Effective Time Series Forecasting supported by Causal Discovery :

An Application to Stock Market

Élèves :
Naël Briand
Quentin De La Chaise

Enseignants :
Vassilis Christophides
Son Vu

16 janvier 2025

Table des matières

1 INTRODUCTION

1.1 Project Context

The rapid advancements in artificial intelligence (AI) and machine learning (ML) have opened new opportunities in the domain of financial forecasting, particularly in predicting stock market behaviors. Accurate forecasts are crucial for investors, portfolio managers, and financial institutions aiming to optimize their strategies and mitigate risks. However, financial time series present significant challenges, including high volatility, complex interdependencies, and the influence of external factors such as macroeconomic events.

Traditional statistical methods, such as ARIMA and GARCH, while effective for stationary series, often struggle to capture the intricate relationships and temporal dynamics present in financial data. To address these limitations, modern AI-based models leverage data-driven approaches to extract patterns and relationships within the data. However, these models often prioritize correlation over causation, which can lead to suboptimal predictions and decision-making.

This project aims to bridge the gap between causality and prediction by integrating causal discovery methods into the time series forecasting pipeline. Specifically, the focus is on combining **causal discovery algorithms**, such as VAR-LiNGAM and PCMCI, with deep learning models (LSTM) to identify meaningful features and improve predictive accuracy. By doing so, we not only aim to enhance the forecasting performance but also provide interpretable insights into the causal relationships driving stock market behaviors.

1.2 Definition of the Problem Studied

Forecasting financial time series, such as stock prices, is a notoriously difficult task. Traditional models often struggle due to several intrinsic and external challenges :

- **Distribution shifts** : Sudden changes in market dynamics caused by major events, such as the 2007-2008 financial crisis or the COVID-19 pandemic (2020-2021), disrupt established patterns and render models less effective.
- **Excessive parameters influencing stock prices** : Stock prices depend on a multitude of factors, including economic conditions, geopolitical events, weather, and even specific calendar dates. This complexity introduces variability and unpredictability.
- **Correlation issues** :
  - **Variation of correlation** : The relationships between variables change over time, making models based on fixed correlations unreliable.
  - **Spurious correlations** : Models may overfit to meaningless correlations that do not represent causal relationships, leading to poor generalization.
- **Noise and randomness** : Financial time series are inherently noisy, with a significant portion of their variability attributed to random, unpredictable events.

- **Model limitations** : Statistical models like ARIMA and GARCH often assume stationarity and linearity, which are inadequate for capturing the complex dynamics of financial data. Similarly, machine learning models like LSTMs excel at pattern recognition but fail to differentiate between correlation and causation, making them less interpretable and robust.

The key challenge lies in addressing the following question :

How can we make these models more robust to distribution shifts while being sensitive to a broader range of influencing parameters ?

To address this challenge, this project explores the integration of causality into forecasting models. By prioritizing **causation over correlation**, we aim to build models that are both robust and interpretable. Specifically, our approach focuses on :

1. **Feature selection** : Leveraging causal discovery methods to identify the most relevant features and eliminate spurious correlations.
2. **Integration of causality into machine learning models** : Developing a pipeline where causal discovery enhances feature selection, and models like LSTMs use these features to improve prediction accuracy and interpretability.
3. **Practical application to stock market data** : Demonstrating the feasibility and benefits of this approach through real-world stock price forecasting.

By addressing these objectives, we seek not only to improve predictive performance but also to provide insights into the causal relationships driving stock market behaviors, paving the way for more robust decision-making.

1.3 Definition of subject terms

For the purposes of this project, a number of key concepts relating to financial time series need to be defined for a better understanding of the subject :

- **Time series** (financial) : A time series is a sequence of data observed and collected at regular intervals over time. In the financial context, these series represent variations in variables such as share prices, trading volumes or economic indicators.
- **Multivariate** : A multivariate time series comprises several interconnected variables that change over time. For example, when analysing shares, variables such as opening and closing prices, highs and lows, or volumes traded, can be studied together to understand the interactions between them.
- **Instantaneous** : This term refers to causal or correlative relationships observed at a given moment in time, without taking time lags into account. For example, the closing price of a share at a given moment may depend on the opening price at the same moment.
- **Acyclic** : A time series is said to be acyclic when it has no causal loops. This means that an event A influences an event B, but the latter does not return to influence A, thus eliminating cycles. For example, a causal sequence such as Open(t) $\rightarrow$

Close(t) → Open(t) constitutes a cyclic loop that does not conform to the acyclicity assumption.

- **Non-Gaussian distribution** : Financial time series are often characterised by a distribution of data that deviates from the normal (Gaussian) curve. This reflects extreme behaviour, such as sudden variations or rare events, which cannot be modelled by conventional distributions.
- **Discrete time series** : A discrete time series is made up of observations collected at well-defined time intervals (e.g. daily, monthly). Unlike continuous series, discrete data makes it easier to process mathematically and statistically within a numerical framework.

2 METHODOLOGY

2.1 Causal discovery method

Causal discovery is an essential approach to data analysis, particularly in complex areas such as finance, where multiple variables interact. It aims to identify cause-and-effect relationships beyond simple statistical correlations, enabling a better understanding of underlying dynamics and more informed decision-making.

Causality VS Correlation

It is important to make a clear distinction between causality and correlation :

- **Causality** : A causal relationship implies that an event A directly causes a change in an event B. For example, an increase in interest rates may lead to a decrease in investment, because the cost of credit becomes higher. This link reflects a direct and verifiable mechanism.
- **Correlation** : A correlation, on the other hand, is simply a statistical association between two variables that appear to vary together. However, this relationship does not necessarily mean that there is a causal link. For example, a correlation could be observed between the size of an individual's feet and their spelling skills. This apparent link is misleading : the real common explanatory variable here is age. Children with small feet are generally less advanced in spelling, whereas adults with larger feet are more proficient in this skill. This situation illustrates what is known as a spurious correlation, where the relationship between the two variables studied is due to a third factor not taken into account.

