ADIA ¦ Lab

A PROTOCOL FOR CAUSAL FACTOR INVESTING

Marcos López de Prado *
Vincent Zoonekynd §

ADIA Lab Research Paper Series, No. 16

First version (v0.1): May 4, 2025
Current version (v1.0): May 31, 2025

* Global Head, Quantitative Research & Development, Abu Dhabi Investment Authority (ADIA); Board Member, ADIA Lab; Professor of Practice, School of Engineering, Cornell University; Research Fellow, Applied Mathematics & Computational Research Department, Lawrence Berkeley National Laboratory. E-mail: marcos.lopezdeprado@adia.ae.
§ Quantitative Research & Development Lead, Abu Dhabi Investment Authority (ADIA); Research Affiliate, ADIA Lab.

The views expressed in this paper are the authors', and do not necessarily represent the opinions of the organizations they are affiliated with. We would like to thank our ADIA colleagues, especially Pascal Blanqué, Alexander Lipton, Anders Svennesen, and Jean-Paul Villain for their suggestions. The paper has also benefited from conversations with Patrick Cheridito (ETH Zürich), Gerald Cubbin (Invesco), Frank Fabozzi (EDHEC), Campbell Harvey (Duke University), Miguel Hernán (Harvard University), Guido Imbens (Stanford University), Alessia López de Prado Rehder (ETH Zürich), Riccardo Rebonato (EDHEC), Alessio Sancetta (Royal Holloway University of London), Luis Seco (University of Toronto), Horst Simon (ADIA Lab), and Josef Teichmann (ETH Zürich).

A PROTOCOL FOR CAUSAL FACTOR INVESTING

ABSTRACT

Factor investing has become a foundational paradigm in quantitative asset management. Yet, despite the proliferation of factors and widespread institutional adoption, many strategies have failed to live up to their in-sample promise. While *p*-hacking and backtest overfitting have received considerable blame, a more insidious source of error is rarely discussed: the uncritical application of associational econometrics that ignores causal structure. This paper introduces the concept of the factor mirage—a factor model that appears statistically valid but is causally misspecified. We show how collider bias and confounder bias, when embedded in the standard regression framework, can yield misleading inferences, poor out-of-sample performance, and misguided investment decisions. To address this, we propose a seven-step protocol for causal factor investing that draws from recent advances in econometrics and machine learning. This protocol is practical, intuitive, and designed to be used by portfolio managers, risk officers, and asset owners. By shifting from associational to causal reasoning, practitioners can build more robust strategies, reduce false discoveries, and restore trust in factor-based approaches.

Keywords: Causal inference, causal discovery, confounder, collider, factor investing, *p*-hacking, underperformance, systematic losses.

JEL Classification: G0, G1, G2, G15, G24, E44.
AMS Classification: 91G10, 91G60, 91G70, 62C, 60E.

1. INTRODUCTION

Factor investing was once heralded as the future of systematic asset management. Academic breakthroughs like Fama and French's three-factor model gave rise to a burgeoning field, with hundreds of anomalies proposed and institutionalized (Harvey et al. [2016]). Multi-factor products now manage trillions in AUM, promising style tilts that enhance returns, reduce risk, or both. Yet, the real-world long-term performance of many such strategies has fallen short of expectations (López de Prado et al. [2024]).

This shortfall has prompted soul-searching across the industry. Critics cite *p*-hacking and backtest overfitting as the main culprits, see Cochrane [2011], Bailey et al. [2014], Harvey and Liu [2020]. Others argue that market participants arbitrage away these opportunities shortly after publication (McLean and Pontiff [2016]). Still some suggest that factors only work in certain regimes, and the recent regime has been unfavorable –a dubious ex-post argument given that the original publications made no regime distinctions.

These explanations may contain elements of truth, but they overlook a deeper problem: the misapplication of econometric tools. The standard canon—linear regression, two-pass estimation, *p*-values, and correlation-based statistics—originated in settings where causality was not the primary concern. Yet investment decisions are inherently causal, because they require the attribution of returns to risk sources. We do not merely want to know that a portfolio composed of high book-to-market stocks delivers positive returns; in order to optimize a portfolio, we need to determine how much of that performance is attributable to (i.e., caused by) the value factor, to the exclusion of (i.e., controlling for) other explanations, and what may disrupt this relationship (i.e., causal mechanism) going forward.

