1 Introduction

Science is more than a collection of observed associations. While the description and cataloging of phenomena play a role in scientific discovery, the ultimate goal of science is the amalgamation of theories that have survived rigorous falsification (Hassani et al. 2018). For a theory to be scientific, it is generally expected to declare the falsifiable causal mechanism responsible for the observed phenomenon (for one definition of falsifiability, see Popper 1963).[1] Put simply, a scientific theory explains *why* an observed phenomenon takes place, where that explanation is consistent with all the empirical evidence (ideally, including experimental results). Economists subscribe to this view that a genuine science must produce refutable implications, and that those implications must be tested through solid statistical techniques (Lazear 2000).

In the experimental sciences (physics, chemistry, biology, etc.), it is relatively straightforward to propose and falsify causal mechanisms through interventional studies (Fisher 1971). This is not generally the case in financial economics. Researchers cannot reproduce the financial conditions of the Flash Crash of May 6, 2010, remove some traders, and observe whether stock market prices still collapse. This has placed the field of financial economics at a disadvantage when compared with experimental sciences. A direct consequence of this limitation is that, for the past fifty years, most factor investing researchers have focused on publishing associational claims, without theorizing and subjecting to falsification the causal mechanisms responsible for the observed associations. In the absence of plausible falsifiable theories, researchers must acknowledge that they do not understand why the reported anomalies (risk premia) occur, and investors are entitled to dismiss their claims as spurious. The implication is that the factor investing literature remains in an immature, phenomenological stage.

From the above, one may reach the bleak conclusion that there is no hope for factor investing (or financial economics) to produce and build upon scientific theories. This is not necessarily the case. Financial economics is not the only field of study afflicted by barriers to experimentation (e.g., astronomers produce scientific theories despite the unfeasibility of interventional studies). Recent progress in causal inference has opened a path, however difficult, for advancing factor investing beyond its current phenomenological stage. The goal of this

[1] Strict falsificationism is not widely accepted among philosophers of science, and throughout this Element I do not follow Popper's falsificationist framework. I use the term "falsifiable" as the general requirement that theories must conform to the empirical evidence, without subscribing to a particular definition of what such conformity entails. *Mutatis mutandis*, this Element accommodates, and its results remain valid, under a number of competing accounts of what makes a theory "scientific."

Element is to help factor investing wake up from its associational slumber, and plant the seeds for the new field of "causal factor investing."

In order to achieve this goal, I must first recite the fundamental differences between association and causation (Section 2), and why the study of association alone does not lead to scientific knowledge (Section 3). In fields of research with barriers to experimentation, like investing, it has become possible to estimate causal effects from observational studies, through natural experiments and simulated interventions (Section 4). After laying out this foundation, I turn the reader's attention to the current state of causal confusion in econometrics (Section 5) and factor investing studies (Section 6). This state of confusion easily explains why factor investing remains in a phenomenological stage, and the proliferation of hundreds of spurious claims that Cochrane (2011) vividly described as the "factor zoo"[2] (Section 7). The good news is, once financial economists embrace the concepts described in this Element, I foresee the transformation of factor investing into a truly scientific discipline (Section 8).

This Element makes several contributions. First, I describe the logical inconsistency that afflicts the factor investing literature, whereby authors make associational claims in denial or ignorance of the causal content of their models. Second, I define the two different types of spurious claims in factor investing, type-A and type-B. These two types of spurious claims have different origins and consequences, hence it is important for factor researchers to distinguish between the two. In particular, type-B factor spuriosity is an important topic that has not been discussed in depth until now. Type-B spuriosity explains, among other literature findings, the time-varying nature of risk premia. Third, I apply this taxonomy to derive a hierarchy of empirical evidence used in financial research, based on the evidence's susceptibility to being spurious. Fourth, I design Monte Carlo experiments that illustrate the dire consequences of type-B spurious claims in factor investing. Fifth, I propose an alternative explanation for the main findings of the factor investing literature, which is consistent with type-B spuriosity. In particular, the time-varying nature of risk premia reported in canonical journal articles is a likely consequence of under-controlling. Sixth, I propose specific actions that academic authors can take to rebuild factor investing on the more solid scientific foundations of causal inference.

