entropy

Article

Conducting Causal Analysis by Means of Approximating Probabilistic Truths

Bo Pieter Johannes Andrée [1,2]

[1] Analytics and Tool Unit, Development Economics Data Group, World Bank, 1818 H St NW, Washington, DC 20433, USA; bandree@worldbank.org or b.p.j.andree@vu.nl
[2] Department of Spatial Economics, School of Business and Economics, VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

Simple Summary: The current paper develops a probabilistic theory of causation and suggests practical routines for conducting causal inference applicable to new machine learning methods that have, so far, remained relatively underutilized in this context.

Abstract: The current paper develops a probabilistic theory of causation using measure-theoretical concepts and suggests practical routines for conducting causal inference. The theory is applicable to both linear and high-dimensional nonlinear models. An example is provided using random forest regressions and daily data on yield spreads. The application tests how uncertainty in short- and long-term inflation expectations interacts with spreads in the daily Bitcoin price. The results are contrasted with those obtained by standard linear Granger causality tests. It is shown that the suggested measure-theoretic approaches do not only lead to better predictive models, but also to more plausible parsimonious descriptions of possible causal flows. The paper concludes that researchers interested in causal analysis should be more aspirational in terms of developing predictive capabilities, even if the interest is in inference and not in prediction per se. The theory developed in the paper provides practitioners guidance for developing causal models using new machine learning methods that have, so far, remained relatively underutilized in this context.

Keywords: causality; Bitcoin; inflation; yield spreads; approximation theory; Hellinger distance; Kullback–Leibler divergence; correct specification; misspecified models

check for updates

Citation: Andrée, B.P.J. Conducting Causal Analysis by Means of Approximating Probabilistic Truths. *Entropy* **2022**, *24*, 92. https://doi.org/10.3390/e24010092

Academic Editor: Kateřina Hlaváčková-Schindler

Received: 22 September 2021
Accepted: 27 December 2021
Published: 6 January 2022

1. Introduction

Philosophers have debated at length whether causality is a subject that should be treated probabilistically or deterministically. This resulted in the development of different inferential systems and views on reality. Pure logic dealt with inferences about deterministic truths [1,2]. Probabilistic reasoning has been developed to allow for uncertainty in inferences about deterministic truths [3,4], to make inferences about probabilistic truths [5,6], or to imply the existence of associated deterministic truths [7–11]. Probabilistic theories about causality were developed throughout the 20th century, with notable contributions by Reichenbach, Good, and Suppe [12]. At the same time, however, the classical model of physics maintained its position as a role model for other sciences, which led researchers, including those concerned with human behavior and economic systems, to reject ideas about probabilistic causation, opting, often, to reason probabilistically about deterministic truths.

In modern physics, the standard equations of quantum mechanics suggest that reality is, in fact, better described by probability laws [13]. The outcome of the Bohr–Einstein debates settled on the assertion that these probability laws are a result of a real indeterminacy and that reality itself is probabilistic (One may also argue that this is simply a correct *exposition of the theory* and not necessarily of the physical world, as more complete theories may yet be discovered). Ref. [14] provides an alternative interpretation of quantum physics in which the probability laws are statistical results of the development of completely determined, but hidden, variables. At a macroscopic level, deterministic laws and contingencies induce associated probabilistic laws (Contingencies is a term used by Ref. [14] to refer to independent factors that may exist outside the scope of what is treated by the laws under consideration, and which do not follow necessarily from anything that may be specified under the context of these laws). In particular, by broadening the context of the processes under consideration, new laws that govern some of the contingencies can be found. This inevitably leads to new contingencies: a process that repeats indefinitely. For this reason, any theory about reality that embraces either of deterministic law or chance, to the exclusion of the other, is inherently incomplete. Regardless of one's position on real indeterminism, it holds, according to this logic, that any natural process that arises deterministically must also satisfy statistical laws that are more general, and so any complete theory about interesting real-world phenomena must be probabilistic.

In a probabilistic view of reality cause and consequence are related by probability laws rather than laws of logical truths. A theory about probabilistic causality can, therefore, be stated in terms of the properties of the *true* measure that describes a process stochastically. The theory of causation developed here is that a causal relationship exists if there exists a *true* probability measure that produces a non-empty stochastic sequence that describes the directly caused effects from perturbations in one variable in terms of the responses in another. The paper shows that ideas about causality, including the direction, statistical significance, and economic relevance of effects, may be tested by formulating a statistical model that correctly describes observed data, and evaluating its dynamic properties. In practice, this means that the inference is conducted with a best approximation of the true probability measure. It is the position of the paper that in order to demonstrate that causality runs from a potential causal variable to the target variable, one requires developing the best approximation of the true probability measure using the potential causal variable and a best approximation of the true probability measure without the potential causal variable. The analysis should then (1) conclude whether the first modeled measure is closer to the true measure, and (2) test that the two modeled measures are not equivalent. Practical routines to do so shall be discussed and an example is provided using random forest (RF) regressions and daily data on yield spreads. The application tests how uncertainty around short- and long-term inflation expectations interact with spreads in the daily Bitcoin price, a digital asset with a predetermined finite supply that has been characterized as a new potential inflation hedge. The results are contrasted with those obtained with standard linear Granger causality tests. It is shown that the suggested approaches do not only lead to better predictive models, but also to more plausible parsimonious descriptions of possible causal flows.

The focus on approximating a correct stochastic representation of the DGP (data generating process) as a means of learning about true causal linkages is different from the approaches that try to simulate laboratory conditions by testing for statistical differences in control groups, such as described by [15,16]. The focus on obtaining a correct functional representation of the data is also different from attributing the presence of causal relationships directly to the values of parameters representing averages in treatment groups, see for instance [17–19] on this approach. Placing emphasis on the need for accurate statistical models for the full data distribution when conducting causal analysis introduces an obvious weakness: it is generally accepted that all empirical models will be mis-specified to a certain degree and that empirical models are likely never correctly specified. The *true* process, after all, is unknown in practice. This is the reason to conduct analyses in the first place. The aim to develop correct models can therefore be seen as an idealistic idea that is difficult to put into practice. However, it is still valuable to understand the role of the correct-specification assumption in causal analysis. It is commonly taught that mis-specification leads to residual dependencies that violate the assumptions made by general central limit theorems needed to obtain correct standard errors, see for example chapter 2 in [20]. However, more general estimation theory for dependent processes, as those developed and discussed for instance by [21–25], may help correct standard error estimation but do not remedy the issue that the structural response of the model is incorrect [26]. These are theories to correct the variance estimator when the underlying model is wrong, and do not address the issue that the structural response of the model does not correctly describe the data.

The paper builds on contributions of others in the following lines of research. The views on causality developed in the paper are related to the information theoretic view on testing causal theories, as discussed by [27–30], which, as here, emphasizes model parsimony. The line of reasoning is inspired by the work of [31,32], who emphasized the importance of a probabilistic formulation of economic theories and warned against the use of statistical methods without any reference to a stochastic process. The paper also emphasizes the importance of the overall model response, and, thus, on focusing on system behavior, rather than on isolated parameters that make no reference to a wider economic system. This has previously been advocated by [33]. The main result of the paper is that convincing statements about partial causal linkages must be underpinned by an accurate model of broader reality, even if the interest is in inference and not prediction per se. In order to do so, researchers must, as shall be discussed, pay due attention to distinguishing between direct causal impacts and system memory and take note of developments in the field of predictive modeling.

The plan of the paper is as follows. Section 2 develops definitions for probabilistic causality in terms of true probability measures using a flexible type of dynamical system that covers many processes observed in economics, physics, finance, and related fields of study. Section 3 discusses approximating this true probability measure as an act of minimizing divergence between the modeled probability measure and the true probability measure, while section 4 forges the link between statistical divergence and distance. This draws the connections between distance-minimization and the use of maximum likelihood criteria. Section 5 provides practical considerations and applies the theory. Finally, Section 6 concludes. Proofs are provided in the Appendix A.

2. Causality in Terms of True Probability Measures

Notation will be as follows.

Notation 1. *$\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{R}$, respectively denote the sets of natural, integer, and real numbers. If $\mathcal{A}$ is a set, $\mathfrak{B}(A)$ denotes the Borel-σ algebra over $\mathcal{A}$, and $\times_{t=1}^{t=T}\mathcal{A}$, alternatively denoted as $\mathcal{A}_T$, is the Cartesian product of T copies of $\mathcal{A}$. Definitional equivalence is denoted $:=$, which is to be distinguished from $\equiv$ denoting equivalence, for example in the functional sense. For two maps, f and g, their composition arises from their point-wise application and is denoted $f \circ g := f(g)$ and f^{-1} is the inverse function of f. The tensor product is denoted $\otimes$. The notation $\mu \ll \nu$ is used to indicate that μ is absolutely continuous with respect to ν, i.e., if μ and ν are two measures on the same measurable space $(X, \mathcal{A})$, μ is absolutely continuous with respect to ν if $\mu(A) = 0$ for every set A for which $\nu(A) = 0$, or, as an example, if ν is the counting measure on$[0,1]$ and μ is the Lebesgue measure, then $\mu \ll \nu$. It is also said that ν is* dominating *μ when $\mu \ll \nu$, see for instance ([34] p. 574). Finally, the empty set $\emptyset$ is also used in the context of an empty sequence, which sometimes would be notated as $()$ in the literature.*

Directional causality is interesting when at least two sequences are considered. Specifically, when the focus is on a T-period sequence $\{\mathbf{x}_t(\omega)\}_{t=1}^{T}$, that is a subset of the realized path of the $n_{\mathbf{x}}$-variate stochastic sequence $\mathbf{x}(\omega) := \{\mathbf{x}_t(\omega)\}_{t\in\mathbb{Z}}$ for events in the event space $\omega \in \Omega$. (That is, $\mathbf{x}_t(\omega) \in \mathcal{X} \subseteq \mathbb{R}^{n_{\mathbf{x}}}\ \forall\ (\omega, t) \in \Omega \times \mathbb{Z}$. The random sequence $\mathbf{x}(\omega)$ is a Borel-σ $\mathcal{F}/\mathfrak{B}(\mathcal{X}_\infty)$-measurable map $\mathbf{x} : \Omega \to \mathcal{X}_\infty \subseteq \mathbb{R}^{n_{\mathbf{x}}}_\infty$. In this, $\mathbb{R}^{n_{\mathbf{x}}}_\infty := \times_{t=-\infty}^{t=\infty}\mathbb{R}^{n_{\mathbf{x}}}$ denotes the Cartesian product of infinite copies of $\mathbb{R}^{n_{\mathbf{x}}}$ and $\mathcal{X}_\infty = \times_{t=-\infty}^{t=\infty}\mathcal{X}$ with $\mathfrak{B}(\mathcal{X}_\infty) := \mathfrak{B}(\mathbb{R}^{n_{\mathbf{x}}}_\infty) \cap \mathcal{X}_\infty$, and $\mathfrak{B}(\mathbb{R}^{n_{\mathbf{x}}}_\infty)$ denotes the Borel-σ algebra on the finite dimensional cylinder set of $\mathbb{R}^{n_{\mathbf{x}}}_\infty$, see Theorem 10.1 of [35], p. 159). As always, the complete probability space of interest is described by a triplet $(\Omega, \mathcal{F}, \mathbb{P})$, with $\mathcal{F}$ as the σ-field defined on the event space. $\mathbb{P}$ is used here informally as a placeholder for a collection of probability measures, as we shall introduce the exact probability measures of interest shortly.

If $\mathbf{x}$ is considered as a univariate sequence independent from causal drivers, then for every event $\omega \in \Omega$, the stochastic sequence $\mathbf{x}_t(\omega)$ would live on the probability space $(\mathcal{X}_\infty, \mathfrak{B}(\mathcal{X}_\infty), P^{\mathbf{x}})$ where $P^{\mathbf{x}}$ assigns probability to all elements of $\mathfrak{B}(\mathcal{X}_\infty)$. In a similar fashion, one can consider $\{\mathbf{y}_t(\omega)\}_{t=1}^{T}$ as the subset of the realized path of the $n_{\mathbf{y}}$-variate stochastic sequence $\mathbf{y}(\omega) := \{\mathbf{y}_t(\omega)\}_{t\in\mathbb{Z}}$ indexed by identical t for events $\omega \in \Omega$ (i.e., $\mathbf{y}_t(\omega) \in \mathcal{Y} \subseteq \mathbb{R}^{n_{\mathbf{y}}}$ $\forall$ $(\omega, t) \in \Omega \times \mathbb{Z}$ and the random sequence $\mathbf{y}(\omega)$ is a Borel-σ $\mathcal{F}/\mathfrak{B}(\mathcal{Y}_\infty)$-measurable map $\mathbf{y} : \Omega \to \mathcal{Y}_\infty \subseteq \mathbb{R}^{n_{\mathbf{y}}}_{\infty}$.) If $\mathbf{y}$ would live similarly isolated from outside influence, then for every $\omega \in \Omega$, the stochastic sequence $\mathbf{y}_t(\omega)$ would operate on a space $(\mathcal{Y}_\infty, \mathfrak{B}(\mathcal{Y}_\infty), P^{\mathbf{y}})$ where $P^{\mathbf{y}}$ assigns probability to all the elements of $\mathfrak{B}(\mathcal{Y}_\infty)$. We have a system of two unrelated sequences (This naturally covers to most common auto-regression case, only stated for $\mathbf{y}_t$ here, $\mathbf{y}_t = f^{\mathbf{yy}}(\mathbf{y}_{t-1}) + \boldsymbol{\varepsilon}_t$, where $\boldsymbol{\varepsilon}_t$ is unobserved. The linear auto-regression case is obtained when $f^{\mathbf{yy}}$ is a scaled identity function.):

$$\begin{array}{c} \mathbf{x} := \{\mathbf{x}_t = f^{\mathbf{xx}}(\mathbf{x}_{t-1}), t \in \mathbb{Z}\} \\ \mathbf{y} := \{\mathbf{y}_t = f^{\mathbf{yy}}(\mathbf{y}_{t-1}), t \in \mathbb{Z}\} \end{array} . \tag{1}$$

As we shall see, an important aspect of causal analysis is to rule out that the observed data is not generated by Equation (1). As such, it is important to comment on a number of properties. First, in this system of equations, the functions $f^{\mathbf{xx}}$ and $f^{\mathbf{yy}}$ are intentionally not indexed by t. This does not imply that these functions cannot posses complex time-varying properties; it only limits the discussion to observation-driven models (to the exclusion of parameter-driven models), in which time-varying parameters arise as nonlinear functions of the data. An example would be the threshold models considered by [36,37], in which parameter values are allowed to differ across regimes in the data. The choice to restrict the discussion is made because it is intuitively easier to conceive of causal effects in an observation-driven context where observations represent verifiable values describing different states of real-world phenomena. At the same time, it has been shown that parametric observation-driven models can produce time-varying parameters of a wide class of nonlinear models [38] and that the forecasting power of such models may be on-par with parameter-driven models, even if the latter are correctly specified [39]. Moreover, Refs. [20,40,41] show how observation-driven models may be used to not only investigate how observations impact future observations, but also future parameter values, which may empirically be interesting if those parameters carry an economic interpretation. Finally, many popular machine learning algorithms, such as neural networks, can be reduced to equations that show how parameter values change according to levels in the data [42].

