696906.137000	691702.863000
0	1379.299391	11677	999.000000	2344	2341656.000000	1673380.635000	668275.364700
0	1234.199268	11871	999.000000	2412	2409588.000000	1707252.614000	702335.386000

3.1 Data Properties

We employed several statistical methods to identify data properties, including:

Basic Data Characteristics

The shape of the data, variable types, and the presence of missing values were assessed directly from the DataFrame. In contrast, properties such as time-series structure and heterogeneity were inferred with LLM based on user queries and DataFrame.

Linearity Testing

We conducted the Ramsey's RESET test to assess linearity between each pair of variables. When the total number of possible variable pairs was fewer than 100, all pairs were tested. If the number exceeded 100, a random subset of 100 pairs was selected for testing to ensure computational feasibility. To account for multiple testing, we employed the Benjamini and Yekutieli procedure, which is robust when dealing with dependent or correlated data. The linearity assumption was considered satisfied only if all tested pairs exhibited linearity; otherwise, it was considered violated.

Normality of Residuals

The assumption of Gaussian (normally distributed) noise was assessed using the Shapiro-Wilk test. The testing approach depended on the outcome of the linearity evaluation. If linearity was satisfied, we fitted ordinary least squares (OLS) models for each variable pair and extracted the residuals for testing. If linearity was not satisfied, we used a flexible non-parametric method—locally weighted scatterplot smoothing (LOWESS)—to model the relationships and obtain residuals. The Benjamini and Yekutieli correction was again applied to control for false discovery under multiple testing.

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
(365, 8)	Time-series	False	False	False	True	False

3.2 Correlation Analysis

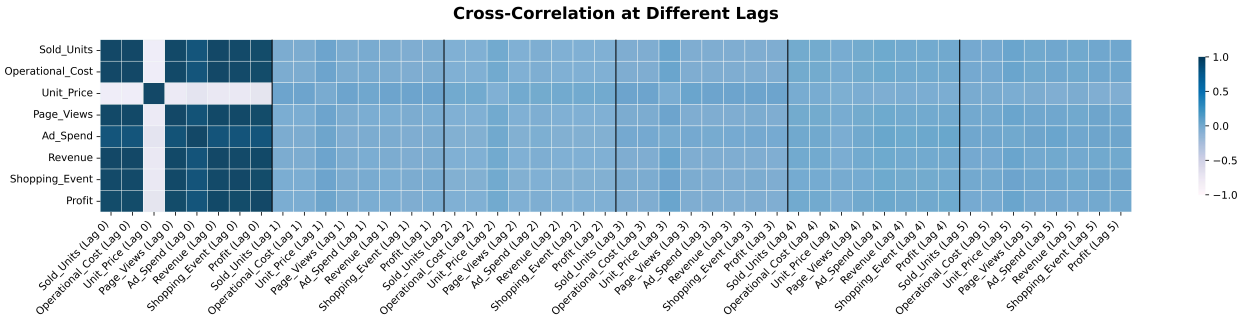


Figure 2: Heatmap of Time-Lagged Correlations Among Variables

We calculated both time-lagged and instantaneous correlation coefficients between variables. Here we list some variable pairs which have large absolute lagged correlation coefficients.

- Ad Spend → Profit at Lag 4: +0.064
- Unit Price → Ad Spend at Lag 5: -0.062
- Profit → Ad Spend at Lag 4: +0.059
- Unit Price → Profit at Lag 5: -0.058
- Unit Price → Shopping Event at Lag 5: -0.056

