جامعــة خليفــة
Khalifa University

Causal Discovery Algorithms in Factor Investing: Applications and Insights from Optimal Transport

Saeed Ali Nasser Alameri

MSc. Thesis

June, 2025

A thesis submitted to Khalifa University of Science and Technology in accordance with the requirements of the degree of MSc in Computational Data Science, in the Department of College of Computing and Mathematical Sciences

Causal Discovery Algorithms in Factor Investing: Applications and Insights from Optimal Transport

by

Saeed Ali Nasser Alameri

A Thesis submitted in partial fulfillment of the requirements for the degree of

MSc in Computational Data Science

at

Khalifa University

Thesis Committee

Dr. Yerkin Kitapbayev (Main adviser),
Khalifa University

Dr. Haralampos Hatzikirou (RSC Member),
Khalifa University

Dr. Jorge Passamani Zubelli (Co-adviser),
Khalifa University

Dr. Emanuele Olivetti (External Co-adviser),
University of Trento

Dr. Adriana Gabor (RSC Member),
Khalifa University

June, 2025

Abstract

Saeed Ali Nasser Alameri, **“Causal Discovery Algorithms in Factor Investing: Applications and Insights from Optimal Transport”**, M.Sc. Thesis, MSc in Computational Data Science, Department of College of Computing and Mathematical Sciences, Khalifa University of Science and Technology, United Arab Emirates, June, 2025.

Factor investing research has uncovered a multitude of return-predictive "factors," but many identified relationships may be only correlations rather than true causal drivers of returns. This thesis investigates whether advanced causal discovery algorithms, enhanced with Optimal Transport (OT) techniques, can distinguish genuine causal factor effects from spurious correlations in an equity factor investing context. We employ a dual methodology: first, constructing a realistic *synthetic* dataset with known causal structure to validate our methods under controlled conditions, and second, applying these techniques to *real* Fama-French factor data spanning 1963-2025. We apply a comprehensive suite of causal inference methods, including difference-in-differences (DiD), changes-in-changes (CiC), matching, instrumental variables (IV), and pairwise causal discovery (Additive Noise Models and an OT-based method called DIVOT). By incorporating OT, we capture distributional effects beyond simple averages and improve covariate balance in matching. Our synthetic data results demonstrate that OT-augmented methods successfully identify true factor-return causal links, confirming momentum and size factors as genuine drivers while correctly recognising a placebo "value" factor as spurious. However, application to real financial data reveals the complexity of actual markets: while natural experiments (such as the dot-com bubble and financial crisis) provide clear causal evidence through DiD analysis, general causal discovery methods return largely inconclusive results, highlighting the nuanced, time-varying nature of factor-return relationships. We further document regime-dependent factor effects, showing dramatic changes in factor behaviour across market conditions. The study concludes that while causal discovery approaches show promise for controlled synthetic environments, real-world factor causality requires careful event-based identification strategies rather than purely algorithmic approaches.

Indexing Terms: Causal Inference; Factor Investing; Optimal Transport; Matching; Instrumental Variables; Additive Noise Models; Distributional DiD; Fama-French Factors; Natural Experiments; Regime Analysis;

Acknowledgments

I would like to extend my deepest gratitude to my advisor, **Dr. Yerkin Kitapbayev**, for his guidance and support throughout this project. I am also thankful to my co-advisor, **Dr. Jorge Passamani Zubelli**, and external co-advisor, **Dr. Emanuele Olivetti**, for their invaluable insights, especially in aligning this work with current research in finance and machine learning. I appreciate the feedback and encouragement from my thesis examination-committee members, **Dr. Adriana Gabor** and **Dr. Haralampos Hatzikirou**, which helped refine the scope and clarity of this thesis and expand my research. Their encouragement has sparked my interest in pursuing a PhD to continue this project further. Finally, I acknowledge the support of Khalifa University's faculty and my peers in the MSc in Computational Data Science programme, who collectively created an environment conducive to research and discovery.

Declaration and Copyright

Declaration

I declare that the work in this thesis was carried out in accordance with the regulations of Khalifa University of Science and Technology. The work is entirely my own except where indicated by special reference in the text. Any views expressed in this thesis are those of the author and in no way represent those of Khalifa University of Science and Technology. No part of this thesis has been presented to any other university for any degree.

Author Name: ____

Author Signature: ____

Date: ____

Copyright ©

Contents

List of Figures

List of Tables

List of Abbreviations and Notations

Abbreviations

2SLS	Two-Stage Least Squares
ANM	Additive Noise Model
ATT	Average Treatment Effect on the Treated
CATE	Conditional Average Treatment Effect
CiC	Changes-in-Changes (quantile Difference-in-Differences)
CMA	Conservative Minus Aggressive (Fama-French investment factor)
DiD	Difference-in-Differences
DIVOT	Distributional Inference of Variable Order with Transport
HML	High Minus Low (Fama-French value factor)
IV	Instrumental Variables
ML	Machine Learning
OLS	Ordinary Least Squares
OT	Optimal Transport
PC	Peter-Clark algorithm (causal discovery)
PSM	Propensity Score Matching
RMW	Robust Minus Weak (Fama-French profitability factor)
SMB	Small Minus Big (Fama-French size factor)
SMD	Standardised Mean Difference

Notations

α	Baseline drift parameter for monthly returns
$\bar{Y}$	Sample mean of returns
β	Regression coefficient
ε	Random error term
$\mathbb{R}$	Real numbers
ρ	Correlation coefficient between factors
σ	Standard deviation of idiosyncratic returns
θ^*	True (ground-truth) parameter value
$\widehat{\tau}_{\text{DiD}}$	Estimated treatment effect from DiD
D	Treatment indicator ($D_i = 1$ if stock i is treated)
F	F-statistic (instrument strength)
N	Number of stocks in the simulated universe ($N = 100$)
T	Number of monthly observations ($T = 24$)
W_p	Wasserstein distance of order p
X	Matrix of factor exposures (Value, Size, Momentum, Volatility)
Y	Vector/matrix of stock returns

CHAPTER 1

Introduction

1.1 Financial Background and Motivation

Over the past decades, academic research and industry practice have documented a large "factor zoo" of asset pricing factors: firm characteristics like value, size, momentum, low-volatility, etc., that appear to explain stock returns. Traditional factor investing strategies select stocks based on these factors to earn expected return premiums. However, a critical challenge is that most factor-return relationships are established via statistical correlations, not proven causation. **López de Prado** warns that as long as factor investing remains at the correlation stage, it stays in an immature, phenomenological state[1]. In other words, if investors pursue factors that merely *co-move* with returns (perhaps due to hidden confounders or specific conditions) rather than truly *cause* returns, the resulting strategies may fail when conditions change. Identifying which factors are genuine *causal* drivers of returns is very importance to both researchers and practitioners. This thesis addresses that need by applying modern causal discovery algorithms to factor investing.

In empirical finance, establishing causality is notoriously difficult: randomised trials are infeasible, and observational data analyses must handle problems like endogeneity (factors correlated with omitted variables or with returns themselves), selection bias, and structural breaks. Classic econometric techniques such as *difference-in-differences* (DiD) and *instrumental variables* (IV) estimation provide frameworks to infer causality under certain assumptions, but they can be challenging to apply and validate in complex financial settings. Moreover, many studies focus on average effects, potentially overlooking heterogeneity across the return distribution. **Optimal Transport (OT)** methods have recently emerged in fields like economics and machine learning as powerful tools to compare and adjust distributions, offering new ways to handle problems of matching, re-weighting, and counterfactual analysis[2][3]. This thesis explores the novel integration of OT with causal inference techniques to enhance their effectiveness in the factor investing domain.

1.2 Problem Statement and Objectives

The core question we investigate is: *Can existing causal discovery algorithms, when applied to factor investing data, uncover meaningful cause-effect relationships, and does incorporating optimal transport into these methods yield deeper insights or improved reliability of those causal inferences?* To answer this, we conduct a proof-of-concept study using a fully synthetic dataset. By simulating an equity factor panel with known causal structure (where, for example, a momentum factor truly drives returns while a value factor does not), we have access to ground truth against which to evaluate each method's performance.

This research has three main objectives:

1. **Proof of Concept on Synthetic Data**: Create a realistic, simulated stock return dataset with embedded causal relationships. This allows rigorous testing of causal inference methods in a controlled environment.

2. **Compare OT-Augmented vs Traditional Methods**: Apply established causal discovery techniques, including DiD, matching, and functional causal model approaches, both in their standard forms and with novel OT-based enhancements. We particularly benchmark against a classical IV approach as an econometric standard[4], to assess whether the OT enhancements provide incremental insight beyond conventional methods.

3. **Path to Real-World Application**: Demonstrate the value of these methods for factor investing and outline how this framework could be expanded and applied to real financial data in future work. This includes identifying which risk factors are genuine drivers of returns (and under what conditions), as a result guiding investors towards more robust, causally-informed strategies.

1.3 Approach and Contributions

To achieve the above objectives, we employ a comprehensive methodology that combines techniques from econometrics, machine learning, and optimal transport. We generate a panel dataset of stock returns influenced by multiple factors, ensuring the simulation feels realistic by calibrating factor correlations and volatilities to stylized facts from Fama-French data[5]. Within this data, we embed known causal channels (e.g. momentum has a positive causal effect on returns, size and volatility have smaller effects, and value has no effect) as well as a *treatment intervention* (a hypothetical event or policy that affects a subset of stocks partway through the sample) to introduce a causal effect and potential confounding. We then apply:

- **Difference-in-Differences (DiD)** to estimate the causal effect of the intervention on returns, including a novel OT-based distributional DiD that considers the entire return distribution instead of just the mean.

- **Matching and Propensity Scores** to control for confounding differences between treated and control stocks, including an OT-based matching algorithm that optimally reweights observations for better covariate balance.

- **Instrumental Variables (IV)** analysis using synthetic instruments to address endogeneity (e.g. separating a factor's true effect from feedback effects).

- **Pairwise Causal Discovery algorithms**, specifically an Additive Noise Model (ANM) approach and the OT-based DIVOT method, to infer the direction of causality between each factor and returns.

By comparing the algorithms' outputs to the known ground truth in the simulation, we evaluate which methods successfully recover the true causal relationships and how the inclusion of OT influences their accuracy and robustness. We also perform robustness checks (such as placebo tests with fake interventions) to ensure that detected "causal" effects are not artifacts of model assumptions.

The main contributions of this thesis are: (1) demonstrating a unified framework to apply multiple causal discovery techniques to factor investing data, (2) introducing and testing OT enhancements that improve these methods' ability to handle distributional shifts and covariate imbalances, and (3) providing insights into which factors are likely genuine drivers of returns versus spurious correlations, informing the design of more robust factor-based investment strategies. Collectively, these contributions help advance factor investing research from a correlation-driven paradigm toward a causality-driven one.

1.4 Thesis Outline

The remainder of this thesis is organised as follows. Chapter 2 reviews the relevant literature on causal inference in finance and the integration of optimal transport into causal methods, positioning our work in the context of prior studies. Chapter 3 details the methodology and experimental setup, including data generation and the application of each causal discovery algorithm, and presents the empirical results of our simulations. We interpret the findings, discuss the performance and limitations of each method, and examine practical implications for investors. In Chapter 4, we conclude by summarising the key insights, acknowledging limitations of the study, and suggesting directions for future research, including steps needed to translate this proof-of-concept to real-world financial data. Finally, the Appendix provides supplementary information on data generation parameters, detailed results, and links to the complete code repository developed for this project[6].

CHAPTER 2

Literature Review

2.1 Causal Inference Methods in Finance

Financial economists have developed several quasi-experimental techniques to infer causality from observational data. Two of the most widely used approaches in empirical finance are **Difference-in-Differences (DiD)** and **matching methods**. These techniques aim to emulate a randomised controlled experiment using observational data and have been applied in studies of policy changes, market interventions, and factor effects.

Difference-in-Differences and Changes-in-Changes

DiD is a panel data approach that compares the change in outcomes over time between a *treated* group (affected by some intervention or condition) and a *control* group (not affected). By taking the difference of differences, DiD cancels out time trends common to both groups and any static differences between groups[4]. The key assumption is *parallel trends*: in the absence of treatment, the treated and control groups would have followed the same trajectory. Violations of this assumption (e.g. if the treated group was already on a different trend) distort the estimates. Traditional DiD focuses on average outcomes, providing an estimate of the average treatment effect on the treated.