While the dynamics in Equation (1) may be nonlinear, the notation is too restrictive to nest long-memory processes. In particular, the state at time t is only a function of the previous state at time $t-1$, or $t-p$ if the model would be generalized to p-order lags, but not of the full history. Vanishing dependence, implied under contraction conditions [43], is often key to verifying irreducibility and continuity [44] and proving the ergodicity of time series [45]. Proving the ergodicity of a model is needed to obtain an estimation theory under an assumption of correct specification [20,24]. Later, multivariate models will be considered, in which case long-memory properties may arise, for example, when time-varying parameters in one of the functions are a function of past data as well as of past values of those time-varying parameters.

If interrelated stochastic sequences are at the center of inference, additional building blocks are required to describe the processes. This increases the potential complexity of $P^{\mathbf{x}}$ and $P^{\mathbf{y}}$, but it also allows to distinguish between causality, non-causality, and feedback. Consider the stochastic system:

$$\begin{array}{c} \mathbf{x} := \{\mathbf{x}_t = f^{\mathbf{xx}}(\mathbf{x}_{t-1}) + f^{\mathbf{xy}}(\mathbf{y}_{t-1}), t \in \mathbb{Z}\} \\ \mathbf{y} := \{\mathbf{y}_t = f^{\mathbf{yx}}(\mathbf{x}_{t-1}) + f^{\mathbf{yy}}(\mathbf{y}_{t-1}), t \in \mathbb{Z}\} \end{array} . \tag{2}$$

In this multivariate context, $f^{\mathbf{xy}}$ and $f^{\mathbf{yx}}$ will be referred to as the direct causal maps, while $f^{\mathbf{xx}}$ and $f^{\mathbf{yy}}$ control the memory properties within each channel.

When $\mathbf{x}$ and $\mathbf{y}$ are analyzed individually, the properties of $f^{\mathbf{xx}}$ and $f^{\mathbf{yy}}$ are of key interest. They carry information on the future positions of $\mathbf{x}_{t+1}$ and $\mathbf{y}_{t+1}$, and provide predictability without considering outside influence directly. However, correct causal inference around the interdependencies of $\mathbf{x}$ and $\mathbf{y}$ may be preferred over developing predictive capabilities that can result from many configurations within the parameter space that are associated with untrue probability measures. The properties of $f^{\mathbf{xy}}$ and $f^{\mathbf{yx}}$ determine the direction in which effects move. Verifying their properties is central to causality studies. The functions $f^{\mathbf{xx}}$ and $f^{\mathbf{yy}}$, on the other hand, play a central role in the system's responses to external impulses by shaping memory of the causal initial impact of a sequence of interventions, even after that sequence turns inactive.

The functions that control memory properties within channels in some sense determine how the past reverberates into the future, and specifying correct empirical equivalents to $f^{\mathbf{xx}}$ and $f^{\mathbf{yy}}$ is as crucial to the inference about the causal interdependencies as is specifying mechanisms for the action of interest (it would be more general to write Equation (2) with $\mathbf{x} := \{\mathbf{x}_t = f^{\mathbf{xx}}(\mathbf{x}_{t-1}; \mathbf{w}_{t-1}) + f^{\mathbf{xy}}(\mathbf{y}_{t-1}; \mathbf{w}_{t-1}), t \in \mathbb{Z}\}$ and $\mathbf{y} := \{\mathbf{y}_t = f^{\mathbf{yx}}(\mathbf{x}_{t-1}; \mathbf{w}_{t-1}) + f^{\mathbf{yy}}(\mathbf{y}_{t-1}; \mathbf{w}_{t-1}), t \in \mathbb{Z}\}$ and with $\mathbf{w}_t = (\mathbf{x}_t, \mathbf{y}_t)$. In this case, for instance, the dependence of $\mathbf{x}_t$ on its own past, $\mathbf{x}_{t-1}$, is allowed to vary based on the levels in past data. However, under this notation, one could at any point in time, decompose the change in one variable into effects attributed to memory and outside influence separately, which the simplified notation in Equation (2) is intended to focus on). In fact, as Ref. [46] point out, systems may be dominated by memory and the influence of the causal components may be small on the overall process in which case predictive power can be obtained without specifying any causal maps and focusing solely on memory. Inversely, this also suggests that one must obtain a model for the memory process to isolate the causal impacts themselves, suggesting that long-memory applications in which causal inference is of interest must develop a high degree of predictive power, even if prediction is not needed for policy purposes. This can be made more clear by considering the following:

$$\begin{array}{l} \mathbf{x}^0 := \{\mathbf{x}_t^0 = f^{\mathbf{xy}}(\mathbf{y}_{t-1}), t \in \mathbb{Z}\} \\ \mathbf{y}^0 := \{\mathbf{y}_t^0 = f^{\mathbf{yx}}(\mathbf{x}_{t-1}), t \in \mathbb{Z}\} \end{array}, \tag{3}$$

with $\mathbf{x}^0$ and $\mathbf{y}^0$ defined as $\mathbf{x}_t^0 = \mathbf{x}_t - f^{\mathbf{xx}}(\mathbf{x}_{t-1})$ and $\mathbf{y}_t^0 = \mathbf{y}_t - f^{\mathbf{yy}}(\mathbf{y}_{t-1})$. Given the realized sequences $\mathbf{y}(\omega)$ and $\mathbf{x}(\omega)$ generated by Equation (2), the sequential system of Equation (3) moves forward in time as the one-step-ahead directly caused parts of $\mathbf{y}$ and $\mathbf{x}$ that are filtered from the reverberating effects of $f^{\mathbf{xx}}$ and $f^{\mathbf{yy}}$. More specifically, while $\mathbf{y}$ partially consists of memory, there is a part, $\mathbf{y}^0$, that, at any point, is directly mapped from the previous state of $\mathbf{x}$, while, at the same time, $\mathbf{x}$ consists partially of memory and a part $\mathbf{x}^0$ directly generated from the last position of $\mathbf{y}$. In this view, directional causality can be stated in terms of whether (3) produces any values, i.e., diagnosing if there is any statistically significant signal from initial causal impulses left after all memory properties have been stripped from the data. Importantly, the system reveals that by the definitions of $\mathbf{x}_t^0$ and $\mathbf{y}_t^0$, obtaining appropriate estimates for $f^{\mathbf{xy}}$ and $f^{\mathbf{yx}}$ involves $f^{\mathbf{xx}}$ and $f^{\mathbf{yy}}$ being modeled correctly as $\mathbf{x}_t^0$ and $\mathbf{y}_t^0$ are not observed and only result as functions from the observable processes $\mathbf{y}$ and $\mathbf{x}$. Moreover, if $\mathbf{y}(\omega)$ and $\mathbf{x}(\omega)$ are triggered by an event, then it is possible, by process of infinite backward substitution, to write Equation (3) as an infinite chain initialized in the infinite past. Plugging in the equalities $\mathbf{x}_t = \mathbf{x}_t^0 + f^{\mathbf{xx}}(\mathbf{x}_{t-1})$ and $\mathbf{y}_t = \mathbf{y}_t^0 + f^{\mathbf{yy}}(\mathbf{y}_{t-1})$ and defining the random functions $f_{\mathbf{y}}^0(\mathbf{y}_t^0, \mathbf{y}_{t-1}) = f^{\mathbf{xy}}\big(\mathbf{y}_t^0 + f^{\mathbf{yy}}(\mathbf{y}_{t-1})\big)$ and $f_{\mathbf{x}}^0(\mathbf{x}_t^0, \mathbf{x}_{t-1}) = f^{\mathbf{yx}}\big(\mathbf{x}_t^0 + f^{\mathbf{xx}}(\mathbf{x}_{t-1})\big)$, one can write

$$\begin{array}{l} \mathbf{x}^0 := \{\mathbf{x}_t^0 = f_{\mathbf{y}}^0(\mathbf{y}_{t-1}^0, \mathbf{y}_{t-2}), t \in \mathbb{Z}\} \\ \mathbf{y}^0 := \{\mathbf{y}_t^0 = f_{\mathbf{x}}^0(\mathbf{x}_{t-1}^0, \mathbf{x}_{t-2}), t \in \mathbb{Z}\} \end{array}. \tag{4}$$

Repeating infinitely, and extending infinitely in the direction $T \to \infty$,

$$\begin{array}{l} \mathbf{x}^0 := \{\mathbf{x}_\infty^0 = (f_{\mathbf{y}}^0)^\infty(\mathbf{y}_1^0, \mathbf{y}_1), t \in \mathbb{Z}\} \\ \mathbf{y}^0 := \{\mathbf{y}_\infty^0 = (f_{\mathbf{x}}^0)^\infty(\mathbf{x}_1^0, \mathbf{x}_1), t \in \mathbb{Z}\} \end{array}. \tag{5}$$

$(f^0_{\mathbf{y}})^\infty$ and $(f^0_{\mathbf{x}})^\infty$ are the maps that generate $\mathbf{y}^0$ and $\mathbf{x}^0$ infinitely after $\mathbf{y}$ and $\mathbf{x}$ have been generated into infinity. Subscript $_1$ has been used, here, to mark the initialization points. This shows that $\mathbf{x}^0$ can be written as a sequence of iterating functional operations that are all defined on $\mathbf{y}$, and $\mathbf{y}^0$ defined on $\mathbf{x}$ in a similar way (Equation (5) reveals that the sequences that constitute the directly caused parts of $\mathbf{x}$ and $\mathbf{y}$ are ultimately dependent on the values at which the observable process has been initialized. That is, the entire causal pathway depends on the initial impact. In practice, one cannot observe all impacts—including those that occurred in the infinite past—and assurance is required that the initialization effect of the causal pathway must, asymptotically, be irrelevant). For ease of notation, let us write

$$\begin{aligned} \mathbf{x}^0 &:= \{\mathbf{x}^0_t = \mathbf{f}^0_{\mathbf{y}}(\mathbf{y}_{-\infty:t}), t \in \mathbb{Z}\} \\ \mathbf{y}^0 &:= \{\mathbf{y}^0_t = \mathbf{f}^0_{\mathbf{x}}(\mathbf{x}_{-\infty:t}), t \in \mathbb{Z}\} \end{aligned} . \tag{6}$$

where bold-faced $\mathbf{f}^0$ is used to refer to the entire sequence of functional operations f^0 up to t, starting in the infinite past $t = -\infty$. This highlights that generating the unobserved quantities $\mathbf{x}^0$ and $\mathbf{y}^0$ from the observed quantities $\mathbf{x}$ and $\mathbf{y}$ by back substitution eventually involves the unobserved quantities $\mathbf{x}_1$ and $\mathbf{y}_1$. This means that some feasible form of approximation is needed, since time series data in practice area almost never recorded since the beginning of the process.

Note first that $\mathbf{f}^0_{\mathbf{y}} : \mathcal{Y} \to \mathcal{X} \subseteq \mathbb{R}$ is a $\mathfrak{B}(\mathcal{Y})/\mathfrak{B}(\mathcal{X})$-measurable mapping, and $\mathbf{f}^0_{\mathbf{x}} : \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}$ is a $\mathfrak{B}(\mathcal{X})/\mathfrak{B}(\mathcal{Y})$-measurable mapping. The sequence $\mathbf{x}^0$ thus lives on $(\mathcal{X}_\infty, \mathfrak{B}(\mathcal{X}_\infty), P^{\mathbf{x}}_0)$, where $P^{\mathbf{x}}_0$ is induced according to $P^{\mathbf{x}}_0(B_{\mathbf{x}}) = P^{\mathbf{y}} \circ (\mathbf{f}^0_{\mathbf{y}})^{-1}(B_{\mathbf{x}})\ \forall\ B_{\mathbf{x}} \in \mathfrak{B}(\mathcal{X}_\infty)$, and $\mathbf{y}^0$ lives on $(\mathcal{Y}_\infty, \mathfrak{B}(\mathcal{Y}_\infty), P^{\mathbf{y}}_0)$, where $P^{\mathbf{y}}_0$ is induced according to $P^{\mathbf{y}}_0(B_{\mathbf{y}}) = P^{\mathbf{x}} \circ (\mathbf{f}^0_{\mathbf{x}})^{-1}(B_{\mathbf{y}})\ \forall\ B_{\mathbf{y}} \in \mathfrak{B}(\mathcal{Y}_\infty)$, see [47] p. 118 and [48] p. 115. The notation shows that the probability measures underlying the stochastic causal sequences result from the functional behavior of the entire system. In particular, the causal sequences can be written as recursive direct effects from another variable that itself consists of memory and causal effects, and the probability measures underlying the causal sequences are thus induced by the functional relationships that describe all dynamical dependencies. This is important to the extent that many causal studies focus on one single marginal dependency, while, from the measure-theoretic perspective developed here, the wider system within any one single process operates, is of importance to the analysis. This suggests that researchers must pay attention to referencing the workings of a broader system when designing their models for inference, something [33] has also argued. Moreover, it has been argued (see [49] for discussion) that probabilistic definitions of causality are not strictly causal in the sense that they do not provide insight in the origin of the probability law that regulates the process of interest, and that a (correct) time-series model only describes (correctly) the probabilistic behavior as the outcome of that unknown causal origin. The notation, here, shows, however, explicitly the relation between the functional behavior of a system and its induced probability measure that assigns probability to all possible outcomes. This suggests that such critiquing views, rather, relate to disagreements around the level of detail in the structure of a model, which in turn would be guided by the research question of interest and the availability of detailed data. Particularly, dynamical systems in economics are often modeled using aggregate macro-economic data that do not have the same granularity as micro-economic data containing information about the behaviors of individual economic agents.