The distinction between association and causation goes beyond semantics. Some associational models may produce good forecasts, at the expense of offering no risk and return attribution, thus exposing investors to unknown or unwanted risks. Without accounting for causal structure, models are likely to be biased, unstable, and unprofitable out-of-sample.

2. FROM THE "FACTOR ZOO" TO THE "FACTOR MIRAGE"

Cochrane's "factor zoo" metaphor captured the explosion of empirical findings in asset pricing. Hundreds of published anomalies now compete for attention, yet most fail to survive replication or implementation. The literature has responded with tools to mitigate data-snooping, such as corrections for multiple testing, deflated Sharpe ratios, Bayesian shrinkage, and out-of-sample tests.

While useful, these techniques address the selection of models, not their specification. A model can be *p*-hacking-free and still misspecified. In this paper, we introduce the concept of the "factor mirage": a model that appears sound by conventional statistical standards but is structurally invalid because it misrepresents the causal relationships among variables, leading to a biased risk and return attribution.

Factor mirages arise from two common specification errors: (1) Confounder bias – failing to control for variables that are *causes* of both an independent variable (factor) and the dependent variable (returns); and (2) collider bias – controlling for variables that are *consequences* of both an

independent and dependent variables, introducing non-causal correlations. In econometric terms, these biases distort coefficient estimates. In financial terms, they lead to inefficient investments.

Unlike brute-force *p*-hacking, which relies on exhaustive search, the factor mirage is subtler. It arises from practices that are widely taught, widely applied, and rarely questioned.

3. WHERE THE CANON FAILS: ECONOMETRICS WITHOUT CAUSALITY

The econometric methods most widely used in empirical finance—OLS regressions, stepwise model selection, and significance testing—were developed for associational inference, not for causal discovery. They are based on the assumptions that the data-generating process is stationary and linear, and that the regression model is correctly specified. However, these assumptions are often violated in financial applications, particularly in asset pricing models.

A case in point is the assumption of correct model specification. In asset pricing, the standard practice for identifying a factor is to follow a two-pass regression approach: First, run time series regressions of assets excess returns on a set of factors to estimate factor loadings (exposures). Second, run cross-sectional regressions of assets excess returns on the estimated factor exposures to estimate factor premia. The choice of specification is typically driven by associational power maximization, not causal considerations.

A confounder is a variable that is a cause to both an independent variable and the dependent variable. Confounder bias arises when the model's specification does not control for a confounder.[1] If leverage, for instance, influences both book-to-market and returns, and is not included in the model, the estimate on book-to-market may be biased in magnitude and sign.

A collider is a variable that is causally downstream of both an independent variable and the dependent variable. Controlling for a collider introduces a non-causal correlation, which inflates the adjusted R-squared and lowers *p*-values. For example, if quality is influenced by both book-to-market and returns, including it as a control will bias the coefficient on book-to-market. This collider bias is subtle: it does not cause multicollinearity and it reduces standard errors, however it systematically distorts inference. What makes colliders particularly dangerous is that they often change the sign of estimated coefficients, thus inducing investors to buy securities that should be sold, and to sell securities that should be bought.

These are not hypothetical concerns. Shanken [1992] discussed the consequences of estimation error in factor betas. Giglio and Xiu [2021] showed that many popular factors are likely mispriced due to omitted variables, and introduced an intermediate principal components analysis (PCA) step for identifying potential latent confounders. While these considerations partially address concerns regarding confounder bias, the problem of collider bias has received far less attention in the finance literature.

The standard two-pass or three-pass regression procedures are particularly vulnerable to collider bias. The inclusion or exclusion of controls is typically based on statistical criteria (e.g., increasing R-squared) rather than causal logic, see Fama and French [1993, 2015]. Similarly, the three-pass factor regression approach in Giglio and Xiu [2021] applies PCA to the matrix of asset returns to

[1] This is sometimes also referred to as "omitted variable bias" in the econometrics literature.

extract latent variables—i.e., directions of maximal variance in returns unexplained by observed factors—, assuming without causal evidence that those latent variables must be confounders. The problem is, PCA attempts to maximize explained variance without differentiating between confounders and colliders. As a result, models may look compelling in-sample while encoding fundamentally flawed relationships.