These spurious correlations are particularly problematic in the analysis of large and complex data, where misleading relationships can appear simply by chance. This reinforces the need for robust tools to distinguish truly causal relationships from purely statistical associations.

To overcome these limitations, several causal discovery models have been developed, each with its own specific advantages. Among them, **Multivariate Granger Causality** analyses whether one variable predicts another in a time series, thereby capturing temporal causality. **Dynotears** is an approach optimised for dynamic and complex causal

relationships. However, two tools in particular stand out in the context of financial data : **VAR-LiNGAM** and **PCMCI**.

In this study, we will use VAR-LiNGAM and PCMCI. VAR-LiNGAM was chosen because of its ability to model linear causal relationships and to identify structural dependencies in financial series that are often influenced by exogenous factors. As for PCMCI, its robustness in the face of false positives and its effective management of auto-correlations make it particularly well suited to exploring complex causalities in variable-rich datasets. Together, these two tools will maximise the accuracy and robustness of predictive models, while reducing the biases associated with spurious correlations.

2.2 VAR-LiNGAM method

2.2.1 Definition

VAR-LiNGAM (Vector AutoRegressive Linear Non-Gaussian Acyclic Model) is a hybrid method designed to indentify complex causal relationships in multivariate time series. It combines two complementary approaches, each specialized in a particular type of dependency : VAR and LiNGAM.

VAR (Vector AutoRegressive) model is a statistical method used to model and analyze multivariate time series. It is particularly useful for studying dynamic relationships between several interconnected time variables. This is a generalization of the autoregressive model (AR) for several time variables. It assumes that each variable in a time series depends not only on its own past values, but also on the past values of the other variables in the system.

LiNGAM (Linear Non-Gaussian Acyclic Model) is a statistical method used to identify causal relationships in systems where variables follow an acyclic and linear structure, while respecting an assumption of non-Gaussianity of residuals. Unlike traditional models, such as VAR, which focus primarily on correlations, LiNGAM aims to infer causality by exploiting specific properties of non-Gaussian distributions.

2.2.2 Model assumptions

Linearity : The model assumes that the relationships between variables are linear, whether they are temporal dependencies in the VAR or contemporaneous relationships in the LiNGAM. This simplifies the modelling and allows a clear analytical representation, but limits the application to systems where non-linear relationships are negligible. The residuals must be analysed to check that there is no systematic structure.

Non-Gaussianity : The contemporaneous residuals, E_t, must follow a non-Gaussian distribution. This assumption is essential if LiNGAM is to correctly identify contemporary causal relationships, as Gaussian distributions render these relationships indeterminate. It can be verified by normality tests (Shapiro-Wilk, Kolmogorov-Smirnov).

Acyclicity : Contemporary causal relationships must be acyclic, i.e. without direct feedback within the same period. This ensures that the relationships can be represented by an acyclic directed graph.

Absence of latent variables : All relevant variables must be observed. The presence of latent, unobserved variables could bias the causal relationships identified. Although difficult to verify directly, methods such as LiNGAM with latent variables or principal component analysis can help to detect hidden influences.

2.2.3 Operating principle

VAR-LiNGAM works in two stages : The VAR model is used to explain each variable of interest in terms of the past values of all the variables. The equation for n variables $X_1, X_2, ..., X_n$ is :

$$X_t = \sum_{p=1}^{n} B_p x(t-p) + \epsilon(t)$$

Where,

- x(t) : Vector of variables observed at time t.
- B_p : Matrix of coefficients describing the effect of variables at time t-p on variables at t.

$$b_{i,j}^{(p)}, \text{ coefficient of } B_p, \text{ represent the influence of } X_j(t-p) \text{ sur } X_i(t)$$

- $\epsilon(t)$: Vector of residual error terms at t, assumed to be non-Gaussian, independent and without common causes.

Then we apply the LiNGAM model. Once the $\epsilon(t)$ extracted from the VAR model, the LiNGAM model is applied to identify causal relationships between variables. LiNGAM assumes that e follows a linear, non-Gaussian acyclic structure :

$$\epsilon(t) = B\epsilon(t) + \eta(t)$$

Where,

- B : Matrix of causal relationships between contemporary variables.
- η_t : Independent, non-Gaussian error term.

2.2.4 Strengths and Limitations

Strengths

- **Comprehensive Causal Analysis** : VAR-LiNGAM effectively combines temporal analysis (VAR) and contemporaneous causal inference (LiNGAM), allowing it to model both lagged and instantaneous relationships in multivariate time series.
- **Causal Inference Beyond Correlation** : Unlike traditional models focused on correlation, VAR-LiNGAM identifies true causal relationships by leveraging non-Gaussianity and acyclicity. This makes it particularly valuable in systems where spurious correlations can lead to incorrect conclusions.

- **Adaptability to Financial and Economic Data** : Many financial datasets exhibit non-Gaussian distributions and complex interdependencies. VAR-LiNGAM is well-suited for such contexts, providing insights into the dynamics of highly interconnected variables.
- **Interpretability** : The method produces directed acyclic graphs (DAGs) that represent causal relationships between variables. These visualizations enhance the interpretability of results, making them accessible to both technical and non-technical stakeholders.
- **Validation of Assumptions** : The model includes tools to test its key assumptions, such as the Shapiro-Wilk or Kolmogorov-Smirnov tests for non-Gaussianity and checks for acyclicity. This adds rigor to the analysis and ensures the reliability of causal inferences.