[2] A more appropriate name might have been "factor bestiary," because a zoo is populated only by real animals, while a medieval bestiary described in great detail real (e.g., lions, leopards, and elephants) as well as mythical animals (e.g. chimeras, griffins, and harpies), with equal conviction regarding the existence of both.

2 Association vs Causation

Every student of statistics, and by extension econometrics, learns that association does not imply causation. This statement, while superficially true, does not explain why association exists, and its relation to causation. Two discrete random variables X and Y are statistically independent if and only if $P[X = x, Y = y] = P[X = x]P[Y = y], \forall x, y$, where $P[.]$ is the probability of the event described inside the squared brackets. Conversely, two discrete random variables X and Y are said to be statistically associated (or codependent) when, for some (x, y), they satisfy that $P[X = x, Y = y] \neq P[X = x]P[Y = y]$. The conditional probability expression $P[Y = y|X = x] = P[X = x, Y = y]/P[X = x]$ represents the probability that $Y = y$ among the subset of the population where $X = x$. When two variables are associated, observing the value of one conveys information about the value of the other: $P[Y = y|X = x] \neq P[Y = y]$, or equivalently, $P[X = x|Y = y] \neq P[X = x]$. For example, monthly drownings (Y) and ice cream sales (X) are strongly associated, because the probability that y people drown in a month conditional on observing x ice cream sales in that same month does not equal the unconditional probability of y drownings in a month for some (x, y). However, the expression $P[Y = y|X = x] \neq P[Y = y]$ does *not* tell us whether ice cream sales cause drownings. Answering that question requires the introduction of a more nuanced concept than conditional probability: an intervention.

A data-generating process is a physical process responsible for generating the observed data, where the process is characterized by a system of structural equations. Within that system, a variable X is said to cause a variable Y when Y is a function of X. The structural equation by which X causes Y is called a causal mechanism. Unfortunately, the data-generating process responsible for observations is rarely known. Instead, researchers must rely on probabilities, estimated on a sample of observations, to deduce the causal structure of a system. Probabilistically, a variable X is said to cause a variable Y when *setting* the value of X to x increases the likelihood that Y will take the value y. Econometrics lacks the language to represent interventions, that is, setting the value of X (Chen and Pearl 2013). To avoid confusion between conditioning by $X = x$ and setting the value of $X = x$, Pearl (1995) introduced the do-operator, $do[X = x]$, which denotes the intervention that sets the value of X to x. With this new notation, causation can be formally defined as follows: $X = x$ causes $Y = y$ if and only if $P\big[Y = y|do[X = x]\big] > P[Y = y]$.[3]

[3] At first, it may seem counterintuitive that causality is defined in terms of a strict inequality (" $>$"), in contrast to the difference ("$\neq$") used to define association. The reason is, there is no need to consider the "$<$" case, due to complementary probabilities. For example, let $X = 1$ represent

For example, setting ice cream sales to x will not make y drownings more likely than its unconditional probability for any pair (x, y), hence ice cream sales are not a cause of drownings. In contrast, smoking tobacco is a cause of lung cancer, because the probability that y individuals develop lung cancer among a collective where the level of tobacco smoking is set to x (through an intervention) is greater than the unconditional probability of y individuals developing lung cancer, for some pair (x, y).[4]

Variables X and Y may be part of a more complex system, involving additional variables. The causal structure of a system can be represented through a directed acyclic graph, also denoted a causal graph.[5] While a causal graph does not fully characterize the data-generating process, it conveys topological information essential to estimate causal effects. Causal graphs declare the variables involved in a system, which variables influence each other, and the direction of causality (Pearl 2009, p. 12). Causal graphs help visualize do-operations as the action of removing all arrows pointing toward X in the causal graph, so that the full effect on Y can be attributed to setting $X = x$. This is the meaning of the *ceteris paribus* assumption, which is of critical importance to economists.