3.3 Time Series Stationarity Analysis

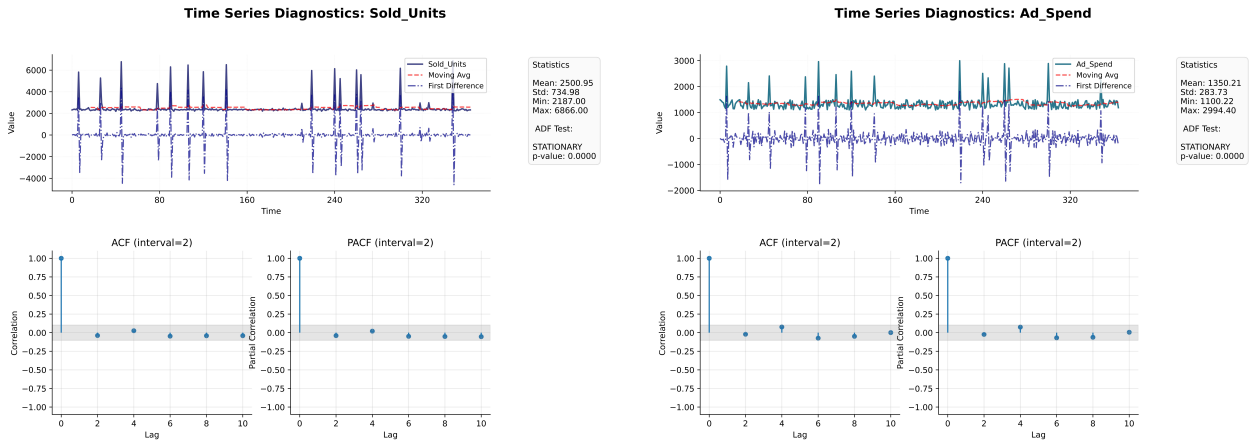


Figure 3: Time Series Diagnostics for two selected variables. Each plot shows the variable with moving average, first difference transform, and ACF/PACF plots.

4 Causal 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.

4.1 Data Preprocessing

In this initial step, we preprocessed the data and examined its statistical characteristics. This process involved data cleaning, handling missing values, and performing exploratory data analysis to examine variable distributions and inter-variable relationships.

4.2 Algorithm Recommendation assisted with LLM

Following 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:

- **PCMCI**:
  - **Description**: Peter and Clark algorithm with Momentary Conditional Independence, optimized for time-series data with flexible handling of noise and relationships.
  - **Justification**: PCMCI is specifically designed for time-series data, which matches the dataset's structure. It is efficient on CPU, handles non-linear relationships and non-Gaussian noise well, and provides a DAG output, which is acceptable to the user. Its strong empirical performance and scalability to medium-to-large-scale datasets make it a suitable choice for this dataset.
- **DYNOTEARS**:
  - **Description**: Dynamic Nonlinear Optimization with Trace Exponential and Augmented lagRangian for Structure learning, designed for time-series data with a focus on non-linear relationships.
  - **Justification**: DYNOTEARS is a continuous-optimization method that is well-suited for time-series data. It handles non-linear relationships and non-Gaussian noise effectively, and provides a DAG output. Its strong empirical performance and moderate efficiency on CPU make it a good fit for the dataset's characteristics, especially given the non-linear relationships and non-Gaussian errors.

Considering data properties, algorithm capability and user's instruction, the final algorithm we choose is PCMCI.

4.3 Hyperparameter Values Proposal assisted with LLM

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

- **Independence Test**:
  - **Value**: parcorr
  - **Explanation**: User specified
- **Minimum Time Lag**:
  - **Value**: 0
  - **Explanation**: Allows the algorithm to consider immediate causal effects, which is a common practice in time-series analysis.
- **Maximum Time Lag**:
  - **Value**: 10
  - **Explanation**: Captures a broad range of temporal dependencies while maintaining computational efficiency.
- **Significance Level of PC Algorithm**:
  - **Value**: 0.1
  - **Explanation**: Balances the risk of false positives with the need to detect causal relationships in a smaller sample size.
- **Significance Level for Graph Thresholding**:
  - **Value**: 0.05
  - **Explanation**: Provides a balanced approach to graph sparsity and accuracy, suitable for the dataset's characteristics.

4.4 Graph Tuning with Bootstrap and LLM Suggestion

In the final step, we performed graph tuning with suggestions provided by the Bootstrap and LLM.

We first applied the Bootstrap method to estimate the confidence level associated with each edge in the initial graph. Specifically:

- If an edge not present in the initial graph exhibited a Bootstrap confidence greater than 90%, we added it to the graph.
- Conversely, if an existing edge had a confidence lower than 10%, we removed it.
- For edges with moderate confidence (between 10% and 90%), we consulted the LLM to assess their validity and directionality, drawing on its extensive background knowledge.