Limitations

- **Linearity Assumption** : VAR-LiNGAM assumes that relationships between variables are linear, which simplifies modeling but limits its applicability to systems with significant non-linear interactions, such as certain biological or financial processes.
- **Dependence on Non-Gaussianity** : The method relies on the residuals being non-Gaussian to identify contemporaneous causal relationships. If the residuals are close to Gaussian, the model's ability to infer causality diminishes, making it unsuitable for certain datasets.
- **Acyclicity Constraint** : VAR-LiNGAM assumes that contemporaneous relationships are acyclic, which excludes feedback loops within the same time period. While this simplifies the causal graph, it can oversimplify systems where such feedback exists.
- **Sensitivity to Latent Variables** : The method requires all relevant variables to be observed. Unobserved or latent variables can bias the causal relationships identified, potentially leading to incomplete or inaccurate models.
- **Computational Complexity** : As the number of variables and time lags increases, the computational demands grow significantly, making the method less practical for large-scale or high-dimensional datasets.
- **Data Quality Dependence** : VAR-LiNGAM is sensitive to missing values, noise, and non-stationarity. Careful preprocessing and data validation are essential to ensure the accuracy of results, which can increase the time and effort required for analysis.

2.3 PCMCI Method

2.3.1 Definition

PCMCI (Peter and Clark Momentary Conditional Independence) is a causal discovery algorithm tailored for time series data. It combines the strengths of the PC

algorithm and Momentary Conditional Independence (MCI) tests to identify both lagged and contemporaneous causal relationships. PCMCI is particularly effective in handling high-dimensional datasets and time series with complex dependencies.

PC Algorithm is a constraint-based causal discovery method that identifies an initial causal graph. It iteratively performs conditional independence tests to remove statistically insignificant edges (potential causal relationships) while respecting temporal constraints. For time series data, the PC algorithm adapts to include lagged dependencies.

Momentary Conditional Independence (MCI) tests refine the initial graph by evaluating the significance of causal links at specific time lags. Using p-values, MCI tests assess whether a causal relationship exists while controlling for other variables and lags, ensuring that spurious correlations are minimized.

2.3.2 Model assumptions

Causal sufficiency : The algorithm assumes that all relevant variables are observed. However, PCMCI can help detect hidden variables by identifying patterns of dependency that cannot be explained by the observed data alone.

Linearity : PCMCI assumes that the relationships between variables are linear. This allows for straightforward statistical testing but may limit the model's applicability to systems with strong nonlinear dynamics.

Stationarity : The algorithm assumes that the statistical properties of the time series do not change over time. This assumption is common in many time series analysis methods but can be relaxed with advanced adaptations.

Control of p-values : PCMCI uses p-values from statistical tests to determine the significance of causal relationships. These p-values are often adjusted using methods such as the False Discovery Rate (FDR) to minimize the risk of false positives.

2.3.3 Operating principle

PCMCI works in two main stages :

1. **PC algorithm** : The algorithm identifies an initial causal structure by testing conditional independence. This step focuses on removing unnecessary edges (potential causal links) based on statistical evidence, producing a preliminary causal graph.
2. **MCI tests** : The MCI tests refine the graph by analyzing the significance of each causal link at specific time lags. These tests condition on other variables and lags to account for temporal dependencies and reduce spurious correlations.

2.3.4 Strengths and Limitations

Strengths :

- Capable of identifying both lagged and contemporaneous causal relationships in time series data.
- Controls false positive rates by leveraging p-values and corrections like FDR.

- Can detect hidden variables by analyzing patterns of unexplained dependencies.
- Well-suited for high-dimensional datasets, where the number of variables may exceed the number of observations.

Limitations :

- Computationally intensive for large datasets with many variables and long time lags.
- Assumes linearity and stationarity, which may not hold for all real-world datasets.
- Sensitive to the choice of conditional independence test and statistical thresholds.

2.4 Forecasting Models

2.4.1 Random Forest

Random Forest (RF) is an ensemble learning method based on decision trees. It constructs multiple decision trees during training and outputs the mean prediction of the individual trees for regression tasks. Its key features include :

- **Strengths :** Random Forest excels in handling high-dimensional datasets, capturing non-linear relationships, and providing robust predictions with minimal overfitting.
- **Limitations :** However, RF does not inherently account for temporal dependencies in time series data, as it treats each data point independently.

2.4.2 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) specifically designed to handle sequential data. They are capable of learning long-term dependencies by maintaining a memory cell that preserves context over time. Key characteristics of LSTM include :

- **Strengths :** LSTMs are particularly suited for time series forecasting as they capture temporal patterns, trends, and seasonality in the data.
- **Memory Mechanism :** The memory cell and gating mechanisms allow LSTMs to selectively retain or forget information, enabling them to adapt to varying time horizons.
- **Robustness to Temporal Shifts :** LSTMs are effective in scenarios with shifting data distributions, such as those observed in financial crises or volatile markets.

2.4.3 Transition from Random Forest to LSTM

The transition from Random Forest to LSTM was motivated by the sequential nature of financial time series data.

Key Reasons for Transition :

- **Temporal Data Suitability :** Financial time series data exhibit sequential dependencies, trends, and patterns that Random Forest cannot fully capture.

- **Forecasting Accuracy :** LSTM's ability to model temporal dynamics leads to improved forecasting accuracy, especially for longer prediction horizons.
- **Scalability :** LSTM can scale to handle more complex patterns in the data, such as seasonality and lagged dependencies, which are crucial in financial forecasting.

The transition highlights the importance of selecting models that align with the inherent properties of the data. By leveraging LSTM's strengths, the study aims to achieve more accurate and robust forecasts.

3 DATA DESCRIPTION

Dataset description The dataset used in this study, entitled Stock Chart, contains discrete multivariate time series from Yahoo Finance. It covers a period from 1 January 2010 to 31 December 2023, with a total of 3,522 daily observations. The dataset contains 20 time series, representing various financial metrics related to the stock market performance of major companies such as Apple, Google, Microsoft and Amazon. The data is expressed in US dollars (USD).

