## Theory and Application of Dynamic Spatial Time Series Models

Andree, B.P.J.

2020

***document version***
Publisher's PDF, also known as Version of record

# Chapter 7

# Probability and Causality in Spatial Time Series


### Chapter Summary

The current paper discusses approximating a correct theory of cause and effect by minimizing distance to its associated probability measure in a space of measures in which each element is associated with a stochastic representation of a candidate theory. The discussion encourages researchers to use flexible dynamical models to model and discover the *true* quantitative relationships that may be hidden in interrelated stochastic data. The argument is based on the use of a decision criterion that scales to a metric that measures distance between any given measure. When this is the case, a metric space can be considered in which equivalences can be established by partitioning into classes of zero-distance points. Equivalence to the *true* measure, that is associated with the *true* frequencies in Markov chains of iterated causes and effects, is established by reaching zero distance in that space. When the hypothesis space is incorrectly constructed, equivalence is established with respect to a pseudo-*true* measure that by definition is closest to the correct hypothesis across all considered hypotheses. The specific case of Maximum Likelihood is further discussed. In particular, squared Hellinger distance marks a lower bound of Kullback-Leibler divergence. This implies that maximizing complexity penalized likelihood minimizes distance toward the *true* probability measure. As such, it is an objective that approximates the correct causal structure from interrelated stochastic data that are observed and modeled sequentially over time.[1]


[1]This chapter is based on *"Probability, Causality and Stochastic Formulations of Economic Theory"*, available on the Social Science Research Network. The reference is (Andree, 2019).

## 7.1 Introduction

The 20th century has seen much work done on establishing the statistical properties of estimators widely used in econometrics. Notably, Kolmogorov (1933) laid out the axiomatic foundations of modern probability theory and, one year later, Doob (1934) proved the law of large numbers using a probabilistic interpretation of Birkhoff's ergodic theorem. Doob then used this to prove theorems of Fisher (1922, 1925) and Hotelling (1930) on estimating a parameter of a distribution by method of Maximum Likelihood, establishing both the Consistency and Normality of the MLE. Wald (1949) provided a proof for the multi-parameter case with greater generality. Earlier assertions on the efficiency of the MLE by Fischer, and importantly Cramer (1946), were eventually substantiated by rigorous proof by Rao (1962) resulting in what is now known as the Cramer-Rao bound. Generalizations that cover the nonlinear case were developed, initially with difficult to verify conditions (Le Cam, 1953) and (Kraft, 1955), but later for the general case of stationary Markov processes (Roussas, 1965) which in fact was a result that extended the theory by Wald (1949). It took several decades, but eventually the nonlinear Least Squares case was tackled (Jennrich, 1969; Malinvaud, 1970) which set the basis to a general asymptotic theory of extremum estimators. In the decades that followed, asymptotic properties of extremum estimators have covered multivariate dynamic settings, miss-specified models, heterogeneity, and dependence of the data. A good modern review is Pötscher and Prucha (1997).

While the important early statisticians Pearson and Fischer were primarily biometricians, their statistical methods for data analysis were eagerly integrated into economics. Wald, who played a crucial role in developing the Consistency and Normality results, spent much of his time with econometricians and he produced economic theories of his own. Arguably, the most notable contribution in integrating probabil-

ity into econometrics is, however, by Haavelmo (1944) *"The Probability Approach in Econometrics"*. In a less well-known paper that was published one year before, Haavelmo (1943) already provided the basis for stochastic formulation of economic theories and the integration of error terms into regressions. Good historical accounts on how Haavelmo's work shaped modern econometrics are by Spanos (1989) and Bjerkholt (2007), and a more general reconstruction of the interaction between early econometricians and statisticians is provided by Aldrich (2010).