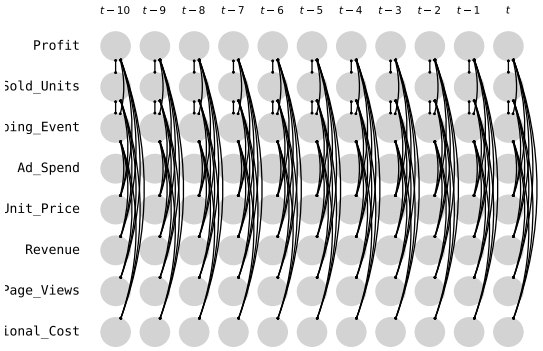
The LLM contributed by:

- Reintroducing plausible edges that may have been overlooked by statistical methods;
- Removing or redirecting edges that appeared statistically valid but were conceptually implausible.

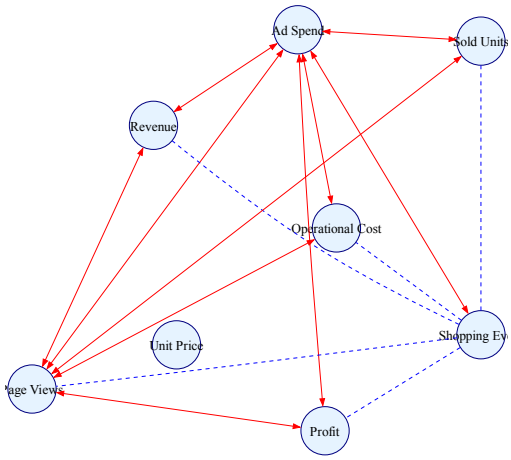
To improve the robustness of LLM-generated suggestions, we employed a voting mechanism. Importantly, LLM recommendations were not allowed to override high-confidence decisions made by the Bootstrap procedure. By integrating insights from both of Bootsratp and LLM to refine the causal graph, we can achieve improvements in graph's accuracy and robustness.

5 Causal Graph Estimation Summary

5.1 Causal Graph Discovered by the Algorithm



(a) Time Lag Graph Discovered by the Algorithm



(b) Summary Graph Discovered by the Algorithm. Solid lines represent causal edges identified by the algorithm, while dashed lines indicate strong correlations without inferred causality.

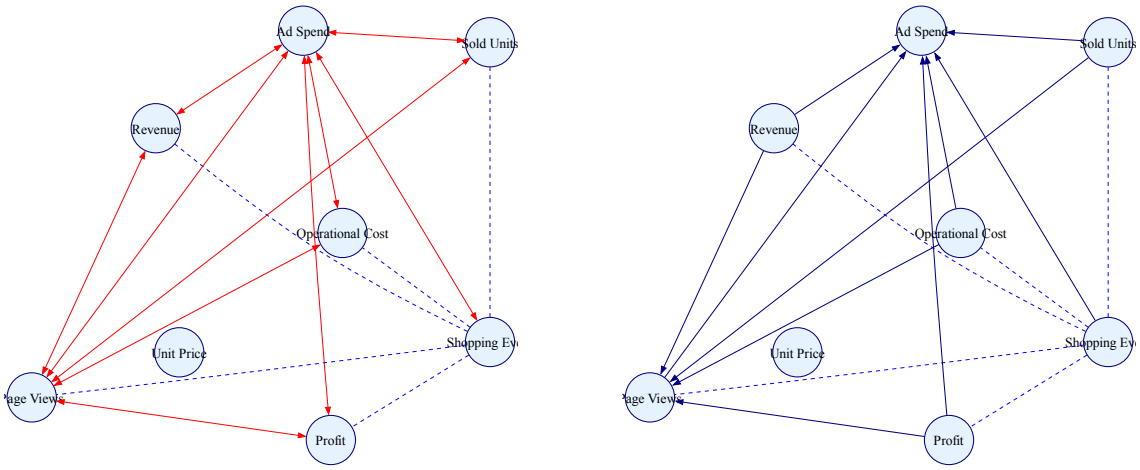
Figure 4: Graphs Discovered by the Algorithm.