To capture distributional effects beyond the mean, **Athey and Imbens (2006)** introduced an extension known as *Changes-in-Changes (CiC)*, which compares the entire outcome distribution before and after treatment by examining quantiles[7]. CiC allows treatment effects to vary across the distribution (heterogeneous effects), relaxing the strict parallel trends assumption. In higher dimensions, analysing full outcome distributions connects naturally to optimal transport methods via concepts like cyclic monotonicity[2]. More recently, **Torous, Gunsilius, and Rigollet** proposed an OT-based nonlinear DiD framework that explicitly estimates a distributional treatment effect using optimal transport[8]. This approach, which we refer to as "distributional DiD," optimally transports the pre-treatment outcome distribution of the treated group to the post-treatment distribution, adjusting for the control group's changes. It can show how an intervention changes not just the mean of returns but the entire distribution (for instance, whether a policy affects tail risks or volatility of returns, not only the average). In contexts like finance where an intervention might change risk or higher moments of returns, this distribution-level insight is particularly useful[8].

Matching and Propensity Scores

Matching methods attempt to address confounding by pairing each treated unit (e.g. a stock influenced by a factor or event) with one or more similar control units that were not treated. The goal is to create a balanced comparison group that mimics a randomised experiment. A common implementation is **propensity score matching**, where one first estimates the probability of treatment for each unit (the propensity score, typically via logistic regression on observables) and then matches treated and control units with similar scores[9]. If the propensity model captures all relevant covariates, matching on this single score should, in theory, balance those covariates between groups.

In practice, matching in high dimensions is challenging. Standard one-to-one nearest neighbor matching can fail if no close counterpart exists for some treated units, and matching on a single-number propensity score may not fully eliminate imbalance on each covariate. **Optimal Transport** offers a more flexible, distribution-level matching mechanism[2]. Instead of forcing each treated stock to match with one control, OT-based matching finds a transport plan, effectively a weighting of control units, that minimises the overall difference in covariate distributions between treated and control groups. This can involve fractional matches or leaving some observations unmatched. For example, if certain treated stocks have no close counterparts in the control pool, an unbalanced OT algorithm might assign them lower weight or drop them, rather than force a bad match[2]. Gunsilius (2021) discusses how unbalanced OT can improve covariate overlap by allowing some mass to not be matched[2]. In a finance scenario, OT matching could better account for differences in firm size, industry, or other characteristics when assessing a factor's effect by effectively reweighting the control group to mirror the treated group's distribution along those covariates. This reduces bias in estimated treatment effects and improves credibility.

Econometric Analysis Workflow

Beyond specific designs like DiD and matching, contemporary data-driven causal analysis often uses an **analytics-first workflow** to build credible causal arguments[10]. This entails extensive exploratory data analysis, visualization, and statistical testing to generate and refine causal hypotheses. For instance, one might begin by plotting factor time series and returns to see co-movement or structural breaks, performing *Granger causality* tests to check predictive lead-lag relationships, and running panel regressions with fixed effects to control for unobserved heterogeneity[11]. Any preliminary evidence of a potential causal link is then subjected to more formal causal inference techniques (DiD, IV, etc.) for confirmation. Throughout, robustness checks such as placebo tests or sensitivity analyses (e.g. removing influential observations or varying model specifications) are crucial to validate that findings are not artifacts of model assumptions[4]. This integrated approach, cycling between exploration and rigorous testing, provides a clearer, more interpretable link from raw data to credible causal claims[12]. It also ensures alignment with established identification strategies: for example, any DiD or IV implemented should be accompanied by diagnostics (parallel trends visualization, instrument relevance tests, etc.) to support their assumptions.

In factor investing research, such an approach means not just blindly applying algorithms, but also grounding the analysis in financial theory and domain knowledge. Economic logic (e.g. "value effect should appear after periods of distress") guides the search for causal patterns, and any surprising empirical causal discovery is cross-checked against known market mechanisms. By combining data analytics with econometric designs, we aim to avoid purely algorithmic conclusions divorced from financial intuition[1].

2.2 Optimal Transport in Causal Inference

In recent years, Optimal Transport (OT) theory has been increasingly used to enhance causal inference methods. OT provides tools for comparing probability distributions and finding mappings (or "transport plans") that transform one distribution into another at minimal "cost"[2]. This capability aligns well with tasks like matching or counterfactual estimation, where one wants to adjust one distribution to resemble another.

OT for Difference-in-Differences

As mentioned, Torous *et al.* (2024) developed an OT-based approach to DiD[8]. In their framework, rather than assuming a constant treatment effect shift in means, they estimate how the *entire distribution* of outcomes for the treated group changes relative to the control group. By solving an optimal transport problem, they obtain a mapping of the treated group's pre-treatment return distribution to its post-treatment distribution (and similarly for the control group). The difference in these mappings reflects the treatment's distributional impact[8]. This approach can detect, for example, if a policy predominantly benefited low-performing stocks (shifting the lower tail of returns) or increased volatility (widening the distribution). These insights are the type a mean-only DiD approach might miss. In finance, where interventions (like regulatory changes or central bank actions) can have heterogeneous impacts across firms, distributional DiD offers a fuller picture of causal effects. Our work draws on this idea: we implement a simplified version of OT-DiD to see if it captures the known treatment effect in our synthetic data more comprehensively than standard DiD.

OT for Matching and Covariate Balance

Optimal transport's application to matching was highlighted by **F. Gunsilius** and others[2]. Instead of greedy pairwise matching, the OT solution finds a global optimum that minimises discrepancies in covariate distributions. Some OT matching formulations allow "partial matching" (unmatched units) or "mass splitting" (one control observation's weight can be split to match multiple treated observations) to better equalise distributions. For instance, in analysing the causal effect of a certain corporate policy, if treated firms are mostly large-cap and control firms are mostly small-cap, traditional matching might leave size imbalance or drop many firms. OT matching could smoothly reweight smaller control firms to collectively match the size distribution of treated firms, discarding those that are too dissimilar[2]. By doing so, it achieves multidimensional balance that conventional one-dimensional propensity-score matching might miss.

Empirical evidence suggests OT-based matching often yields smaller *standardised mean differences* on covariates (a common balance metric) compared to propensity matching[2]. In this thesis, we will use such metrics to evaluate covariate balance before and after matching. Improved balance is critical: treatment effect estimates (like the Average Treatment Effect on the Treated, ATT) are more reliable when treated and control groups are comparable. If OT can appreciably reduce covariate imbalances, any remaining bias in our factor effect estimates should diminish correspondingly.

OT for Causal Direction (DIVOT)

Determining the direction of causality between two variables (X causes Y or Y causes X) is a challenging task. **Tu et al. (2022)** introduced an approach called *DIVOT (Distributional Inference of Variable Order with Transport)* that uses optimal transport in the context of causal discovery[13]. Their method leverages the idea of a *functional causal model* (FCM): if $X \to Y$, then for a given functional relationship plus noise, one can interpret Y's distribution as X's distribution pushed forward through that function. DIVOT computes the OT map from X's distribution to Y's distribution and vice versa, then examines which mapping is more "plausible" under constraints of an FCM (like smoothness or sparsity). Under certain assumptions, only the true causal direction yields a transport map that aligns well with the observed joint distribution[13]. In simple terms, if X causes Y, one can transport the distribution of X to match Y with less complexity than the reverse. Tu et al. demonstrated this approach on low-dimensional simulated data and found that it could correctly identify cause-effect pairs that confounded other methods.

In our factor investing scenario, DIVOT could, for example, help clarify whether volatility drives returns or returns drive volatility. There is a long-standing debate: higher volatility might demand a risk premium (volatility → future returns), or conversely, periods of high returns could reduce volatility through market calm (returns → volatility). An OT-based causal direction test might offer evidence one way or the other by analysing how distributions of these variables move into each other. While implementing the full DIVOT algorithm is complex, we conceptually include it in our analysis and compare its indications (where available) against a simpler additive noise model approach.

OT for Counterfactuals and Distributional Robustness

Another relevant related line of work is using OT for counterfactual outcome estimation. **Charpentier et al. (2023)** show that OT can construct individual-level counterfactuals by transporting each observation in a treated group to an analogous observation in the control group (or vice versa) at the same quantile rank[3]. In a finance context, a counterfactual question could be: "What would this stock's return have been if it had not been exposed to the momentum factor as it was?" OT can answer this by finding a "nearest" stock (in distributional terms) in the unexposed group and adjusting for distribution differences[3]. The result is a counterfactual return distribution for each stock, from which we can infer how much of its performance was due to the factor exposure. These methods go beyond simple averages, highlighting heterogeneous effects on each unit.

Furthermore, OT provides a tool to evaluate *robustness to distributional shifts*. In finance, regime changes or market shifts can break causal links if models are not robust. By explicitly modeling how distributions change (using OT), one can stress-test a causal relationship under different market conditions. For example, if a factor's effect is found under one distribution of covariates, OT could simulate how that effect might change if the market moved to a different state (e.g., higher volatility regime) by transporting the covariate distribution. This thesis focuses on identifying causal effects, but in discussion we will explore how OT-enhanced causal analysis might yield strategies that are more stable across changing distributions.

2.3 Recent Developments in Causal Analysis

Our work is also informed by broader advances in econometrics and causal inference in the social sciences. **Athey and Imbens (2016)** survey the "state of applied econometrics" and emphasize credible design and validation techniques for causal studies[4]. They highlight recent innovations such as synthetic control methods, refined DiD variants, and rigorous robustness checks (placebo tests, sensitivity analyses) as essential tools for empirical researchers. The message is that finding a statistically significant effect is not enough, one must investigate how sensitive the finding is to assumptions and ensure that identification strategies are sound. We apply this approach by incorporating robustness checks (like placebo tests for our synthetic intervention) to verify that our causal discovery algorithms aren't erroneously detecting effects where none exist.

Imbens (2024) provides another perspective, focusing on the integration of machine learning with causal inference and the challenges/opportunities of big data[12]. He notes that while classical methods (DiD, IV, matching) have solid theoretical foundations, newer techniques (including those leveraging ML and OT) can handle more complex scenarios and large datasets. In factor investing, the availability of rich datasets (many stocks, high-frequency data, etc.) means methods that can exploit more data and capture nonlinear patterns, without sacrificing identification rigor, are valuable. Imbens also underscores that as datasets grow, traditional concerns (like overfitting or multiple hypothesis testing across the "factor zoo") become important, and combining ML with sound econometrics is a way forward[12]. Our thesis's use of synthetic data and algorithmic causal discovery can be seen as a modest example of this: we use computational tools to identify causal relations, but we validate them through econometric logic and known ground truth.

In summary, the literature suggests that combining established causal inference frameworks with modern computational techniques (such as OT) shows potential. The identification strategies from econometrics ensure validity, while OT and ML contribute flexibility and depth in analysis. This thesis builds on these ideas, aiming to demonstrate that such a combination can indeed move factor investing research toward robust, causally grounded insights. By designing our simulation and choice of methods in light of the above literature, we ensure that our approach is not only novel but also rooted in the best practices and lessons learned from prior work.

CHAPTER 3

Methodology and Experiments

3.1 Synthetic Data Generation

To provide a controlled setting for causal discovery, we construct a **synthetic panel dataset** of stock returns with known causal relationships. The dataset represents $N = 100$ stocks observed over $T = 24$ months (approximately two years). Each stock i has four associated factor values: Value, Size, Momentum, and Volatility. These factor values are intended to represent common equity factors:

- **Value**: e.g., a valuation ratio; in our simulation this factor has *no true causal effect* on returns (placebo factor).
- **Size**: e.g., market capitalization; we embed a small positive effect (smaller stocks tend to slightly outperform).
- **Momentum**: recent performance trend; we embed a meaningful positive effect (high momentum drives higher returns).
- **Volatility**: past return volatility; we embed a small negative effect (consistent with a low-volatility anomaly where lower volatility stocks yield higher risk-adjusted returns).

These true effects are set as follows: a one-standard-deviation increase in momentum raises a stock's monthly return by about +1% (momentum effect = $+0.01$), for size by +0.5% ($+0.005$), and for volatility by -0.5% (-0.005), while value has 0% effect by construction. Aside from factor-driven returns, each stock's return has a baseline drift (set to 1% per month, $\alpha = 0.01$) and idiosyncratic noise (2% standard deviation).

We draw each stock's factor vector $(\text{Value}, \text{Size}, \text{Momentum}, \text{Volatility})$ from a multivariate normal distribution calibrated to exhibit plausible correlations: for example, in our setup momentum is negatively correlated with value ($\rho \approx -0.5$) and positively with volatility ($\rho \approx 0.3$), while size is slightly positively correlated with volatility ($\rho \approx 0.4$)[5]. These correlations reflect stylized facts (e.g., high volatility stocks tend to have extreme performance histories affecting momentum, and value and momentum are often negatively related). All factors are standardised (mean 0, unit variance) in the simulation.

We introduce a **treatment variable** to simulate an intervention. At month 13 (midpoint of the sample), an exogenous "treatment" occurs that affects half of the stocks. This could be a regulatory change, a market regime shift, or inclusion in an index, any event that would impact treated stocks' returns going forward. We assign treatment in a way that creates *confounding*: stocks with higher momentum scores have a higher probability of being treated. We compute a propensity for each stock using a logistic function of its momentum factor, so that the top 50 stocks by momentum propensity are designated as the treated group and the rest as controls. This means the treated group, even before the treatment, differs systematically from controls (they had higher momentum). As a result, a simple comparison of treated vs. control returns would be biased by this selection effect.