Though the probabilistic view laid out by Haavelmo to model economic theories has largely been embraced by many applied economists, it seems that more mechanic definitions of causality are largely preferred over probabilistic ones. Particularly Rubins' viewpoint (Rubin, 1974), which originated from studies on human psychology, has been widely embraced as a model for causality. The core idea behind the Rubins' approach to identification is that treatment groups of populations with otherwise equal properties can be used to isolate a treatment effect, as no other factor can otherwise be attributed to account for the differential in an observed outcome. This view on causality implies that treatments must, in a deterministic manner, cause outcomes to occur. In fact, Pearl (2000) states that cause and effect relations are fundamentally deterministic, explicitly excluding quantum mechanical phenomena from his concept of cause and effect but mentioning that causal analysis involves probability language (see also the review by Neuberg et al. (2003)). The probabilistic approaches to causality, such as laid out by Granger (1969, 1980); Covey and Bessler (1992) that involve contrasting the probabilistic forecasting performance of a univariate and bivariate specification, are done away by Pearl (2000). In particular, Pearl (2000) makes explicit mention that this is not causality, and that the concepts of "strong exogeneity" (Engle et al., 1983) and Granger-causality are only statistical concepts. His view on causality is purely mechanical. At the same time, there are many examples in physics, the study that was born out of classical mechanics,

that approach causality from the probabilistic angle. On one hand this may relate to the fact that branches in both physics and economics evolve around models of dynamical systems in which the control-theoretic concept of a non-anticipative system is the basis for causal relationships, see for example Liang (2016); Harnack et al. (2017); Krakovska et al. (2018) for examples of recent causal studies in physics that work around estimating the time dependencies in dynamical systems. On the other hand, it may be related to developments in quantum mechanics that suggest that reality itself is probabilistic in nature, a profound notion that arguably still has not been fully clarified since the Bohr-Einstein debates. This would, at some level of abstraction, turn the universe non-causal under Pearl's view, which may be a philosophically difficult proposition. For example, in *"Causality and Chance in Modern Physics"*, Bohm (1999), one year before Pearl, argues that any theory about reality that embraces either one of causality and chance, to the exclusion of the other, is inherently incomplete.

The conflicting views might seem an inconsistency, and the mechanical approach to causality that is widely used in economics seems in stark contrast to the viewpoint presented in Haavelmo (1943) that turned econometrics into a probabilistic study. Particularly, Haavelmo's core argument was that it is the very nature of economic behavior itself, that implies the necessity of stochastic formulations of economic theory and the inclusion of error terms in otherwise exact relationships to make simplifications of reality elastic enough for application. Moreover, Kalman (1983) definitively argues that the classical model of reality developed in mechanical physics is simply inapplicable to the problems of economics. It is certainly interesting that, after a century of probability work by statisticians and econometricians that led statistics to be accepted as the leading model for inference, the working definition of causality used by many economists is deterministic in nature, while physicists are open to work with Granger's definition.

Efforts to reunite the conflicting views have recently begun to produce interesting results. New developments began by noting that questions about important concepts in economics, such as choice and uncertainty, can even in very simplistic settings not be answered within Pearl's framework. White and Chalak (2009); White et al. (2014) extend Pearl's causal model to include optimization, equilibrium, learning concepts, and choice that are integral parts of economics and game theory, or social systems in which agents act and react under uncertainty. Under the extended causal framework, White and Lu (2010) forge the previously missing link between Granger causality and structural causality by showing that, given a corresponding conditional form of exogeneity, Granger causality holds *if and only if* a corresponding form of structural causality holds, and Eichler and Didelez (2010) provide conditions under which Granger non-causality implies that an intervention has had no effect. White et al. (2011) show that tests for Granger causality can be used to test for direct causality in sequential systems, and Lu et al. (2017) produce tests for cross-section and panel data valid in a general case that does not assume linearity, monotonicity in observables or unobservables, or separability between observed and unobserved variables in the structural relations. White and Pettenuzzo (2014) show that instead of relying on exogeneity (weak, strong, or super) conditional on the model or *Data Generating Process* (DGP), causal effects can also be consistently estimated by relying on correct specification of the conditional mean sequence. This highlights the importance of knowledge regarding the important features of the DGP. Specifically, economic theory may suggest which variables are meaningful, while the functional form (numbers of lags, cointegration, or structural shifts), may be resolved directly from the data.