In many cases, a researcher is not able to observe all the relevant variables. When a third, possibly unobserved external variable, $\mathbf{z}$, with effect $f^{\mathbf{z}}(\mathbf{z})$, is considered, the researcher is confronted with the situation that

$$\begin{aligned} \mathbf{x} &:= \{\mathbf{x}_t = f^{\mathbf{xx}}(\mathbf{x}_{t-1}) + f^{\mathbf{xy}}(\mathbf{y}_{t-1}) + f^{\mathbf{xz}}(\mathbf{z}_{t-1}), t \in \mathbb{Z}\} \\ \mathbf{y} &:= \{\mathbf{y}_t = f^{\mathbf{yx}}(\mathbf{x}_{t-1}) + f^{\mathbf{yy}}(\mathbf{y}_{t-1}) + f^{\mathbf{yz}}(\mathbf{z}_{t-1}), t \in \mathbb{Z}\} \end{aligned} . \tag{7}$$

If $\mathbf{z}$ is unobserved, it can still be approximated as a difference combination of $\mathbf{x}$ and $\mathbf{y}$. To obtain an approximated sequence of the *true* $\mathbf{z}$ sequence to condition empirical counterparts for $f^{\mathbf{xz}}$ and $f^{\mathbf{yz}}$ on, one can work with:

$$\begin{array}{c}\mathbf{z} := \{\mathbf{z}_t = f^{\mathbf{z}|\mathbf{xy}}(\mathbf{x}_{t+1} - (f^{\mathbf{xx}}(\mathbf{x}_t) + f^{\mathbf{xy}}(\mathbf{y}_t))), t \in \mathbb{Z}\} \\ \mathbf{z} := \{\mathbf{z}_t = f^{\mathbf{z}|\mathbf{yx}}(\mathbf{y}_{t+1} - (f^{\mathbf{yx}}(\mathbf{x}_t) + f^{\mathbf{yy}}(\mathbf{y}_t))), t \in \mathbb{Z}\}\end{array} . \tag{8}$$

Equation (8) suggests to write Equation (7) in terms of $\mathbf{y}$ and $\mathbf{x}$ only by defining $\mathbf{z}$ as a difference combination of $\mathbf{x}$ and $\mathbf{y}$ (Apart from stability conditions imposed on the endogenous process, one requires also that the exogenous impacts enter the system in some suitable manner, which, for example, requires that $f^{\mathbf{xz}}$ and $f^{\mathbf{yz}}$ are appropriately bounded. Following the same arguments that resulted in Equation (5), the initialization of the exogenous impacts $\mathbf{z}_1$ should similarly not carry information influential in the empirical estimates of $f^{\mathbf{xy}}$ and $f^{\mathbf{yx}}$, conditional on partial information). This allows us to define the spaces and measures in terms of $\mathbf{x}$ and $\mathbf{y}$ when the multivariate process includes further variables, in this case, $\mathbf{z}$. If the process is invertible, one can write, by aggregating the functions:

$$\begin{array}{c}\mathbf{x} := \{\mathbf{x}_t = f^{\mathbf{xx}}(\mathbf{x}_{t-1}) + f^{\mathbf{xy}}(\mathbf{y}_{t-1}) + f^{\mathbf{xz}}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{y}_{t-1}), t \in \mathbb{Z}\} \\ \mathbf{y} := \{\mathbf{y}_t = f^{\mathbf{yx}}(\mathbf{x}_{t-1}) + f^{\mathbf{yy}}(\mathbf{y}_{t-1}) + f^{\mathbf{yz}}(\mathbf{y}_t, \mathbf{x}_{t-1}, \mathbf{y}_{t-1}), t \in \mathbb{Z}\}\end{array} . \tag{9}$$

$$\begin{array}{c}\mathbf{x} := \{\mathbf{x}_t = f^{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}), t \in \mathbb{Z}\} \\ \mathbf{y} := \{\mathbf{y}_t = f^{\mathbf{y}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}), t \in \mathbb{Z}\}\end{array} . \tag{10}$$

$$\begin{array}{c}\mathbf{x} := \{\mathbf{x}_t = f^{\mathbf{x}}(\mathbf{w}_{t-1}), t \in \mathbb{Z}\} \\ \mathbf{y} := \{\mathbf{y}_t = f^{\mathbf{y}}(\mathbf{w}_{t-1}), t \in \mathbb{Z}\}\end{array} . \tag{11}$$

For every $t \in \mathbb{Z}$, the map $f^{\mathbf{x}} \circ (\mathbf{y}_{t-1}, \mathbf{x}_{t-1}) : \Omega \to \mathcal{X}$ is $\mathcal{F}/\mathfrak{B}(\mathcal{X})$-measurable and $\mathbf{x}(\omega)$ lives on the space $(\mathcal{X}_\infty, \mathfrak{B}(\mathcal{X}_\infty), P^{\mathbf{x}})$ where the probability measure $P^{\mathbf{x}}$ is induced by $f^{\mathbf{x}}$ on $\mathfrak{B}(\mathcal{X}_\infty)$ according to the point-wise application of $P^{\mathbf{w}}$ and the inverse of $f^{\mathbf{x}}$. ($P^{\mathbf{x}}(B_{\mathbf{x}}) = P^{\mathbf{w}} \circ (f^{\mathbf{x}})^{-1}(B_{\mathbf{x}}) \,\forall\, (B_{\mathbf{x}}) \in \mathfrak{B}(\mathcal{X}_\infty)$). Similar arguments follow for $P^{\mathbf{y}}$. This tells us that, in the general case of multivariate dependencies and in the presence of possibly unobserved variables, the probability measures underlying the individual sequences are possibly a result of those of the other sequences. This means the space of empirical candidates for the probability measure $P^{\mathbf{w}}$ that underlies the joint process $\mathbf{w} := \{\mathbf{w}_t = (\mathbf{y}_t, \mathbf{x}_t), t \in \mathbb{Z}\}$ operates on $(\mathcal{W}_\infty, \mathfrak{B}(\mathcal{W}_\infty), P^{\mathbf{w}})$. (The sequence realizes under the events $\omega \in \Omega$, $\mathbf{w}_t(\omega) \in \mathcal{W}$, where $\mathcal{W} := \mathcal{Y} \times \mathcal{X}$ and $\mathbf{w}(\omega) \in \mathcal{W}_\infty$, with $\mathcal{W}_\infty := \mathcal{Y}_\infty \times \mathcal{X}_\infty \subseteq \mathbb{R}_\infty^{n_{\mathbf{x}}+n_{\mathbf{y}}} := \times_{t=-\infty}^{t=\infty} \mathbb{R}^{n_{\mathbf{x}}+n_{\mathbf{y}}}$, and the probability measure of the joint process $P^{\mathbf{w}}$ is thus defined on the product σ-algebra $\mathfrak{B}(\mathcal{W}_\infty) = \mathfrak{B}(\mathcal{X}_\infty \times \mathcal{Y}_\infty) = \mathfrak{B}(\mathcal{X}_\infty) \otimes \mathfrak{B}(\mathcal{Y}_\infty) := \mathcal{W}_\infty \cap \mathfrak{B}(\mathbb{R}_\infty^{n_{\mathbf{x}}+n_{\mathbf{y}}})$ (see, [47] p. 119)).

Regardless, the measure $P^{\mathbf{w}}$ is induced by functional relations of Equation (2), which, as was shown, can be decomposed into memory and causal subsystems. One can thus state causality conditions, based on the measures that describe the directly caused effects represented by Equation (6). In particular, one can keep the focus on $P_0^{\mathbf{x}}$ and $P_0^{\mathbf{y}}$, bearing in mind that they are lower-level constituents of $P^{\mathbf{w}}$ on which, in turn, the complete estimation objective will be defined.

Definition 1 (Non-causality). *The stochastic sequences $\mathbf{x}(\omega)$ and $\mathbf{y}(\omega)$ are not causally related if $P_0^{\mathbf{x}}$ and $P_0^{\mathbf{y}}$ are null measures, such that $\mathbf{x}^0(\omega) \in \emptyset \,\forall\, (\omega, t) \in \Omega \times \mathbb{Z}$ and $\mathbf{y}^0(\omega) \in \emptyset \,\forall\, (\omega, t) \in \Omega \times \mathbb{Z}$.*

Definition 2 (Uni-directional Causality). *Causality runs uni-directionally from the stochastic sequence $\mathbf{x}(\omega)$ to another stochastic sequence $\mathbf{y}(\omega)$ (visa versa), if $P_0^{\mathbf{x}}$ is a null measure, and $P_0^{\mathbf{y}}$ is a non-null measure, such that $\mathbf{x}^0(\omega) \in \emptyset \,\forall\, (\omega, t) \in \Omega \times \mathbb{Z}$ and $\mathbf{y}^0(\omega) \in \mathcal{Y} \,\forall\, (\omega, t) \in \Omega \times \mathbb{Z}$ (visa versa).*

Definition 3 (Bi-directional Causality). *The stochastic sequence $\mathbf{x}(\omega)$ is causal with respect to $\mathbf{y}(\omega)$ and $\mathbf{y}(\omega)$ is causal with respect to $\mathbf{x}(\omega)$, if $P_0^{\mathbf{x}}$ and $P_0^{\mathbf{y}}$ are both non-null measures, such that $\mathbf{x}^0(\omega) \in \mathcal{X} \,\forall\, (\omega, t) \in \Omega \times \mathbb{Z}$ and $\mathbf{y}^0(\omega) \in \mathcal{Y} \,\forall\, (\omega, t) \in \Omega \times \mathbb{Z}$.*

Respectively, conditioning on impacts in $\mathbf{x}$, these probabilistic causality definitions can thus be understood broadly as:

1. Whenever an intervention in $\mathbf{x}$ occurs, there is no chance that $\mathbf{y}^0$ reacts as a result of that.
2. Whenever an intervention in $\mathbf{x}$ occurs, there is positive chance that $\mathbf{y}^0$ reacts as a result of that.
3. Whenever an intervention in $\mathbf{x}$ occurs, there is positive chance that $\mathbf{y}^0$ reacts as a result of that. Subsequently there is positive chance that $\mathbf{x}$ reacts to this initial reaction, a probabilistic process that repeats recursively.

Remark 1. *With null-measures, it is meant that the stochastic sequence describing the directly caused effects from one variable to the other takes values in the empty set with probability 1. This is because the functions that induce the probability measure cancel out, hence, they can be removed from the equations resulting in a probability measure that is not induced by any remaining rule or relationship. In practice, one can test whether $P^{\mathbf{x}}|f^{\mathbf{xx}} \equiv P^{\mathbf{x}}|f^{\mathbf{xx}}f^{\mathbf{xy}}$ or $P^{\mathbf{x}}|f^{\mathbf{xx}} \not\equiv P^{\mathbf{x}}|f^{\mathbf{xx}}f^{\mathbf{xy}}$, where $P^{\mathbf{x}}|f^{\mathbf{xx}}$ here denotes the probability measure induced by the functional relationships in Equation (1) and $P^{\mathbf{x}}|f^{\mathbf{xx}}f^{\mathbf{xy}}$ denotes the probability measure induced by the functional relationships in Equation (2), to test whether $P_0^{\mathbf{x}}$ exists. A practical test is a Kolmogorov–Smirnov-type test.*

3. Limit Divergence on the Space of Modeled Probability Measures

The definitions of causality, in terms of the lower-level components of $P^{\mathbf{w}}$, suggest that correct causal statements can be obtained empirically by extracting relevant counterparts to $P_0^{\mathbf{x}}$ and $P_0^{\mathbf{y}}$ from a relevant counterpart to $P^{\mathbf{w}}$, and investigating the stochastic sequences produced by these modeled measures. For such an approach to be of relevance in an empirical context, one must ensure that the concepts introduced adequately transfer over from the *true* measure $P^{\mathbf{w}}$ to a modeled measure $P^{\hat{\mathbf{w}}}$. The focus is therefore shifted towards detailing how $P^{\hat{\mathbf{w}}}$ can be approximated as a minimally divergent measure relative to $P^{\mathbf{w}}$, and draw on approximation theory to construct equivalence around the *true* measure under an axiom of correct specification.