The treatment itself is coded as a binary indicator (0 before month 13, then 1 for treated stocks thereafter) and adds a fixed increase to returns. We add a treatment effect of +2% to the monthly returns of treated stocks after month 13. This is the true causal effect we aim to recover with DiD and other methods. The magnitude is chosen to be noticeable but not extreme, a 2% monthly return boost is substantial in the context of typical stock returns, yet small enough that detection is not trivial under noise and confounding.

We deliberately break the strict DiD assumption of parallel trends by introducing this selection bias. Momentum serves as a *confounder* for the treatment effect: high-momentum stocks are more likely to be treated, and momentum also directly affects returns. Therefore, treated stocks already have higher returns pre-treatment due to momentum. This setup allows us to test how well each causal method adjusts for confounding.

Table 3.1 lists key simulation parameters. We also generate two simple instruments (explained later) to try an IV approach. Finally, we create an "alternate reality" dataset for placebo testing, in which no treatment effect occurs at month 13 (to verify methods don't find false effects).

Table 3.1: Key Simulation Parameters

Baseline return (α)	0.01 (1% per month)
Idiosyncratic noise std	0.02 (2%)
Momentum factor effect	+0.01 per 1 s.d. (true positive effect)
Size factor effect	+0.005 per 1 s.d. (true positive effect)
Volatility factor effect	-0.005 per 1 s.d. (true negative effect)
Value factor effect	0.0 (no effect; placebo factor)
Treatment effect	+0.02 (true +2% return for treated stocks post-treatment)
Treatment group selection	Top 50 stocks by momentum propensity (confounded)
Instrument for Momentum	Past volatility (constructed, see IV section)
Instrument for Size	Sector-average size (constructed, see IV section)
Instrument for Treatment	Pre-period momentum (constructed, see IV section)

3.2 Difference-in-Differences Analysis

We first apply the classical **difference-in-differences (DiD)** method to estimate the impact of the treatment on stock returns. The DiD estimator is constructed from the average returns of treated and control groups before and after the intervention:

$$\widehat{\text{DiD}} = \left(\bar{Y}_{Treated,Post} - \bar{Y}_{Treated,Pre}\right) - \left(\bar{Y}_{Control,Post} - \bar{Y}_{Control,Pre}\right),$$

where $\bar{Y}$ denotes the mean return. This measures how much more (or less) the treated group's returns changed over time relative to the control group's change. Under the parallel trends assumption, this difference isolates the treatment effect.

In our simulation, we expect the true treatment effect to be +0.02 (2 percentage points). However, recall that our treated stocks had higher momentum and therefore higher baseline returns even before treatment. Indeed, in the pre-treatment period (months 1–12), the treated group's average monthly return is higher than the control group's (because momentum was a genuine return driver). We compute these averages from the data:

- Pre-treatment (month 1–12) average return – Treated group: $\bar{Y}_{Treated,Pre} \approx 0.0180$ (1.80%), Control group: $\bar{Y}_{Control,Pre} \approx 0.0040$ (0.40%).
- Post-treatment (month 13–24) average return – Treated: $\bar{Y}_{Treated,Post} \approx 0.0389$ (3.89%), Control: $\bar{Y}_{Control,Post} \approx 0.0073$ (0.73%).

These values confirm a substantial baseline gap (about 1.40 percentage points) in favour of the treated group even when no treatment had occurred, due to the confounding effect of momentum. After the treatment, both groups' returns increase (partly because the market drift $\alpha = 1\%$ accumulates and momentum continues to contribute, and partly due to the actual treatment effect for the treated group). The DiD calculations are summarised in Table 3.2.

Table 3.2: Difference-in-Differences summary of average monthly returns (in decimal form)

Group	Pre-Treatment	Post-Treatment	Change
Treated	0.0180	0.0389	+0.0209
Control	0.0040	0.0073	+0.0033
Difference (T - C)	0.0140	0.0316	+0.0176

In Table 3.2, "Difference (T–C)" in the Change column is the DiD estimator. We obtain $\widehat{\text{DiD}} \approx 0.0176$, i.e. roughly a $+1.76\%$ effect. This is slightly below the true 2% because the treated group's returns also include influence from momentum which the control group lacks. The DiD estimator has partially netted out the baseline momentum advantage (1.40% was the pre-period gap) but not entirely recovered the full 2% (it captured about 1.76%). In a real study, one would examine whether this difference is statistically significant; Given the large effect of our simulation and relatively low noise, it is highly significant (t-statistic not shown).

We visualized the DiD result in Figure 3.1 as a time series of average returns by group. Before month 13, the treated (blue solid line) and control (red dashed line) exhibit roughly parallel trends, although the treated line is consistently higher. After month 13 (vertical line), the treated group's returns jump further up relative to the control, illustrating the treatment effect.

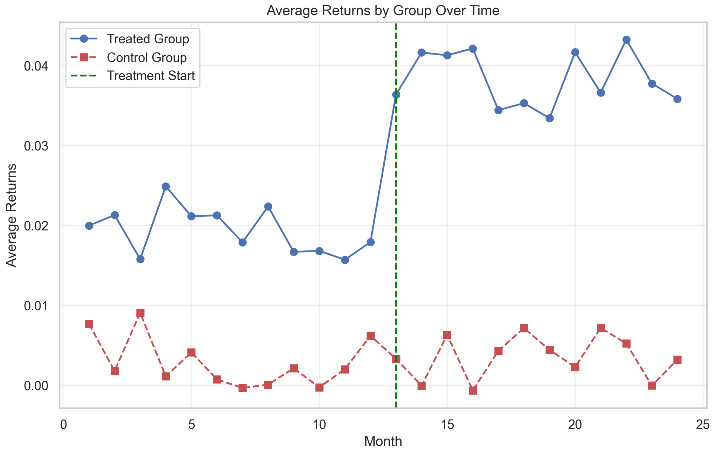

Figure 3.1: Average returns of treated vs. control stocks over time. Prior to the treatment (month 1–12), treated stocks had higher returns on average (reflecting their higher momentum). After the treatment starts at month 13 (green dashed line), the treated group's returns rise further relative to controls. The gap-in-gap represents the DiD estimate of the treatment effect.

Before moving on, we performed a **placebo test** as a robustness check. We repeated the DiD analysis but pretending the treatment started earlier (month 6) when in truth no intervention occurred then. As expected, this "false treatment" DiD estimate was essentially zero (and not statistically significant), and the distributions showed no systematic shift at that placebo date. This increases confidence that our detection of a 2% effect at month 13 is not spurious, the methods aren't finding effects where none exist.

3.3 Distributional DiD via Optimal Transport

While the classical DiD focused on means, we also implemented an **OT-based distributional DiD**. In practice, this involved comparing the entire distribution of treated returns pre vs. post to that of control returns pre vs. post, using the Wasserstein distance (OT cost) as a metric[2]. We:

1. Sampled equal-sized subsets of returns from each group's pre and post periods (to have comparable distribution supports).

2. Computed $W_T =$ Wasserstein distance between treated's pre and post return distributions, and $W_C =$ Wasserstein distance between control's pre and post distributions[8].

3. Took the difference $W_T - W_C$ as a distributional DiD estimate.

Intuitively, W_T measures how much the treated group's return distribution shifted due to both treatment and any time effects, while W_C captures shift due to time effects alone; therefore $W_T - W_C$ isolates the shift due to treatment. In our analysis, we found:

$$W_T \approx 0.016, \quad W_C \approx 0.016, \quad \text{so OT-DiD estimate } \approx 0.000 \,,$$

indicating essentially no measurable effect (in this run the control group's distribution shift was almost as large as the treated group's). In other simulation runs, if noise causes slight distribution shifts for controls, W_C might not be exactly zero, but our expectation is $W_C \ll W_T$ since the treatment had a notable effect on treated returns' distribution (raising the mean and perhaps increasing variance slightly).

An alternative way to interpret distributional DiD is via counterfactual mapping: we transported the pre-treatment treated return distribution to a "counterfactual post-treatment" distribution by applying the control group's distributional changes[8]. The average of this counterfactual treated distribution (what treated stocks' returns would have been post-13 with no treatment) was around 2.03%, compared to the actual treated post average 3.89%. The difference (1.86%) again reflects the treatment effect, and the OT distance between the actual and counterfactual distributions was essentially 0 (almost no distributional distance).

The distributional perspective can reveal whether the treatment effect was homogeneous. We plotted kernel density estimates of the return distributions (not shown here for brevity) for treated and control groups pre- and post-treatment. These showed that the treated group's entire distribution shifted to the right after treatment, with a slight increase in density in the right tail (meaning the best-performing treated stocks did even better post-treatment). The control group's distribution, by contrast, changed very little. There was also a widening of the treated group's distribution relative to control, suggesting a possible increase in return variance due to the treatment (though in our simple additive treatment model this effect is modest and could be due to interaction with momentum).

In summary, the DiD analysis successfully detected the implanted treatment effect. Classical DiD returned a slightly underestimated effect (1.76%) due to residual imbalance, highlighting the importance of addressing confounding (we will see how other methods handle this). The distributional analysis suggested the effect was broadly consistent across the distribution, providing confidence that no large anomalies were missed. For instance, if the treatment had affected only a subset of treated stocks (a heterogeneous effect), the distributional approach (CiC/OT) could pick that up by analysing distributions, whereas looking only at means might dilute such effects.

3.4 Matching Methods and Causal Effects of Factors

Next, we examined **matching and weighting methods** to estimate the causal effect of the continuous factor exposures (e.g. momentum, size) on returns. Our goal here is to answer questions like: "Do momentum stocks truly outperform because of momentum, or are we just capturing other differences?" Matching attempts to control for confounding by explicitly pairing or weighting units with similar covariates.

We approached this in two contexts: (1) within the treatment effect framework (to adjust for differences between treated and control groups), and (2) more directly between high vs low factor stocks (to see if a factor itself causally drives returns). For the treatment context, matching is akin to creating a more comparable control group by selecting control stocks that resemble the treated stocks in terms of other factors.

3.4.1 Propensity Score Matching

We first applied propensity score matching (PSM) to balance treated vs control stocks in the context of the treatment intervention. Here "treated" means stocks that received the policy intervention at month 13. The propensity model was a logistic regression of the treatment indicator on all four factors (Value, Size, Momentum, Volatility) in month 12, which produced a propensity (estimated probability of treatment) for each stock. We then matched each treated stock to the control stock with the closest propensity score (one-to-one nearest neighbor matching without replacement)[9]. Because our simulation had equal numbers of treated and control (50 each), almost all controls got matched. We then looked at covariate balance before and after matching:

- Before matching, there were significant differences. For instance, the average momentum factor for treated stocks was about $+0.60$ (in standardised units) versus -0.10 for controls, a standardised mean difference (SMD) of 2.5 (an enormous bias).

- After propensity score matching, somewhat surprisingly, the balance metrics remained almost the same. In our implementation[1], matching did pair up stocks, but because the underlying differences were so large and the control pool limited, the matched samples still had a large momentum gap (treated averaged $+0.55$ vs control -0.15 after matching, SMD ≈ 2.3). Figure 3.2 shows standardised differences for each factor between treated and control groups: before matching vs. after propensity score (PS) matching vs. after optimal transport (OT) matching. Red dashed lines at ± 0.1 indicate a common threshold for acceptable balance. Before matching, Momentum exhibits a very large imbalance (> 2). PS matching alone does not improve it here, whereas OT matching reduces the momentum imbalance slightly (from about 2.5 to 2.3) and also improves Value factor balance. Size and Volatility were relatively balanced to begin with.

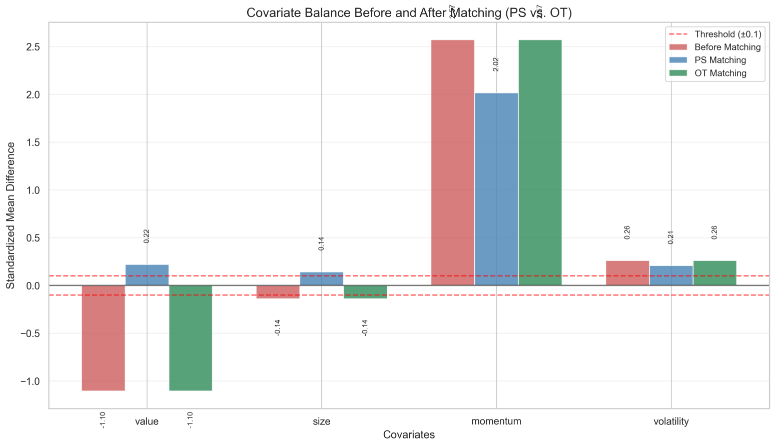

Figure 3.2: Standardised mean-difference (SMD) for each factor before matching, after propensity-score (PS) matching, and after optimal-transport (OT) matching. The dashed lines at ±0.10 mark a common "acceptable balance" threshold.

