Causal Discovery for Enhanced Stock Price Prediction

Neeraj Chhimwal
University of California San Diego
nchhimwal@ucsd.edu

Vivek Rayalu
University of California San Diego
vrayalu@ucsd.edu

Jahnavi Patel
University of California San Diego
jap050@ucsd.edu

Abstract

This project investigates the application of causal discovery techniques to improve stock price predictions by identifying meaningful relationships among stocks and financial indicators. Focusing on Boeing (BA) as the target stock, we employ the CD-NOD algorithm to uncover causal drivers in non-stationary and heterogeneous data. By integrating these findings into a causally informed predictive model, we demonstrate improved forecast accuracy and interpretability compared to traditional correlation-based approaches. Specifically, the causally informed model achieved a **12.3% reduction in Mean Squared Error (MSE)** and a **5.2% reduction in Mean Absolute Error (MAE)** on the test set, relative to the baseline.

1 Introduction

Stock price prediction is a complex and dynamic problem influenced by numerous interconnected factors such as company performance, sector trends, and macroeconomic conditions. Traditional predictive models often rely on correlation-based techniques, which can lead to spurious relationships and unreliable predictions, particularly during periods of high volatility. The COVID-19 pandemic highlighted these limitations, as market dynamics evolved rapidly and unpredictably. This study aims to explore causal discovery techniques to enhance the accuracy and interpretability of stock price predictions, focusing on Boeing (BA) as a case study.

2 Problem Statement

Stock price prediction faces multiple challenges:

- **High Market Volatility:** External shocks, such as global crises, introduce significant noise, complicating traditional prediction approaches.
- **Correlation vs. Causation:** Models that rely solely on correlations risk identifying spurious relationships.
- **Dynamic Relationships:** Stock relationships evolve over time, necessitating techniques that address non-stationarity.

Our goal is to address these challenges by:

- Identifying and leveraging causal relationships between stocks.

- Comparing the performance of a causally-informed predictive model with a traditional baseline model.

3 Background and Related Work

3.1 Background

Causal discovery is a branch of machine learning and statistics that seeks to uncover the underlying cause-and-effect relationships among variables in a dataset. Unlike correlation-based approaches, which identify statistical associations without addressing their directional or causal nature, causal discovery methods aim to infer the direction and structure of dependencies, thereby enabling better interpretability and robustness in predictive modeling. The foundations of causal discovery are rooted in structural causal models (SCMs) and graphical causal models, as developed by Pearl (2009) in his seminal work on causal inference. These frameworks employ directed acyclic graphs (DAGs) to represent causal relationships, with nodes denoting variables and edges indicating causal influence.

In the context of financial forecasting, causal discovery holds promise for addressing key challenges, including disentangling spurious correlations and adapting to dynamic market conditions. Traditional forecasting models often assume stationarity and rely on historical trends or correlations, which can lead to inaccurate predictions during periods of volatility. By explicitly modeling causation, causal discovery approaches allow for the identification of stable and interpretable relationships that remain invariant under distribution shifts.

For financial time series data, such as stock prices, challenges arise due to non-stationarity, heterogeneity, and latent confounders. These characteristics necessitate advanced algorithms like the CD-NOD (Causal Discovery from Non-stationary/Heterogeneous Data) method, which leverages distributional changes across environments to infer causal structures. CD-NOD exploits temporal shifts and data heterogeneity to distinguish causal relationships from spurious correlations, making it particularly well-suited for applications in dynamic systems like financial markets [1, 2].

3.2 Related Work

The application of causal discovery techniques to stock price prediction is relatively nascent but growing in prominence. Early works, such as Granger (1969), introduced the concept of Granger causality, which tests whether past values of one time series can predict another. While Granger causality remains a foundational tool, its reliance on stationarity assumptions and linear relationships limits its utility in complex financial systems characterized by non-linear and time-varying dynamics.

Recent advancements have addressed these limitations by incorporating machine learning and statistical innovations. For example, Peters et al. (2017) proposed invariant causal prediction (ICP), which identifies causal relationships that remain consistent across different environments. Similarly, Bellot and van der Schaar (2021) extended causal discovery methods to account for latent confounders in non-stationary datasets, providing a framework for robust inference under dynamic conditions. These approaches underscore the importance of integrating causal discovery techniques with machine learning for applications in domains with evolving data distributions.