For some event $\omega \in \Omega$, a realized T-period sequence $\mathbf{w}_T(\omega) := (\mathbf{y}_T(\omega), \mathbf{x}_T(\omega))$ consisting of sequences $\{\mathbf{y}_t(\omega)\}_{t=1}^{t=T}$ and $\{\mathbf{x}_t(\omega)\}_{t=1}^{t=T}$ can be observed. The *true* function $f^{\mathbf{w}}$, consists of our main functions of interest $f^{\mathbf{x}}$ and $f^{\mathbf{y}}$ that in turn are composed of $f^{\mathbf{xy}}$ and $f^{\mathbf{yx}}$ that are of particular interest to the researcher focused on causality, but possibly also functions $f^{\mathbf{xx}}$ and $f^{\mathbf{yy}}$ that shape the responses of an initial causal effect. The exact properties are generally unknown to the observer, but one can design a parameterization mapping that learns the behavior of $f^{\mathbf{x}}$ and $f^{\mathbf{y}}$ when exposed to sufficient data. To learn from the data an approximation of $f^{\mathbf{x}}$ and $f^{\mathbf{y}}$, one can postulate a model

$$\hat{\mathbf{w}} := \{\hat{\mathbf{w}}_t = f(\mathbf{w}_{t-1}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta, t \in \mathbb{Z}\}, \tag{12}$$

with $f{:}\mathcal{W} \times \Theta \to \mathcal{W}$ as our postulated model function and $\hat{\mathbf{w}}$ as the modeled data. In the context of parametric inference, the parameter space Θ is of finite dimensionality, but also in the nonparametric case, the vector $\boldsymbol{\theta} \in \Theta$ indexes parametric models nested by the nonparametric model, each inducing its own probability measure, and Θ indexes families of parametric models, each inducing a space of parametric functions generated under Θ. In this discussion a compact set of potential hypotheses is considered, limiting the inference to parametric models. The arguments can be extended to the nonparametric case, by focusing on a compact subset $\Theta_s \subset \Theta$ of solutions (For example, by letting Θ_s grow as $T \to \infty$, hence focusing on the case $\Theta_{s1} \subset \Theta_{s2} ... \subset \Theta_{s\infty} \subseteq \Theta$, see for example [50]). For example, by using priors or penalties that discard $\Theta \setminus \Theta_s$ such that any solution of the criterion necessarily falls within a compact subset space, see [20] p. 210 and [24]. Let f be $\mathfrak{B}(\mathcal{W})$-measurable $\forall\ \boldsymbol{\theta} \in \Theta$ so that $f(\mathbf{w}_t; \boldsymbol{\theta}) : \Omega \to \mathcal{W}$ is $\mathcal{F}/\mathfrak{B}(\mathcal{W})$-measurable $\forall\ \boldsymbol{\theta} \in \Theta$ and $t \in \mathbb{Z}$. $F_\Theta := \{f(\cdot; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ is our space of parametric functions defined on $\mathcal{W}$ generated under Θ under the injective $f_{\mathcal{W}} : \Theta \to F_\Theta(\mathcal{W})$ where $f_{\mathcal{W}}(\boldsymbol{\theta}) := f(\cdot; \boldsymbol{\theta}) \in$

$F_{\Theta}(\mathcal{W})\ \forall\ \boldsymbol{\theta} \in \Theta$. Under any *true* probability measure $P^{\mathbf{w}}$, every potential parameter vector included in the parameter space $\boldsymbol{\theta} \in \Theta$ induces a probability measure $P^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}}$ indexed by $\boldsymbol{\theta}$ on $\mathfrak{B}(\mathcal{W}_{\infty})$, according to $P^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}}(B_{\mathbf{w}}) = P^{\mathbf{w}} \circ f^{-1}(B_{\mathbf{w}}, \boldsymbol{\theta})\ \forall\ (B_{\mathbf{w}}, \boldsymbol{\theta}) \in \mathfrak{B}(\mathcal{W}_{\infty} \times \Theta)$. Thus, for every potential parameter vector included in the parameter space $\boldsymbol{\theta} \in \Theta$, there is a triplet $(\mathcal{W}_{\infty}, \mathfrak{B}(\mathcal{W}_{\infty}), P^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}})$ that describes the probability space of modeled data under $\boldsymbol{\theta}$. The triplet $(\mathcal{W}_{\infty}, \mathfrak{B}(\mathcal{W}_{\infty}), P^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}})$ is, thus, itself an element of the measure spaces indexed by $\boldsymbol{\theta}$ across all Θ. Given the *true* probability measure $P^{\mathbf{w}}$ on $\mathfrak{B}(\mathcal{W})$, this process is summarized by a functional $\mathfrak{P} : F_{\Theta}(\mathcal{W}) \to \mathcal{P}^{\hat{\mathbf{w}}}_{\Theta}$, that maps elements from the space of parametric functions generated by the entire parameter space $F_{\Theta}(\mathcal{W})$, onto the space $\mathcal{P}^{\hat{\mathbf{w}}}_{\Theta}$ of probability measures defined on the sets of $\mathfrak{B}(\mathcal{W}_{\infty})$ generated by Θ through $f(\cdot;\boldsymbol{\theta})$.

Now, $f^{\mathbf{w}}$ is generally not only unknown, but for a finite Θ there is no guarantee that $\exists \boldsymbol{\theta}_0 \in \Theta : P \circ f_{\mathcal{W}}(\boldsymbol{\theta}_0) = P^{\mathbf{w}}$, implying that, in many empirical applications, one is concerned with the situation where $P^{\mathbf{w}} \notin \mathcal{P}^{\hat{\mathbf{w}}}_{\Theta}$. However, if $\exists P^{\mathbf{w}} \in \mathcal{P}^{\hat{\mathbf{w}}}_{\Theta}$, one can learn all about $P^{\mathbf{w}}$ by uncovering the properties of f, given that a sufficient amount of observations is available. (As discussed in the literature on miss-specification, even when the axiom of correct specification is abandoned, f may converge to a function that produces the optimal conditional density, which may reveal properties of $f^{\mathbf{w}}$). Let

$$\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta} \in \Theta} Q_T(\mathbf{w}_T; \boldsymbol{\theta}), \tag{13}$$

$\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$, be the extremum estimate for $\boldsymbol{\theta}_0$ as judged by the criterion $Q_T : \mathcal{W}_T \times \Theta \to \mathbb{R}$. Trivially, $\mathcal{W}_T := \mathcal{Y}_T \times \mathcal{X}_T$ and $\mathbf{w}_T(\omega) \in \mathcal{W}_T$. To see that under correct specification it is possible to approximate the *true* function $f^{\mathbf{w}}$ in terms of equivalence (in the sense of function equivalence [51] p. 288), one can write the criterion function also as a function of the *true* function and the postulated model $Q_T(f^{\mathbf{w}}(\mathbf{w}_T), f(\mathbf{w}_T; \boldsymbol{\theta}))$ in which it is made use of the fact that $f^{\mathbf{w}}(\mathbf{w}_T) := \{f^{\mathbf{w}}(\mathbf{w}_t)\}_{t=1}^{T} := \mathbf{w}_T$ and $f(\mathbf{w}_T; \boldsymbol{\theta}) := \{f(\mathbf{w}_t; \boldsymbol{\theta})\}_{t=1}^{T} := \hat{\mathbf{w}}_T$.

The discussion further evolves toward showing that the element in $\mathcal{P}^{\hat{\mathbf{w}}}_{\Theta}$ that is closest to $P^{\mathbf{w}}$ minimizes a divergence metric that results from a transformation of the limit criterion that measures the divergence between the *true* density and the density implied by the model. Note that $\mathcal{P}^{\hat{\mathbf{w}}}_{\Theta}$ is induced by the proposed candidates for $P^{\mathbf{w}}$; studies on causality thus rely on flexible model design as the researcher determines which hypotheses are considered in a study by exerting control over Θ. Naturally, if $\Theta_1 \subset \Theta_2$, then Θ_2 produces a larger $\mathcal{P}^{\hat{\mathbf{w}}}_{\Theta_2} \supset \mathcal{P}^{\hat{\mathbf{w}}}_{\Theta_1}$. This suggests that minimizing this divergence metric over a large as possible $\mathcal{P}^{\hat{\mathbf{w}}}_{\Theta}$ results in selecting $P^{\hat{\mathbf{w}}}$ at a point in $\mathcal{P}^{\hat{\mathbf{w}}}_{\Theta}$ that attains equivalence to $P^{\mathbf{w}}$ only when Θ is large enough to produce a correctly specified hypothesis set. Note that the definition of $F_{\Theta} := \{f(\cdot;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$, as our space of parametric functions generated under Θ, under the injective $f_{\mathcal{W}} : \Theta \to F_{\Theta}(\mathcal{W})$ and the functional $\mathfrak{P} : F_{\Theta}(\mathcal{W}) \to \mathcal{P}^{\hat{\mathbf{w}}}_{\Theta}$ that induces the space of probability measures, is defined on the sample space $\mathcal{W}$. This highlights that the correct specification argument, $P^{\mathbf{w}} \in \mathcal{P}^{\hat{\mathbf{w}}}_{\Theta}$, not only stresses flexible parameterization in the sense that parameterized dependencies can take on many values, but also in the sense of using correct data (Indeed, the potential parameters that would interact with data that is not used are essentially treated as zero, so the focus on using correct data is implicitly already contained in the standard statements of correct specification that focus directly on the dimensions of Θ. The distinction is nevertheless useful because nonparametric models are often popularized as methods to reduce miss-specification bias as Θ becomes infinite dimensional, but this does not imply that $P^{\mathbf{w}} \in \mathcal{P}^{\hat{\mathbf{w}}}_{\Theta}$ if important data is missing). When little is known about f, one is thus not only concerned with flexibility in terms of the type of parametric functions generated under Θ, but also the variables on which the modeled measures are defined. When these concerns are appropriately addressed, testing for causality is deciding based on the approximation $P^{\hat{\mathbf{w}}}$ whether the best approximation of the *true* model suggests (1) that $\mathbf{x}$ and $\mathbf{y}$ live in isolation, (2) unidirectional causality, or (3) that $P^{\mathbf{w}}$ produces feedback.

To turn this problem into a selection problem that can be solved by divergence minimization w.r.t. the *true* measure, first introduce the limit criterion by taking $T \to \infty$ and working with the modeled data as the minimizer of the criterion. Specifically, let the limit criterion be $Q_\infty(\boldsymbol{\theta}) := Q_T(f^{\mathbf{w}}(\mathbf{w}_T), f(\mathbf{w}_T; \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{w}_T;\boldsymbol{\theta})))$ evaluated at $T \to \infty$ with $Q_\infty : \Theta \to \mathbb{R}$ and $Q_\infty(\boldsymbol{\theta}) = Q^{\mathcal{P}}_\infty(P^{\mathbf{w}}; P^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}})\ \forall\ \boldsymbol{\theta} \in \Theta$ with the criterion $Q_\infty(\boldsymbol{\theta}) = Q^{\mathcal{P}}_\infty$ as a measure of divergence $d_{\mathcal{P}}$ on the *true* probability measure and the modeled measure. More specifically, $d_{\mathcal{P}} \equiv Q^{\mathcal{P}}_\infty : \mathcal{P}^{\hat{\mathbf{w}}}_\Theta \times \mathcal{P}^{\hat{\mathbf{w}}}_\Theta \to \mathbb{R}_{\geq 0}$. By definition of $Q^{\mathcal{P}}_\infty$ as a divergence on the space that contains $P^{\mathbf{w}}$ and $P^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}}\ \forall\ \boldsymbol{\theta} \in \Theta$, the element $\boldsymbol{\theta}_0$ is thus the minimizer of that divergence.

Moreover, arg min in the parameter sense, arg min in the function sense (in terms of a divergence metric on the *true* function), and arg min in the measure sense (in terms of a divergence metric on the *true* probability measure), are equivalent limits under the same consistency result. To see this, it is convenient to focus once more on the target and write $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} Q^{\mathcal{P}}_\infty \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} Q^F_\infty(f^{\mathbf{w}}, f_{\mathcal{W}}(\boldsymbol{\theta}))$, with $Q^F_\infty : F(\mathcal{W}) \times F(\mathcal{W}) \to \mathbb{R}_{\geq 0}$, to make clear that the criterion establishes a divergence d_F on $F(\mathcal{W}) \times F(\mathcal{W})$, which is, in turn, induced by $d_{\mathcal{P}}$ through $\mathfrak{P}$ according to $d_F(f^1, f^2) = d_{\mathcal{P}}(P(f^1), P(f^2))\ \forall\ (f^1, f^2) \in F(\mathcal{W}) \times F(\mathcal{W})$. This ensures that our statement on the probability measure is relevant under standard consistency results that are focused on the convergence of an estimated parameter vector toward $\boldsymbol{\theta}_0$, while, equivalently, the impulse response functions (IRFs) converge to the *true* IRFs at $\boldsymbol{\theta}_0$. This implies that deciding between Definitions 1–3 can be read from the responses produced by the IRF that minimizes divergence w.r.t. the *true* IRF

Not necessary, but convenient for a proof that holds easily in practical situations, is to assume the existence of a strictly increasing function $r : \mathbb{R} \to \mathbb{R}_{\geq 0}$ that ensures the existence of a transformation of the limit criterion into a metric, $d^*_{\mathcal{P}} \equiv r \circ d_{\mathcal{P}}$, with r being a continuously and strictly increasing function. For convenience, all assumptions are summarized in Assumption 1.

Assumption 1. *For a limit criterion $Q_\infty : \Theta \to \mathbb{R}$ of the form $Q_\infty(\boldsymbol{\theta}) \equiv Q^{\mathcal{P}}_\infty(P^{\mathbf{w}}, P^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}})\ \forall\ \boldsymbol{\theta} \in \Theta$, $d_{\mathcal{P}} \equiv Q^{\mathcal{P}}_\infty : \mathcal{P}^{\mathbf{w}} \times \mathcal{P}^{\mathbf{w}} \to \mathbb{R}_{\geq 0}$ is a divergence. Assume there exists a continuous strictly increasing function $r : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that $d^*_{\mathcal{P}} \equiv r \circ d_{\mathcal{P}}$ is a metric. The functional $f_{\mathcal{W}} : \Theta \to F_\Theta(\mathcal{W})$ is injective and $\boldsymbol{\theta}_0 \in \Theta$.*

Proposition 1. *Assume 1, then the following are equivalent limits:*

1. $\boldsymbol{\theta}_0$,
2. $\arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty(\boldsymbol{\theta})$,
3. $\arg\min_{\boldsymbol{\theta}\in\Theta} d^*_F(f^{\mathbf{w}}, f^{\hat{\mathbf{w}}}(\cdot, \boldsymbol{\theta}))$,
4. $\arg\min_{\boldsymbol{\theta}\in\Theta} Q^{\mathcal{P}}_\infty(P^{\mathbf{w}}, P^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}})$,
5. $\arg\min_{\boldsymbol{\theta}\in\Theta} d^*_{\mathcal{P}}(P^{\mathbf{w}}, P^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}})$.

Remark 2. *Dropping the axiom of correct specification implies $\hat{\boldsymbol{\theta}}_\infty \neq \boldsymbol{\theta}_0$, hence, the equivalences of 3–5 are now w.r.t. item 2.*

The equivalences in Proposition 1 not only ensure that for a correctly specified model $\exists \boldsymbol{\theta}_0 \in \Theta$, the element $\boldsymbol{\theta}_0$ results in functional equivalence between the model and the *true* model (item 3), but also in zero divergence between the probability measures $P^{\mathbf{w}}$ and $P^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}}$ (item 4). Moreover, it follows that at $\boldsymbol{\theta}_0$, the empirically estimated probability measure $P^{\hat{\mathbf{w}}}$ is equivalent to $P^{\mathbf{w}}$ in the sense that there is zero distance between the two (item 5).