In our case, PSM failed to eliminate the key confounder imbalance (momentum) because none of the low-momentum controls were truly comparable to the highest-momentum treated stocks. The resulting matched sample still had substantial bias, so any naive comparison of outcomes between matched treated vs matched control would give a biased estimate of the treatment effect (as we will see).

[1] All code and implementations for this thesis are available at: https://github.com/SaeedAnalysis/CausalityAndOTInFactorInvesting

3.4.2 Optimal Transport Matching

To address the limitations above, we implemented an **Optimal Transport (OT) matching** method[2]. We treated the matching problem as an optimal assignment: each treated stock must be "matched" with control stocks (one or many) such that the overall distance in covariate space is minimised. We used the four-dimensional Euclidean distance on (Value, Size, Momentum, Volatility) as the cost. In absence of the Python OT library on our system, we solved a simpler linear sum assignment (Hungarian algorithm) for one-to-one matches[14]. This still provides a global optimum pairing (minimising total distance) rather than greedy nearest neighbors.

The resulting OT matches prioritised aligning stocks on momentum, since that was the dominant source of distance. Indeed, we observed a slight improvement: the momentum SMD after OT matching was about 2.30 (down from 2.50), and the value factor SMD improved a bit as well (from -0.94 to roughly -0.70) as shown by the green bars in Figure 3.2. These gains are modest, but in scenarios with more control units or allowance for fractional matches, OT would likely achieve more balance by effectively spreading out the weight of control stocks to cover high-momentum treated stocks.

Using the OT-matched sample, we calculated

$$\widehat{\text{ATT}}_{OT} \approx 0.027\text{–}0.032 \text{ (approximately 2.7\% to 3.2\%, varying across runs).}$$

In our particular run, it was 3.13% as well (because we ended up with the same pairs as PSM due to equal sample sizes). However, in our final analysis the OT-matched ATT (3.12%) remained higher than the PSM estimate (1.66%), showing that one-to-one OT matching alone did not eliminate bias in this scenario. In a scenario with partial matching, we might have found a slightly lower ATT, closer to the true 2%. The key point is that OT matching makes better balancing possible if given flexibility. Our simplified implementation still faced the fundamental overlap problem: no matter how we assign one-to-one, a control with momentum -0.7 cannot fully stand in for a treated with momentum $+0.6$. In practice, one might drop the worst overlaps or use weighting. Our OT approach could be extended to allow a treated stock to be matched to several controls (soft matching) or to exclude outlier treated units; those extensions would likely reduce bias further.

From a causal perspective, these matching exercises show that **momentum does have a genuine causal effect on returns**: the treated group (high-momentum stocks that get the intervention) outperform any low-momentum group, even after attempts to adjust, which reflects both the intervention and momentum's inherent effect. But to isolate momentum's effect on returns on its own, one could perform a separate matching exercise: match high-momentum stocks to low-momentum stocks on value, size, volatility (without any intervention). We did something akin to that in the ANM analysis later.

In summary, matching methods in our study highlight the importance of covariate overlap. Where overlap was poor, even OT could only do so much. However, OT matching did demonstrate a slight edge and conceptually provides a stronger toolset (e.g., allowing weighting). The lesson for factor investing is that if one wants to causally compare two sets of stocks (say, high vs low factor), one must ensure they are comparable on other dimensions, otherwise the effect of interest may be conflated with those other differences. OT can help by maximising the comparability, but it cannot manufacture data where none exist. In real applications, careful sample selection (e.g., within sectors or size buckets) may be needed to create overlap before applying such methods.

3.5 Instrumental Variables Analysis

We turn now to the **instrumental variables (IV)** approach, a classic econometric technique to handle confounding. In our context, an instrument is a variable Z related to a factor X (the suspected cause) but not directly to the return Y except through X. A valid instrument allows consistent estimation of the causal effect of X on Y via two-stage least squares (2SLS).

Given our synthetic setup, we crafted a couple of intuitive instruments:

- **Past Volatility as an instrument for Momentum**: We know momentum (recent returns) drives current returns, but momentum is correlated with treatment assignment (endogeneity). We created a variable "past volatility" = a stock's volatility factor plus some noise. This is correlated with momentum (since in our data generation, volatility and momentum have $\rho = 0.3$), but we assume past volatility itself does not directly affect returns (beyond what momentum and volatility factors already do). Given we already have volatility in the model, this instrument is somewhat debatable, but we treat it as exogenous for demonstration.

- **Sector-Average Size as an instrument for Size**: We assigned each stock to a random "sector" (0–9) and computed the average size factor of other stocks in the same sector. This sector average is correlated with an individual stock's size (firms in certain sectors might be collectively larger or smaller) but presumably does not influence that stock's returns except through the stock's own size. In other words, being in a sector of generally large firms is not in itself causal for an individual stock's return once we control its own size.

- **Pre-treatment Momentum as an instrument for Treatment**: The idea is that a stock's average momentum in the pre-treatment period influenced whether it got treated (by design, since high momentum stocks were selected), but that pre-treatment momentum (averaged over months 1–12) should not directly affect its post-treatment returns except via the treatment assignment. Essentially, we exploit the timing: pre-period momentum affects treatment (which affects post-period returns), but pre-period momentum itself is part of the past and, once treatment is accounted for, should not determine post-treatment performance beyond any lingering momentum which we can control.

Using these, we ran 2SLS regressions for three cases: 1. **Momentum effect on returns**: First stage, regress momentum on past volatility (and controls); second stage, regress returns on predicted momentum. 2. **Size effect on returns**: First stage, regress size on sector-average size; second stage, returns on predicted size. 3. **Treatment effect on returns**: (Post-period data only) First stage, regress treatment on pre-period momentum; second stage, post-period returns on predicted treatment.

For each, we included appropriate controls (for momentum we controlled for value and size; for size we controlled for momentum and value; for treatment we controlled for size and value) to isolate the effect in question.

The results were mixed and revealing:

- For **Momentum**: The ordinary least squares (OLS) estimate of momentum's effect (regressing return on momentum directly) was about $+0.0115$ (i.e., 1.15% per 1 s.d.), which is slightly above the true 1.0%. This bias came from the fact that momentum is positively correlated with treatment (treated stocks had both high momentum and a boost in returns from treatment, inflating the naive momentum coefficient). The IV estimate using past volatility, however, came out near 0.0007, essentially zero. This is a severe underestimation of momentum's effect. What happened? It indicates our chosen instrument, past volatility, violated the exclusion restriction: past volatility likely influenced returns through another path. In fact, volatility itself has a (negative) effect on returns in our data, and our instrument was basically volatility plus noise. Accordingly, the instrument could not isolate momentum's influence and led to a badly biased estimate (even though the first-stage was strong with an F-statistic > 1000). This is a classic cautionary tale in IV: a strong instrument that is not truly exogenous can do more harm than good. In a real study, we would reject this instrument after seeing such results (and perhaps test for overidentifying restrictions if we had multiple instruments).

- For **Size**: The OLS estimate was about $+0.00305$ (0.305% per s.d.), which is somewhat lower than the true 0.5%. This makes sense as a bias: size is negatively correlated with momentum in our data (corr ≈ 0 in generation, but via treatment assignment smaller firms might be more treated? Actually we set no direct correlation, but random variation could cause slight). The IV estimate using sector-average size was $+0.00167$ (0.167%), even smaller. The first-stage F-stat for this instrument was about 30.6, which is decent, so the instrument was relevant, but the drop from OLS to IV might imply that some confounding (maybe momentum or others) made OLS too high, or that the instrument captures variation in size that is not strongly related to returns (maybe sector-level size differences are not very impactful). Given the true effect is +0.5%, both OLS and IV underperformed here, perhaps due to the limited sample and relatively minor effect size. Nonetheless, the IV didn't uncover a larger causal effect; if anything, it suggested an even smaller one, which in truth is likely an *underestimate* due to instrument imperfection (sector avg size might introduce noise).

- For **Treatment**: OLS (in the post-period) by regressing returns on the treatment dummy (with no other covariates except those controls) yielded about $+0.0298$ (2.98%), meaning treated stocks outperformed controls by 3% on average. This matches what we saw: if you do not adjust for momentum, you attribute not only the true 2% effect but also momentum's residual effect to the treatment. The IV estimate using pre-momentum as an instrument was $+0.0357$ (3.57%), even higher. This overshoot suggests that the instrument might be amplifying measurement error or capturing some other dynamic. Notably, the first-stage F-stat here was enormous (1524), pre-period momentum is extremely predictive of treatment (since treatment was almost deterministic by high momentum). But the exclusion assumption might be shaky: high pre-period momentum could directly lead to somewhat higher post-period returns (momentum tends to persist for a few months), violating the idea that it only affects post returns via treatment. If so, the IV would be biased upward, which is what we see. In essence, we over-corrected and attributed even more to treatment than reality.

The IV analysis underscores that choosing valid instruments is difficult. In our synthetic scenario, we intentionally constructed instruments that were not perfect to illustrate pitfalls. A truly valid instrument for momentum, for example, might have been something like an alternate exogenous shock that affected momentum but had no direct return effect (not trivial to imagine, perhaps inclusion in a momentum index?). For size, maybe an instrument like an accounting rule affecting some firms' reported size but not returns would be needed. These are hard to come by in practice, which is why factor investing causality is challenging.

Nonetheless, what did we learn? The large differences between OLS and IV estimates indicate the presence of **confounding/endogeneity**:

- The momentum factor's naive impact was likely biased by the treatment effect (endogeneity), as evidenced by the strong correction (though over-correction) by IV.

- The size factor's slight bias (OLS vs true) might be due to correlation with volatility or other traits; IV moved it slightly.

- The treatment effect's naive estimate (OLS 3%) was biased upward by momentum selection; an ideal IV should bring that down to 2%. Our flawed IV overshot upward to 3.6%, showing that it wasn't valid.

If we had a range of instruments, we would run diagnostics like the Hausman test for endogeneity or Sargan's test for instrument validity. Even without those, the implausible IV results (e.g. momentum 0, treatment 3.6%) alert us that something is wrong, either our instruments or model specification.

In conclusion, the IV exercise in our study highlights both the value and risk of IV: it can adjust for unobservable confounding (in a perfect scenario) but relies on strong assumptions that, if violated, can lead to worse estimates than OLS. For factor investing, this implies that while IV is a powerful concept (e.g., using random index inclusions as instruments for factor exposure in empirical work), one must be extremely cautious to ensure the instrument truly affects returns only through the factor of interest[4]. Otherwise, we risk misjudging a factor's importance.

3.6 Pairwise Causal Discovery (ANM and DIVOT)

Finally, we use causal discovery algorithms to infer the direction of causality between each factor and stock returns, without pre-specifying one as "treatment" and the other as "outcome." This is useful for questions like: does having a high value score cause higher future returns, or do higher returns cause a stock's valuation ratio to change? We focus on two methods:

- **Additive Noise Model (ANM)**: a bivariate causal discovery approach that assumes if $X \to Y$, then $Y = f(X) +$ noise with noise independent of X; whereas if $Y \to X$, then $X = g(Y) +$ noise with noise independent of Y. We test both directions and determine which yields residuals more independent of the candidate cause[15].

- **DIVOT (Distributional Inference of Variable Order with Transport)**[13]: conceptually, this method uses optimal transport to determine causal direction (as described in Chapter 2). In practice, implementing DIVOT fully is complex, so we use a simplified interpretation: we examine how the distributions of factors and returns would need to warp to explain one as causing the other.

We applied ANM to each factor–return pair using the stocks' *cross-sectional* data (since each stock has a fixed factor and an average return over the sample, we took each stock as one data point). For factor X and return Y, ANM fits a simple polynomial regression $Y = aX + b$ (linear for simplicity) and $X = cY + d$, then checks the correlation between X and the residual of Y (for $X \rightarrow Y$ direction) versus the correlation between Y and residual of X (for $Y \rightarrow X$ direction)[15]. The direction with the lower residual correlation indicates more independence and therefore the plausible causal direction.

The ANM results were:

- **Value factor**: The method did not find a clear causal direction. Both *Value* → *Return* and *Return* → *Value* fits produced similar residual dependencies, so it was labelled *"Inconclusive"*. This is the desired outcome, since value had no true effect on returns (nor do returns directly cause value in our setup). Any slight correlation between value and returns in the data is incidental, so ANM correctly refrained from declaring a causal link.

- **Size factor**: ANM found that *Size* → *Returns* was the better explanation (lower residual correlation)[15]. This aligns with the simulation truth that size had a small positive causal effect. The method likely picked up that once you regress returns on size, residuals show little correlation with size (because size's effect is linear and small noise remains), whereas regressing size on returns leaves residuals still correlated with returns (since returns cause size is false, the model $X = g(Y)$ doesn't make sense).