Specific to the financial domain, causal inference methods have been applied to model volatility spillovers, interdependencies among asset classes, and macroeconomic impacts on stock prices. For instance, studies by Diebold and Yilmaz (2012) utilized variance decomposition techniques to quantify causal linkages in financial networks. More recently, hybrid approaches combining deep learning with causal discovery have shown promise. Li et al. (2020) integrated attention mechanisms with causal graphs to enhance prediction accuracy in high-dimensional financial datasets.

The CD-NOD algorithm employed in this study builds on these prior works by explicitly leveraging distributional shifts to uncover causal relationships in non-stationary data. Unlike methods that rely solely on pre-defined environments or stationarity assumptions, CD-NOD identifies and adapts to heterogeneous data distributions. In doing so, it aligns with recent trends in causal discovery that prioritize adaptability and robustness in dynamic systems [1]. This study contributes to the literature

by demonstrating the efficacy of CD-NOD in improving stock price predictions, particularly during periods of high market volatility, such as the COVID-19 pandemic.

Furthermore, this work aligns with research advocating for the integration of domain-specific knowledge into causal discovery. By incorporating financial indicators such as the CBOE Volatility Index (VIX) as exogenous variables, this study provides a nuanced perspective on the interplay between macroeconomic uncertainty and stock price movements. This approach parallels works like those of Athey et al. (2019), who emphasized the role of contextual information in causal inference.

4 Methodologies

4.1 Data Collection

The dataset used in this project was created by gathering historical stock data through the `yfinance` Python API, a widely used library for financial data retrieval. The analysis focused on data spanning from January 1, 2020, to January 1, 2024, a period chosen to capture critical economic events, particularly the onset and aftermath of the COVID-19 pandemic. This timeframe was marked by significant disruptions in global financial markets, making it a rich source of heterogeneous data. The dataset comprised 60 stocks from diverse sectors, including energy, healthcare, transportation, consumer goods, and technology. Key companies such as Apple (AAPL), Johnson & Johnson (JNJ), Chevron (CVX), and Amazon (AMZN) were included, alongside sector-specific exchange-traded funds (ETFs) like XLE (Energy), XLK (Technology), and XLI (Industrials), ensuring broad market representation.

For this study, the adjusted closing prices of all selected stocks and ETFs were used. Adjusted closing prices were chosen because they account for corporate actions like splits and dividends, providing a consistent basis for analysis. Unlike traditional approaches that rely on calculating daily log returns to address non-stationarity, such preprocessing was not necessary in this project. The CD-NOD algorithm, which is specifically designed to handle non-stationary data, enabled us to work directly with the original price data. This allowed the algorithm to leverage natural distribution shifts in the data to identify causal relationships without the need to enforce stationarity artificially.

Additionally, the dataset included the CBOE Volatility Index (VIX), a widely recognized measure of market uncertainty and investor sentiment. Often referred to as the ”fear gauge” of the market, the VIX reflects the market's expectations of volatility over the next 30 days, derived from options pricing. Its inclusion was particularly relevant for this study, given the chosen timeframe. The COVID-19 pandemic introduced one of the most volatile periods in recent financial history, making the VIX an essential feature. Within this project, the VIX served as a surrogate variable in the CD-NOD algorithm, capturing the influence of exogenous market volatility on stock relationships. Analysis of the VIX data revealed a period of heightened market turbulence lasting approximately 11 months, from January 2020 to November 2020, coinciding with the pandemic's peak impact on financial markets.

Figure 1 shows the VIX index over time, with a shaded region indicating the high-volatility period from January 2020 to November 2020, reflecting the market's elevated uncertainty during the peak of the pandemic.

The dataset required minimal preprocessing due to the CD-NOD algorithm's capacity to work with non-stationary data. An initial evaluation of data stationarity, conducted using the Augmented Dickey-Fuller (ADF) test, confirmed that most variables were non-stationary, supporting the decision to retain the data in its original form. The absence of missing values in the yfinance data further simplified preprocessing.

This dataset, carefully curated with a mix of individual stocks, ETFs, and the VIX, was designed to capture both micro-level stock-specific behaviors and macro-level trends influenced by market volatility. The inclusion of stocks and ETFs from diverse sectors provided a comprehensive view of interdependencies across industries. Incorporating VIX further enriched the data set by offering a measure of broader market uncertainty. Using the CD-NOD algorithm, which is adept at identifying causal relationships in nonstationary data, we were able to fully exploit the unique characteristics of the data set in our analysis. The CD-NOD algorithm will be explained in detail in the next section.