Standard model evaluation metrics—such as adjusted R-squared, AIC, BIC, and t-statistics—reward misspecification and penalize parsimony, even when the extra variables introduce collider bias. In a world of limited data and noisy signals, these practices create an illusion of robustness. It is important to recognize that multiple testing adjustments do not correct for model misspecification: a model can be misspecified after a single test.

This is the essence of the factor mirage: a model that seems plausible by conventional standards but fails to capture the underlying causal structure. It performs well in backtests and cross-validation but delivers disappointing results out-of-sample or in production.

4. EXAMPLE: COLLIDERS AMONG BARRA FACTORS

To illustrate the above points, we apply the PC causal discovery algorithm of Spirtes et al. [2000] to the time series of daily returns for the risk factors of 85 Barra risk models. Figure 1 aggregates the resulting causal graphs by retaining the edges present in at least a third of the graphs.

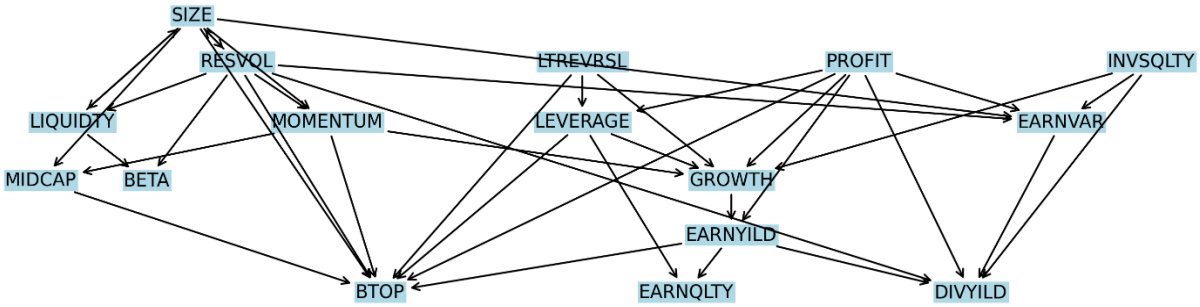

Figure 1 – Aggregate causal graph discovered through the PC algorithm

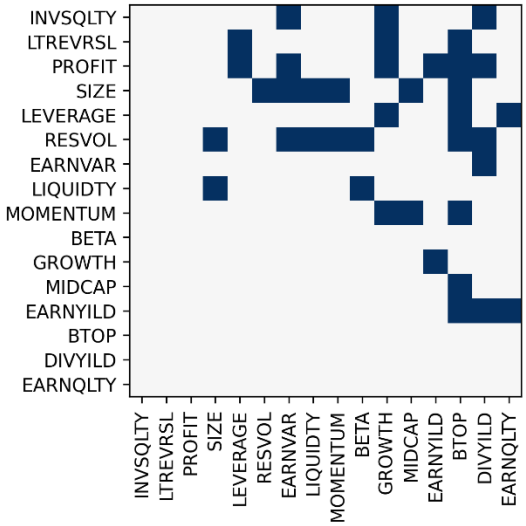

Figure 2 – Adjacency matrix

Figure 2 derives the corresponding adjacency matrix. The two entries below the main diagonal indicate for what edges the PC algorithm was not able to determine the direction of the causal relationship, namely ("size", "liquidity") and ("size", "resvol").

With the help of this discovered graph, an investor can formulate the correct model specification. In particular, an investor willing to invest in one of those factors should condition on the parents of that factor, and not on its descendants, to avoid the risk of controlling for a collider. For instance, to invest on "growth", we should control for "momentum", "leverage", "long-term reversal", "profit" and "investor quality", but not on value factors ("earnings yield", "book-to-price", "earnings quality", "dividend yield"). Figure 3 highlights in green the correct controls, in red the controls to be avoided, and in grey irrelevant variables.