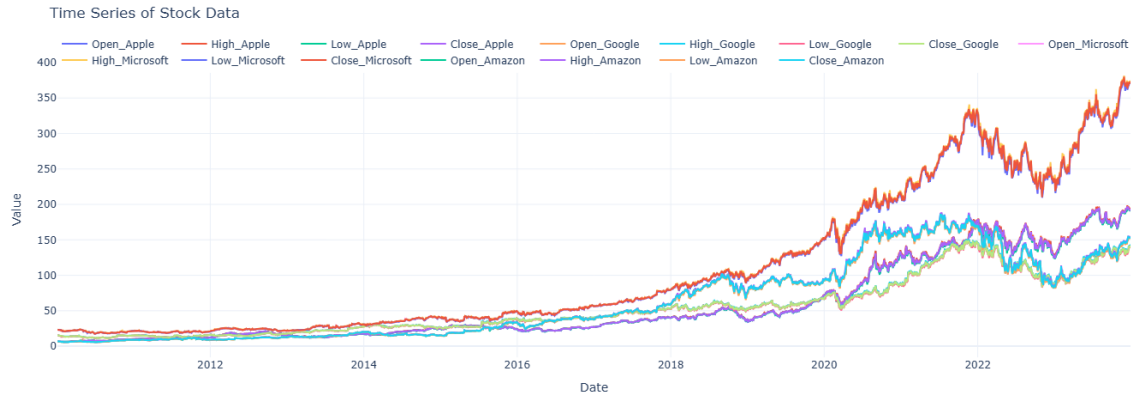

FIGURE 1 – Stock Chart

The variables studied include :

- **Open** : Opening price of shares at the start of the day.
- **High** : Highest share price during the day.
- **Low** : Lowest share price during the day.
- **Close** :Closing share price at the end of the day.
- **Volume** : Total volume of shares traded during the day.

The target time series for the prediction is Close_Google, representing the daily closing price of Google shares. This variable will serve as the main dependent variable in the analysis.

The prediction horizon is set at 1 month, with the aim of predicting the future values of Close_Google over this period. The potential explanatory variables include all the observed metrics (Open, High, Low, Close and Volume) of the other stocks, excluding Close_Google itself in order to avoid any bias or data leakage.

The dataset covers a period rich in major economic events, offering a vision of recent economic cycles and their impact on financial markets :

- **Post-2008 financial crisis** : The sample begins shortly after the global financial crisis, a period marked by a gradual market recovery and increased volatility.
- **Inflation** : The period studied includes several phases of inflation, notably the rise in inflation following the COVID-19 pandemic, which had a major impact on investor decisions and asset valuations.
- **Economic cycles** : The dataset captures several economic cycles, including phases of growth, recession and economic slowdowns due to global events (pandemics, geopolitical tensions).
- **COVID-19 pandemic** : In early 2020, markets experienced historic swings due to global uncertainty, followed by a rapid recovery driven by accommodative monetary policies and economic stimulus.

4 EXPERIMENTAL ENVIRONMENT

4.1 Objective and Hypothesis

The main objective of the study is to evaluate whether integrating causality into forecasting models leads to more robust and accurate predictions by capturing the fundamental relationships between variables. Specifically, the hypothesis posits that models incorporating causal information outperform traditional models in terms of prediction accuracy and robustness to distributional changes.

4.2 Pipeline Overview

The experimental framework is divided into three main steps, as illustrated in Figure 1 :

1. **Causal Discovery :** Identification of causal relationships between variables using VAR-LiNGAM and PCMCI. This step produces a causal graph and a table of causality scores, which serve as the foundation for feature selection.
2. **Causality-Based Forecasting :** Selection of causally relevant features to improve forecasting models. The primary idea is to leverage causally significant features to guide the prediction process. Two prediction approaches are explored :
    - **Baseline Models :** Random Forest, a traditional machine learning algorithm known for its robustness and interpretability.

- **Advanced Models :** Long Short-Term Memory (LSTM), a deep learning model specifically designed for sequential data.

3. **Evaluation and Comparison :** Assessment of the models' performance evaluation under different configurations to measure the contribution of our methodology using standard error metrics on forecasting results :
   - **Mean Squared Error (MSE) :** Measures the average squared difference between predicted and actual values.
   - **Mean Absolute Error (MAE) :** Evaluates the average magnitude of errors, offering a more interpretable metric for practitioners.

4.3 Data Source and Preprocessing

The data used in this study are financial time series obtained from Yahoo Finance. The preprocessing pipeline includes :

- **Data Extraction :** Retrieval of historical data for selected financial instruments and relevant macroeconomic indicators.
- **Data Cleaning :** Handling of missing values, outliers, and inconsistencies in the time series data.
- **Feature Engineering :** Creation of lagged variables, normalization, and transformation of variables to ensure stationarity.

Additionally, the preprocessed data are split into training and testing datasets to evaluate the forecasting models under realistic conditions.

4.4 Experimental Configurations

The experiments are conducted under the following configurations :

- **With/Without Causality :** Models are evaluated both with and without the integration of causality-driven feature selection to assess the added value of causal information.
- **Temporal Horizons :** Short-term and long-term forecasting horizons are tested to examine the robustness of the models under different prediction timeframes.

4.5 Software and Tools

The experimental framework is implemented using Python, leveraging the following libraries and tools :

- **Causal Discovery :** tigramite (for PCMCI) and custom implementations for VAR-LiNGAM.
- **Prediction Models :** scikit-learn (Random Forest) and TensorFlow/Keras (LSTM).
- **Data Preprocessing :** pandas, numpy, and statsmodels.
- **Evaluation Metrics :** sklearn.metrics for error computation.