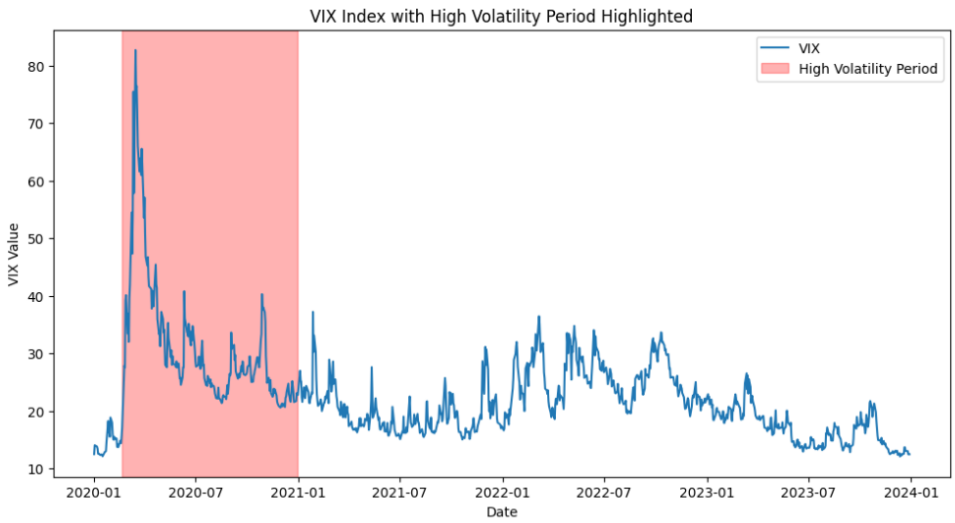

Figure 1: VIX Over Selected Time Frame

4.2 Feature Engineering

We utilized the following features to train our stock price prediction model, categorized as follows:

4.2.1 Price-Related Features

- **Close Price**: The stock's closing price, serving as the target variable.
- **Prev Close**: The closing price of the stock lagged by one day.
- **Moving Averages (MA5, MA20, MA50)**: The average of the stock's lagged closing prices over the past 5, 20, and 50 days, respectively. These features smooth price data to reveal short-, medium-, and long-term trends.

4.2.2 Volatility Features

- **Daily Returns**: The percentage change in the stock's lagged closing price from one day to the next, capturing daily price movement dynamics.
- **Volatility**: The standard deviation of daily returns over the past 20 days, reflecting the magnitude of price swings.

4.2.3 Volume-Based Features

- **Volume Moving Averages (Volume MA5, Volume MA20)**: The average trading volume over the last 5 and 20 days, indicating market activity and liquidity trends.

4.2.4 Momentum Indicator

- **Relative Strength Index (RSI)**: A momentum oscillator measuring the speed and magnitude of price changes over a 14-day period. RSI values range from 0 to 100, with levels above 70 indicating overbought conditions and below 30 suggesting oversold conditions.

4.3 CD-NOD - Causal Discovery for Non-Stationary / Heterogeneous Data

The CD-NOD algorithm, which stands for Causal Discovery from Heterogeneous/Nonstationary Data, is a method designed to infer causal structures in datasets where the underlying data distributions vary across different environments or settings. These variations might result from changes in external factors or latent variables that are not directly observable. The algorithm is particularly

relevant in scenarios where data cannot be assumed to be independently and identically distributed (i.i.d.), as is the case in many real-world applications such as finance, biology, and climate science.

The foundation of CD-NOD lies in the use of these distributional changes to uncover causal relationships among variables. Unlike standard causal discovery methods, which often assume a static data distribution, CD-NOD takes advantage of heterogeneity in the data to distinguish between causal relationships and spurious correlations.