The causal graph in Figure 1 tells us that Z causes X, and Z causes Y. In the language of causal inference, Z is a confounder, because this variable introduces

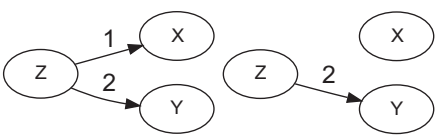

Figure 1 Causal graph of a confounder (Z), before (left) and after (right) a do-operation

receiving a vaccine against COVID-19, and $Y = 1$ represent developing COVID-19. For an effective vaccine, two causal statements are true. First, $P\left[Y = 1|do[X = 1]\right] < P[Y = 1]$, which means that receiving the vaccine ($X = 1$) reduces the likelihood of developing the disease ($Y = 1$). Second, $P\left[Y = 0|do[X = 1]\right] > P[Y = 0]$, which means that receiving the vaccine ($X = 1$) increases the likelihood of not developing the disease ($Y = 0$). One statement cannot be true without the other, and the redundancy is resolved by picking the latter.

[4] A variable X may be a necessary cause of Y, a sufficient cause of Y, a necessary-and-sufficient cause of Y, or neither a necessary-nor-sufficient cause of Y (also known as a contributory cause). I do not explain the difference in this Element because it is not required for the discussion that follows.

[5] Acyclic graphs have the advantage of allowing the factorization of the joint probability as a product of conditional probabilities between ancestors and descendants only. However, cyclic graphs may be preferred for representing bidirectional causality. Representing bidirectional causal relationships with acyclic graphs requires explicit temporal modeling and duplication of the graph over multiple time steps. Neither representation (cyclic or acyclic) is better, and it depends on the modeler's objectives. This Element focuses on the treatment of acyclic graphs, without dismissing the usefulness of cyclic graphical models.

an association between X and Y, even though there is no arrow between X and Y. For this reason, this type of association is denoted noncausal. Following with the previous example, weather (Z) influences ice cream sales (X) and the number of swimmers, hence drownings (Y). The intervention that sets ice cream sales removes arrow (1), because it gives full control of X to the researcher (X is no longer a function of Z), while keeping all other things equal (literally, *ceteris paribus*). And because X does not cause Y, setting $X = x$ (e.g., banning the sale of ice cream, $X = 0$) has no effect on the probability of $Y = y$. As shown later, noncausal association can occur for a variety of additional reasons that do not involve confounders.

Five conclusions can be derived from this exposition. First, causality is an extra-statistical (in the sense of beyond observational) concept, connected to mechanisms and interventions, and distinct from the concept of association. As a consequence, researchers cannot describe causal systems with the associational language of conditional probabilities. Failure to use the do-operator has led to confusion between associational and causal statements, in econometrics and elsewhere. Second, association does not imply causation, however causation does imply association because setting $X = x$ through an intervention is associated with the outcome $Y = y$.[6] Third, unlike association, causality is directional, as represented by the arrows of the causal graph. The statement "X causes Y" implies that $P\left[Y = y|do[X = x]\right] > P[Y = y]$, but not that $P\left[X = x|do[Y = y]\right] > P[X = x]$. Fourth, unlike association, causality is sequential. "X causes Y" implies that the value of X is set first, and only after that Y adapts. Fifth, the *ceteris paribus* assumption simulates an intervention (do-operation), whose implications can only be understood with knowledge of the causal graph. The causal graph shows what "other things" are kept equal by the intervention.

[6] Here I am referring to direct causes (a single link in the causal graph). There are causal structures where one cause may cancel another, resulting in *total* causation without association.

3 The Three Steps of Scientific Discovery

Knowing the causes of effects has long been a human aspiration. In 29 BC, ancient Roman poet Virgil wrote "happy the man, who, studying Nature's laws, / thro' known effects can trace the secret cause" (Dryden 1697, p. 71). It was not until around the year 1011 that Arab mathematician Hasan Ibn al-Haytham proposed a scientific method for deducing the causes of effects (Thiele 2005; Sabra 1989).

Science has been defined as the systematic organization of knowledge in the form of testable explanations of natural observations (Heilbron 2003). Mature scientific knowledge aims at identifying causal relations, and the mechanisms behind them, because causal relations are responsible for the regularities in observed data (Glymour et al. 2019).

The process of creating scientific knowledge can be organized around three critical steps: (1) the phenomenological step, where researchers observe a recurrent pattern of associated events, or an exception to such a pattern; (2) the theoretical step, where researchers propose a testable causal mechanism responsible for the observed pattern; and (3) the falsification step, where the research community designs experiments aimed at falsifying each component of the theorized causal mechanism.