The causal relationship analysis among the variables reveals a complex interplay where various factors influence each other in a potentially cyclical manner. Notably, Ad Spend has a pivotal role as it not only drives the occurrence of Shopping Events but is also influenced by a variety of other variables such as Revenue, Operational Cost, and Profit. This interdependence suggests a feedback loop where increased Ad Spend can enhance Page Views, Sold Units, and ultimately affect Revenue and Profit. However, changes in Sold Units and Revenue also stimulate Ad Spend, indicating a dynamic response system. Thus, the analysis can be summarized as follows:

- Ad Spend drives both Shopping Events and Page Views, indicating its role in generating customer engagement and sales.
- Shopping Events and Sold Units have reciprocal effects on Ad Spend, suggesting that promotions can lead to increased spending.
- Page Views, as a measure of online engagement, significantly influences Ad Spend, reinforcing the value of attracting potential customers.
- Revenue and Profit are both outcomes that affect and are affected by Ad Spend, indicating a relationship where spending affects financial performance and vice versa.
- Operational Costs show a similar pattern, impacting and being impacted by Ad Spend, highlighting the cost considerations that accompany advertising strategies.

In conclusion, the analysis underscores the necessity for a balanced approach to Ad Spend, where careful consideration of its reciprocal relationships with variables like Revenue and Operational Cost is vital for maximizing profitability and efficiency in advertising efforts.

5.2 Result Graph Comparision



(a) Initial Result Graph of PCMCI

(b) Initial Result Graph of VARLiNGAM

Figure 5: Result Graph Comparision of Different Algorithms

In comparing the causal graphs from the PCMCI and VARLiNGAM algorithms, we observe a number of differences in the directed edges identified. PCMCI identifies a larger set of causal relationships, including the bidirectional edge between Ad_Spend and Shopping_Event, highlighting a potential feedback loop. Additionally, PCMCI lists edges from Ad_Spend to multiple other variables such as Sold_Units, Revenue, Operational_Cost, and Profit, as well as from Page_Views to these variables. In contrast, VARLiNGAM captures a more streamlined view with unidirectional edges, particularly noting that Shopping_Event influences Ad_Spend, and it specifies edges from Sold_Units, Revenue, Operational_Cost, and Profit towards Page_Views, but does not identify the reciprocal effects or the extensive causal web seen in PCMCI.

Common edges between the two algorithms include the directed edges from Page_Views, Sold_Units, Revenue, Operational_Cost, and Profit towards Ad_Spend, which indicates a consensus on these relationships. Both algorithms agree on the influence of Page_Views and Sold_Units on Ad_Spend, along with the higher-level relationships of profit and operational costs feeding back into the advertising spend.

However, the edges identified by PCMCI, particularly those suggesting feedback loops such as the bidirectional edge between Ad_Spend and Shopping_Event, can be viewed as less reliable given that VARLiNGAM does not support such complexity and focuses instead on unidirectional influences. This suggests that VARLiNGAM may be incorporating stronger assumptions about the directionality of causality based on temporal ordering or other criteria, potentially leading to more stable and interpretable results. In contrast, the extensive and interconnected edges identified by PCMCI, while they offer richer information, may incorporate more uncertainty, making them less reliably indicative of causal direction. Thus, the edges from VARLiNGAM could be considered more reliable due to their consistency and simplicity, whereas PCMCI edges may reflect more complex and less certain causal dynamics.

5.3 Conclusion

In this comprehensive causal discovery report on an online shopping dataset from 2021, we explored significant variables such as Shopping_Event, Ad_Spend, Page_Views, Sold_Units, Revenue, Operational_Cost, and Profit to unveil complex interdependencies and feedback mechanisms influencing consumer behavior and business performance. Utilizing advanced methodologies including data preprocessing, algorithm recommendation aided by a large language model (LLM), and the PCMCI algorithm, our analysis revealed intricate causal relationships that culminate in cyclical dynamics, primarily highlighting Ad_Spend as both a driver and a dependent variable within the system.

The results indicate that Ad_Spend significantly influences Shopping Events and Page Views, while simultaneously being impacted by Sold Units and Revenue, establishing a feedback loop that is crucial for strategic decision-making.

Our findings contribute essential insights into the causal structure governing online shopping behaviors, emphasizing the need for careful management of Ad_Spend in relation to Operational Costs and Revenue to optimize profitability. This analysis not only advances our understanding of consumer engagement in e-commerce but also sets a foundation for future research leveraging causal discovery techniques to enhance operational strategies in online retail contexts.