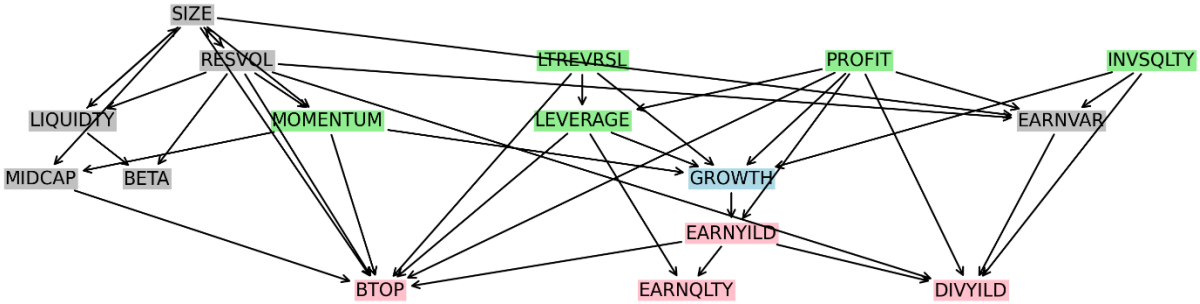

Figure 3 – Correct and incorrect controls of a Growth factor model

If we take one of those risk models, say USE4L, we can try to forecast "growth" from the other factors, selecting them to maximize the adjusted R-squared. The left plot in Figure 4 shows the adjusted R-squared of models that include all descendants, where a greedy algorithm adds parents –in the y-axis– one by one. The full model, with all descendants and the parents listed in the y-axis has an adjusted R-squared of approximately 8.5%. These models are misspecified, as descendants should not have been included as control variables. The right plot in Figure 4 shows models without descendants, where a greedy algorithm adds parents –in the y-axis– one by one. The full model, with all parents but no descendants, has an adjusted R-squared of approximately 7.8%, which is "worse" than the adjusted R-squared of the misspecified model.

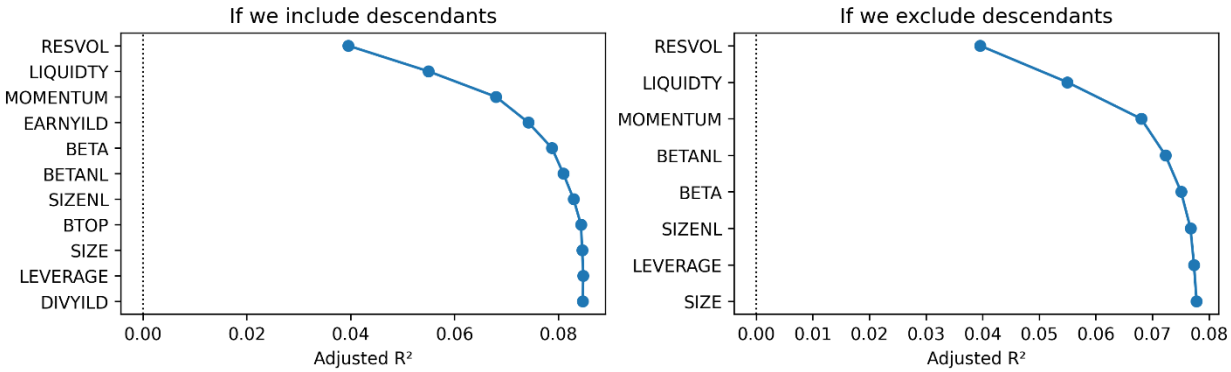

Figure 4 – Adjusted R-squared for correctly and incorrectly specified factor models

5. A SEVEN-STEP PROTOCOL FOR CAUSAL FACTOR INVESTING

To reduce the risk of misspecification and introduce causal reasoning into factor model research, we outline a practical seven-step protocol.[2] This framework leverages recent advances in machine learning, causal discovery, and causal inference, while remaining accessible and applicable in real-world investment contexts.

Step 1: Variable Selection

The first step involves identifying variables potentially associated with current or future returns, depending on the goal of the factor model (attribution vs. risk premia harvesting). Researchers should employ powerful machine learning techniques capable of capturing nonlinear relationships and interaction effects without assuming a particular model specification. Three categories of tools assist in this task: (1.a) non-parametric methods that detect mutual information between predictors and returns; (1.b) in-sample measures of variable importance—such as Shapley values or mean decrease impurity (MDI); and (1.c) feature importance metrics derived from cross-validation—such as mean decrease accuracy (MDA). See López de Prado [2018, 2020] for further details.