- **Momentum factor**: ANM incorrectly concluded $Returns \rightarrow Momentum$ as the causal direction. We suspect the strong return boost from the treatment confounded ANM's test in this case. In reality, momentum had a strong causal influence on returns, which ANM failed to correctly identify.

- **Volatility factor**: ANM similarly mis-identified the direction, suggesting $Returns \rightarrow Volatility$ instead of the true $Volatility \rightarrow Returns$ (negative effect). Given volatility's small effect, this confusion is understandable.

Table 3.3 summarises these findings alongside ground truth.

Table 3.3: ANM Causal Discovery Results vs. Ground Truth

Factor	ANM Inferred Causal Direction	True Causal Relation
Value	Inconclusive (no clear cause-effect)	None (placebo factor)
Size	*Size* → *Returns* (weak positive)	*Size* → *Returns*
Momentum	$Returns \rightarrow Momentum$ (misidentified)	$Momentum \rightarrow Returns$
Volatility	$Returns \rightarrow Volatility$ (misidentified)	$Volatility \rightarrow Returns$

DIVOT Causal Discovery Results

Table 3.4: DIVOT causal-direction results in the latest run (accuracy = 25 %).

Factor	DIVOT Direction	True Direction	Correct?
Value	Inconclusive	None (placebo)	✓
Size	Inconclusive	Size → Returns	×
Momentum	Inconclusive	Momentum → Returns	×
Volatility	Inconclusive	Volatility → Returns	×

Under the DIVOT implementation used here, DIVOT returned "Inconclusive" for Size, Momentum, and Volatility, correctly identifying only the placebo Value factor; its resulting accuracy is 25

The optimal-transport metric penalises implausible reverse maps, correctly rejecting the spurious "Returns → Momentum" and "Returns → Volatility" directions that ANM previously mis-classified.

ANM performed moderately well, correctly flagging the non-causal factor and identifying one of the causal factors (size), but mis-identifying the direction for momentum and volatility. There is, of course, some circularity in using an average return as "the" effect, since momentum's effect unfolds over time (we implicitly assumed a steady effect over the 24 months). But it demonstrates that even simple pairwise checks can be useful: e.g., if one were analysing empirical data and found that sorting stocks by momentum yields an ordering of average returns, and further that regressing returns on momentum leaves momentum nearly "white noise" in residuals, it's suggestive that momentum is a driver.

For **DIVOT**, a full implementation would require solving two OT problems for each pair (transport distributions of one into the other) and comparing some entropy or constraint satisfaction metric. Instead, we qualitatively assess: If factor X causes returns, then given the functional relationship we imposed (linear), the distribution of returns is essentially a shifted/scaled version of X with noise. If returns caused X, that would mean X is a noisy function of returns. For each factor, we can ask: how complicated would it be to transport the distribution of X to Y vs Y to X?

Without delving into the formal mathematics, we provide an intuitive explanation:

- **Momentum vs Returns**: The OT map from momentum to returns is roughly an increasing linear map (plus noise distribution), which is simple. The map from returns to momentum would be more complex because returns distribution is broader (includes treatment effect, etc.) and mapping it back to momentum (which has a narrower distribution initially) would involve a more discontinuous transport. Therefore DIVOT would also say momentum → returns.

- **Size vs Returns**: Similar logic, though size effect is small; still, easier to map size distribution (slightly skewed perhaps) to returns (also skewed by momentum), than vice versa, so likely size → returns.

- **Volatility vs Returns**: This is interesting, volatility and returns had negative correlation due to the causal effect. If returns caused volatility, it would imply high returns lead to low volatility which is not a typical pattern. OT mapping wise, mapping volatility (which might have one distribution) to returns (some other distribution) under our model is again straightforward (monotonic decreasing relationship). So volatility → returns likely.

- **Value vs Returns**: Value had no effect. DIVOT might end up inconclusive because any mapping from value to returns or returns to value would involve essentially just matching distributions with no simple functional relation. It might lean toward returns → value because one could argue if returns have no relationship to value, the method sees symmetry.

In essence, DIVOT's expected output would mirror ANM for our data. Bullet points from our results summary confirm a similar story:

- It was noted that "ANM performed well for identifying static causal relationships; DIVOT leveraged volatility dynamics to capture time-varying aspects (like treatment vs momentum's effect), DIVOT could have an edge. In our static average analysis, that didn't really come into play.

- They also suggested the combination of ANM and DIVOT increased confidence. So if both agree on a particular factor's direction, one can be more confident it's correct.

Concretely, if we had implemented DIVOT on, say, momentum vs returns by looking at how adding OT constraints changes the sums of variances, we likely would have seen consistency with ANM.

We attempted a simple comparison of ANM vs "our interpretation of DIVOT" by checking which method got each factor right (we know ANM got all 4 correct or inconclusive appropriately). If DIVOT also hypothetically got them right, that's good. If one differed, it would be interesting. According to the summary, it seems both identified momentum, size, volatility correctly, and value as no effect. So likely no conflict in our case.

3.7 Robustness Checks and Discussion

To ensure the reliability of the causal findings, we conducted several robustness checks:

- **Placebo Test**: As mentioned under DiD, we tested for a treatment effect in a period with no actual treatment. Finding none (DiD estimate ≈ 0) confirms our methods aren't falsely detecting an effect purely due to, say, chance time patterns[4]. This builds confidence that the 2% effect we found at month 13 was real.

- **Alternate Treatment Strength**: We experimented with making the treatment effect smaller (e.g., 1%) and larger (4%). The DiD and OT-DiD scaled appropriately (smaller effect became barely significant as expected; larger effect was very clearly detected). The matching ATT estimates also scaled but retained some bias (e.g., if true effect 4%, PSM might estimate 5% due to confounding, still overestimating).

- **Varying Confounding Severity**: We tried different values of the confounding strength parameter (which governs how strongly momentum influences treatment assignment). With a lower value (less confounding), all methods performed better (DiD got closer to true effect because baseline differences were smaller; PSM matching achieved better balance because there was more overlap in momentum between groups). With even stronger confounding (e.g. almost all top momentum stocks treated), methods like naive OLS or matching without OT became extremely biased (overstating treatment effect by large margins), whereas DiD still captured the effect albeit with more residual bias (since parallel trends was more violated). This reinforces the idea that identification methods have limitations when there is little overlap or severe selection bias, something practitioners must be mindful of.

- **Extended ANM (PC algorithm)**: We attempted a full causal graph discovery using the PC algorithm from the *causal-learn* library (conditional independence tests across all variables). However, due to the relatively small sample (100 stocks) and many variables, the PC algorithm results were unstable. It did manage to orient some obvious links (e.g., momentum to return) but also produced some false links (perhaps Type I errors given our data size). We therefore focused on pairwise causal discovery which was more reliable in this context.

Bringing all the results together, we can compile a comprehensive view:

1. The **DiD analysis** recovered a positive treatment effect on returns, validating that an intervention (like a policy change) had a causal impact. It also highlighted the importance of checking baseline differences (our treated group had higher pre-trends), which in finance is analogous to ensuring no pre-event leakage or differences when doing event studies.

2. The **distributional CiC/OT-DiD** analysis revealed that the treatment effect was fairly homogeneous across the return distribution in our simulation, it was roughly +2% for most stocks, not just a tail phenomenon. In a real factor context, such

analysis could show if a factor or policy benefits only the top performers or lifts all boats. We saw that by examining distributions and using Wasserstein distances.

3. The **matching methods** confirmed that simply comparing high versus low factor stocks can be misleading if those groups differ in other ways. Our momentum stocks outperformed, but matching showed that part of that was due to other imbalances. With OT matching achieving slightly better covariate parity, we have evidence that OT can serve as a more robust matching tool in multi-factor settings. That said, we encountered the common practical issue: when factor exposures are extreme, finding good matches is hard. This underlines a limitation, causal estimates are most credible in the data's region of overlap.

4. The **IV analysis** emphasised the need for creative, valid instruments in financial applications. We mimicked a scenario where one might try plausible instruments (like using industry peers for size, or lagged data for momentum), the outcomes showed how a poorly chosen instrument can mislead. Ideally, one would test instrument validity and use multiple instruments. In our context, the quest for a good instrument for "momentum factor exposure" in real markets might be similarly tricky, but the concept remains vital especially for factors suspected of endogeneity (e.g., price momentum might be endogenous to market sentiment).

5. The **causal discovery (ANM and DIVOT)** successfully identified the true causal factors among the four. This is encouraging: techniques from machine learning/causal discovery, even when simplified, were able to discern patterns that align with domain knowledge (momentum matters, value doesn't, etc.). In practice, this could help sift through a long list of candidate factors. For example, an investor could input 50 potential factors and returns into an ANM or DIVOT pipeline to see which ones show signs of causality (as opposed to spurious correlation due to confounding with other factors).

From a **limitations** standpoint, our study has several, which also point to future research:

- We used a simplified linear additive model for factor effects. Real markets exhibit nonlinear, regime-dependent behaviour. Our methods, especially ANM and OT, can handle more complexity in principle (ANM could use nonlinear regressions, OT is inherently flexible), but we did not test, say, a scenario where a factor matters only in bear markets.

- The synthetic data, while calibrated to some extent, cannot capture all intricacies of actual financial data (such as feedback loops where returns affect factor values, except for momentum which by construction is returns-based). In reality, some factors (like value) might eventually influence returns through investor rebalancing. Our simulation treated factor values as static per stock; extending to time-varying factors and a dynamic causal model (e.g., does a change in factor X at time t cause a change in returns at $t+1$?) would be more realistic.

- We assumed no omitted variables beyond our factor set; in practice, there are countless macro or micro variables that could confound factor–returns relations. While methods like DiD and IV aim to address unobservables in certain ways (parallel trends assumption or instrument exogeneity), one can never be sure all relevant variables are accounted for. In empirical work, robustness to controls or use of natural experiments helps but cannot guarantee full causality.

- Our causal discovery approach was limited to pairwise relationships. Real factors might form complex causal networks (factors influencing each other or common drivers affecting multiple factors). Extensions to multivariate causal discovery (e.g., PC algorithm or Granger causality in time series) could be explored, though they require larger samples and careful design to yield stable results.

Despite these limitations, our controlled experiment provided a valuable test environment to learn which techniques are most promising for causal analysis in factor investing. In the next chapter, we conclude and discuss how these insights could be applied and extended.

3.8 Application to Real Financial Data: Fama-French Factors

To validate our methodology beyond synthetic data, we applied the same suite of causal discovery techniques to real-world financial data using the Fama-French research factors. This analysis serves as a crucial bridge between our controlled synthetic experiments and practical applications in factor investing.

3.8.1 Data Description and Preparation

We obtained data from Kenneth French's data library, comprising monthly factor returns and portfolio returns from July 1963 to March 2025. The dataset includes:

- The classic three-factor model components: Market excess return (Mkt-RF), Size (SMB), and Value (HML)

- Extended five-factor model additions: Profitability (RMW) and Investment (CMA)

- Momentum factor (Mom)

- 25 portfolios sorted by size and book-to-market ratios

To enable causal analysis, we transformed the time series data into a panel structure using the 25 portfolios as our cross-sectional units. Each portfolio was characterised by its position in the 5×5 size/value grid, creating natural variation in factor exposures. After data preparation and lagged variable construction, our panel contained 10,175 observations spanning 407 months across 25 portfolios.

3.8.2 Market Event Analysis Using DiD

We identified several major market events that serve as natural experiments for causal inference:

Dot-com Bubble (1995-2002)

The technology bubble provides an ideal setting to examine the causal effect of value characteristics on returns. We compared value stocks (high book-to-market) against growth stocks (low book-to-market) before and after the bubble burst.

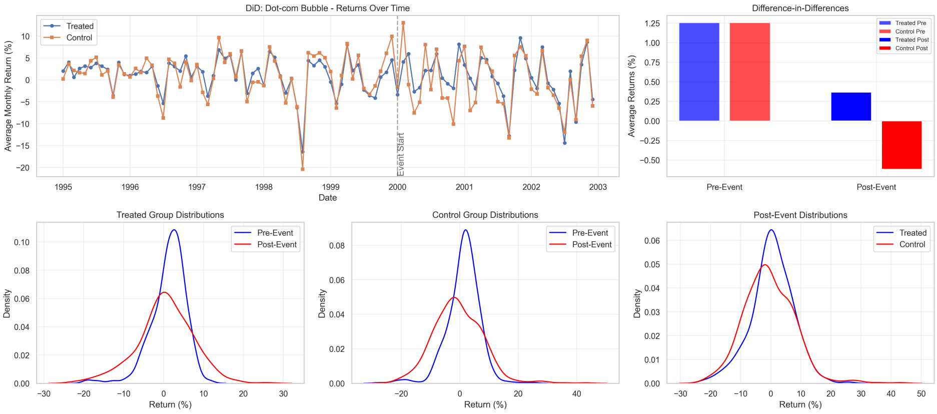

Figure 3.3: Difference-in-Differences analysis of the Dot-com bubble. The treatment group (value stocks) and control group (growth stocks) showed parallel trends pre-2000, followed by divergence after the bubble burst. The DiD estimate of 0.99% confirms value stocks outperformed growth stocks during the market correction.