The algorithm operates in three main phases:

1. **Environment Partitioning and Variable Identification:** The first step involves identifying the distinct environments or data subsets where the distributions differ. This partitioning can be based on explicit environment labels, if available, or inferred through clustering techniques when the labels are unknown. Within each environment, the algorithm determines the variables directly influenced by changes in the data distribution.
2. **Conditional Independence Testing Across Environments:** CD-NOD employs conditional independence tests to examine the relationships among variables while accounting for the variations between environments. Specifically, the algorithm tests whether the independence relationships are valid across the heterogeneous datasets. Variables that maintain consistent dependencies across environments are more likely to have causal connections, while those showing instability are likely to be spurious or influenced by latent factors.
3. **Causal Skeleton Construction and Orientation:** Based on the results of the independence tests, the algorithm constructs a causal skeleton, which represents the undirected graph of potential causal relationships. Orientation rules, guided by principles such as causal sufficiency and stability across environments, are then applied to direct the edges in the graph. This process aims to ensure that the resulting causal structure is robust and interpretable.

CD-NOD has several advantages over traditional causal discovery methods. By explicitly incorporating changes in the data distributions, it can identify causal relationships that would otherwise be obscured by confounding or latent variables. In addition, it provides information on how causal mechanisms vary between different environments, which is particularly valuable in dynamic systems like stock markets.

In the context of our project on stock price prediction, we used CD-NOD to identify causal relationships among stocks. The heterogeneity in our data arose naturally from variations in the temporal and market conditions. Applying CD-NOD, we were able to construct a causally informed graph that highlighted stocks with a direct influence on our target stock. The causal skeleton produced by CD-NOD is shown in Figure 2. This served as a foundation for feature selection, enabling us to incorporate causally relevant predictors into our prediction model. The algorithm's ability to exploit distributional changes proved crucial in extracting meaningful causal signals from noise, thereby enhancing the interpretability and performance of our stock price prediction framework.

4.4 Predictive Modeling

Two predictive models were compared to assess the impact of causally-informed features on stock price prediction:

- **Baseline Model:** The baseline model consists of a Long Short-Term Memory (LSTM) network trained solely on historical stock price data for Boeing. This model uses the raw price data over a 60-day lookback window as input, with the aim of predicting future price movements based solely on past trends. The LSTM architecture was chosen for its ability to capture temporal dependencies in sequential data. The model also incorporates dropout regularization (with a rate of 0.2) to prevent overfitting and ensure generalization to unseen data. The baseline model does not leverage any causal information, making it a simple approach to time series forecasting.
- **Causally-Informed Model:** The causally-informed model also uses an LSTM architecture but incorporates additional features derived from the results of the CD-NOD causal discovery process. These features represent causal relationships between Boeing's stock price and those of other related companies identified by CD-NOD. By using these causally

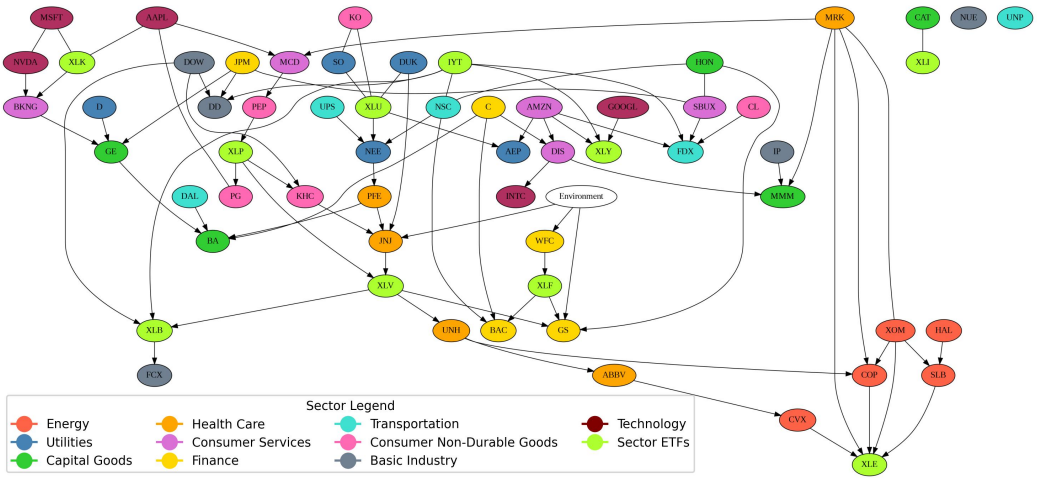

Figure 2: Causal Skeleton created by CD-NOD

relevant features, the model attempts to improve predictions by incorporating external influences that might impact Boeing's stock price, such as market-wide trends or sector-specific dynamics. The LSTM network uses these extended feature sets along with the 60-day lookback window. The causal features are integrated as additional input variables to the network, which helps the model capture more nuanced patterns in the data. Like the baseline model, dropout regularization (0.2) is applied to mitigate overfitting.