Step 2: Causal Discovery

Next, the researcher estimates a causal graph that captures the dependency structure among the variables identified in Step 1, using causal discovery algorithms (see Glymour et al. [2019]), economic theory, and expert judgment. This graph may represent: (2.a) investor beliefs based on prior knowledge (Laudy et al. [2022]); or (2.b) empirically observed relationships (Olivetti et al. [2025], Sadeghi et al. [2024], Gu et al. [2023]). Algorithms like PC and LiNGAM are valuable even if not fully accurate, as they help eliminate theoretical causal graphs that are inconsistent with the observations. From the refined set of plausible structures, the researcher selects the one most aligned with economic logic, clearly stating any assumptions. Even if the graph is ultimately incorrect, the process ensures transparency and falsifiability (López de Prado [2023]).

Step 3: Causal Adjustment Set

Do-calculus is then applied to the selected graph from Step 2 to determine the appropriate set of control variables—those that block all backdoor paths while preserving the causal link (Pearl [2009], Pearl et al. [2016]). This requires adjusting for confounders while avoiding colliders. Specific approaches for identifying an adjustment include: (3.a) backdoor; (3.b) front-door; and (3.c) instrumental variables. Each control variable must have a sound economic and causal rationale to avoid mechanical overfitting.

Step 4: Causal Explanatory and Predictive Power

The researcher estimates causal effects, e.g., applying double machine learning (Fuhr et al. [2024]). The model's generalization error can be assessed in terms of: (4.a) the probability distribution of discretized returns (a classification task); (4.b) the rank ordering of returns (also a classification task); and (4.c) the magnitude of returns (a regression task). A model that excels in just one of these areas may still be profitably deployed, and this analysis informs how to use that model.

Step 5: Causal Portfolio Construction

Using forecasts or risk attributions from the causal model, the researcher constructs a portfolio that reflects targeted exposures while minimizing unintended risks. Unlike traditional methods based

[2] For a more complete description, see López de Prado and Zoonekynd [2024].

on correlational insights, this step emphasizes: (5.a) position sizing aligned with causal effects—allocating capital proportionally to causal impact, not to coefficients from misspecified regressions; (5.b) risk control via the causal graph—ensuring neutrality towards variables outside the causal pathway (e.g., colliders), or hedging undesired causal exposures; (5.c) alignment with economic rationale—positions should reflect structural relationships (e.g., long high book-to-price stocks if causally justified), and avoid spurious correlations; (5.d) stress testing for causal fragility—simulating changes to the graph or parameter values to test robustness against model misspecification; (5.e) incorporation of realistic constraints and transaction costs, using cost-aware optimization that preserves causal integrity (e.g., avoid noise trading); and (5.f) measurement of strategy distortion by computing the transfer coefficient between the optimal causal portfolio and the implemented strategy.

López de Prado et al. [2025] show that portfolio optimization based on misspecified factor models leads to severe investment inefficiencies. They demonstrate that the efficient frontier cannot be estimated accurately without a causal model, and that associational approaches may even lead investors to buy assets they should be selling, even if investors had perfect information about means and covariances.

Step 6: Backtest

The performance of the systematic strategy is then evaluated through backtesting, with three principal methods: (6.a) walk-forward testing; (6.b) resampling; and (6.c) Monte Carlo simulation.

Backtest (6.a) assumes the future will mirror the past exactly. It suffers from several limitations: it represents only one possible path of realizations of the data-generating process; it may use data inefficiently due to burn-in requirements; and it offers no insight into the underlying process—making it impossible to identify failure conditions if the process changes (López de Prado [2018], chapter 11).

Backtest (6.b) addresses the first two limitations by generating alternative future paths via deterministic (jackknife, cross-validation) or random (bootstrap, subsampling) resampling. These methods generate a range of scenarios consistent with the observed data. For instance, combinatorial purged cross-validation (CPCV) enables bootstrapping of Sharpe ratios, offering more informative and robust metrics than single-path estimates (see López de Prado [2018], chapter 12). Still, these methods are constrained by the finite nature of historical data.