4.6 VAR-LiNGAM Experimentation

4.6.1 Parameter Settings for VAR-LiNGAM

Experimentation with the VAR-LiNGAM model is based on a specific configuration of parameters in order to optimise the identification of causal relationships in the multivariate time series studied. The parameters are defined as follows :

- **Lags** : The model was configured with a maximum lag of 30 days. This parameter allows the model to capture the temporal dependencies between variables over an extended period, taking into account the effect of past values up to 30 days on current observations.
- **Criterion** : The criterion used to optimise the model is the BIC (Bayesian Information Criterion). This criterion is particularly suitable for selecting one model from several, penalising complex configurations while favouring those that effectively explain the observed data. It guarantees a balance between goodness of fit and parsimony.

4.6.2 Causal Analysis Results

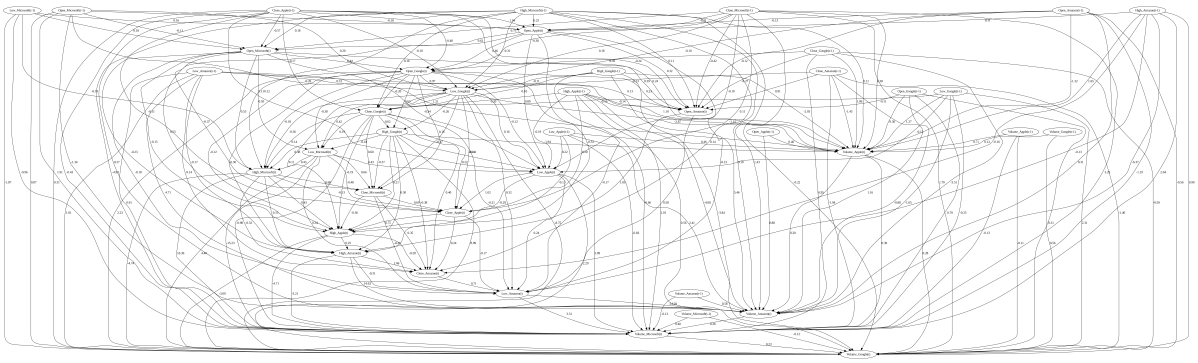

FIGURE 2 – Causal graph resulting from VAR-LiNGAM training

This network reveals complex interactions between different financial time series, highlighting direct and indirect influences between variables. However, this global graph can be difficult to interpret due to its density.

Therefore, in the following figure, we will isolate and represent only the causal relationships involving Close_Google. This will give a better understanding of which variables have a direct or indirect influence on the target series, while simplifying the analysis for better interpretation.

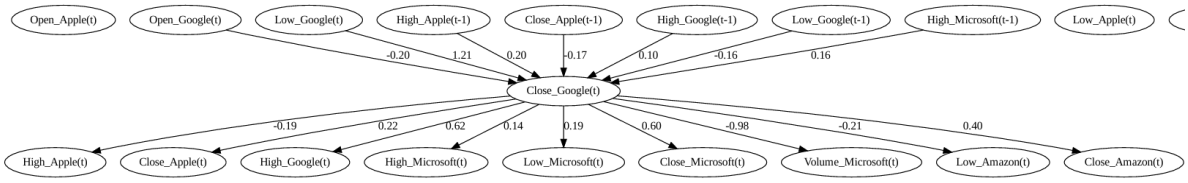

FIGURE 3 – Graph showing only causal relationships Close_Google

The graph presented illustrates the direct and indirect causal relationships involving the target variable Close_Google(t), identified by the VAR-LiNGAM model. These relationships are represented by directed arrows, with coefficients indicating the intensity of influence between the variables.

On the basis of this causal analysis, the following variables were selected as explanatory features for the prediction of Close_Google :

- High_Apple
- Close_Apple
- High_Google
- Low_Google
- High_Microsoft

Variables such as Open_Google(t) and Low_Google(t), although related to Close_Google, have not been selected due to the risk of **data leakage**. Their inclusion could lead to biased results by giving the model an artificial advantage (over-optimisation), which would compromise the robustness and generalisability of the predictions.

4.6.3 Checking the validity of the initial hypotheses

In order to guarantee the validity of the results obtained with the VAR-LiNGAM model, several fundamental assumptions made with VAR-LiNGAM must be verified.

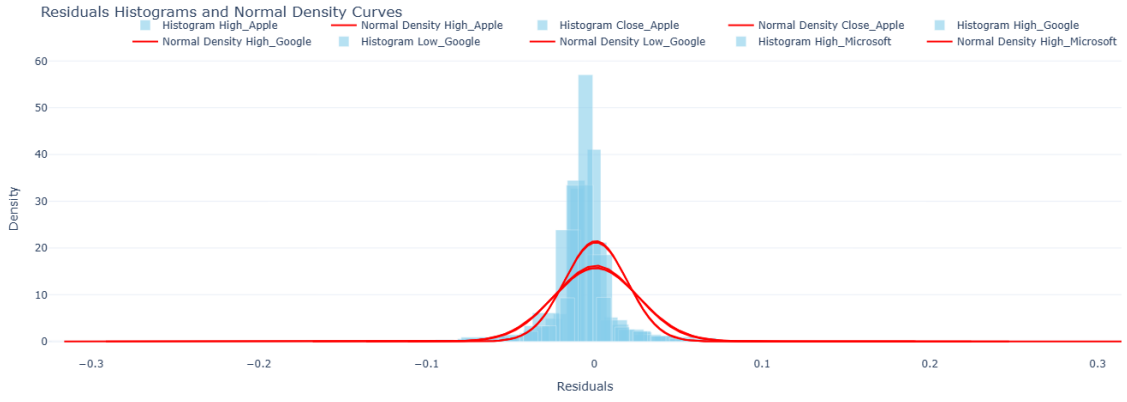

FIGURE 4 – Residuals Histograms and Normal Density Curves

The histogram of the residuals, together with the normal density curves, shows that they do not conform to a Gaussian distribution. This result verifies the hypothesis of non-Gaussianity.