Remark 3. *Proposition 1 is applicable to a large class of extremum estimators, even those not initially conceived as minimizers of distance. In particular it is often possible to find a divergence on the space of probability measures. For example, method of moments estimators are naturally defined in terms of features of the underlying probability measures. In Section 4 and example is given, using Kullback–Leibler divergence, for which penalized likelihood is an estimator. In this case squared Hellinger distance can be shown to be a lower bound.*

Corollary 1 now delivers that our definitions, set on the *true* measures, transfer to modeled probability measures in the limit for correctly specified cases. It is well-known that standard consistency proofs apply also to approximate extremum estimators, therefore, assuming additionally that $\sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\mathbf{w}_T;\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})| \to 0$ *a.s.*, is sufficient for a consistency result together with the uniqueness of $\boldsymbol{\theta}_0$ within the compact hypothesis space Θ (Note that, under the axiom of correct-specification, consistency results require suitable forms of stability defined on the process rather than the data. While we have loosely remarked on the fact that the non-parametric case of an infinite dimensional Θ is easily allowed, stability of highly nonlinear multivariate time series is a difficult separate topic. Regardless, Refs. [44,45] provide Ergodicity results for a large class of nonlinear time series that include non-parametric ones. The conditions require the nonlinearities to be sufficiently smooth. Specific stability results have also been established for certain neural network models, for example by [52]). This implies that our causality conditions on the *true* measures do not only transfer to the approximate in the limit, but also for large T under standard regularity conditions. Essentially, this is the setting considered by Ref. [11]. Summarized:

Corollary 1. *Given a* true *probability measure $P^{\mathbf{w}}$, and an equivalent modeled probability measure $P^{\hat{\mathbf{w}}}$ in the sense that $d^*_{P^{\hat{\mathbf{w}}}} = r \circ d_{\mathcal{P}}(P^{\mathbf{w}}, P^{\hat{\mathbf{w}}}_{\theta}) \sim 0$, there are four possibilities for causality:*

1. *There is no causation if $P^{\hat{\mathbf{x}}}_0$ and $P^{\hat{\mathbf{y}}}_0$ adhere to Definition 1.*
2. **x** *causes* **y** *if the probability measure $P^{\hat{\mathbf{y}}}_0$ adheres to Definition 2.*
3. **y** *causes* **x** *if the probability measure $P^{\hat{\mathbf{x}}}_0$ adheres to Definition 2.*
4. *There is bi-directional causality if $P^{\hat{\mathbf{x}}}_0$ and $P^{\hat{\mathbf{y}}}_0$ adhere to Definition 3.*

Finally, in the case of a miss-specified model, Proposition 2 implies that the divergence between the optimal probability measure as judged by the criterion and the *true* probability measure attains a minimum at a strictly positive value $d^*_{P^{\mathbf{w}}} > 0$. In this case, the quantity $d^*_{P^{\hat{\mathbf{w}}}}$ determines how "close" the empirical claim is to the *true* hypothesis about causality. While it is difficult to make claims about this quantity, it is evident that minimizing $d^*_{P^{\hat{\mathbf{w}}}}$ may involve widening $\mathcal{P}^{\hat{\mathbf{w}}}_{\Theta}$ in the direction of $P^{\mathbf{w}}$ by increasing the dimensionality of Θ and allow flexibility while investigating a wide range of data. Disregarding the value of $d^*_{P^{\hat{\mathbf{w}}}}$, the following holds.

Proposition 2. *If $\boldsymbol{\theta}_0 \notin \Theta$, then $P^{\mathbf{w}} \notin \mathcal{P}^{\hat{\mathbf{w}}}_{\Theta}$. However, $\hat{\boldsymbol{\theta}}_\infty$ is still the pseudo-*true *parameter that minimizes $r \circ d_{\mathcal{P}}(P^{\mathbf{w}}, P^{\hat{\mathbf{w}}}_{\theta})$ over Θ. Therefore, $P^{\hat{\mathbf{w}}}$ is the probability measure minimally divergent from $P^{\mathbf{w}}$ within $\mathcal{P}^{\hat{\mathbf{w}}}_{\Theta}$. As such, it follows that, from all the potential probability measures in $\mathcal{P}^{\hat{\mathbf{w}}}_{\Theta}$, the measure closest to $P^{\mathbf{w}}$ is supportive of one out of $1-4$ in corollary 1 based on the properties of $P^{\hat{\mathbf{x}}}_0$ and $P^{\hat{\mathbf{y}}}_0$ as the best approximations. $P^{\hat{\mathbf{w}}}$ provides the best approximation of the* true *causal measure across all the hypotheses considered.*

This leads to the following collection of results.

Corollary 2. *Given a* true *probability measure $P^{\mathbf{w}}$, and a non-equivalent, but pseudo-*true *modeled probability measure, $P^{\hat{\mathbf{w}}}$, in the sense that $d^*_{P^{\mathbf{w}}} = r \circ d_{\mathcal{P}}(P^{\mathbf{w}}, P^{\hat{\mathbf{w}}}_{\theta})$ has attained a non-zero minimum, there are four possible optimal hypotheses about causality, as judged by the criterion:*

1. *There is no causation if $P^{\hat{\mathbf{x}}}_0$ and $P^{\hat{\mathbf{y}}}_0$ adhere to Definition 1.*
2. **x** *causes* **y** *if the probability measure $P^{\hat{\mathbf{y}}}_0$ adheres to Definition 2.*
3. **y** *causes* **x** *if the probability measure $P^{\hat{\mathbf{x}}}_0$ adheres to Definition 2.*
4. *There is bi-directional causality if $P^{\hat{\mathbf{x}}}_0$ and $P^{\hat{\mathbf{y}}}_0$ adhere to Definition 3.*

Respectively, conditioning on interventions in **x**, the results can be understood as:

1. Whenever an intervention in **x** occurs, our best hypothesis is that there is no chance that **y** reacts as a result of that.

2. Whenever an intervention in $\mathbf{x}$ occurs, our best hypothesis is that there is positive chance that $\mathbf{y}$ reacts as a result of that.
3. Whenever an intervention in $\mathbf{x}$ occurs, our best hypothesis is that there is positive chance that $\mathbf{y}$ reacts as a result of that, and these interactions continue to repeat with positive probability.

4. Limit Squared Hellinger Distance

Both Corollaries 1 and 2 assume that an appropriate transformation of the limit criterion exists that provides us with a metric or norm. This assumption allows us to make use of the classical theorems on existence and uniqueness of best approximations that have been naturally obtained for metric, normed, and inner product spaces [53]. While this retains the simplicity of the argument, it also shows that a direct interpretation of Corollaries 1 and 2 can be obtained within the framework of maximum likelihood. Let us first define the criterion function as the maximum likelihood estimator:

$$\arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{w}_T;\boldsymbol{\theta}) := \arg\max_{\boldsymbol{\theta}\in\Theta} \sum_{t=1}^{T} \ln p_t(\mathbf{w}_t|\boldsymbol{\theta}). \tag{14}$$

Note that this is conforming to $Q_\infty(\boldsymbol{\theta}) := Q_T(f^{\mathbf{w}}(\mathbf{w}_T), f(\mathbf{w}_T; \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{w}_T;\boldsymbol{\theta})))$ with $T \to \infty$ and $Q_\infty : \Theta \to \mathbb{R}$. It can be shown that, under this definition with $Q_\infty(\boldsymbol{\theta}) = Q_\infty^{\mathcal{P}}(P^{\mathbf{w}}; P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})\ \forall\ \boldsymbol{\theta} \in \Theta$, the criterion $Q_\infty(\boldsymbol{\theta}) = Q_\infty^{\mathcal{P}}$ is a measure of divergence $d_{\mathcal{P}}$ on the *true* probability measure and the modeled measure. Specifically, we can introduce a divergence $d_{\mathcal{P}} \equiv Q_\infty^{\mathcal{P}} : \mathcal{P}^{\mathbf{w}} \times \mathcal{P}^{\mathbf{w}} \to \mathbb{R}_{\geq 0}$ as follows. Let $p^{\mathbf{w}}(\mathbf{w}_t|\boldsymbol{\theta}_{\mathbf{w}})$ and $p^{\hat{\mathbf{w}}}(\mathbf{w}_t|\boldsymbol{\theta}_{\hat{\mathbf{w}}})$ be, respectively, the *true* density evaluated under the *true* parameter and a modeled density at $\hat{\boldsymbol{\theta}}$, evaluated under the estimated parameter, both at time t, with respect to the Lebesque measure (such that they are probability density functions); then the following is a divergence from the true probability measure to the modeled probability measure (Kullback–Leibler divergence, see [54]):

$$\begin{aligned} &KL\big(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})||P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})\big) = \\ &\qquad \begin{cases} \int_{-\infty}^{\infty} p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) \ln \dfrac{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})}{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})} d\mathbf{w} & \forall\ p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) \ll p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}}) \\ \infty & \text{otherwise} \end{cases}. \end{aligned} \tag{15}$$

Naturally, $KL\big(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})||P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})\big) \geq 0$ with equality if and only if $p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) = p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})$ almost everywhere, i.e., when the probability measures are the same (this is known as Gibb's inequality and can be verified by applying Jensen's inequality).

Kullback–Leibler divergence is not a distance metric, as was used in Corollaries 1 and 2 to establish equivalences by partitioning into classes of zero-distance points. In particular, it is asymmetric

$$KL\big(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})||P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})\big) \neq KL\big(P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})||P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})\big), \tag{16}$$

and the triangle inequality is also not satisfied. However, it has the product–density property

$$KL(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})||P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})) = \sum_{t}^{T} \ln KL(p_t^{\mathbf{w}}(\mathbf{w}_t|\boldsymbol{\theta}_{\mathbf{w}})||p_t^{\hat{\mathbf{w}}}(\mathbf{w}_t|\boldsymbol{\theta}_{\hat{\mathbf{w}}})), \tag{17}$$

for $p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) = p_1^{\mathbf{w}}(\mathbf{w}_1|\boldsymbol{\theta}_{\mathbf{w}}) \cdot p_2^{\mathbf{w}}(\mathbf{w}_2|\boldsymbol{\theta}_{\mathbf{w}}) \dots p_T^{\mathbf{w}}(\mathbf{w}_T|\boldsymbol{\theta}_{\mathbf{w}})$, and $p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})$ defined similarly. Hence, the MLE is an unbiased estimator of minimized Kullback–Leibler divergence:

$$\begin{aligned} \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{w}_T;\boldsymbol{\theta}) &:= \arg\max_{\boldsymbol{\theta}\in\Theta} \sum_{t=1}^{T} \ln \frac{p^{\mathbf{w}}(\mathbf{w}_t|\boldsymbol{\theta}_{\mathbf{w}})}{p^{\hat{\mathbf{w}}}(\mathbf{w}_t|\boldsymbol{\theta}_{\hat{\mathbf{w}}})} \\ &= \arg\min_{\boldsymbol{\theta}\in\Theta} KL\big(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})||P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})\big). \end{aligned} \tag{18}$$

Note that under standard assumptions, a law of large numbers can be applied to obtain the convergence, hence, by maximizing log likelihood, we minimize Kullback–Leibler divergence. Now, we need to either find a continuously scaling function, r, to ensure that it also minimizes distance between the *true* measure and the modeled measure so that we may reach zero at $d^*_{P\hat{\mathbf{w}}} = r \circ d_{\mathcal{P}}(P^{\mathbf{w}}, P^{\hat{\mathbf{w}}}_{\theta}) \sim 0$. Alternatively, we find the distance metric directly. We argued above that Kullback–Leibler divergence is not a proper distance (in particular, it is not symmetric and does not satisfy the triangle inequality). However, notably useful is specifying $d^*_{P\hat{\mathbf{w}}}$ directly as the Hellinger distance between a modeled probability measure and the true probability measure [55]:

$$H\Big(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}), P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})\Big) = \sqrt{\frac{1}{2}\int\Big(\sqrt{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})} - \sqrt{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}\Big)^2 d\mathbf{w}}. \quad (19)$$

Specifically, the squared Hellinger distance provides a lower bound for the Kullback–Leibler divergence. Therefore, maximizing log likelihood implies minimizing Kullback–Leibler divergence, which implies minimizing the Hellinger distance. This is easily seen by the following:

Proposition 3. *The squared Hellinger distance provides a lower bound to Kullback–Leibler divergence:*

$$\Big(H\big(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})||P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})\big)\Big)^2 \leq KL\big(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})||P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})\big).$$

Remark 4 below highlights that these notions do not just apply to the standard real-valued time series settings considered by Granger, but can apply to the explicit probability modeling of binary outcomes as well. Remark 4 further clarifies a result that has so far only been presented implicitly—that the probabilistic truth identified at the discussed zero-distance point may allow for a base level of entropy to exist even when all functional relationships in the process have been accounted for in a model.

Remark 4. *While the paper has implicitly alluded to modeling continuous real-valued processes though the notational conventions, the connections between true probability and modeled probability are also easily made by focusing on an explicit binary outcome problem. Define cross-entropy for two discrete probability distributions p and q with the same support $\mathcal{X}$:*

$$H(p,q) = \mathbb{E}_p[-\ln q] = H(p) + \mathcal{D}_{KL}(p||q) = -\sum_{x\in\mathcal{X}} p(x)\ln q(x),$$

in which $\mathcal{D}_{KL}$ is Kullback–Leibler divergence, or the relative entropy of q with respect to p, and $H(p)$ is the entropy of p. Now if $p \in \{y, 1-y\}$ and $q \in \{\hat{y}, 1-\hat{y}\}$, we can rewrite cross-entropy:

$$H(p,q) = -\sum_{x\in\mathcal{X}} p_x \ln q_x = -y\ln\hat{y} - (1-y)\ln(1-\hat{y}),$$

or, for predictions generated under a set of parameters $\boldsymbol{\theta}$ and a predictor x, as

$$H(y,x;\boldsymbol{\theta}) = -\sum_{t=1}^{T} y_t \ln p_{\boldsymbol{\theta}}(y|x_{t-1}) - (1-y_t)\ln(1-p_{\boldsymbol{\theta}}(y|x_{t-1})).$$

Remember that the maximum likelihood estimator maximizes the likelihood of the data under some probabilistic model. The correct likelihood in the case of binary classification is Bernoulli:

$$p(y|\pi) = \Pi_{t=1}^{T}\pi_t^{y_t}(1-\pi_t)^{1-y_t},$$

which results in the likelihood function

$$p(y|x;\boldsymbol{\theta}) = \Pi_{t=1}^{T} p_{\boldsymbol{\theta}}(y|x_{t-1})^{y_t}(1-p_{\boldsymbol{\theta}}(y|x_{t-1}))^{1-y_t}.$$

Taking logs then gives the following log likelihood function

$$L(\boldsymbol{\theta}; x, y) = \sum_{t=1}^{T} y_t \ln p_{\boldsymbol{\theta}}(y|x_{t-1}) + (1 - y_t) \ln(1 - p_{\boldsymbol{\theta}}(y|x_{t-1})).$$

This shows that negative log likelihood is proportional to Kullback–Leibler divergence and differs by the basic entropy in the data, which is constant. Maximizing the likelihood of a binary model can, thus, be understood as minimizing statistical distance toward a true probability measure; the minimum value is determined by the entropy in the observed data.