3.1 The Phenomenological Step

In the phenomenological step, researchers observe associated events, without exploring the reason for that association. At this step, it suffices to discover that $P[X = x, Y = y] \neq P[X = x]P[Y = y]$. Further, a researcher may model the joint distribution $P[X = x, Y = y]$, derive conditional probabilities $P[Y = y|X = x]$, and make associational statements of the type $E[Y|X = x] = y$ (an associational prediction) with the help of machine learning tools. Exceptionally, a researcher may go as far as to produce empirical evidence of a causal effect, such as the result from an interventional study (e.g., Ohm's law of current, Newton's law of universal gravitation, or Coulomb's law of electrical forces), but without providing an explanation for the relationship. The main goal of the phenomenological step is to state "a problem situation," in the sense of describing the observed anomaly for which no scientific explanation exists (Popper 1994b, pp. 2–3). At this step, inference occurs by *logical induction*, because the problem situation rests on the conclusion that, for some unknown reason, the phenomenon will reoccur.[7]

[7] Reasoning by induction occurs when, given some premises, a probable conclusion is inferred non-reductively, by generalizing or extrapolating from specific cases to a general rule. The evidence to support this extrapolation may come from a large number of cases (enumerative induction) or a wide range of cases (variative induction). See Gensler (2010, pp. 80–117).

For instance, a researcher may observe that the bid-ask spread of stocks widens in the presence of imbalanced orderflow (i.e., when the amount of shares exchanged in trades initiated by buyers does not equal the amount of shares exchanged in trades initiated by sellers over a period of time), and that the widening of bid-ask spreads often precedes a rise in intraday volatility. This is a surprising phenomenon because under the efficient market hypothesis asset prices are expected to reflect all available information at all times, making predictions futile (Fama 1970). The existence of orderflow imbalance, the sequential nature of these events, and their predictability point to market inefficiencies, of unclear source. Such associational observations do not constitute a theory, and they do not explain why the phenomenon occurs.

3.2 The Theoretical Step

In the theoretical step, researchers advance a possible explanation for the observed associated events. This is an exercise in *logical abduction* (sometimes also called retroduction): Given the observed phenomenon, the most likely explanation is inferred by elimination among competing alternatives. Observations cannot be explained by a hypothesis more extraordinary than the observations themselves, and of various hypotheses the least extraordinary must be preferred (Wieten et al. 2020). At this step, a researcher states that X and Y are associated because X causes Y, in the sense that $P\big[Y = y|do[X = x]\big] > P[Y = y]$. For the explanation to be scientific, it must propose a causal mechanism that is falsifiable, that is, propose the system of structural equations along the causal path from X to Y, where the validity of each causal link and causal path can be tested empirically.[8] Physics Nobel Prize laureate Wolfgang Pauli famously remarked that there are three types of explanations: correct, wrong, and not even wrong (Peierls 1992). With "not even wrong," Pauli referred to explanations that appear to be scientific, but use unfalsifiable premises or reasoning, which can never be affirmed nor denied.

A scientist may propose a theory with the assistance of statistical tools (see Section 4.3.1), however data and statistical tools are not enough to produce a theory. The reason is, in the theoretical step the scientist injects extra-statistical information, in the form of a subjective framework of assumptions that give meaning to the observations. These assumptions are unavoidable, because the simple action of taking and interpreting measurements introduces subjective choices, making the process of discovery a creative endeavor.

[8] Following on the earlier examples, in the year 1900, Paul Drude was the first to offer a falsifiable explanation to Ohm's law of 1827; in the year 1915, Albert Einstein offered a falsifiable explanation for Newton's law of gravitation of 1687, and so on.

If theories could be deduced directly from observations, then there would be no need for experiments that test the validity of the assumptions.