Normality Test Results

Variable	Shapiro-Wilk Statistic	Shapiro-Wilk p-value	Kolmogorov-Smirnov Statistic	Kolmogorov-Smirnov p-value
High_Apple	0.7587	0.0000	0.2078	0.0000
Close_Apple	0.7826	0.0000	0.1945	0.0000
High_Google	0.7326	0.0000	0.1922	0.0000
Low_Google	0.7671	0.0000	0.1837	0.0000
High_Microsoft	0.7432	0.0000	0.2124	0.0000

FIGURE 5 – Results of normality tests

After observing the histograms of the residuals, we carried out normality tests to validate the hypothesis of non-Gaussianity, a key point in guaranteeing the reliability of the model. The results show that the p-values of the Shapiro-Wilk and Kolmogorov-Smirnov tests are all less than 0.05, leading to a systematic rejection of the hypothesis of normality of the residuals. These results confirm that the distributions of the residuals differ significantly from those of a Gaussian distribution, thus validating the non-Gaussianity hypothesis.

The other assumptions of the VAR-LiNGAM model are also validated. The linearity assumption is respected, as the relationships between variables, whether temporal or contemporaneous, are well modelled by linear dependencies with no systematic residuals. Acyclicity is guaranteed by the structure of the model, which imposes the absence of direct feedback in the contemporaneous relationships, resulting in a directed and acyclic causal graph. Finally, the absence of latent variables is ensured by the nature of the dataset, where all relevant variables have been observed and included in the model, minimising the risk of hidden influences. These checks confirm that the model meets the necessary conditions for a reliable causal analysis.

4.6.4 Model Performance Comparison

With Causality

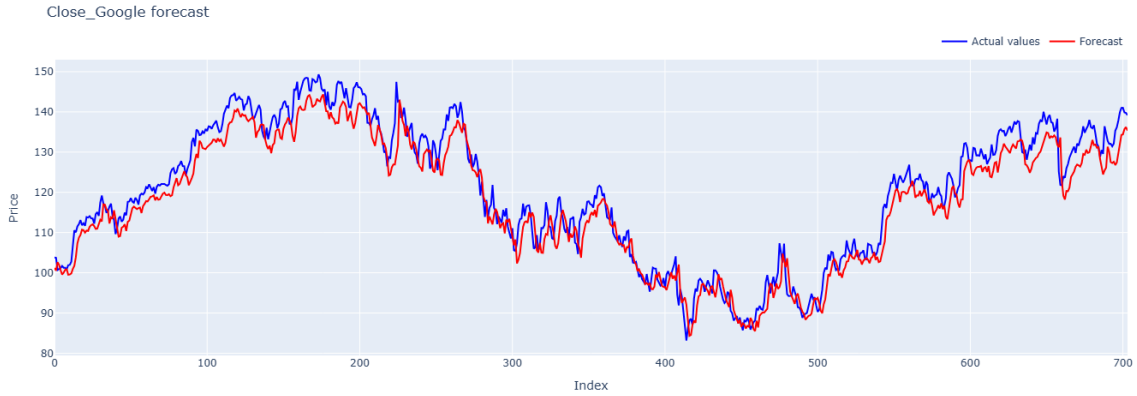

FIGURE 6 – Forecasting with causality

Training an LSTM model combined with causal discovery using VAR-LiNGAM effectively predicts Close_Google values. The graph shows a strong correspondence between the actual values and the predicted values, indicating the accuracy of the model.

With a root mean square error (RMSE) of less than 4 dollars, the model shows an ability to capture the complex dynamics of financial time series. This low margin of error underlines the effectiveness of integrating the causal relationships identified by VAR-LiNGAM into the explanatory variable selection process.

FIGURE 7 – Forecasting without causality

The training of an LSTM model without causal discovery was carried out by selecting features solely on the basis of their apparent correlation with the Close_Google target, without taking account of the causal relationships identified. The variables chosen, namely Low_Apple(t-1), Open_Microsoft(t-1), Open_Amazon(t-1), High_Amazon(t-1) and Low_Microsoft(t-1), were selected because they have no direct causal relationship with Close_Google, as shown in the causal graph (Figure 3). These features come from the right-hand side of the graph, where relationships with Close_Google are absent.

This approach, although consistent in a classical selection framework, shows limited performance. The graph reveals a significant divergence between the actual values and the predictions, with a root mean square error of 22 dollars. This large error contrasts with that of the model incorporating causal discovery, which achieves an RMSE of 3.77 dollars, representing an 82 percentage improvement.

These results highlight that the absence of causal relationships in the selection of explanatory variables significantly compromises the accuracy of the model. The inclusion of causal relationships is therefore essential to reduce bias, even when apparent correlations seem promising.

4.7 PCMCI Experimentation

4.7.1 Parameter Settings for PCMCI

The PCMCI method was applied to analyze causal relationships within the financial dataset. The key parameters used for the experiment are as follows :

- **Maximum Lag (τ_{max}) :** 30 days. This value defines the maximum delay between variables considered for causal relationships.
- **Dependency Metric :** *ParCorr* (Partial Correlation) with an analytic approach. ParCorr measures linear dependencies between variables, adjusted for the influence of other variables.
- **Significance Threshold (p-value) :** 0.05. This threshold determines whether a causal link is statistically significant.

4.7.2 Causal Analysis Results

The PCMCI analysis identified a reduced subset of features that were deemed significant for causal inference. Starting with an initial set of 20 features, the following 5 features were retained after applying PCMCI :

- **Low_Apple :** Represents the lowest stock price of Apple during the analyzed period.
- **Open_Microsoft :** Indicates the opening stock price of Microsoft.
- **Open_Amazon :** Represents the opening stock price of Amazon.
- **High_Amazon :** Reflects the highest stock price of Amazon.
- **Low_Microsoft :** Indicates the lowest stock price of Microsoft.

This reduction in features highlights the ability of PCMCI to filter out irrelevant variables, focusing on those with the most significant causal impact. Notably, volume-related variables were excluded from the final selection.