5. Application

5.1. Practical Considerations

We continue this section first with some notes on practical considerations. Let $L_T(\boldsymbol{\theta})$ denote the sample log likelihood at $\boldsymbol{\theta} \in \Theta$. Naturally, if $\Theta_s \subset \Theta$, it follows that $\mathcal{P}_{\Theta}^{\hat{\mathbf{w}}} \supset \mathcal{P}_{\Theta_s}^{\hat{\mathbf{w}}}$. In the limit, this means that maximizing likelihood minimizes Hellinger distance over both $\mathcal{P}_{\Theta}^{\hat{\mathbf{w}}}$ and $\mathcal{P}_{\Theta_s}^{\hat{\mathbf{w}}}$. Following Corollary 1, if $\boldsymbol{\theta} \in \Theta_s$, this results in selecting $P^{\hat{\mathbf{w}}}$ at a point in $\mathcal{P}_{\Theta_s}^{\hat{\mathbf{w}}}$ that attains equivalence to $P^{\mathbf{w}}$. In practice, when finite data is used, two different points, one in $\mathcal{P}_{\Theta}^{\hat{\mathbf{w}}} \setminus \mathcal{P}_{\Theta_s}^{\hat{\mathbf{w}}}$ and one in $\mathcal{P}_{\Theta_s}^{\hat{\mathbf{w}}}$, may be obtained because the finite sample log likelihoods $L_T(\hat{\boldsymbol{\theta}}_{sT})$ and $L_T(\hat{\boldsymbol{\theta}}_T)$ that are available are both asymptotically biased estimators of the expected log likelihood $\mathbb{E}L_T(\boldsymbol{\theta}_0)$. This is easily shown by using a quadratic expansion [20,40]

$$\lim_{T\to\infty} \mathbb{E}\Big(L_T(\hat{\boldsymbol{\theta}}_T) - \mathbb{E}L_T(\boldsymbol{\theta}_0)\Big) = \lim_{T\to\infty} \mathbb{E}\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0)' \frac{1}{T} L_T''(\boldsymbol{\theta}_T) \sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \neq 0. \quad (20)$$

Under considerably restrictive conditions, the original work by [56,57] showed that the right hand-side approaches the dimension of $\hat{\boldsymbol{\theta}}_T$ and, hence, an asymptotically unbiased estimator of $\mathbb{E}\ell_t(\boldsymbol{\theta}_0)$ is given by $\frac{1}{T}\sum_{t=2}^{T} \ell_t(\hat{\boldsymbol{\theta}}_T) - k$. Akaike also proposed the well-known AIC given by AIC$= 2T\big(k - \frac{1}{T}\sum_{t=2}^{T} \ell_t(\hat{\boldsymbol{\theta}}_T)\big)$. Several authors have shown that the AIC can be used to consistently rank models according to Kullback–Leibler divergence in considerably more general settings, including the mis-specified case and have suggested further finite sample improvements [58–60]. The AIC is also valid to decide between economic theories for which no test statistics can be found [27]. This highlights that, while maximizing log likelihood over Θ is not the same objective as minimizing Kullback–Leibler divergence in finite samples, working with a complexity-penalized log likelihood (i.e., minimizing the AIC) does select the model that attains the lowest KL-bound of all considered models generated under Θ. Hence, in practice, a researcher can minimize the AIC as the practical objective to minimize Hellinger distance, and use specification tests to diagnose which of Corollaries 1 and 2 is more relevant. Since in-sample fits typically overfit data, a form of regularization would usually allow better out-of-sample results; see, for instance the (supplementary) discussion of [61] or the work of [62,63].

The challenge remains, however, that the AIC cannot be computed for all models as the degrees of freedom used in the correction is generally not a well-defined quantity for non-parametric models. As opposed to relying on in-sample corrections, cross-validation may instead be used to obtain unbiased estimates of $\mathbb{E}\ell_t(\boldsymbol{\theta}_0)$ in a setting that is more attuned to machine learning approaches, see for example [64]. Tests have been developed by [20,40,65] by following the general strategy of [66] adapted to the log likelihood case. The work has shown that choosing the model with the highest out-of-sample log likelihood equals choosing the model configuration that has achieves the highest probability of being the model that has lower Kullback–Leibler divergence. As the training T and validation data $\tilde{T}$ grows $T, \tilde{T} \to \infty$, this strategy chooses the model that has achieved the lowest Kullback–Leibler divergence, with probability converging to one.

5.2. Application to Treasury Yield Spreads and Bitcoin Spreads

The developed theory is now put into practice using daily data from short-term and long-term Treasury yield spreads and Bitcoin spreads. This is an interesting problem because each of these three assets has an important relation to inflation expectations. Rising inflation is also an acute problem, see [67,68].

The empirical strategy is as follows. First, standard linear Granger causality tests are performed as a benchmark. Next, non-parametric models will be fit in an effort to obtain an accurate-as-possible description of the true probability measure. The focus will be on maximizing out-of-sample log likelihood to minimize KL-divergence. Finally, Definitions 1 to 3 show that our conclusions about causality should be supported by a study of the probability measure that describes the causal effects. In particular, it must be decided whether this measure is a null-measure or produces real-valued data. This will be done by taking the best approximation of the true probability measure using the potential causal variable and the best approximation of the true probability measure without the potential causal variable, and (1) concluding whether the first achieves a lower KL-bound, and (2) testing whether the first is not stochastically equivalent to the latter. Section 5.2.1 first describes the data.

5.2.1. Data

Dynamic interactions between spreads in short-term and long-term bond yields can naturally be expected to occur in the data. In the absence of any credit risk, the net value of future bond payments is a function of the return required based on the inflation expectation used to discount the cash stream. Each of the Treasury securities typically caries a different yield, depending on maturity, the ratio between short and long-term treasury yields signals how investors feel about the economy in the short versus long term. If the yields vary substantially throughout the day, the market is uncertain about its expectations. Investigating the flow of causality between long-term and short-term yields and the interactions with other variables has been the objective of a large number of studies. To name a few, refs. [69,70] investigate causality between bonds and credit default swaps, while [71–75] investigate how financial distress propagates throughout connected bond markets.

Proponents of Bitcoin have argued that it is an important hedge due to its predetermined finite supply. While Bitcoin, as an asset class, has only recently attracted the public attention of large institutional investors, many researchers have already analyzed the time-series behavior of Bitcoin prices. An overview of recent developments and more discussion on forecasting Bitcoin prices is by [76]. They investigate a large set of covariates that cover nearly all important classes of financial assets, except bonds. They conclude that the intra-day distribution of daily returns follows a nonlinear memory process better captured by machine learning methods than conventional econometric models, which is further supported by a large body of literature that has documented related modeling exorcises [77–83].

If investors treat Bitcoin as an inflation hedge, then the spreads may causally interact with the U.S. yield spreads. Moreover, spreads in U.S. Treasury yields will arise predominantly from uncertainty in the expectations about the U.S. economy. Bitcoin, on the other hand, as a global asset that can be exchanged peer-to-peer by individuals without the need of a financial intermediary, might react to economic uncertainty in non-U.S. economies that may have the potential to spill over. Bitcoin also trades 24 h a day, every day of the year, and so may react to turmoil that happens outside U.S. trading hours and pass it on when the markets open. At the same time, Bitcoin is a relatively small market and the large institutional investors that dominate the bond market may not be active in the Bitcoin market. Causality from Bitcoin to the bond market could, then, be unlikely. Similarly, since Bitcoin trades non-stop, information assimilates rapidly, and so it may be likely that there is no causal influence of bond spreads at the daily time frame. The different hypotheses about the causal flows will be tested first using standard Granger causality tests.

5.2.2. Estimation Results

The following general system will be considered.

$$\begin{aligned} s(\mathbf{T}_t) &= f^1(L(\mathbf{T}_t, \mathbf{Q}_t, \mathbf{B}_t, \mathbf{S}_t)) \\ s(\mathbf{Q}_t) &= f^2(L(\mathbf{T}_t, \mathbf{Q}_t, \mathbf{B}_t, \mathbf{S}_t)) \\ s(\mathbf{B}_t) &= f^3(L(\mathbf{T}_t, \mathbf{Q}_t, \mathbf{B}_t, \mathbf{S}_t)) \end{aligned} \tag{21}$$

In which L is a lag operator, s is a function that calculates the spread between daily highs (h_t) and lows (l_t) as the log difference $o(\log(1+h_t) - \log(1+l_t))$ where 1 is added to account for negative rates. The function o is a simple outlier replacement function that replaces the largest observed spread (the Corona-crash) with the second largest value. The matrices $\mathbf{T}_t$, $\mathbf{Q}_t$, $\mathbf{B}_t$ are, respectively, the daily data of the ten-year bond, Quarterly bond, and Bitcoin price at time t, and $\mathbf{S}_t$ is SP500 price data used as a control. The data used in the analysis runs from 1 January 2017 to 20 December 2021 and were obtained from Yahoo finance using ticker symbols ^TNX, ^IRX and BTC-USD and ^GSPC.

First, a linear VAR model is considered with lags selected using the AIC. All of the maximums of 10 considered lags were selected, and stability was confirmed by verifying that the largest eigenvalue of the companion matrix remained below 1 (The largest eigenvalue was approximately 0.95, indicating that the process was stable but strongly dependent. Results were also generated using differenced data, which resulted in stronger causal linkages. Results are implemented in the code available with the paper but not shown here for compactness. see Supplementary Materials). Conditional Granger tests for causality are calculated by applying an F-test to the squared residuals of the model with and without the lags of a variable of interest in the presence of the autoregressive lags and the other control variables. The table below reports the p-values.

There are two important results in Table 1. First, the AIC, as an in-sample estimator of KL-divergence, selects a very large number of lags. The BIC is not an estimator of KL-divergence, see [84], but is a closely related Bayesian alternative to the AIC that is widely used. It places a larger penalty on the number of parameters and, as such, behaves somewhat similar to the corrected AIC in finite samples. The table shows that with this alternative criterion, a vastly different model is chosen. As Equation (20) showed, and the discussion after mentioned, the in-sample estimator of log likelihood is a biased estimator of expected log likelihood and, in practice, it is difficult to determine the appropriate penalty. In Table 1, two vastly different results are obtained. In both cases, however, the p-values of all causality tests are small. Both models suggest that there are strong causal linkages between spreads in all three markets. The statistical significance is somewhat dubious: the VAR(AIC) suggests that the causal flow of financial distress spills over in all directions. Moreover, Table 1 shows that, by adding more lags the significance of the causality tests increases, while it is likely that with 10 lags the model is trying to approximate a nonlinear process and the extremely high number of parameters involved in this approximation are likely over-fitting the data.

Table 1. p-values for Granger causality tests using VAR methods. Columns indicate the dependent variables, rows correspond to exogenous lags tested for causality. Each linkage is tested in the presence of lagged SP500 spreads as a control. Note that the BIC is not an estimator of KL-divergence, but it is widely used as a Bayesian alternative that places a higher penalty on dimensionality. Blank entries are intentionally left so, as they refer to endogenous linkages.

	AIC (lags = 10)			**BIC (lags = 3)**		
	$s(\mathbf{T}_t)$	$s(\mathbf{Q}_t)$	$s(\mathbf{B}_t)$	$s(\mathbf{T}_t)$	$s(\mathbf{Q}_t)$	$s(\mathbf{B}_t)$
$L(s(\mathbf{T}_t))$		0	0.0225		0	0.2207
$L(s(\mathbf{Q}_t))$	0		0.0021	0.0183		0.0450
$L(s(\mathbf{B}_t))$	0.0142	0.0083		0.1093	0.0635	

The section will now use an RF model to better approximate (f^1, f^2, f^3). The implementation used is that of [85], all possible tuning parameters are considered. The consistency of the RF in a time-series context under the assumption of data generated by a nonlinear autoregressive process is developed by [86]. As the previous sections detailed, the out-of-sample estimate of log likelihood is proportional to KL-divergence but RF models are typically not estimated using an in-sample log likelihood approach. A log likelihood function can nevertheless still be specified for out-of-sample predictions. To retain simplicity of the example, the commonly used Gaussian formulation is used:

$$\ell(v_t, \mu_t, \sigma_t) = \sum_t^T \frac{1}{2}(2\pi\sigma_t^2) - \frac{(v_t - \mu_t)^2}{2\sigma_t^2} \tag{22}$$

In this function, v_t are holdout validation samples at time t and μ_t is the mean parameter, which will be substituted by the conditional means predicted on the holdout data by the model. Note that σ_t, the variance parameter, is allowed to be time-varying. This is important because spread data is not homoskedastic, and the variance varies over the time dimension [87,88]. The log likelihood function thus allows for heteroskedasticity, the standard literature is followed and σ_t estimated using an ARMA-GARCH model. (The algorithm is as follows. Consider the time-varying density $F_t = (\mu_t, \sigma_t, \vartheta)$, where μ_t is a conditional mean process. For simplicity, it is defined as an ARMA $(1,1)$ process

$$\mu_t = c + \phi\mu_{t-1} + \theta\varepsilon_{t-1} + \varepsilon_t, \tag{23}$$

and the conditional variance, again for simplicity, is specified as a GARCH process of order $(1,1)$:

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{24}$$

with σ_t^2 as the conditional variance, ω an intercept, and L the back-shift operator. The vector ϑ specifies any remaining parameters of the distribution, in this case, the log likelihood is estimated using the Gaussian distribution in line with the validation criterion).