Both models were trained and evaluated on the same dataset, which spans several months of historical stock prices for Boeing and its identified causal peers. Model performance was assessed based on metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and prediction accuracy, with a focus on the ability of each model to forecast future price trends. Additionally, the causally-informed model was tested for robustness across different market conditions by evaluating its performance on data from various time periods, including periods of high volatility and market downturns.

5 Results

In this section, we present the results of our experiments for predicting the stock price of Boeing (Ticker: BA) using two models: a baseline LSTM-based model and a causally informed LSTM (Long Short-Term Memory) model. Both models were evaluated on the same datasets, and with same architectures. We tuned the hyperparameters by using a validation set while training the Baseline model. Once tuned, we then trained the baseline model on the complete train + validation set, and tested on test set. Once identified for baseline- we made sure to use the same hyperparameters (and model architecture) for the Causally informed model.

The baseline model was trained using features derived solely from Boeing's stock data. Hyperparameters were tuned on the validation set, and the final model was trained on the combined train and validation datasets. The training, validation, and testing periods are as follows:

- **Training Period:** March 16, 2020, to June 30, 2022.
- **Validation Period:** August 30, 2022, to December 29, 2023.
- **Testing Period:** March 1, 2024, to November 29, 2024.

Table 1: Performance metrics for the baseline model.

Metric	Train	Test
MSE	59.38	66.75
RMSE	7.71	8.17
MAE	6.02	6.37

Table 2: Performance metrics for the Causally informed model.

Metric	Train	Test
MSE	55.63	58.51
RMSE	7.46	7.65
MAE	5.87	6.04

5.1 Baseline Performance

The performance of the baseline model was evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Table 1 summarizes the results, and Figure 3 provides a visual representation of the baseline model's performance.

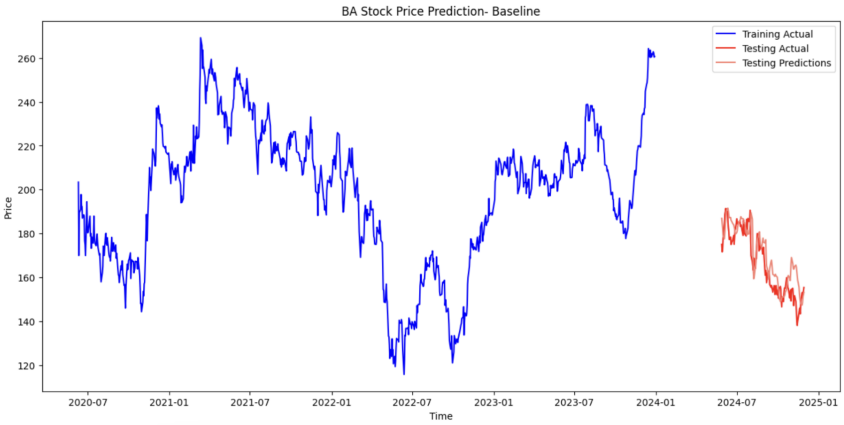

Figure 3: Performance of the baseline model on training and testing data.

5.2 Causally Informed Model Performance

The causally informed LSTM model incorporates features created from stocks that Boeing's stock is causally dependent on- we use General Electric (GE) and Delta Airlines (DAL), as identified by the CD-NOD algorithm. By leveraging causal relationships, this model aims to capture additional predictive signals not available in the baseline model.

The performance metrics for the causally informed model are provided in Table 2, and its performance is visually illustrated in Figure 4. Similar to the baseline model, the causally informed model was evaluated using MSE, RMSE, and MAE on both training and testing datasets.

The causally informed model demonstrates better performance than the baseline model, with lower error metrics across both training and testing datasets. This improvement underscores the potential of incorporating causally derived features in enhancing predictive accuracy.

