Saeed Alameri
21 December, 2025

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

Introduction

Causal discovery refers to identifying cause-and-effect relationships rather than mere correlations. In finance, it means determining whether factors (e.g. value, momentum) actually cause changes in returns or risk, rather than simply correlating with them. This distinction is especially important in factor investing, where numerous "factor premiums" have been proposed - many may be spurious if only discovered through correlations [1]. This proliferation is sometimes called the "factor zoo." As López de Prado warns, failing to identify causal relationships can lead to an immature, purely phenomenological stage for factor investing [1]. Hence, employing causal discovery methods is essential to differentiate genuine, robust factor drivers from coincidental patterns.

A key emerging tool for causal analysis is optimal transport (OT), a mathematical framework for comparing and transforming probability distributions. It has proven valuable for handling distributional differences and heterogeneous effects in observational data [2]. In econometrics, OT has been adopted to generalize common causal inference methods- such as difference-in-differences, synthetic controls, and matching - beyond strict average-based assumptions. By "aligning" treated and control groups in a principled way, OT can help construct counterfactual scenarios with minimal distortion.

In the context of factor investing, combining OT with causal discovery could yield deeper insights. For instance, one could improve matching across groups of stocks or measure how an "intervention" (e.g. a factor index inclusion) shifts the entire distribution of returns. This thesis proposes to apply existing causal discovery algorithms to financial data and investigate whether OT-based enhancements offer more robust insights about genuine factor effects.

Literature Review

Causal Discovery Methods in Finance

Two widely used approaches in finance for inferring cause-and-effect are difference-in-differences (DiD) and matching.

- **Difference-in-Differences:** DiD compares outcomes before vs. after a treatment for a treated group versus a control group. The crucial assumption is parallel trends: that in the absence of treatment, the two groups would follow parallel paths. However, DiD typically focuses on average outcomes. To capture distributional shifts, Athey and Imbens introduced changes-in-changes (CiC), which uses quantile functions to see how an entire outcome distribution changes [3]. CiC relaxes the standard DiD assumption by allowing heterogeneous treatment effects. In higher dimensions, this connects naturally to optimal transport, via a concept called cyclic monotonicity.

- **Matching:** Matching attempts to mimic a randomized experiment by pairing each treated unit with a comparable untreated unit based on observed covariates (firm size, industry, etc.). Standard implementations can fail in high dimensions or when the control group lacks good "matches." Optimal transport can remedy this by allowing for a more flexible, distribution-level matching. Instead of forcing one-to-one matches, OT-based matching can split or discard some observations, reducing bias and focusing on areas of overlap [2].

Machine Learning-Based Causal Discovery

Beyond these econometric designs, functional causal models (FCMs) assume each effect is a function of its direct causes plus noise. A classic example is additive noise models (ANMs), where $Y = f(X) + \varepsilon$ ($\varepsilon \perp X$) if X causes Y.

Such structural assumptions can reveal causal directions in observational data. Methods like LiNGAM or constraint-based searches can discover causal graphs, but must cope with the noise and potential nonstationarity in financial time series. While economists often prefer structured designs (e.g. DiD, instrumental variables), FCM approaches can still help reveal causal relationships among factors.

Optimal Transport in Causal Inference

A growing body of work uses OT to enhance or extend standard causal methods:

- OT for Difference-in-Differences: Torous, Gunsilius, and Rigollet propose a nonlinear DiD that estimates a full distributional effect [3]. By optimally mapping the treated group's pre-treatment distribution to its post-treatment distribution (and accounting for the control group's evolution), their framework captures richer heterogeneity than standard DiD. This can be especially relevant in finance if a policy or factor intervention alters risk or higher moments, not just average returns.
- OT for Matching: Instead of matching each treated unit to a single control, OT finds a weighted transport plan that "balances" covariate distributions. Some variants allow partial matching (unbalanced OT), so if certain treated units have no close match, they can be dropped or down-weighted [2]. In finance, this approach could better reflect practical constraints when analyzing, say, the causal effect of a corporate action on stock returns.
- OT for Causal Direction: Tu et al. develop a framework (DIVOT) that interprets cause-effect pairs as a dynamical system, using OT to map the distribution of one variable into the other [4]. Under functional causal model constraints, the unique OT map can reveal which variable is cause

vs. effect. While not yet widely deployed in factor investing, the approach could clarify whether, for example, volatility leads returns or vice versa, by examining how distributions must “flow.”