In this paper, we continue the debate focusing on the application strategy to estimate causal relationships, taking a general, data-driven, stand. Specifically, we assume that a theorized causal relationship between two economic variables in the possible presence of unobserved factors leads to

a probability law that regulates the transitions from one phase to another in Markov chains of iterated processes of causes and effects. This is particularly relevant given the arguments of Haavelmo (1943) and others discussed, that are in favor of formulating economic theories stochastically. In this probabilistic setting, the properties of the extremum estimators introduced earlier provide a natural interpretation of an estimated result regardless of whether correct specification is assumed. Specifically, under simple conditions that are often guaranteed by the design of standard estimation problems, the limit result is the closest to the correct hypothesis about causality out of all considered hypotheses. Assuming correct specification, ensures naturally that minimal distance is zero, which corresponds to the setting of White and Pettenuzzo (2014). When the hypothesis space is incorrectly constructed, equivalence is established with respect to the pseudo-*true* measure that by definition, again, is closest to the correct hypothesis out of all considered hypotheses. If the hypothesis space is sufficiently large to ensure a small divergence between the *true* causal probability law and the closest possible modeled measure, then the limit result should naturally capture important aspects of the *true* causal probability law even under miss-specification. This suggests that in the absence of clear economic theories to guide model specification, a researcher can still focus on ensuring that the parameter space is able to produce as much hypotheses about causality as possible and proceed with a general estimation method that penalizes model complexity.

The core of the argument is based on the use of a decision criterion that scales to a metric measuring distance between any two probability measures. When this is the case, a metric space can be considered in which equivalences can be established by partitioning into classes of zero-distance points. Equivalence to the *true* measure, that correctly describes the *true* frequencies in Markov chains of iterated processes of causes and effects, is established by reaching zero distance in that metric space. These type of decision criteria are common in econometrics and an example is

provided in the context of Maximum Likelihood. We show that squared Hellinger distance marks the lower bound of Kullback-Leibler divergence, implying that minimizing information loss using the AIC, or another suitable penalized Likelihood variation that ranks hypotheses according to their Kullback-Leibler divergence, minimizes a distance metric toward the *true* probability measure. As such, minimizing AIC is a theoretically sound objective to uncover the correct causal structure from interrelated stochastic data that are observed and modeled sequentially over time.

The remainder of this paper is structured as follows. Section 7.2 introduces definitions of causality in terms of probability measures, section 7.3 discusses the divergence between modeled measures and the causal measure, section 7.4 discusses the particular case of the Maximum Likelihood estimator and squared Hellinger distance. Section 7.5 ends with concluding remarks.

## 7.2 Causality and probability

Cause and effect in deterministic settings involve propositions along the lines of "*if $X$ occurs then $Y$ must occur*". That deterministic definition of causality is difficult to reconcile with probability. Causality statements in a statistical context often spur a great deal of discussion among researchers. In fact, while many researchers meet the concept of causality early in their career, few eventually agree on what it truly means and how it should be approached in an empirical context. To introduce a concept of causality appropriate in a probabilistic setting, let us consider a simple game of chance; a dice. Throwing a dice does not cause a certain outcome. In that sense, one cannot say that "*if $X$ occurs then $Y$ must occur*" with $X$ being a throw, and $Y$ being the outcome of a throw. In fact, the outcome is one of seven, six being one to six eyes, and seven being no outcome at all. Each outcome occurs with a certain probability,

the latter being zero. The measure that assigns probability to each of the outcomes describes the *true* probabilistic property of the dice. In that sense, a faulty dice may *cause* a certain outcome to occur with higher probability *truly*. One can say that a faulty dice is characterized by a probability measure that leads to outcomes with a certain outcome having a higher probability assigned to it, such that the expected value of a throw minus the expected value of a throw of a non-faulty dice is nonzero. In this sense, one can describe the causal effect of being faulty, in terms of probability. Specifically, " $\nabla Y$ *must occur with probability* $P \geq 0$ *if* $X$ *has occurred*" with $\nabla Y$ being a non-zero difference value between throws of the faulty dice and a correct dice.