4.7.3 Observed Limitations of PCMCI

While PCMCI demonstrated effectiveness in reducing the feature set and identifying causal relationships, certain limitations were observed :

- **Increased Complexity :** The computational complexity of PCMCI increases significantly with the number of variables and the maximum lag (τ_{max}). This poses challenges for large datasets with high-dimensional variables or long time lags.

4.7.4 Model Performance Comparison

The impact of applying PCMCI on model performance was evaluated through 2 dataset. Firstly we try with a simple univariate dataset (Dataset 1) and then with a multivariate one (Dataset 2) to show the dependencies. Results were compared for models trained with and without PCMCI-based feature selection and are in percentage due to normalization.

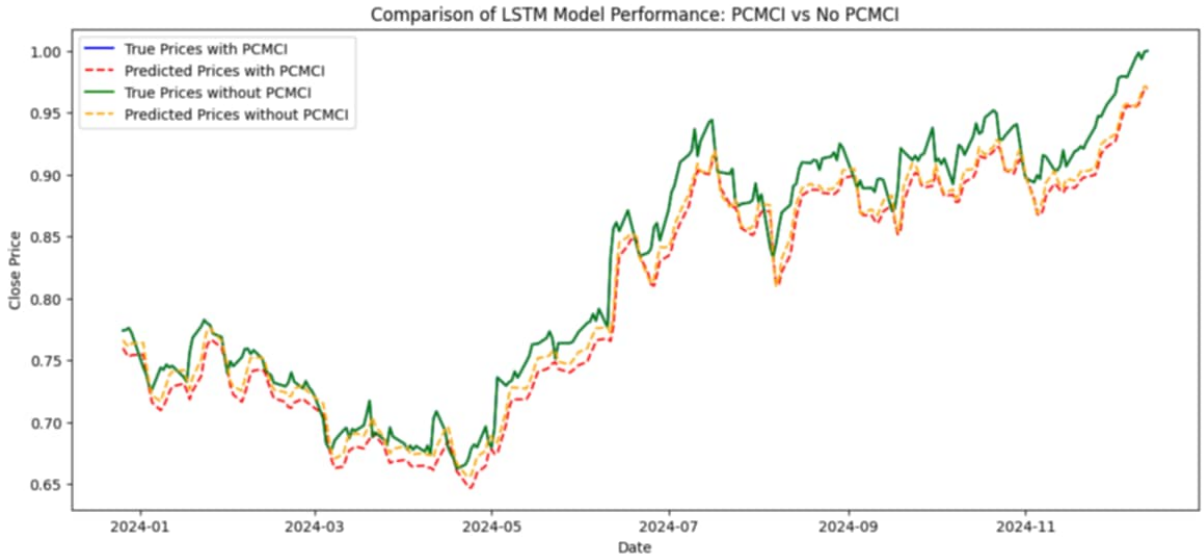

FIGURE 8 – Univariate series

Results with Dataset 1 :

- **With PCMCI :**
  - MAE : 0.017365
  - RMSE : 0.020349
- **Without PCMCI :**
  - MAE : 0.019173
  - RMSE : 0.023842

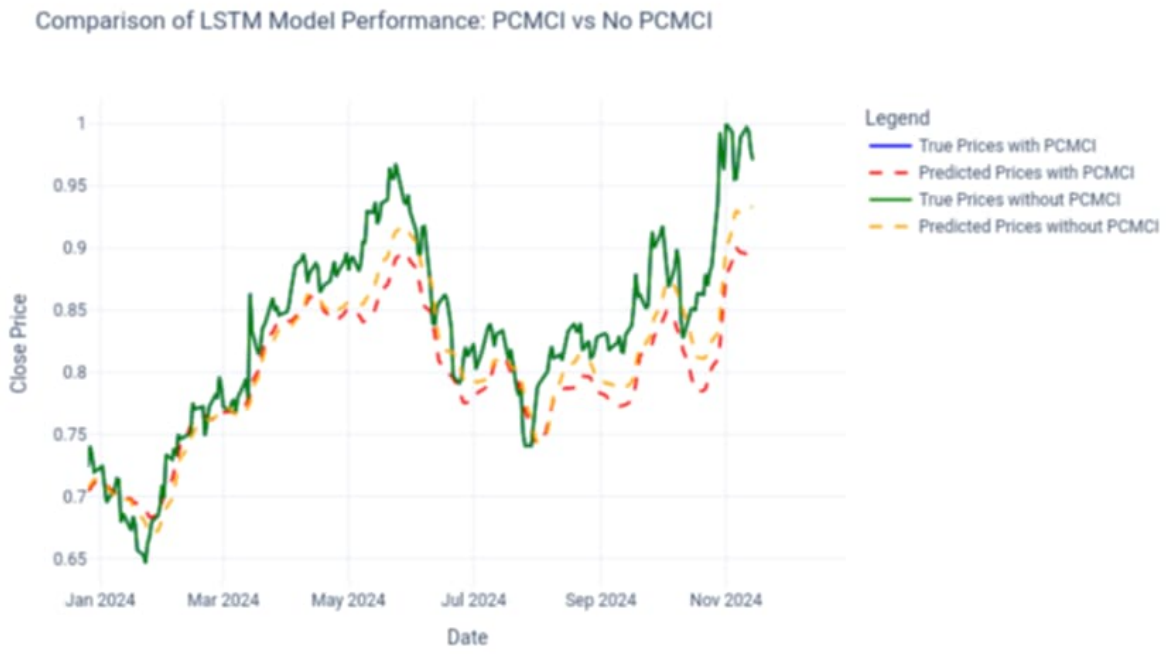

FIGURE 9 – Multivariate series

Results with Dataset 2 :

- **With PCMCI :**
  - MAE : 0.018412
  - RMSE : 0.023992
- **Without PCMCI :**
  - MAE : 0.026074
  - RMSE : 0.032915

The results indicate that applying PCMCI improves forecasting accuracy by reducing both MAE and RMSE. This demonstrates the effectiveness of causal feature selection in enhancing model performance, particularly in complex financial datasets even if with a long horizon like 30 days the difference is minimal (around 1 percent of improvement).