Backtest (6.c) addresses the above limitations by using Monte Carlo simulation to generate synthetic paths based on an explicit data-generating process. This process may be informed by empirical analysis or theoretical models (e.g., market microstructure, institutional dynamics, economic mechanisms). For instance, if theory indicates that two variables are cointegrated, simulations can vary the cointegration vector within its estimated range, producing richer data than resampling past observations (López de Prado [2018], chapter 13; Joubert et al. [2024]).

Step 7: Multiple Testing Adjustments

Finally, the researcher accounts for the elevated risk of false discoveries due to multiple hypothesis testing. Two primary approaches are: (7.a) p-value adjustments—such as Holm [1979], Hochberg [1988], or Benjamini and Hochberg [1995]; and (7.b) Sharpe ratio adjustments—such as the

Deflated Sharpe Ratio (DSR), which corrects for the variance in Sharpe estimates and adjusts for the effective number of trials. Since tests are typically correlated, the effective number of trials is less than the total number conducted. See López de Prado [2020] for details.

6. BEST PRACTICES FOR PROFESSIONALS AND ASSET OWNERS

The seven-step protocol is not just for academic researchers or portfolio managers. It is also relevant for allocators, risk managers, consultants, and investment committee members who evaluate factor strategies proposed by external managers or internal teams.

To help translate the protocol into an actionable evaluation tool, we propose the checklist in the appendix. These questions can form part of a due diligence questionnaire (DDQ), investment memo, or strategy approval process.

By insisting on these standards, asset owners and supervisors can protect capital, enhance accountability, and improve alignment between strategy design and economic intent.

7. THE ECONOMIC COST OF CAUSAL NEGLECT

Misapplying associational tools to causal problems is not just a technical misstep—it has real-world consequences. Portfolios are constructed, risks are hedged, and billions of dollars are deployed based on models that may be misspecified.

The economic costs of causal neglect fall into several categories:

- Capital misallocation: Investors allocate to strategies that appear statistically significant but are not economically meaningful. This misdirection persists until performance disappoints or capital is withdrawn (López de Prado et al. [2025]).
- Hidden leverage and risk stacking: When multiple models share similar specification errors, portfolios may unknowingly stack risk exposures. For example, many value strategies may be exposed to the same macroeconomic confounder.
- Excessive turnover: Spurious signals lead to unnecessary trades, increasing transaction costs, bid-ask spreads, and slippage. This further erodes alpha.
- Lack of persistence: Models built on non-causal relationships often fail to persist when economic conditions change or new data becomes available. Investment models based on causal relationships can be profitable even if parameters shift (López de Prado and Zoonekynd [2024]).
- Loss of trust: When backtests consistently outperform live performance, clients begin to lose trust in the scientific validity of systematic investing. This reputational damage is difficult to reverse (López de Prado [2015]).

In short, causal neglect leads to inefficient investment decisions, unrewarded risks, and poor stewardship of capital. The current state of factor investing—marked by underwhelming performance, crowded trades, and skepticism—can be traced, in part, to these methodological shortcomings.

8. FROM MIRAGE TO METHOD

Investment factors exist, but the way we build and evaluate them is flawed. The reliance on associational econometrics has led to a proliferation of anomalies, many of which fail to hold up

under scrutiny or deliver in practice. This phenomenon, which we call the factor mirage, reflects the consequences of misspecification—particularly collider bias and confounder bias—within canonical estimation frameworks.

This paper has offered both a diagnosis and a remedy. Drawing from the causal inference literature, we propose a seven-step protocol that is practical, transparent, and grounded in economic logic. By clearly stating hypotheses, drawing causal diagrams, selecting controls intentionally, applying robust estimators, testing out-of-sample, and documenting assumptions, practitioners can materially improve their models.

Our goal is not to impose a rigid orthodoxy but to enable better reasoning. Causal methods are not new; they are simply underused in finance. Fields like medicine, policy evaluation, and macroeconomics have already embraced them with measurable benefits. It is time for asset management to evolve. The transition from associational to causal modeling will not be easy. It requires unlearning old habits, revisiting familiar practices, and retooling teams. But the rewards are worth it: more stable strategies, better risk control, clearer communication, and ultimately more trustworthy products.

In a world increasingly skeptical of backtests and statistical alchemy, causal factor investing offers a credible path forward. It replaces the illusion of precision with the discipline of structure, helping us see beyond the mirage.