- OT for Counterfactuals: Charpentier et al. show how OT can generate individualized counterfactuals, e.g. “How would a firm’s return change if it didn’t have high exposure to momentum?” [5]. They do so by transporting each observation’s covariates across treatment groups at the same “rank,” yielding a distributionally consistent counterfactual. In factor investing, this could reveal whether an observed premium is actually attributable to factor exposure.

Recent Developments in Applied Econometrics and Social Science Causal Inference

In addition to the methods highlighted above, the broader applied econometrics literature has provided a unifying perspective on how to conduct credible empirical studies using observational and experimental data. Athey & Imbens review emerging methods in program evaluation, including synthetic controls, advanced DiD variants, and regression discontinuity. They emphasize the importance of supplementary analyses such as placebo tests or sensitivity checks to support identification assumptions. [6] This viewpoint complements the OT-based approaches by stressing that well-chosen identification strategies and robust designs ultimately improve causal inference.

Meanwhile, Imbens surveys causal inference in the social sciences, focusing on how the potential-outcomes framework has broadened the scope of econometric research. The discussion covers both classical methods (e.g., instrumental variables, unconfoundedness-based matching) and newer distributional techniques akin to OT. Imbens also underscores how integrating machine learning and large datasets is reshaping causal inference, with particular relevance to finance where diverse factor data and big microstructure data are increasingly available. [7]

These developments in applied econometrics reinforce the thesis’s objective of combining recognized causal discovery approaches (DiD, matching, FCMs) with optimal transport. In essence, a well-chosen identification strategy, supported by advanced econometric diagnostics, can leverage OT’s ability to address heterogeneity and distributional shifts, moving factor investing research closer to robust, truly causal insights.

Research Objectives

1. **Apply Causal Discovery Algorithms**: Implement established causal inference methods: difference-in-differences, matching, and functional causal models - on factor investing data (e.g. Fama-French factors, extended factor libraries). Identify “treatments” (like a major factor-related event) to estimate causal impacts on returns or risk.
2. **Incorporate Optimal Transport**: Enhance each method with OT-based techniques (e.g. distributional DiD, OT matching, DIVOT for causal direction). Assess whether OT addresses biases or reveals deeper effects.
3. **Evaluate Efficacy and Robustness**: Compare OT-augmented approaches vs. traditional ones. Test if certain causal relationships become more evident or stable with OT, and check for improvements in distributional metrics (like tail risk).
4. **Contribute to Financial Causality**: Showcase how robust causal methods can identify which factors truly drive returns. This is crucial to advancing factor investing from correlation-heavy to causality-driven methodology.

In summary, this thesis seeks to answer: *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?* The outcomes are expected to be significant for both academic research in financial economics and the practical design of investment strategies, as they will help identify which risk factors are genuine drivers of returns (and under what conditions), while showcasing novel methodological enhancements through optimal transport.

Methodology

Data Selection and Preparation

- **Data Sources:** Use recognized factor datasets (e.g. Fama-French five factors), plus individual-level data on returns and covariates (firm size, industry, etc.). These data sources will be provided through my industry advisor.
- **Event Identification:** Locate relevant "events" or interventions, e.g. a new factor index inclusion, regulatory changes affecting liquidity, or distinct market regimes.
- **Preprocessing:** Clean outliers, unify time horizons, possibly reduce dimension for high-dimensional factors. If needed, simulate data with known causal structures to validate methods.