Following on the previous example, the Probability of Informed Trading (PIN) theory explains liquidity provision as the result of a sequential strategic game between market makers and informed traders (Easley et al. 1996). In the absence of informed traders, the orderflow is balanced, because uninformed traders initiate buys and sells in roughly equal amounts, hence market impact is mute and the mid-price barely changes. When market makers provide liquidity to uninformed traders, they profit from the bid-ask spread (they buy at the bid price and sell at the ask price). However, the presence of informed traders imbalances the orderflow, creating market impact that changes the mid-price. When market makers provide liquidity to an informed trader, the mid-price changes before market makers are able to profit from the bid-ask spread, and they are eventually forced to realize a loss. As a protection against losses, market makers react to orderflow imbalance by charging a greater premium for selling the option to be adversely selected (that premium is the bid-ask spread). In the presence of persistent orderflow imbalance, realized losses accumulate, and market makers are forced to reduce their provision of liquidity, which results in greater volatility. Two features make the PIN theory scientific: First, it describes a precise mechanism that explains the causal link: orderflow imbalance → market impact → mid-price change → realized losses → bid-ask spread widening → reduced liquidity → greater volatility. Second, the mechanism involves measurable variables, with links that are individually testable. An unscientific explanation would not propose a mechanism, or it would propose a mechanism that is not testable.

Mathematicians use the term theory with a different meaning than scientists. A mathematical theory is an area of study derived from a set of axioms, such as number theory or group theory. Following Kant's epistemological definitions, mathematical theories are synthetic *a priori* logical statements, whereas scientific theories are synthetic *a posteriori* logical statements. This means that mathematical theories do not admit empirical evidence to the contrary, whereas scientific theories must open themselves to falsification.

3.3 The Falsification Step

In the falsification step, researchers not involved in the formulation of the theory independently: (i) deduce key implications from the theory, such that it is impossible for the theory to be true and the implications to be false; and (ii) design and execute experiments with the purpose of proving that the implications are false. Step (i) is an exercise in *logical deduction* because given some theorized premises, a falsifiable conclusion is reached reductively (Gensler 2010, pp. 104–110).

When properly done, performing step (i) demands substantial creativity and domain expertise, as it must balance the strength of the deduced implication with its testability (cost, measurement errors, reproducibility, etc.). Each experiment in step (ii) focuses on falsifying one particular link in the chain of events involved in the causal mechanism, applying the tools of mediation analysis. The conclusion that the theory is false follows the structure of a *modus tollens* syllogism (proof by contradiction): using standard sequent notation, if $A \Rightarrow B$, however $\neg B$ is observed, then $\neg A$, where A stands for "the theory is true" and B stands for a falsifiable key implication of the theory.

One strategy of falsification is to show that $P\left[Y = y|do[X = x]\right] = P[Y = y]$, in which case either the association is noncausal, or there is no association (i.e., the phenomenon originally observed in step (i) was a statistical fluke). A second strategy of falsification is to deduce a causal prediction from the proposed mechanism, and to show that $E\left[Y|do[X = x]\right] \neq y$. When that is the case, there may be a causal mechanism, however, it does not work as theorized (e.g., when the actual causal graph is more complex than the one proposed). A third strategy of falsification is to deduce from the theorized causal mechanism the existence of associations, and then apply machine learning techniques to show that those associations do not exist. Unlike the first two falsification strategies, the third one does not involve a do-operation.

Following on the previous example, a researcher may split a list of stocks randomly into two groups, send buy orders that set the level of orderflow imbalance for the first group, and measure the difference in bid-ask spread, liquidity, and volatility between the two groups (an interventional study, see Section 4.1).[9] In response to random spikes in orderflow imbalance, a researcher may find evidence of quote cancellation, quote size reduction, and resending quotes further away from the mid-price (a natural experiment, see Section 4.2).[10] If the experimental evidence is consistent with the proposed PIN theory, the research community concludes that the theory has (temporarily) survived falsification. Furthermore, in some cases a researcher might be able to inspect the data-generating process directly, in what I call a "field study." A researcher may approach profitable market makers and examine whether

[9] Sophisticated large asset managers routinely conduct so-called algo-wheel experiments to assess broker performance, however the results from these controlled experiments are rarely made public, and are generally unknown to the academic community (López de Prado 2017). See Webster and Westray (2022) for an example of a theoretical framework that covers this kind of execution experiments.

[10] Random spikes in orderflow imbalance allow researchers to observe the reaction of market makers while removing the influence of potential confounders. For the purpose of this experiment, a researcher is interested in orderflow imbalance fluctuations that market makers cannot rule out as random at their onset, however the researcher can determine to have been random (likely ex-post).