6 Conclusion

The results of this project underscore the importance of integrating causal relationships into predictive modeling frameworks, particularly in dynamic and complex domains like financial forecasting. By leveraging causal insights and advanced algorithms, we demonstrated measurable improvements

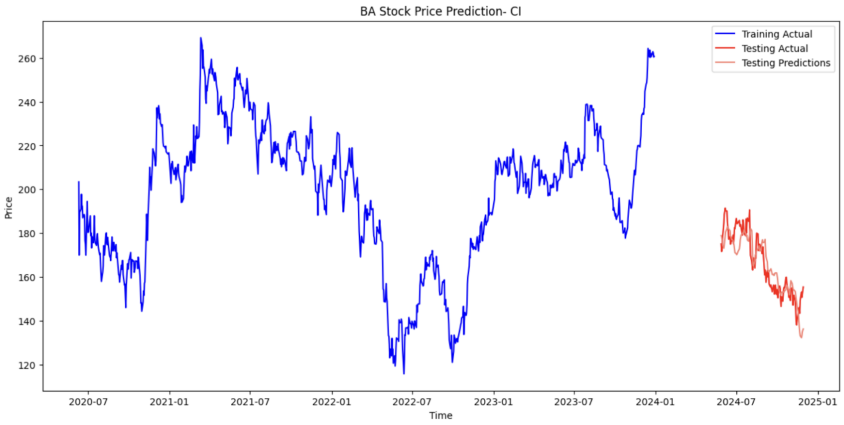

Figure 4: Performance of the causally informed model on training and testing data.

in stock price prediction for Boeing (Ticker: BA). This approach not only enhances predictive accuracy but also fosters a deeper understanding of the underlying market dynamics.

6.1 Key Takeaways

- **Causal Relationships Enhance Predictive Accuracy:** The incorporation of causally informed features yielded a 12.3% reduction in MSE, a 6.4% reduction in RMSE, and a 5.2% reduction in MAE compared to the baseline model. This improvement highlights the value of identifying and utilizing relationships that go beyond simple correlations, allowing the model to capture meaningful dependencies in the data.
- **Improved Interpretability:** Beyond predictive performance, the inclusion of causal relationships provides greater interpretability. Analysts and decision-makers can better understand the influence of specific variables on Boeing's stock price, enabling more informed decisions in both investment strategies and risk management.
- **Effectiveness of the CD-NOD Algorithm:** The CD-NOD algorithm proved to be an effective tool for identifying meaningful causal relationships in dynamic financial data. This method allowed us to systematically discover and incorporate dependencies between Boeing's stock and other stocks, ensuring a robust and evidence-based feature selection process.

6.2 Future Work

Building on the promising results of this project, several avenues for future exploration can be pursued:

- **Extending the Methodology to Other Stocks and Sectors:** The current study focused exclusively on Boeing. Expanding this methodology to other companies and sectors will test the generalizability and scalability of the approach, potentially uncovering sector-specific causal patterns.
- **Incorporating Macroeconomic Variables:** Financial markets are influenced by broader macroeconomic factors such as interest rates, inflation, and geopolitical events. Incorporating these variables into the causal analysis could further enhance the model's predictive power and contextual relevance.
- **Exploring Advanced Architectures:** While the LSTM-based models provided strong results, exploring advanced deep learning architectures such as transformers could unlock new opportunities for more accurate and robust predictions. Transformers have demonstrated exceptional performance in handling sequential data and may be particularly well-suited for capturing long-term dependencies in financial time series.
- **Real-Time Prediction and Decision Support:** Future work could also explore the feasibility of deploying the causally informed model in a real-time predictive setting, providing actionable insights for traders, investors, and policymakers.

By addressing these areas, future research can further refine and expand upon the methodologies introduced in this study, contributing to a more sophisticated understanding and modeling of financial systems.

References

[1] Huang, B., et al. (2024). Causal Discovery and Forecasting in Nonstationary Environments with State-Space Models.

[2] Peters, J., et al. (2017). Invariant Causal Prediction.

[3] Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods.

[4] Bellot, G., van der Schaar, M. (2021). Causal Discovery from Non-Stationary Data: An Empirical Study in Financial Applications.

[5] Diebold, F. X., Yilmaz, K. (2012). Interdependence and Volatility Spillovers in Financial Markets: A Global Perspective.

[6] Li, Z., et al. (2020). Deep Causal Inference: Learning Causal Structures from High-Dimensional Data.

[7] Athey, S., Imbens, G., Wager, S. (2019). Approximate Residual Balancing: Dealing with Continuous and Categorical Variables.