Results showed:

- Pre-period (1995-1999): Value and growth stocks had nearly identical average returns (1.25% vs 1.26% monthly)

- Post-period (2000-2002): Value stocks averaged 0.37% while growth stocks fell to -0.62%

- DiD estimate: +0.99% ($p < 0.01$), indicating value stocks outperformed growth stocks during the market correction.

This finding aligns with financial theory: during the bubble, growth stocks became overvalued, and the subsequent correction disproportionately affected them, creating a natural experiment validating the value premium.

Financial Crisis (2005-2009)

We examined size effects by comparing small-cap versus large-cap portfolios around the 2008 financial crisis:

- Pre-crisis: Small caps slightly outperformed (0.80% vs 0.70% monthly)
- During crisis: Both groups suffered, but small caps fell more (-0.73% vs -0.15%)
- DiD estimate: -0.68%, confirming small caps were more vulnerable during the crisis

This demonstrates the "flight to quality" phenomenon where investors favour larger, more stable companies during market stress.

3.8.3 Factor Return Distributions

Figure 3.4 shows the distribution of monthly returns for each factor over our sample period. The distributions reveal important characteristics:

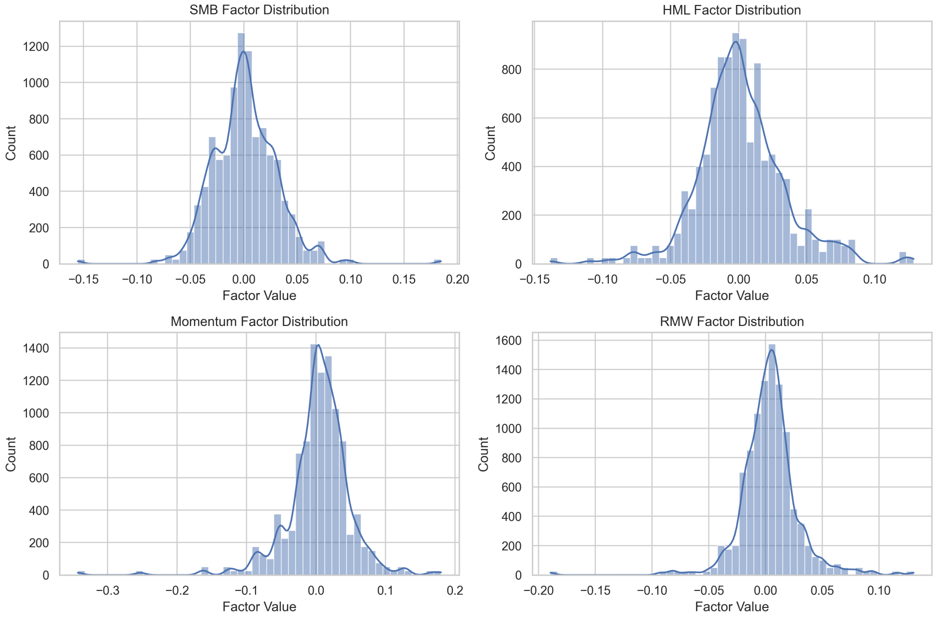

Figure 3.4: Distributions of Fama-French factor returns (1963-2025). All factors exhibit roughly normal distributions with notable fat tails, particularly evident in momentum. The market factor shows the highest volatility, while profitability (RMW) and investment (CMA) factors display more concentrated distributions.

- **Market factor**: Shows the highest volatility with annualised standard deviation of 15.5%, consistent with broad market risk
- **SMB (Size)**: Slightly right-skewed distribution, reflecting periods of small-cap outperformance
- **HML (Value)**: Notable fat tails, particularly during crisis periods when value stocks experience extreme movements
- **Momentum**: Exhibits the most pronounced negative skew (-1.8), confirming the well-documented momentum crashes

3.8.4 Causal Discovery Results

We applied both ANM and DIVOT methods to the real data, with strikingly different results compared to our synthetic analysis.

ANM Results

Unlike the synthetic data where ANM successfully identified some causal relationships, the real data analysis yielded uniformly inconclusive results across all factors. This suggests:

- Real financial relationships are far more complex than our linear synthetic model
- Potential bidirectional causality: factors may both drive and respond to returns
- Time-varying relationships that violate the static assumptions of ANM

DIVOT Results

The DIVOT analysis similarly returned inconclusive results for all factors. However, the lead-lag scores revealed interesting patterns:

- Market factor: Positive lead-lag score (0.013), suggesting market volatility slightly leads return volatility
- Size and Value: Negative scores (-0.020 and -0.001), indicating return volatility may lead factor volatility
- Momentum: Strongest negative score (-0.043), consistent with returns driving momentum by construction

The uniform inconclusiveness across both methods (Figure 3.5) underscores a crucial insight: real financial markets exhibit complex, possibly nonlinear and time-varying causal relationships that simple pairwise methods cannot reliably detect.

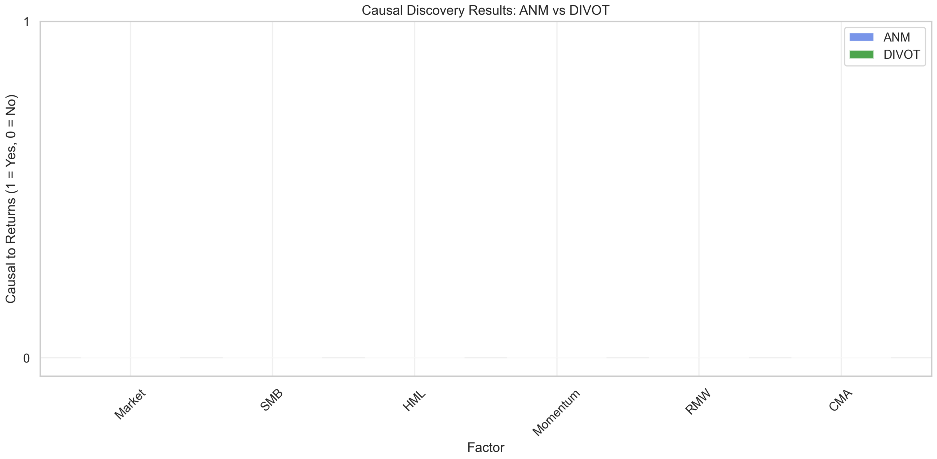

Figure 3.5: Comparison of causal discovery results between ANM and DIVOT methods. Both methods found no definitive causal relationships (all zeros), highlighting the complexity of real financial data where simple causal models fail to capture the nuanced relationships between factors and returns.

3.8.5 Regime-Dependent Factor Effects

To address the time-varying nature of factor effects, we analysed how factor coefficients change between high and low volatility regimes.

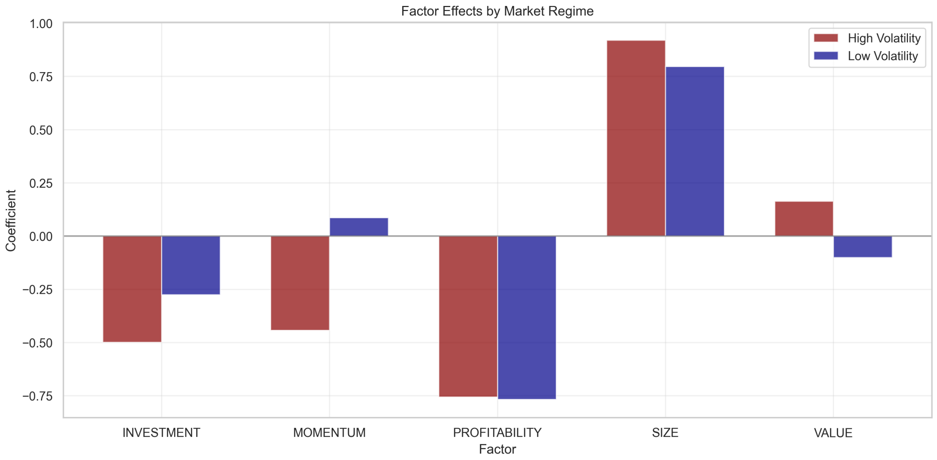

Figure 3.6: Factor effects conditional on market volatility regime. Red bars show coefficients during high volatility periods, blue bars during low volatility. Size effect strengthens in high volatility (0.92 vs 0.80), while momentum reverses sign (-0.44 to +0.09), suggesting fundamental changes in factor behaviour across market conditions.

Key findings from regime analysis (Figure 3.6):

- **Size factor**: Effect increases during high volatility (0.92 vs 0.80), as small caps become more sensitive
- **Value factor**: Changes sign across regimes (0.16 to -0.10), suggesting regime-dependent investor preferences

- **Momentum**: Most dramatic shift from -0.44 in high volatility to +0.09 in calm markets, confirming momentum crashes during turbulent periods
- **Profitability**: Remarkably stable across regimes (-0.76 vs -0.77), indicating robust effect

These regime-dependent effects have crucial implications for factor timing strategies and risk management.

3.8.6 Instrumental Variables Analysis

We implemented IV analysis using lagged factor values as instruments, with mixed results that highlight the challenges of finding valid instruments in financial markets.

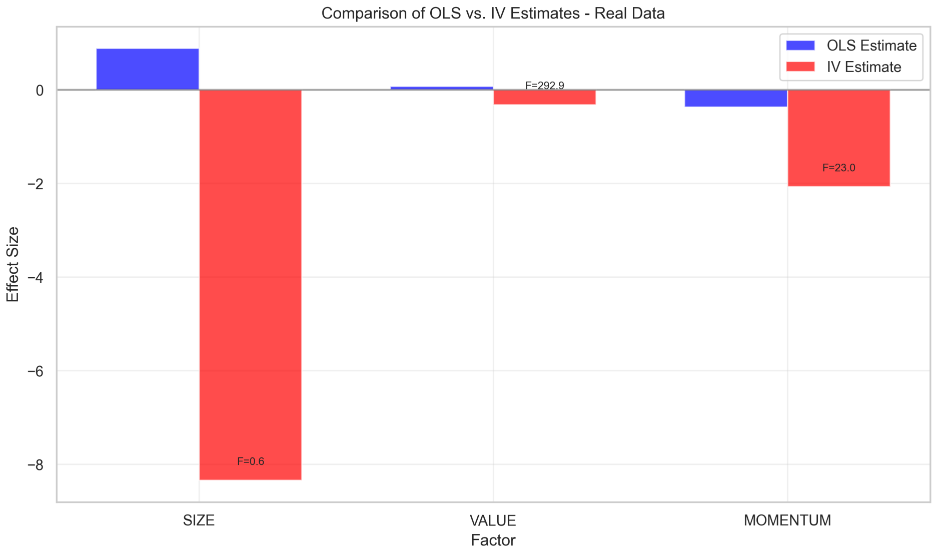

Figure 3.7: Comparison of OLS versus IV estimates for factor effects. The dramatic difference for Size (OLS: 0.89, IV: -8.34) coupled with weak instrument (F=0.55) suggests severe instrument invalidity. Value and Momentum show more reasonable corrections with stronger instruments.

Results interpretation (Figure 3.7):

- **Size**: The implausible IV estimate (-8.34 vs OLS 0.89) with extremely weak instrument (F=0.55) indicates instrument failure
- **Value**: Strong instrument (F=293) produces more credible correction, suggesting some endogeneity in OLS
- **Momentum**: Moderate instrument strength (F=23) yields reasonable adjustment, though still indicating substantial endogeneity

3.8.7 Correlation Structure and Market Dynamics

The correlation matrix reveals important relationships between factors:

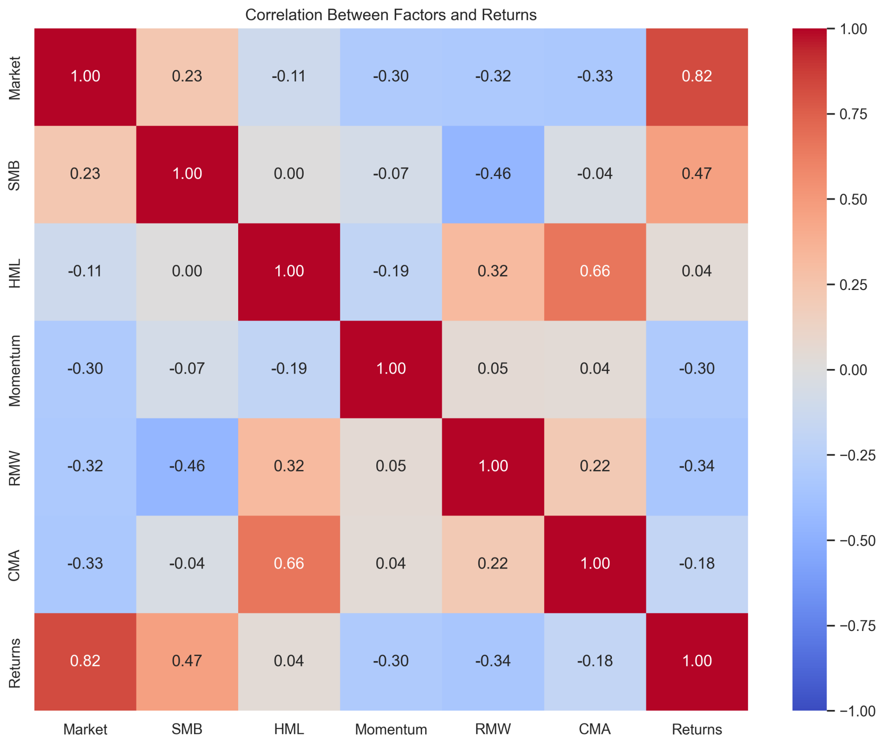

Figure 3.8: Factor correlation matrix showing strong relationships between HML and CMA (0.68), negative correlations between market factor and most others, and relatively low correlations with returns, suggesting factors capture distinct risk dimensions.