5 RESULTS

The experimental results demonstrate the effectiveness of integrating causal inference methods, into financial forecasting models. The analysis focuses on the following aspects : model accuracy, impact of causal feature selection, and comparative performance between random forest and LSTM.

5.1 Impact of VAR-LiNGAM and PCMCI on Feature Selection

Applying VAR-LiNGAM and PCMCI to the dataset resulted in a significant reduction in the number of features from 20 to 5. This reduction highlights the method's ability to identify causally relevant variables while excluding irrelevant ones.

This feature selection led to an improvement in the model's predictive performance, as seen in the reduced MAE and RMSE values compared to models without causal discovery. The following subsections provide detailed results.

5.2 Model Performance with VAR-LiNGAM

Table 2 compares the model performance with and without VAR-LiNGAM-based feature selection for two datasets.

TABLE 1 – Comparison of Model Performance with and without VAR-LiNGAM

Aspect	**With VAR-LiNGAM**	**Without VAR-LiNGAM**
Number of Features	5 (causally relevant)	5 (correlation-based)
RMSE	3.77 dollars	22 dollars
Robustness	Eliminates spurious correlations	Sensitive to variations
Performance Improvement	82% improvement in RMSE	Baseline with lower performance

The results show that models trained on VAR-LiNGAM-selected features consistently outperform those trained on the full feature set. This highlights the importance of causal inference in improving prediction accuracy.

5.3 Model Performance with PCMCI

Table 2 compares the model performance with and without PCMCI-based feature selection for two datasets.

TABLE 2 – Comparison of Model Performance with and without PCMCI

Dataset	**Feature Selection**	**MAE**	**RMSE**
Dataset 1	With PCMCI	0.017365	0.020349
	Without PCMCI	0.019173	0.023842
Dataset 2	With PCMCI	0.018412	0.023992
	Without PCMCI	0.026074	0.032915

The results show that models trained on PCMCI-selected features consistently outperform those trained on the full feature set. This highlights the importance of causal inference in improving prediction accuracy.

6 CONCLUSION AND PROSPECTS

This study highlights the importance of integrating causal inference methods, such as VAR-LiNGAM and PCMCI, into financial forecasting frameworks to improve feature selection and model robustness. By applying these causal discovery, we successfully reduced the number of features from 20 to 5, focusing only on the most causally relevant variables. This reduction significantly simplified the model without sacrificing predictive accuracy.

The forecasting results demonstrate that predictions made on causal-discovery-selected features are comparable to those obtained from univariate datasets over long horizons. More importantly, for short horizons, the causal-discovery-based approach outperformed traditional feature selection methods by delivering superior predictive accuracy.

6.1 Conclusion

The integration of causal inference not only enhances prediction robustness but also addresses critical limitations of classical models, such as sensitivity to spurious correlations and excessive dependence on irrelevant parameters. Key findings include :

- A significant reduction in feature space, enabling more efficient and interpretable models.
- Improved prediction accuracy and robustness to distributional shifts in financial time series data.
- Superior performance on short-horizon forecasting, showcasing the effectiveness of causal feature selection.

In addition, by leveraging advanced temporal forecasting models like LSTM, we demonstrated the ability to capture long-term dependencies, thereby achieving optimal performance for sequential data.

6.2 Prospects

This work opens up several avenues for future research and development :

- **Exploration of Other Causal Inference Methods :** Investigating alternative methods such as Granger causality or causal discovery techniques to complement VAR-LiNGAM and PCMCI.
- **Improved Computational Efficiency :** Developing optimization strategies to address the computational complexity of VAR-LiNGAM and PCMCI when analyzing large datasets with high dimensionality and lag depth.
- **Application to Diverse Domains :** Generalizing the methodology to other fields beyond finance, such as climate modeling or healthcare, where causal inference can significantly improve forecasting accuracy.

The results of this study underline the potential of causal inference in transforming financial forecasting by providing robust, interpretable, and efficient models. This work serves as a foundation for further research aimed at leveraging causality for better decision-making in complex, dynamic environments.

7 Bibliographie

[1] Z. Zhang, S. Zohren et S. Roberts. « DeepLOB : Deep Convolutional Neural Networks for Limit Order Books ». In : *IEEE Transactions on Signal Processing* 67.11 (2019), p. 3001-3012.

[2] Y. Tang, P. Yang et Y. Zhang. « A survey on machine learning models for financial time series forecasting ». In : *Neurocomputing* 518 (2022), p. 425-436.

[3] D. C. Oliveira, Y. Lu et X. Lin. « Causality-Inspired Models for Financial Time Series Forecasting ». In : *arXiv preprint arXiv :2408.09960* (2024).

[4] Y. Zhang et S. Zohren. « Multi-Horizon Forecasting for Limit Order Books : Novel Deep Learning Approaches and Hardware Acceleration using Intelligent Processing Units ». In : *arXiv preprint arXiv :2106.01988* (2021).

[5] L. Arsac et T. Spies. « Causal Discovery for Time Series : PMINE ». In : *Proceedings of the 2021 ACM Conference on Knowledge Discovery and Data Mining (KDD)*. 2021, p. 1234-1242.

[6] S. Börjesson et M. Ul Hassan. « Forecasting Financial Time Series through Causal and Dilated Convolutions ». In : *Entropy* 22.11 (2020), p. 1234.

[7] A. Zaremba et K. Shemer. « Assessing Causality in Financial Time Series ». In : *Journal of Financial Econometrics* 21.2 (2023), p. 345-367.

[8] T. Herbelot. *TD Causalité*. 2021.

[9] S. Shimizu et al. *LiNGAM : Applications and Tailor-Made Methods*. Shimizu Lab. n.d.