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

**Introduction**

Causal discovery seeks to uncover genuine cause‑and‑effect relationships rather than mere correlations. In finance, distinguishing causal factor effects from coincidental patterns is critical because the proliferation of documented “factor premiums” – the so‑called “factor zoo” – can mislead investors if those premiums are only correlation‑based. López de Prado (2023) argues that factor investing will remain in an immature, purely phenomenological stage until it embraces causality.

The present master’s thesis proposes a concise proof‑of‑concept: we will create a fully synthetic, simulated environment that mirrors stylised characteristics of equity‑factor data. Synthetic data allow us to embed known causal structures and distributional shifts, providing ground truth against which to evaluate competing causal inference methods. The primary goal is to demonstrate the feasibility and merit of an Optimal‑Transport‑augmented causal discovery methodology. Because machine‑learning model development and validation would exceed the available four‑month window, we pivot from a machine‑learning emphasis to a focused data‑analysis study that nevertheless lays a foundation for future ML‑based research.

To benchmark our approach against an established standard, we incorporate instrumental‑variables (IV) estimation. IV analysis remains a work‑horse econometric technique for dealing with endogeneity and provides a clear reference point for assessing whether the Optimal‑Transport (OT) enhancements deliver incremental insight.

This master’s‑level project therefore serves three purposes: (i) to establish a credible proof‑of‑concept using synthetic data; (ii) to compare OT‑based causal discovery to conventional IV benchmarks; and (iii) to outline how the framework can be expanded into a more ambitious PhD‑level agenda using real‑world financial data.

**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].

**Data Analytics Driven Causal Discovery**

I will empathise data‑analytics toolkit for uncovering cause‑and‑effect relationships. The workflow begins with rich exploratory analysis - visualizing factor distributions, mapping co‑movements over time, and stress‑testing for structural breaks - to generate causal hypotheses. These hypotheses are then probed with transparent, statistics‑first techniques such as Granger‑causality tests, panel regressions with fixed effects, instrumental‑variables (IV) estimation, and path‑analysis style structural‐equation models. By iteratively cycling between visualization, summary statistics, and targeted econometric tests, data analytics provides a clear, interpretable bridge from raw observations to credible causal claims, while remaining fully compatible with established designs like DiD and IV.

**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 Generation, Selection and Preparation

* **Synthetic factor panel -** Generate a four‑factor panel (e.g., value, size, momentum, low‑volatility) plus stock‑level returns and covariates.
  + Calibrate means, volatilities, and cross‑correlations to match stylised facts from Fama‑French data so the simulation feels realistic.
  + Embed known causal channels (e.g., momentum → returns, value uncorrelated with returns) so ground‑truth is available for validation.
* **Preprocessing:** Standardise variables, align monthly frequency, and (if desired) compress factors with PCA to test low‑dimensional vs. full‑dimensional cases.

**Applying Causal Discovery Algorithms**

1. **Difference-in-Differences (DiD)**
   * Baseline: Conduct standard DiD on a chosen event, computing before-and-after return differences for treated vs. control stocks.
   * 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**
   1. Baseline: Use classical matching (nearest-neighbor, propensity scores) to estimate factor effects (e.g. does “value” cause higher returns?).
   2. 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)**
   1. 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.
   2. 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).
* 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. **Verification of True Causal Factors**
   * Synthetic ground‑truth lets us show momentum genuinely drives returns while value does not, mirroring or contradicting classical wisdom.
2. **Accuracy Gains from OT Enhancements**
   * 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 Playbook for Short‑Cycle Theses**
   * Provide a reproducible template (code + narrative) showing how a master’s‑level project can prove methodological merit with synthetic data in < 4 months, while laying groundwork for a full PhD expansion on real datasets.
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 ßà size). This could reshape how factors are viewed, possibly modeling them in a causal network instead of treating them as separate predictors.
5. **Road-map for PhD-Level Extension**
   * Outline how the same analytics pipeline can scale to real market data, richer factor libraries, and advanced OT formulations once time and data‑access constraints are relaxed.

Overall, this revised methodology replaces machine‑learning‑centric causal discovery with a transparent, statistics‑driven analytics workflow, relies on synthetic data for speed and clarity, benchmarks against instrumental‑variables estimation, and still leverages optimal transport where it adds clear value.

**Work Plan**

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

* **Month 1 (Weeks 1 – 4): Literature Review, Design & Synthetic‑Data Build**
  + Conduct a focused review of causal‑inference and optimal‑transport papers in finance/econometrics.
  + Refine research questions and finalise the analytics‑first methodology.
* **Month 2 (Weeks 5 – 8): Exploratory Analysis & Core Estimation**
  + Clean, standardise and visualise the synthetic dataset; run rolling correlations, structural‑break checks and Granger screens to generate hypotheses.
  + Implement baseline causal estimators:
  + Two‑way‑fixed‑effects Difference‑in‑Differences
  + Nearest‑neighbour / IPTW matching
  + Instrumental‑Variables (2SLS) benchmark
  + Integrate optimal‑transport variants (OT‑DiD, OT‑matching) and compare against baselines.
  + Log run‑times, convergence diagnostics and interim results for each method.
* **Month 3 (Weeks 9 – 12): Validation, Synthesis & Write‑Up**
  + Perform robustness checks: placebo events, subsample splits, OT‑hyperparameter sensitivity, and recovery of planted causal effects.
  + Consolidate findings into tables, figures and concise interpretation notes.
  + Draft the full thesis (introduction, methodology, results, discussion, conclusion).
  + Iterate with advisor feedback, proof‑read, format, and prepare defence slides.
  + Submit final thesis and rehearse oral defence.

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.