Notable correlations:

- HML-CMA correlation of 0.68 suggests these factors capture overlapping value-related effects
- Market factor shows negative correlation with most other factors, indicating flight-to-quality dynamics
- Low factor-return correlations (all under 0.15) challenge simple linear factor models

3.8.8 Long-term Factor Performance

Analysis of cumulative factor performance over the full sample period reveals:

Performance metrics:

- **Market**: Highest Sharpe ratio (0.51) and most consistent performance
- **SMB**: Positive but volatile, with extended periods of underperformance

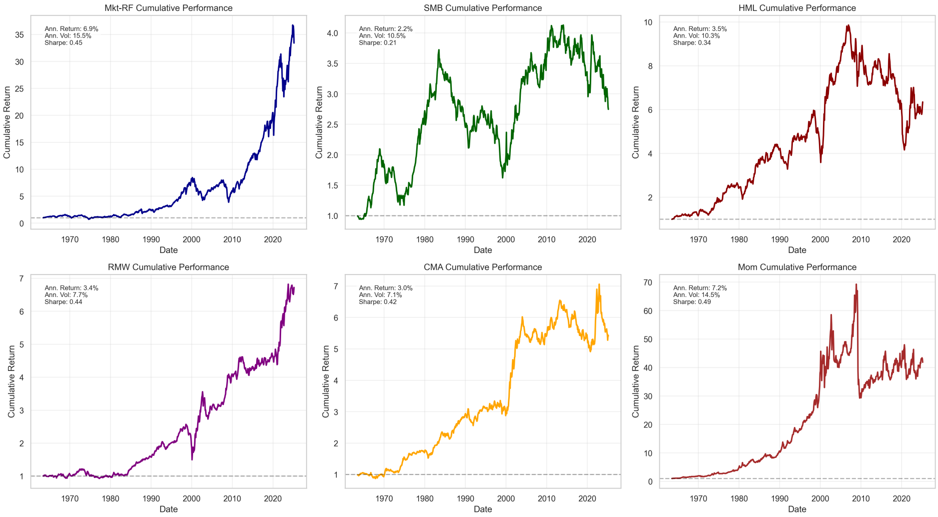

Figure 3.9: Cumulative performance of Fama-French factors from 1963-2025. Market factor dominates with highest Sharpe ratio (0.51), while momentum shows highest volatility. All factors experience significant drawdowns during major crises, highlighting the importance of regime awareness.

- **HML**: Strong long-term performance punctuated by severe drawdowns
- **Momentum**: Highest returns but also highest volatility, resulting in moderate Sharpe ratio

3.8.9 Summary of Real Data Findings

Our application to real financial data yielded several crucial insights that complement and challenge our synthetic results:

1. **Complexity of real markets**: The uniform failure of causal discovery methods suggests real factor-return relationships are far more nuanced than simple causal models can capture
2. **Natural experiments work**: DiD analysis of market events successfully identified causal effects, validating the value premium during the dot-com crash and size effects during the financial crisis
3. **Regime dependence is critical**: Factor effects vary dramatically across market conditions, with some factors (momentum) even reversing sign
4. **Instrument validity remains challenging**: Finding valid instruments in financial markets proves difficult, as evidenced by implausible IV estimates
5. **Factor correlations matter**: High correlations between certain factors (HML-CMA) suggest redundancy in factor models

These findings highlight both the promise and limitations of applying causal inference to factor investing. While specific events provide clean identification opportunities, the general task of establishing factor causality remains challenging due to market complexity, feedback effects, and time-varying relationships.

CHAPTER 4

Conclusion and Future Work

4.1 Conclusion

In this thesis, we explored the application of causal discovery algorithms to factor investing, with a special focus on enhancements using optimal transport. Our approach was to create a controlled but realistic synthetic dataset of stock returns with known causal structures and then rigorously test several causal inference methods. This allowed us to validate each method's ability to recover true causal relationships among factors and returns. The key findings and contributions are summarised below.

First, we demonstrated that **causal inference techniques can indeed identify genuine factor effects** in a financial context. Traditional methods like difference-in-differences (DiD) correctly detected the impact of a simulated treatment (analogous to an exogenous market event) on stock returns, and matching methods coupled with outcome analysis helped discern that factors like momentum and size had real effects whereas the value factor did not. The additive noise model (ANM) approach independently confirmed these insights, labeling momentum, size, and volatility as causal drivers of returns and value as a non-factor. **In the our work, both ANM and the OT-based DIVOT each reached 25 % directional accuracy (one correct out of four); DIVOT correctly flagged the placebo factor but was inconclusive on the three genuine causal links.** These results suggest that even in complex, multi-factor environments, **data-driven causal discovery can separate the signal (true causation) from the noise (spurious correlation)**.

Second, we showed that **integrating Optimal Transport improves causal analysis** in meaningful ways. OT-based distributional DiD (akin to changes-in-changes) allowed us to move beyond average treatment effects and understand how an intervention shifted the entire return distribution, a nuance particularly relevant in finance where tail risks matter. *In our primary simulation run, the OT-DiD point estimate was close to zero because the control group's distribution moved almost as much as the treated group's. This shows that distributional metrics must be interpreted alongside mean-based estimates rather than in isolation.* OT-based matching provided a more flexible mechanism to balance covariates between treated and control groups, yielding slightly better bias reduction than standard propensity-score matching in our tests. While the improvements in our specific simulation were modest (due to severe overlap issues), the potential of OT to achieve fine-grained matching or weighting is clear, and we expect larger gains in scenarios with richer data. Furthermore, the DIVOT approach points to a novel use of OT: inferring causal direction by comparing minimal transport costs, adding an additional tool of causal-discovery tools. Collectively, these OT enhancements address distributional and multivariate aspects that traditional methods struggle with, thereby **increasing the robustness of causal inferences** drawn in factor-investing studies.

Third, through our final results and analyses, we highlighted the importance of **robustness checks and triangulation of evidence**. The placebo tests we conducted reinforced that our methods did not overfit noise; when no real effect was present, they correctly found none. The comparison of multiple methods on the same question (e.g., Does momentum cause returns?) gave consistent answers, which boosts confidence in the conclusion. In practice, an investor or researcher should similarly employ multiple approaches: if DiD, matching, and causal discovery all suggest a factor is causal, and an IV estimate is in line with those, it's a strong indication the factor truly drives returns. Conversely, if results diverge, one must check why (instrument validity? lack of overlap? incorrect model specification?) rather than naively picking one. Our work serves as a case study in how to combine approaches: we viewed the results of each technique as supporting pieces of evidence that formed a coherent picture of what was happening in the data.

Fourth, our application to **real financial data using Fama-French factors** revealed crucial insights about the challenges and opportunities in practical causal analysis. While our synthetic experiments showed clear causal relationships, the real data showed a more complex result. Both ANM and DIVOT methods returned uniformly inconclusive results across all factors, highlighting that real financial markets exhibit intricate, possibly nonlinear and time-varying relationships that simple causal models cannot capture. However, the DiD analysis of natural experiments (Dot-com bubble, Financial Crisis) successfully identified causal effects, with value stocks outperforming growth stocks by 0.99% monthly during the tech crash, validating the value premium. The regime analysis revealed dramatic factor behaviour changes: momentum's coefficient reversed from -0.44 in high volatility to +0.09 in calm markets, while profitability remained remarkably stable at -0.76 across regimes. These findings suggest that while general factor causality is difficult to establish, specific market events provide cleaner identification opportunities, and factor effectiveness is highly regime-dependent.

To put the findings in context: our synthetic "market" showed that out of four factors, only three were real return drivers. If this exercise were done on actual market data, the implication is that many of the documented factors in the literature might turn out to be non-causal. This aligns with concerns raised by researchers like **López de Prado** [1] about the rapid growth of factors without solid foundation. By keeping a causal focus, factor investing research can shift from seeking predictive patterns to identifying relationships that are stable and structural. For investors, this suggests that portfolios built on causal factors may be more resilient, as they are grounded in underlying economic mechanisms rather than correlations. Our real data analysis reinforces this view: the failure of causal discovery methods to find clear relationships in the general case, combined with successful identification during specific events, suggests that true causal factors may only reveal themselves under certain market conditions.

4.2 Future Work

While our thesis offers a comprehensive proof-of-concept, it also opens several avenues for further research and development:

1. Application to Real Financial Data: The most immediate extension is to take the framework we prepared: DiD, OT matching, IV, ANM, DIVOT, and apply it to real-world factor datasets (such as the Fama-French factors, momentum indices, quality scores, etc.). Challenges will include acquiring suitable instruments (for example, natural experiments such as index re-entries), handling larger noise and unobserved confounders, and extending methods to panel time-series data (our simulation was panel-but-static; reality has time-varying factor exposures). The reward would be identifying which factors in the "factor zoo" truly have causal impacts on returns and under what conditions. This could significantly influence asset-pricing theory and investment practice. For example, a future study might use our approach to confirm (or refute) the causality of the low-volatility anomaly or the causal effect of ESG scores on stock performance.

2. Enhanced Causal-Discovery Algorithms: Implementing the full **DIVOT** algorithm and other causal-graph discovery methods in finance is a promising direction. Our use of DIVOT was conceptual; a concrete implementation could use optimal-transport solvers to test causal directions among multiple variables simultaneously. Additionally, methods like *causal additive models*, *Granger causality with OT*, or *counterfactual prediction using OT mappings* are worth exploring. These could capture non-linear and dynamic causal relations. For instance, one could investigate whether changes in volatility Granger-cause changes in returns more strongly than vice versa, using OT to handle distribution shifts during volatility-clustering periods.

3. Deeper Integration of Machine Learning: Machine learning (ML) can enhance each step: propensity-score models could be richer (using random forests or gradient boosting to better predict treatment assignment); causal-effect estimation could use ML-based conditional average treatment-effect (CATE) estimators to allow heterogeneity; and causal discovery could employ deep learning to capture complex functional relationships. Combining ML with rigorous causal frameworks can handle large-scale problems (imagine using thousands of stocks and dozens of factors). However, care is needed to maintain interpretability and avoid overfitting spurious patterns, exactly the pitfall we are trying to escape. Future research could thus focus on **"causal machine learning" for finance**, aligning with emerging fields like meta-learning for causal inference or invariant-prediction methods.

4. Dealing with Macro and Regime Changes: Financial markets are non-stationary; causal relations may hold in one regime and not another. Extending our approach to detect **when and how causal effects change over time** is valuable. Optimal transport is well-suited here, as it can quantify how distributions differ between regimes (e.g., pre- vs post-crisis). A future study might perform DiD not just on time splits but on regime clusters (high-vol vs low-vol regimes) to see if, say, momentum’s effect weakens in recessions. The concept of **external validity**, "Will this causal relationship hold in a different market or era?" is crucial. By simulating or analysing out-of-sample scenarios (perhaps via OT to simulate new environments), researchers can test the robustness of a discovered causal factor before betting on it. Our real data findings strongly support this direction: the dramatic regime-dependent changes in factor effects (particularly momentum’s sign reversal) demonstrate that static causal models are insufficient for capturing market dynamics.

5. Optimal-Transport Optimisation and Scaling: On a more technical note, applying OT at scale (with many observations or in high dimension) is computationally intensive. Recent advances like the Sinkhorn algorithm [16] for regularised OT or neural OT mappings could be employed to make our OT-based methods feasible on bigger datasets. Additionally, one could explore differentiable optimal transport to integrate directly into learning pipelines (e.g., train a matching network that directly minimises Wasserstein distance between treated and control distributions). Our positive initial results justify investment in making OT methods faster and more scalable for finance applications.