The RF models use three lags of the spread data so that the BIC-selected VAR model is nested. Several other features are added that may help describe the long-term dependencies captured by the AIC-selected model more accurately. In particular, a relative strength index (RSI) of all close values, including the SP500 close, is calculated. This is a standard indicator on $[0, 100]$, described in many resources that compare average upward movement to average downward movements over a look-back period. The standard period of 14 days is used along with a look-back of 14 weeks. The latter is also calculated using the spread data. This way, the model may learn different dependencies in periods of sustained decline, increase, or stability, in spreads and prices. The bootstrap sampling algorithm of the RF allows for case weights, effectively increasing the probability that highly weighted cases are over-represented in the random base learners, see [85]. This is exploited; σ_t^2 is standardized in the training data to be used as case-weights so that observations during more volatile periods feature more frequently in the sampling scheme.

The out-of-sample log likelihood is cross-validated using Equation (23), using 20 folds so that each validation sample has approximately 60 observations. The splits are generated using a stratified sampling approach that conditions on the RSI of the SP500. In other words, validation samples are chosen so that each validation sample equally represents days of under-bought, over-bought, and neutral stock market territories. The split is generated once and kept identical for each model so that the results can be directly compared. In total, an out-of-sample log likelihood value is generated for each observation so the sum of the log likelihood is taken to obtain an estimate of total out-of-sample log likelihood.

The results in Table 2 show the following. First, the nonlinear autoregressive models (indicated by the rows that apply a lag operator to the dependent variable listed in each column) all out-compete the VAR model that used all variables. According to the theory of the paper, the causal results obtained using the linear Granger causality tests in Table 1 should thus be discarded in favor of the theory that each variable follows a nonlinear autoregressive process that only makes possible reference to the SP500 but not the other variables of interest. For instance, the VAR of the ten-year Treasury yield spreads reach an out-of-sample log likelihood of 3916.77, while the nonlinear RF model reached a log likelihood of 3932.78 without using the lagged quarterly yield spreads or Bitcoin spreads. The differences in log likelihood are even larger for the models for quarterly Treasuries spreads and Bitcoin spreads.

Table 2 contains only evidence for two possible causal linkages. First, the model for the spreads on the ten-year that reached the lowest KL-bound used the lags of the quarterly yield data. This suggests that causality, in financial distress, may run from the short-term bonds to the long-term bonds. This is sensible; acute economic fears may impact short-term expectations more heavily, and the reaction in the short-term yields may trigger further fears about longer-term economic expectations. The second causal link could run from the Bitcoin market to the quarterly bonds. This is not far-fetched: Bitcoin trades non-stop and so any event globally can impact the Bitcoin market immediately, whereupon the increased fear in the Bitcoin market could then trigger further reactions in the short-term bond market, which would be more susceptible to short-term economic fears. However, the point increase in log likelihood that backs this hypothesis is small compared to the model that only used endogenous lags and control data.

Table 2. Cross-validated log likelihood for different models. Columns indicate the dependent variables, rows correspond to exogenous lagged data that are used by the models in addition to the control data. For each dependent variable, the model that achieved the lowest KL-divergence is marked by *.

	$s(\mathbf{T}_t)$	$s(\mathbf{Q}_t)$	$s(\mathbf{B}_t)$
VAR	3916.77	4084.68	2204.91
RF			
All	3989.14	4239.42	2251.06
$L(s(\mathbf{T}_t))$	3932.78	4230.44	2251.68
$L(s(\mathbf{Q}_t))$	3991.24 *	4240.54	2251.46
$L(s(\mathbf{B}_t))$	3932.90	4242.67 *	2251.84 *

Recall Remark 1: to test whether the evidence for causality is strong enough; it is important to test whether the probability measures that achieved the lowest KL-bound are stochastically different from those that exclude the causal linkages. A Kolmogorov–Smirnov test, under the null of distributional equivalence against a two-sided alternative, is computed. For the ten-year yield spread model, the p-value is 0, so the null is overwhelmingly rejected. The analysis, thus, concludes that the best possible hypothesis is that disruptions in the short-term bond market cause further disruption in the longer-term bond market. The test for distributional equivalence between the model with and without Bitcoin data has a p-value of 0.8591. In other words, the null of equivalence cannot be rejected and, while the model that used Bitcoin data reached the lowest KL-bound, the analysis does not find significant evidence for a causal flow from the Bitcoin market to the short-term Treasuries as the modeled probability measure is not significantly distinguishable from the competing non-causal measure. This suggests that the probability measure that describes the causal effects in Definition 2 is not distinguishable from that of Definition 1, and so Corollary 1 or 2 remain inconclusive. The final conclusion that causal flows are thus parsimonious is far more likely than the result obtained with the VAR, which suggested that causality flows significantly in all directions.

6. Concluding Remarks

This paper has developed a probabilistic theory of causation using measure-theoretical concepts. It discussed how probabilistic truths can be approximated by minimizing distance to the true probability measure over a space of measures in which each element is associated with a probabilistic theory about causation. This notion is flexible and has allowed for a wide range of models to be used for causal inference, including linear and nonlinear dynamical models. The theory has been applied using daily data on yield spreads to test how uncertainty around short-term and long-term expectations about future inflation interact with uncertainty in the daily Bitcoin price. The results were contrasted with those obtained using standard linear Granger causality tests. While linear Granger causality relies on models that assume a constant causal influence from one variable onto another, specified by static parameters, the analysis has shown that time-varying properties of the auto-regressive process provides a better description of the data. While the linear Granger causality tests finds significant causal influence in all directions, the suggested measure-theoretic approach to causality testing, using, in this example, a random forest model, found only one significant causal link that ran from financial distress in the short-term bond market to uncertainty in the long-term bond market.

As with Granger's approach, a convincing theory of how causes produce effect is not necessarily a prerequisite to making correct causal inferences. Clear hypotheses about causal relations may, however, help guide the inference by helping design better models. However, whereas Granger's definition "is based entirely on the predictability of some series" [5], the ideas of the current paper start with the notion that true probabilistic laws exist and can, and should, correctly be approximated to infer causal structures from data. A conclusion from this is that researchers interested in causal analysis should aim to develop strong out-of-sample predictions, as Granger's techniques applied to inaccurate models may provide an overly enthusiastic description of causal linkages.

The general ideas of the paper differ from the linear Granger tests in terms of result, but share a similarity in thought process. Granger's statement about causality followed from the premises that causes occur before effects and that causes contain unique information about their effect, and so that any causal variable must help forecast outcomes after other variables have been used first. For this reason, many refer to Granger causality as predictability. This paper defined causality directly in terms of the probability measures that define a stochastic process. This, in turn, places the emphasis on finding the best approximation of that probability measure. The theory developed here shows that minimizing KL-divergence implies minimizing distance between a model and the true probability measure and shows that maximizing out-of-sample log likelihood implies minimizing KL-divergence. This does not require parametric models or the degrees of freedom to be known. Instead, the KL-ranking of competing models can be directly read from the out-of-sample log likelihood. The stochastic equivalence, or difference, between probability measures that are induced by causal flows, or from autoregressive properties only, can subsequently be tested. The theory provides practitioners guidance for developing causal models using new machine learning methods that have, so far, remained relatively underutilized in this context.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/e24010092/s1.

Author Contributions: The work, and any possible error contained therein, is solely my own. Useful suggestions from two anonymous referees have been integrated into the paper. All authors have read and agreed to the published version of the manuscript.

Funding: The research did not receive any external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Code to reproduce the analysis is available along with the paper or can alternatively be requested from the author. The data is taken from Yahoo, the code calls the data from their API.

Acknowledgments: An earlier version containing sections of this work was shared for discussion [89], parts of which have been improved and replicated here for the special issue on "Causal Inference for Heterogeneous Data and Information Theory". I thank the two anonymous referees for their instructive comments and for help correcting and improving parts of the work. Any remaining errors are my own.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DGP	data-generating process
MLE	maximum likelihood estimator
IRF	impulse response function
VAR	vector auto-regression
RF	random forest
RSI	relative strength index

Appendix A. Proofs

Appendix A.1. Proof for Proposition 1

Proof. By construction of the criterion, as stated in Assumption 1, $\arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty(\boldsymbol{\theta})$ is its minimizer, and, by assuming $\boldsymbol{\theta}_0 \in \Theta$, it is also equal to $\boldsymbol{\theta}_0$. Hence, item 2 is equivalent to item 1 by definition under correct specification.

The equivalence of the deterministic limit criterion (item 2) as a function describing the divergence of the underlying probability measures of $\mathbf{w}$ and $\hat{\mathbf{w}}$ (item 4) is assumed, however, given a limit criterion function $Q_\infty : \Theta \to \mathbb{R}$ and a flexible definition of divergence (e.g., a pre-metric, such as the KL-divergence), it is often possible to find a divergence $d_{\mathcal{P}} : \mathcal{P}_\Theta \times \mathcal{P}_\Theta \to \mathbb{R}_{\geq 0}$ on the space of probability measures satisfying $\arg\min_{\boldsymbol{\theta}\in\Theta} d_{\mathcal{P}}(\mathcal{P}^{\mathbf{w}}, \mathcal{P}^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}}) = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty(\boldsymbol{\theta})$. The KL-divergence example is provided in this paper in the context of the maximum likelihood criterion.

By the assumption that r exists, the deterministic limit criterion that minimizes divergence, is also the minimizer of a distance metric $d^*_{\mathcal{P}}(P^{\mathbf{w}}, P^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}})$, hence item 4 is also equivalent to item 2.

Finally, since $f_{\mathcal{W}} : \Theta \to F_\Theta(\mathcal{W})$ is injective, $(P^{\mathbf{w}}, P^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}}) \equiv d^*_F(f^{\mathbf{w}}, f(\cdot, \boldsymbol{\theta}))\ \forall\ \boldsymbol{\theta} \in \Theta$ and d^*_F is a metric on $F_\Theta(\mathcal{W})$, $\boldsymbol{\theta}_0$ is also the minimizer of $d^*_F(f^{\mathbf{w}}, f(\cdot, \boldsymbol{\theta}))\ \forall\ \boldsymbol{\theta} \in \Theta$ so that item 3 is equivalent to item 2.

□

Appendix A.2. Proof for Proposition 2

Proof. The result follows immediately by the arguments used in proposition 1 dropping only the first equivalence. □

Appendix A.3. Proof for Proposition 3

Proof. First, Hellinger distance is

$$H\big(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}), P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})\big) = \sqrt{\frac{1}{2}\int\left(\sqrt{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})} - \sqrt{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}\right)^2 d\mathbf{w}},$$

hence,

$$\big(H(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}), P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}}))\big)^2 = \frac{1}{2}\int\left(\sqrt{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})} - \sqrt{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}\right)^2 d\mathbf{w}.$$

Now, the R.H.S. can be written as

$$\frac{1}{2}\int p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})d\mathbf{w} + \frac{1}{2}\int p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})d\mathbf{w} - \int\sqrt{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}d\mathbf{w}.$$

The integral of a probability density over its domain equals 1, hence the sum of the first two terms is 1, hence this can be rewritten as

$$1 - \int\sqrt{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})d\mathbf{w}}.$$

This has an upper bound, provided by the inequality

$$1-\int\sqrt{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}d\mathbf{w}\leq-\ln\int\sqrt{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}d\mathbf{w}.$$

Write R.H.S. as $-\ln\int\left[\sqrt{\frac{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})}}p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})\right]d\mathbf{w}$ and to obtain the upper bound

$$-\ln\int\left[\sqrt{\frac{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})}}p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})\right]d\mathbf{w}\leq-\int\left[\ln\sqrt{\frac{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})}}p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})\right]d\mathbf{w},$$

by applying Jensen's inequality, which can be applied to the integral case, since any random variable whose distribution admits a probability density function has the expected value represented by the integral over the full range of the density.