9. REFERENCES

Bailey, D., J. Borwein, M. López de Prado, and J. Zhu (2014): "Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-Of-Sample Performance." *Notices of the American Mathematical Society*, Vol. 61, No. 5, pp. 458-471

Benjamini, Y., and Y. Hochberg (1995): "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B* (Methodological), Vol. 57, No. 1, pp. 289–300.

Cochrane, J. (2011): "Presidential Address: Discount Rates." *Journal of Finance*, Vol. 66, No. 4, pp. 1047–1108.

Fama, E. and K. French (1993): "Common risk factors in the returns on stocks and bonds." *Journal of Financial Economics*, Vol. 33, No. 1, pp. 3-56.

Fama, E. and K. French (2015): "A five-factor asset pricing model." *Journal of Financial Economics*, Vol. 116, No. 1, pp. 1-22.

Fuhr, J., P. Berens, and D. Papies (2024): "Estimating Causal Effects with Double Machine Learning: A Method Evaluation." Working paper. Available at https://arxiv.org/html/2403.14385v1

Giglio, S. and D. Xiu (2021): "Asset Pricing with Omitted Factors." *Journal of Political Economy*, Nol. 129, No. 7, pp. 1947-1990.

Glymour, C., K. Zhang, and P. Spirtes (2019): "Review of Causal Discovery Methods Based on Graphical Models." *Frontiers in Genetics*, Vol. 10, No. 524, pp. 1–15, www.frontiersin.org/articles/10.3389/fgene.2019.00524/full.

Gu, L., H. Zhang, A. Heinz, J. Liu, T. Yao, M. AlRemeithi, Z. Luo, D. Ruppert (2023): "Re-examination of Fama-French factor investing with causal inference methods." Working paper. Available at https://ssrn.com/abstract=4677537

Harvey, C., Y. Liu, and H. Zhu (2016): "... and the Cross-Section of Expected Returns." *Review of Financial Studies*, Vol. 29, No. 1, pp. 5–68.

Harvey, C. and Y. Liu (2020): "False (and Missed) Discoveries in Financial Economics." *Journal of Finance*, Vol. 75, No. 5, pp. 2503-2553.

Hochberg, Y. (1988). "A sharper Bonferroni procedure for multiple tests of significance." *Biometrika*, Vol. 75, No. 4, pp. 800–802.

Holm, S. (1979): "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics*, Vol. 6, No. 2, pp. 65–70.

Joubert, J., D. Sestovic, I. Barziy, W. Distaso, and M. López de Prado (2024): "Enhanced Backtesting for Practitioners." *The Journal of Portfolio Management*, Vol. 51, No. 2, pp. 12-27.

Laudy, O., A. Denev, and A. Ginsberg (2022): "Building Probabilistic Causal Models Using Collective Intelligence." *The Journal of Financial Data Science*, Vol. 4, No. 2, pp. 83 – 109.

López de Prado, M. (2015): "The Future of Empirical Finance." *The Journal of Portfolio Management*, Vol. 41, No. 4, pp. 140-144.

López de Prado, M. (2018): *Advances in Financial Machine Learning*. Wiley, 1st edition.

López de Prado, M. (2020): *Machine Learning for Asset Managers.* Cambridge University Press. https://doi.org/10.1017/9781108883658

López de Prado, M. (2023a): *Causal Factor Investing: Can Factor Investing Become Scientific?* Cambridge University Press. https://doi.org/10.1017/9781009397315

López de Prado, M., and V. Zoonekynd (2024): "Correcting the Factor Mirage: A Research Protocol for Causal Factor Investing." *The Journal of Portfolio Management*, forthcoming. Available in SSRN at https://ssrn.com/abstract=4697929

López de Prado, M., A. Lipton and V. Zoonekynd (2025): "Causal Factor Analysis is a Necessary Condition for Investment Efficiency." *The Journal of Portfolio Management*, forthcoming. Available in SSRN at https://ssrn.com/abstract=5131050

McLean, R. and J. Pontiff (2016): "Does Academic Research Destroy Stock Return Predictability?" *Journal of Finance*, Vol. 71, No. 1, pp. 5–32.