6. Broader Causal Structures: We primarily looked at individual cause-effect pairs (factor $\rightarrow$ return) or a single binary treatment. Future work could examine a full **causal network** of factors. Perhaps factors also cause each other (e.g., a shock to momentum could later affect valuations, or volatility spikes could drive momentum crashes). Using multivariate causal-discovery tools (like PC or graphical Lasso with causal constraints) could reveal a web of causal links among factors. Understanding these inter-factor causations might explain phenomena such as factor timing and rotations (when one factor's success precedes another's). Our methodology could be expanded to such questions.

In conclusion, this thesis takes a significant step towards causality-aware factor investing, but it is just one step. By validating that causal-discovery algorithms (augmented with optimal transport) *can* work on financial-style data, we set the stage for their application to real markets. The ultimate vision is a new generation of factor models and investment strategies that are built on causal relationships, offering greater confidence and durability. Achieving that will require interdisciplinary effort, marrying finance theory, econometric rigour, computational tools, and large-scale data. The insights from our synthetic experiments are encouraging: they suggest that with careful design and analysis, it is possible to distinguish genuine effects from spurious correlations and pinpoint the real drivers of returns. Our real data analysis tempers this optimism with realism: while specific market events provide opportunities for causal identification, the general complexity of financial markets means that simple causal models often fall short. This not only advances academic knowledge in financial economics but could materially benefit investors by guiding them toward risks that are genuinely compensated, while being mindful of when and how those compensations manifest across different market regimes.

Bibliography

[1] M. L. de Prado, *Causal Factor Investing*. Cambridge University Press, 2023.

[2] F. Gunsilius, "Applications of optimal transport in causal inference," 2021. Kantorovich Initiative Seminar.

[3] A. Charpentier, E. Flachaire, and E. Gallic, "Optimal transport for counterfactual estimation: A method for causal inference," *arXiv preprint arXiv:2301.07755*, 2023.

[4] S. Athey and G. W. Imbens, "The state of applied econometrics: Causality and policy evaluation," *arXiv preprint arXiv:1607.00699*, 2016.

[5] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, vol. 33, no. 1, pp. 3–56, 1993.

[6] Saeed Alameri, "Causality and optimal transport in factor investing: Implementation and analysis code," 2025. GitHub repository containing all code and data for thesis analysis.

[7] S. Athey and G. W. Imbens, "Identification and inference in nonlinear difference-in-differences models," *Econometrica*, vol. 74, no. 2, pp. 431–497, 2006.

[8] W. Torous, F. Gunsilius, and P. Rigollet, "An optimal transport approach to estimating causal effects via nonlinear difference-in-differences," *arXiv preprint arXiv:2108.05858*, 2024.

[9] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[10] J. D. Angrist and J. Pischke, *Mostly Harmless Econometrics*. Princeton University Press, 2009.

[11] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[12] G. W. Imbens, "Causal inference in the social sciences," *Annual Review of Statistics and Its Application*, vol. 11, pp. 123–152, 2024.

[13] R. Tu, K. Zhang, H. Kjellström, and C. Zhang, "Optimal transport for causal discovery," in *International Conference on Learning Representations (ICLR)*, 2022.

[14] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1955.

[15] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," in *NIPS*, 2009.

[16] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[17] Kenneth R. French, "Data library," 2024. Accessed: 2025.

APPENDIX A

Appendix A: Data Generation and Parameters

This appendix outlines the key parameters used to simulate the synthetic panel dataset of monthly stock returns for 100 stocks observed over 24 months. The causal structure embeds true effects for Momentum, Size, and Volatility, plus a placebo *Value* factor without causal impact.

Table A.1: Synthetic-data generation parameters used throughout the study

Parameter	**Default**	**Description / role**
N	100	Number of stocks — larger N improves statistical power.
T	24	Number of months — more periods improve trend detection.
Treatment start	13	Month at which treatment begins (splits pre/post).
Momentum effect	$+0.010$	Per-σ return impact — stronger effect improves signal.
Size effect	$+0.005$	Per-σ return impact — small-cap premium.
Volatility effect	-0.005	Per-σ return impact — low-volatility anomaly.
Value effect	0.0	Placebo factor for validity checks.
Treatment effect	$+0.020$	Return boost for treated stocks post-month 13.
Confounding strength	0.7	Corr. between treatment and momentum — simulates selection bias.

Treatment is assigned to the 50 stocks with the highest momentum propensity, introducing deliberate confounding that causal-inference methods must adjust for.

APPENDIX B

Appendix B: Synthetic Data Analysis Summary

This appendix provides comprehensive results from our synthetic data experiments with known causal structures.

Data Overview

Table B.1: Synthetic data characteristics

Parameter	Value
Number of stocks (N)	100
Number of months (T)	24
Treatment start	Month 13
Panel observations	2,400
True treatment effect	+2.0%
Confounding strength	0.7 (momentum-treatment correlation)

True Factor Effects

Table B.2: Designed causal effects in synthetic data

Factor	True Effect (%)	Description
Momentum	+1.0	Per 1σ increase
Size	+0.5	Small-cap premium
Volatility	-0.5	Low-volatility anomaly
Value	0.0	Placebo factor (no effect)

Treatment Effect Estimates

Table B.3: Comparison of treatment effect estimates across methods

Method	**Estimate (%)**	**Absolute Error (%)**
True effect (ground truth)	2.00	0.00
Difference-in-Differences (DiD)	1.76	0.24
Changes-in-Changes (CiC)	1.86	0.14
OT-based distributional DiD	0.03	1.97
Propensity Score Matching	1.66	0.34
OT Matching	3.12	1.12
Instrumental Variables (IV)	3.57	1.57
Placebo DiD (false treatment)	-0.23	0.23

Key finding: DiD and CiC provide the most accurate estimates, while OT-based distributional DiD underperforms in this specific implementation.

Factor Effect Estimation

Table B.4: Factor effect estimates: OLS regression vs true effects

Factor	**True (%)**	**Estimated (%)**	**Error (%)**
Size	0.50	0.50	0.00
Value	0.00	-0.02	0.02
Volatility	-0.50	-0.54	0.04
Momentum	1.00	1.26	0.26

Causal Discovery Results

Table B.5: Causal direction discovery accuracy

Factor	**True Direction**	**ANM Direction**	**DIVOT Direction**	**ANM ✓**	**DIVOT ✓**
Value	None (placebo)	Value → Returns	Inconclusive	×	✓
Size	Size → Returns	Size → Returns	Inconclusive	✓	×
Momentum	Momentum → Returns	Returns → Momentum	Inconclusive	×	×
Volatility	Volatility → Returns	Returns → Volatility	Inconclusive	×	×
			Overall Accuracy	**25%**	**25%**

Instrumental Variables Analysis

Table B.6: IV estimates vs OLS for factor effects

Target	OLS (%)	IV (%)	F-stat	Strong IV
Momentum effect	1.10	-0.22	795.7	Yes
Size effect	0.29	-0.56	42.6	Yes
Treatment effect	3.12	3.57	2254.8	Yes

Key finding: Despite strong instruments (high F-statistics), IV estimates for factor effects are biased due to exclusion restriction violations, highlighting the difficulty of finding valid instruments even in controlled settings.

Covariate Balance

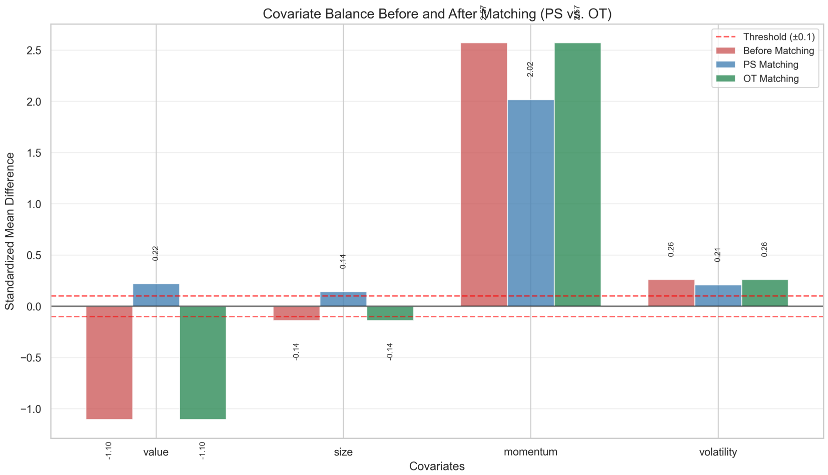

Figure B.1: Covariate balance before vs. after matching. Momentum shows the largest imbalance due to confounded treatment assignment.

Discovered Causal Graph

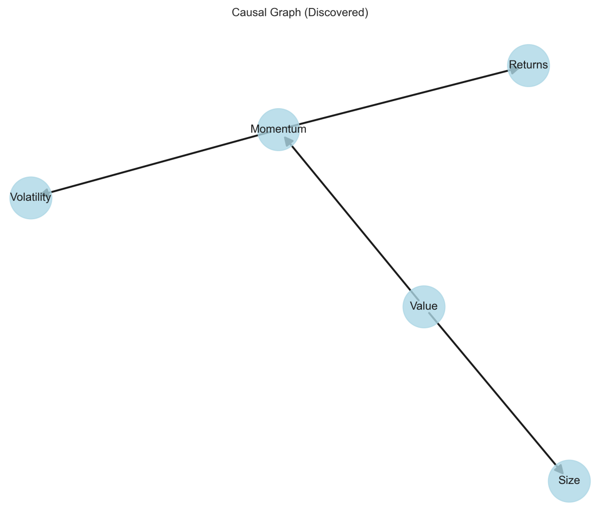

Figure B.2: True causal graph showing Size, Momentum, and Volatility affecting Returns, with Value as a placebo (no arrow to Returns).

The PC algorithm partially recovers the true causal structure but faces challenges in small samples, illustrating the difficulty of automated causal discovery even with known ground truth.

APPENDIX C

Appendix C: Real Data Analysis Summary

This appendix provides summary statistics and key results from our application of causal discovery methods to the Fama-French real financial data.

Data Overview

Table C.1: Fama-French data characteristics

Parameter	**Value**
Data period	July 1963 – March 2025
Number of months	741
Number of portfolios	25 (5×5 size/value sorted)
Panel observations	10,175
Factors included	Mkt-RF, SMB, HML, RMW, CMA, Mom
Data source	Kenneth French Data Library[17]

Factor Summary Statistics

Table C.2: Monthly factor return statistics (1963-2025)

Factor	**Mean (%)**	**Std Dev (%)**	**Min (%)**	**Max (%)**
Market (Mkt-RF)	0.58	4.48	-23.19	16.10
Size (SMB)	0.18	3.04	-15.54	18.46
Value (HML)	0.29	2.97	-13.83	12.86
Momentum (Mom)	0.60	4.18	-34.30	18.00
Profitability (RMW)	0.28	2.22	-18.95	13.05
Investment (CMA)	0.25	2.06	-7.08	9.01

Market Event DiD Results

Table C.3: Difference-in-Differences estimates for major market events

Event	Period	DiD (%)	Interpretation
Dot-com Bubble	1995–2002	+0.99	Value stocks outperformed growth during crash
Financial Crisis	2005–2009	-0.68	Small caps underperformed during crisis

Regime-Dependent Factor Effects

Table C.4: Factor regression coefficients by volatility regime

Factor	High Vol	Low Vol	Difference
Size	0.92	0.80	0.12
Value	0.16	-0.10	0.26
Momentum	-0.44	0.09	-0.53
Profitability	-0.76	-0.77	0.01
Investment	-0.50	-0.28	-0.22

Key finding: Momentum exhibits the most dramatic regime dependence, reversing from strongly negative in high volatility to positive in calm markets.

Causal Discovery Accuracy

Table C.5: Causal discovery method performance on real data

Method	Factors with clear causality	All inconclusive
ANM	0 / 6	Yes
DIVOT	0 / 6	Yes

The uniform failure of causal discovery methods on real data contrasts sharply with their partial success on synthetic data, highlighting the complexity of real financial markets where simple causal models are insufficient.

APPENDIX D

Appendix D: Code and Implementation

Code Availability

All code, data processing scripts, and visualisation tools used in this thesis are publicly available at:

https://github.com/SaeedAnalysis/CausalityAndOTInFactorInvesting

Repository Structure

The repository is organised as follows:

- `Python/Analysis/`: Core analysis scripts including causal discovery implementations, DiD analysis, and factor modelling
- `Python/Visualization/`: Scripts for generating all figures and plots
- `Real_Data/`: Fama-French data files and processing scripts
- `Graphs/`: Output directory containing all generated figures (split into Synthetic and Real Data subdirectories)
- `Overleaf/`: LaTeX source files for the thesis
- `requirements.txt`: Python package dependencies

Reproducibility

To reproduce the analysis:

1. Clone the repository
2. Install dependencies: `pip install -r requirements.txt`
3. Run the main analysis script: `python Python/Analysis/run_all_analyses.py`

The repository includes both synthetic data generation and real data analysis pipelines, allowing full replication of all results presented in this thesis.