Finally, define the R.H.S. as

$$E\int\left[\ln\frac{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})}{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})\right]d\mathbf{w}=-\int\left[\ln\sqrt{\frac{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})}}p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})\right]d\mathbf{w},$$

and conclude that the last expression is equivalent to the Kullback–Leibler divergence by an elementary row operation.

$$E\int\left[\ln\frac{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})}{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})\right]d\mathbf{w}\equiv KL\big(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})||P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})\big).$$

□

References

1. Sundholm, G. A century of judgement and inference,1837–1936: Some strands in the development of logic. In *The Development of Modern Logic*; Oxford University Press: New York, NY, USA, 2009. [CrossRef]
2. Sundholm, G. "Inference versus consequence" revisited: Inference, consequence, conditional, implication. *Synthese* **2012**, *187*, 943–956. [CrossRef]
3. Pearl, J. *Causality: Models, Reasoning, and Inference*; Cambridge University Press: Cambridge, UK, 2000; p. 384.
4. Neuberg, L.G. Causality: Models, Reasoning, and Inference, by Judea Pearl, Cambridge University Press, 2000. *Econom. Theory* **2003**, *19*, 675–685. [CrossRef]
5. Granger, C.W.J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **1969**, *37*, 424. [CrossRef]
6. Granger, C.W. Testing for causality: A personal viewpoint. *J. Econ. Dyn. Control* **1980**, *2*, 329–352. [CrossRef]
7. White, H.; Chalak, K. Settable Systems: An Extension of Pearl's Causal Model with Optimization, Equilibrium, and Learning. *J. Mach. Learn. Res.* **2009**, *10*, 1759–1799.
8. White, H.; Lu, X. Granger Causality and Dynamic Structural Systems. *J. Financ. Econom.* **2010**, *8*, 193–243. [CrossRef]
9. White, H.; Chalak, K.; Lu, X. Causality in Time Series Linking Granger Causality and the Pearl Causal Model with Settable Systems. *JMRL Workshop Conf. Proc.* **2011**, *12*, 1–29.
10. White, H.; Xu, H.; Chalak, K. Causal discourse in a game of incomplete information. *J. Econom.* **2014**, *182*, 45–58. [CrossRef]
11. White, H.; Pettenuzzo, D. Granger causality, exogeneity, cointegration, and economic policy analysis. *J. Econom.* **2014**, *178*, 316–330. [CrossRef]
12. Williamson, J. Probabilistic theories of causality. In *The Oxford Handbook of Causation*; Chapter Probabilistic Theories; Beebee, H., Menzies, P., Hitchcock, C., Eds.; Oxford University Press: Oxford, UK, 2009; pp. 185–212.
13. Bohm, D. *Quantum Theory*; Dover Publications, Inc.: New York, NY, USA, 1951; p. 646.
14. Bohm, D. *Causality and Chance in Modern Physics*; University of Pennslyvania Press: Philadelphia, PA, USA, 1999; p. 170.
15. Rubin, D.B. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **1974**, *66*, 688–701. [CrossRef]
16. Heckman, J.J. Econometric Causality. *Int. Stat. Rev.* **2008**, *76*, 1–27. [CrossRef]
17. Heckman, J.J.; Vytlacil, E. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* **2005**, *73*, 669–738. [CrossRef]

18. Mogstad, M.; Santos, A.; Torgovitsky, A. Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters. *Econometrica* **2018**, *86*, 1589–1619. [CrossRef]
19. Parbhoo, S.; Wieser, M.; Wieczorek, A.; Roth, V. Information Bottleneck for Estimating Treatment Effects with Systematically Missing Covariates. *Entropy* **2020**, *22*, 389. [CrossRef]
20. Andrée, B.P.J. *Theory and Application of Dynamic Spatial Time Series Models*; Rozenberg Publishers and Tinbergen Institute: Amsterdam, The Netherlands, 2020; pp. 1–374.
21. White, H. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* **1980**, *48*, 817. [CrossRef]
22. White, H. Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **1982**, *50*, 1–25. [CrossRef]
23. Domowitz, I.; White, H. Misspecified models with dependent observations. *J. Econom.* **1982**, *20*, 35–58. [CrossRef]
24. Pötscher, B.M.; Prucha, I.R. *Dynamic Nonlinear Econometric Models*; Springer: Berlin/Heidelberg, Germany, 1997. [CrossRef]
25. Driscoll, J.C.; Kraay, A.C. Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data. *Rev. Econ. Stat.* **1998**, *80*, 549–560. [CrossRef]
26. Freedman, D.A. On the So-Called "Huber Sandwich Estimator" and "Robust Standard Errors". *Am. Stat.* **2006**, *60*, 299–302. [CrossRef]
27. Granger, C.; King, M.L.; White, H. Comments on testing economic theories and the use of model selection criteria. *J. Econom.* **1995**, *67*, 173–187. [CrossRef]
28. Hlaváčková-Schindler, K.; Paluš, M.; Vejmelka, M.; Bhattacharya, J. Causality detection based on information-theoretic approaches in time series analysis. *Phys. Rep.* **2007**, *441*, 1–46. [CrossRef]
29. Hlaváčková-Schindler, K. Equivalence of Granger Causality and Transfer Entropy: A Generalization. *Appl. Math. Sci.* **2011**, *5*, 3637–3648.
30. Hlaváčková-Schindler, K.; Plant, C. Heterogeneous Graphical Granger Causality by Minimum Message Length. *Entropy* **2020**, *22*, 1400. [CrossRef]
31. Haavelmo, T. The Statistical Implications of a System of Simultaneous Equations. *Econometrica* **1943**, *11*, 1–12. [CrossRef]
32. Haavelmo, T. The Probability Approach in Econometrics. *Econometrica* **1944**, *12*, 115. [CrossRef]
33. Kalman, R. Identifiability and Modeling in Econometrics. *Dev. Stat.* **1983**, *4*, 97–136. [CrossRef]
34. Schervish, M.J. *Theory of Statistics*; Springer Series in Statistics; Springer: New York, NY, USA, 1995. [CrossRef]
35. Billingsley, P. *Probability and Measure*, 3rd ed.; Wiley Series in Probability and Mathematical Statistics; Wiley-Interscience: New York, NY, USA, 1995.
36. Tong, H. *Threshold Models in Non-Linear Time Series Analysis*; Lecture Notes in Statistics; Springer: New York, NY, USA, 1983; p. 323. [CrossRef]
37. Dijk, D.; Teräsvirta, T.; Franses, P. Smooth transition autoregressive models—A survey of recent developments. *Econom. Rev.* **2002**, *21*, 37–41. [CrossRef]
38. Creal, D.; Koopman, S.J.; Lucas, A. *A General Framework for Observation Driven Time-Varying Parameter Models*; Global COE Hi-Stat Discussion Paper Series; Institute of Economic Research Hitotsubashi University: Tokyo, Japan, 2009.
39. Jan Koopman, S.; Lucas, A.; Schart, M. Predicting time-varying parameters with parameter-driven and observation-driven models. *Rev. Econ. Stat.* **2016**, *98*, 97–110. [CrossRef]
40. Andrée, B.P.J.; Blasques, F.; Koomen, E. *Smooth Transition Spatial Autoregressive Models*; Tinbergen Institute Discussion Paper; Tinbergen Institute: Amsterdam, The Netherlands, 2017. [CrossRef]
41. Blasques, F.; Koopman, S.J.; Lucas, A.; Schaumburg, J. Spillover dynamics for systemic risk measurement using spatial financial time series models. *J. Econom.* **2016**, *195*, 211–223. [CrossRef]
42. Andrée, B.P.J.; Kraay, A.; Chamorro, A.; Spencer, P.; Wang, D. *Predicting Food Crises*; World Bank Policy Research Working Papers; World Bank: Washington, DC, USA, 2020. [CrossRef]
43. Straumann, D.; Mikosch, T. Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. *Ann. Stat.* **2006**, *34*, 2449–2495. [CrossRef]
44. Cline, D.B.H.; Pu, H.M.H. Verifying irreducibility and continuity of a nonlinear time series. *Stat. Probab. Lett.* **1998**, *40*, 139–148. [CrossRef]
45. Cline, D.B.H.; Pu, H.M.H. Geometric Ergodicity of Nonlinear Time Series. *Stat. Sin.* **1999**, *9*, 1103–1118.
46. Amador, L.D.R.; Lovejoy, S. Long-Range Forecasting as a Past Value Problem: Untangling Correlations and Causality with Scaling. *Geophys. Res. Lett.* **2021**, *48*, e2020GL092147. [CrossRef]
47. Dudley, R.M. *Real Analysis and Probability*; Cambridge University Press: Cambridge, UK, 2002; p. 555.
48. Davidson, J. *Stochastic Limit Theory*; Oxford University Press: Oxford, UK, 1994. [CrossRef]
49. Hendry, D.F. Granger Causality. *Eur. J. Pure Appl. Math.* **2017**, *10*, 12–29.
50. Geman, S.; Hwang, C.R. Nonparametric Maximum Likelihood Estimation by the Method of Sieves. *Ann. Stat.* **1982**, *10*, 401–414. [CrossRef]
51. Kolmogorov, A.N.; Fomin, S.V. *Introductory Real Analysis*; Dover Publications: New York, NY, USA, 1975; p. 403.
52. Leisch, F.; Trapletti, A.; Hornik, K. Stationarity and Stability of Autoregressive Neural Network Processes. *Neural Comput.* **2000**, *12*, 2427–2450.
53. Cheney, E.; Respess, J. Best Approximation Problems in Tensor-Product Spaces. *Pac. J. Math.* **1982**, *102*, 437–446.

54. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
55. Hellinger, E. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *J. Reine Angew. Math.* **1909**, *1909*, 210–271. [CrossRef]
56. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. In *Information Theory: Proceedings of the Second International Symposium*; Petrov, B.N., Csaki, F., Eds.; Akadémiai Kiado: Budapest, Hungary, 1973; pp. 267–281.
57. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]
58. Hurvich, C.M.; Tsai, C.L. Regression and time series model selection in small samples. *Biometrika* **1989**, *76*, 297–307. [CrossRef]
59. Hurvich, C.M.; Tsai, C.L. Bias of the corrected AIC criterion for underfitted regression and time series models. *Biomelrika* **1991**, *78*, 499–509. [CrossRef]
60. Sin, C.Y.; White, H. Information criteria for selecting possibly misspecified parametric models. *J. Econom.* **1996**, *71*, 207–225. [CrossRef]
61. Andrée, B.P.J.; Chamorro, A.; Spencer, P.; Koomen, E.; Dogo, H. Revisiting the relation between economic growth and the environment; a global assessment of deforestation, pollution and carbon emission. *Renew. Sustain. Energy Rev.* **2019**, *114*, 109221. [CrossRef]
62. Zou, H. The Adaptive Lasso and Its Oracle Properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
63. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef]
64. Bergmeir, C.; Hyndman, R.J.; Koo, B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Comput. Stat. Data Anal.* **2018**, *120*, 70–83. [CrossRef]
65. Diks, C.; Panchenko, V.; van Dijk, D. Likelihood-based scoring rules for comparing density forecasts in tails. *J. Econom.* **2011**, *163*, 215–230. [CrossRef]
66. Diebold, F.X.; Mariano, R.S. Comparing Predictive Accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263. [CrossRef]
67. Andrée, B.P.J. *Estimating Food Price Inflation from Partial Surveys*; Policy Research Working Paper; World Bank: Washington, DC, USA, 2021; Volume 9886. [CrossRef]
68. Andrée, B.P.J. Monthly food price estimates by product and market. In *WLD_2021_RTFP_v02_M*; Version 2021-12-02; World Bank Microdata Library: Washington, DC, USA, 2021. [CrossRef]
69. Blanco, R.; Brennan, S.; Marsh, I.W. An empirical analysis of the dynamic relation between investment-grade bonds and credit default swaps. *J. Financ.* **2005**, *60*, 2255–2281. [CrossRef]
70. Delis, M.D.; Mylonidis, N. The chicken or the egg? A note on the dynamic interrelation between government bond spreads and credit default swaps. *Financ. Res. Lett.* **2011**, *8*, 163–170. [CrossRef]
71. Matei, I. Contagion and causality: An empirical analysis on sovereign bond spreads. *Econ. Bull.* **2003**, *30*, 1885–1896.
72. Gómez-Puig, M.; Sosvilla-Rivero, S. Granger-causality in peripheral EMU public debt markets: A dynamic approach. *J. Bank. Financ.* **2013**, *37*, 4627–4649. [CrossRef]
73. Gómez-Puig, M.; Sosvilla-Rivero, S. Causality and contagion in EMU sovereign debt markets. *Int. Rev. Econ. Financ.* **2014**, *33*, 12–27. [CrossRef]
74. Corsi, F.; Lillo, F.; Pirino, D.; Trapin, L. Measuring the propagation of financial distress with Granger-causality tail risk networks. *J. Financ. Stab.* **2018**, *38*, 18–36. [CrossRef]
75. Balcilar, M.; Usman, O.; Gungor, H.; Roubaud, D.; Wohar, M.E. Role of global, regional, and advanced market economic policy uncertainty on bond spreads in emerging markets. *Econ. Model.* **2021**, *102*, 105576. [CrossRef]
76. Chevallier, J.; Guégan, D.; Goutte, S. Is It Possible to Forecast the Price of Bitcoin? *Forecasting* **2021**, *3*, 377–420. [CrossRef]
77. Lee, K.; Ulkuatam, S.; Beling, P.; Scherer, W. Generating Synthetic Bitcoin Transactions and Predicting Market Price Movement Via Inverse Reinforcement Learning and Agent-Based Modeling. *J. Artif. Soc. Soc. Simul.* **2018**, *21*, 5. [CrossRef]
78. Pele, D.T.; Mazurencu-Marinescu-Pele, M. Using High-Frequency Entropy to Forecast Bitcoin's Daily Value at Risk. *Entropy* **2019**, *21*, 102. [CrossRef]
79. Cohen, G. Forecasting Bitcoin Trends Using Algorithmic Learning Systems. *Entropy* **2020**, *22*, 838. [CrossRef] [PubMed]
80. Kim, Y.B.; Kim, J.G.; Kim, W.; Im, J.H.; Kim, T.H.; Kang, S.J.; Kim, C.H. Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PLoS ONE* **2016**, *11*, e0161197. [CrossRef]
81. Valencia, F.; Gómez-Espinosa, A.; Valdés-Aguirre, B. Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning. *Entropy* **2019**, *21*, 589. [CrossRef]
82. Lahmiri, S.; Bekiros, S. Randomness, Informational Entropy, and Volatility Interdependencies among the Major World Markets: The Role of the COVID-19 Pandemic. *Entropy* **2020**, *22*, 833. [CrossRef]
83. García-Medina, A.; Luu, T.; Huynh, D.; Schinckus, C.; Stanley, H.E. What Drives Bitcoin? An Approach from Continuous Local Transfer Entropy and Deep Learning Classification Models. *Entropy* **2021**, *23*, 1582. [CrossRef]
84. Burnham, K.P.; Anderson, D.R. Multimodel inference: Understanding AIC and BIC in Model Selection. *Sociol. Methods Res.* **2004**, *33*, 261–304. [CrossRef]
85. Wright, M.N.; Ziegler, A. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **2017**, *77*, 1–17. [CrossRef]
86. Davis, R.A.; Nielsen, M.S. Modeling of time series using random forests: Theoretical developments. *Electron. J. Stat.* **2020**, *14*, 3644–3671. [CrossRef]

87. Clark, E.; Baccar, S. Modelling credit spreads with time volatility, skewness, and kurtosis. *Ann. Oper. Res.* **2018**, *262*, 431–461. [CrossRef]
88. Kim, J.M.; Kim, D.H.; Jung, H. Estimating yield spreads volatility using GARCH-type models. *N. Am. J. Econ. Financ.* **2021**, *57*, 101396. [CrossRef]
89. Andrée, B.P.J. Probability, Causality and Stochastic Formulations of Economic Theory. 2019. Available online: https://ssrn.com/abstract=3422430 (accessed on 21 September 2021). [CrossRef]