Applying Causal Discovery Algorithms

1. **Difference-in-Differences (DiD)**
   - o Baseline: Conduct standard DiD on a chosen event, computing before-and-after return differences for treated vs. control stocks.
   - o OT-Based: Implement the nonlinear DiD from Torous et al. [3], solving a Wasserstein distance minimization to capture how the entire distribution changes. Compare to baseline results.
2. **Matching and Propensity Score Approaches**
   - a. Baseline: Use classical matching (nearest-neighbor, propensity scores) to estimate factor effects (e.g. does "value" cause higher returns?).
   - b. OT-Based: Adopt an OT matching scheme that reweights the control distribution to mirror the treated group, or discards poorly matched units [2]. Compare effect estimates, checking bias reduction or distributional balance.
3. **Causal Graph Discovery (FCM-Based)**
   - a. Pairwise Direction: Apply ANM or DIVOT [4] to see if factor X causes factor Y (e.g. momentum vs. volatility). DIVOT solves a constrained OT problem to detect direction.
   - b. Multivariate: Explore a constraint-based or score-based search for a broader factor network. Optionally integrate OT for preprocessing or handling data shifts.

Throughout, we will perform robustness checks:

- Placebo tests (using pseudo-events where no real treatment exists).
- Subsample analyses by time period or region to see if effects persist.
- Simulations with known causal directions, verifying that OT-based methods better recover them when standard methods fail under distributional shifts.

Analytical Tools: Python libraries for causal inference and OT. We will track computational efficiency (regularization, dimensionality reduction, etc.) given potentially large datasets.

By following these steps, we aim to estimate causal effects (with or without OT) and identify factor relationships. We will then contextualize results considering known economic logic to avoid purely algorithmic conclusions.

Expected Results and Contributions

1. **Identification of True Causal Factors**
   - We expect some classic factors (e.g. momentum) to consistently show a causal impact once we control for confounders, while others may disappear if driven purely by correlation.
2. **Improved Accuracy Through OT**
   - OT-based DiD should uncover distributional effects that the average-focused analysis would miss (e.g. changes in volatility or downside risk).
   - OT matching is likely to yield smaller bias and more reliable estimates by weighting the sample rather than forcing one-to-one matches.
3. **Methodological Guidelines**
   - We will provide documentation on applying OT in factor-based causal inference - covering how to choose cost functions or how to interpret transport plans in a financial context.
4. **Insights into Factor Interactions**
   - By using FCM-based or pairwise approaches with OT (like DIVOT), we may discover new directions of influence among factors (e.g. liquidity ←→ size). This could reshape how factors are viewed, possibly modeling them in a causal network instead of treating them as separate predictors.

Overall, the expected outcome is that the thesis will not only answer whether optimal transport enhances causal discovery in factor investing (we expect to show it does in meaningful ways), but also produce concrete examples of causal findings in finance. The contributions can be summarized as: (a) providing evidence and methodology for *causal factor investing*, an emerging paradigm envisioned by López de Prado (transforming factor research from correlation-driven to causation-driven), and (b) extending the application of optimal transport methods into a new domain (financial economics), demonstrating their value in a practical, high-stakes setting like investment decision-making.

Work Plan

The research will be carried out over the course of the upcoming year according to the following timeline:

- **Months 1-2: Literature Review and Proposal Refinement -** Perform an in-depth review of academic literature on causal inference techniques and their applications in economics/finance, including the provided primary sources. Develop a strong theoretical foundation and refine the research questions. Identify relevant datasets and acquire any necessary data (e.g. downloading factor return series, stock data). By the end of Month 2, finalize the detailed research design and data to be used.
- **Months 3-4: Data Preparation and Exploratory Analysis -** Clean and preprocess the data for analysis. This includes constructing factor portfolios or groupings, labelling treatment and control observations for study (e.g. defining event dates, treated stocks, etc.), and ensuring data quality. Conduct exploratory data analysis to understand distributions, trends, and correlations among factors and returns. If needed, generate *synthetic data* with known causal structure for later validation tests.
- **Months 5-6: Implement Baseline Causal Method**s **-** Apply traditional causal discovery methods to the data. In Month 5, perform initial difference-in-differences analyses for selected events and matching analyses for selected factor effects without incorporating optimal transport. In Month 6, apply causal graph discovery algorithms (e.g. PC or ANM-based methods) on factor relationships. Document the results of these baseline methods, and identify any inconsistencies or limitations (such as evidence of bias or weak causal signals) that motivate the use of optimal transport.
- **Months 7-8: Integrate Optimal Transport Techniques -** Develop and implement the optimal transport enhancements. In Month 7, focus on the OT-DiD approach: write code to estimate counterfactual distributions using OT (possibly starting with available code from the literature or adapting OT libraries) and apply it to the event studies from earlier. In Month 8, implement OT-based matching for the chosen case studies and run the optimal transport causal discovery (DIVOT) for factor pairs. This phase will likely involve iteration and troubleshooting to get the algorithms working properly on the data. By the end of Month 8, obtain results from all OT-augmented analyses.
- **Month 9:** Validation and Robustness Checks **-** Test the results. Conduct placebo tests (e.g. shifting event dates to when no event occurred), sensitivity analyses (how results change with

different hyperparameters like the OT cost function or regularization strength), and compare findings across subsamples (different time periods, or subgroups of stocks). If simulation data was created, run the methods on it to verify that known causal effects are accurately recovered by the OT methods and baseline methods. Refine the analysis as needed based on these checks to ensure the conclusions are well-supported.

- **Month 10:** Synthesis of Results - Analyze and interpret the outcomes of the empirical work. Summarize key findings: which factors show evidence of causality, how optimal transport changed the insight, etc. Create tables and figures (if applicable) to illustrate the effects and comparisons. This month will focus on distilling the technical results into clear messages for the thesis.
- **Months 11-12:** Writing and Revision - Draft the full thesis document. In Month 11, write the bulk of the thesis, including methodology (documenting the techniques and any novel modifications), results, and discussion sections. Ensure that the writing ties back to the research objectives and highlights the contributions. In Month 12, refine the draft by incorporating feedback from the advisors and peers, proofreading for clarity and correctness, and finalizing the formatting. Prepare any additional materials for the thesis defense (presentation slides, etc.). By the end of Month 12, the thesis should be ready for submission and defense.

Regular meetings with my thesis supervisor will be scheduled to report progress and resolve any issues. Meeting will also be scheduled with my industry advisor. The work plan is designed to allocate time for understanding the material (which is focused on building a base skillset), the implementation of advanced methods (which may be technically challenging) and the interpretation of results (which is crucial for deriving meaningful conclusions in finance). This timeline also leaves some buffer toward the end for unexpected delays or additional analyses that may be needed.

References

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

[2] F. Gunsilius, “Applications of Optimal Transport in Causal Inference,” Kantorovich Initiative Seminar (presentation), 2021.

[3] 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.

[4] R. Tu, K. Zhang, H. Kjellström, and C. Zhang, “Optimal Transport for Causal Discovery,” in Proc. International Conference on Learning Representations (ICLR), 2022.

[5] A. Charpentier, E. Flachaire, and E. Gallic, “Optimal Transport for Counterfactual Estimation: A Method for Causal Inference,” arXiv preprint arXiv:2301.07755, 2023.4

[6] S. Athey and G. W. Imbens, “The State of Applied Econometrics: Causality and Policy Evaluation,” arXiv preprint arXiv:1607.00699, 2016.

[7] G. W. Imbens, “Causal Inference in the Social Sciences,” Annual Review of Statistics and Its Application, vol. 11, pp. 123–52, 2024.