Olivetti, E., V. Zoonekynd, P. Yam, M. López de Prado, G. Imbens, and M. Hernán (2025): "ADIA Lab Causal Discovery Challenge." *The Journal of Financial Data Science*, forthcoming.

Pearl, J. (2009): *Causality: Models, Reasoning and Inference*. Cambridge, 2nd edition.

Pearl, J., M. Glymour, and N. Jewell (2016): *Causal Inference in Statistics: A Primer*. Wiley, 1st edition.

Sadeghi, A., A. Gopal, and M. Fesanghary (2024): "Causal discovery in financial markets: A framework for nonstationary time-series data." Working paper. Available at https://arxiv.org/abs/2312.17375

Shanken, J. (1992): "On the Estimation of Beta-Pricing Models." *Review of Financial Studies*, Vol. 5, No. 1, pp. 1–33.

P. Spirtes, C. Glymour, and R. Scheines (2000): Causation, Prediction, and Search. MIT Press, 2nd edition.

APPENDIX

Here we propose a due diligence questionnaire that practitioners and managers can use to assess whether a factor investing proposal is supported by causal evidence.

Step 1: Variable Selection

- What is the intended purpose of the factor model—risk attribution or risk premia harvesting? Are the selected variables consistent with this purpose?
- How were the candidate variables initially selected?
- Were non-parametric or machine learning methods used to detect relationships?
- Were Shapley values, mean decrease impurity, or feature importance used?
- Were domain-specific constraints applied to exclude spurious or uninterpretable variables?

Step 2: Causal Discovery

- Did the researcher construct a causal graph to represent the structure of the problem?
- Were causal discovery algorithms (e.g., PC, LiNGAM) used? If so, which ones?
- What economic rationale or domain expertise supports the chosen graph?
- Have alternative causal graphs been considered and ruled out? On what basis?
- Are the causal graph's assumptions clearly documented and available for review?

Step 3: Causal Adjustment Set

- What method was used to identify the adjustment set (e.g., backdoor, front-door, IV)?

- Which variables are being controlled for, and why?
- Did researchers confirm the validity of the adjustment set using do-calculus software, such as Daggity (http://www.dagitty.net/) or Microsoft's DoWhy?
- Were any known colliders included in the model? How were they identified and excluded?
- Are all control variables economically interpretable and justifiable?

Step 4: Causal Explanatory and Predictive Power
- Does the model allow the simulation of controlled experiments? Does it answer counterfactual questions?
- How was the model's generalization error estimated? Is this approach realistic, given the causal graph?
- Was the model's performance assessed in terms of: (a) Probability of return sign? (b) Ranking of returns? (c) Magnitude of returns?
- Were the findings robust across multiple validation techniques or subsamples?
- Does the model show signs of overfitting to any specific performance metric?

Step 5: Causal Portfolio Construction
- How are causal effects translated into portfolio weights?
- Does the strategy remain neutral or agnostic to non-causal factors?
- Is the causal graph used to hedge unwanted exposures or to guide hedging?
- Are stress tests performed on the causal model (e.g., DAG perturbation)?
- What portfolio constraints and transaction costs are considered, and how are they incorporated?
- Is the transfer coefficient between the ideal and actual portfolio computed and reported?

Step 6: Backtesting Methodology
- Which backtest methods were used (walk-forward, resampling, Monte Carlo)?
- Was the causal graph used to simulate scenarios?
- Are limitations of historical resampling or burn-in periods acknowledged?
- Are multiple scenarios or paths evaluated, not just one historical realization?
- Was combinatorially-purged cross-validation or a similar technique used to estimate a distribution of Sharpe ratios?

Step 7: Multiple Testing Adjustment
- How many models or hypotheses were tested during research development?
- Was the effective number of tests estimated, accounting for test correlation?
- Were p-value adjustments (e.g., Holm, Hochberg, Benjamini-Hochberg) applied?
- Was the Deflated Sharpe Ratio (DSR) reported? If so, how was it computed?
- Are the statistical significance levels adjusted for the model selection process?

Final Assessment
- Is the entire research process transparent, documented, and reproducible?
- Are assumptions stated explicitly and subject to falsifiability or peer-review?
- Has the model been stress-tested against changes in its structure or estimation method?
- Does the investment team demonstrate familiarity with causal inference concepts and limitations?