In this probabilistic view, a theory about causality is a statement about the properties of the *true* measure that describes a process stochastically. Specifically, a causal relationship can be described in terms of whether the *true* probability measure produces a non-empty stochastic sequence describing the directly caused effects from one variable to the other. Or, equivalently, whether the *true* probability measure is associated with a non-empty stochastic sequence of differences between the process that is driven by causes that produce real-valued effects from one variable to the other and the process that does not react to the causes. This is somewhat different than attributing the presence of causal relationships directly to the values of the parameters in a mathematical model of reality, though, as shall be discussed, the definition based on the probability measure equivalently produces statements about parameters or functions. Drawing on Approximation Theory, one can transfer the measure theoretical definitions of *true* causality, to a modeled probability measure in the limit based on an equivalence argument. The modeled measure, in turn, is naturally associated with parameters that determine the functional behavior. Due to well-known results on consistency for approximate extremum estimates, the approximation of the *true* measure eventually thus provides valid descriptions of causality based on empirically modeled

data when a sufficient amount of observations has been collected.

In an empirical sense, stating that a dice is faulty, or equivalently saying that a certain outcome occurs with higher probability, is a statement about the *true* probabilistic property of that dice and such conclusions may result from a modeled probability measure that best confirms to many observed outcomes. Similarly, saying that a modeled dice is faulty, or equivalently saying that a modeled outcome has higher probability than assigned by the *true* probability measure, is a statement about the probabilistic properties of the modeled dice. Such conclusions may result from observing many modeled dices, and many outcomes, and comparing observed outcomes to modeled outcomes repeatedly and selecting the model that best resembles reality. Most estimators by design select from a set of hypothetical realities by some process of divergence minimization w.r.t. the *true* measure. If that decision process is exhaustive across all divergences between possible measures and the *true* measure, then the closest possible measure will be chosen. If the *true* measure is included in all considered measures, then the decision process will end by selecting that measure. When the axiom of correct specification is abandoned, and the correct probability measure is not included in the set of modeled measures, the *true* measure is replaced by a pseudo-*true* measure. This measure by definition still minimizes divergence w.r.t. the *true* measure. The interpretation that a pseudo-*true* measure caries is that, after observing the data and considering all the measures that are induced under all the possible parameter vectors, the pseudo-*measure* probability measure bests confirms to the *true* probability measure. In this case, the decision process thus ends with accepting the best approximation of the correct hypothesis as the one from which to derive causal claims, as no better hypothesis about reality can be constructed until a larger set of hypotheses formally comes under review. This is a stronger result than the common statement that $X$ only helps predicting $Y$ with the arrow of time as the indicator of the direction of effects.

Improved predictability can be a local result within the space of potential hypotheses. This reveals the intrinsic relationship between the size of the parameter space that is set when the regression is specified, and the empirical claim that results from estimating that regression, suggesting that the quality of causal inference depends on the flexibility of the model to produce a wide variety of potentially (in)correct structures.

Let us now more formally express these thoughts. Notation is as follows, $\mathbb{N}$, $\mathbb{Z}$ 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-$\sigma$ 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 $f \circ g := f(g)$. The tensor product is denoted $\otimes$. Finally, the empty set $\emptyset$ is also used in the context of an empty sequence, that sometimes would be notated as $()$ in 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-$\sigma$ $\mathcal{F}/\mathfrak{B}(\mathcal{X}_\infty)$-measurable map $\mathbf{x} : \Omega \to \mathcal{X}_\infty \subseteq \mathbb{R}_\infty^{n_x}$. In this, $\mathbb{R}_\infty^{n_{\mathbf{x}}} := \times_{t=-\infty}^{t=\infty}\mathbb{R}^{n_{\mathbf{x}}}$ 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 $\mathfrak{B}(\mathcal{X}_\infty) := \mathfrak{B}(\mathbb{R}_\infty^{n_{\mathbf{x}}}) \cap \mathcal{X}_\infty$, and $\mathfrak{B}(\mathbb{R}_\infty^{n_{\mathbf{x}}})$ denotes the Borel sigma algebra on the finite dimensional cylinder set of $\mathbb{R}_\infty^{n_{\mathbf{x}}}$, see Billingsley (1995), 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 $\sigma$-field defined on the event space. $\mathbb{P}$ is used here as a placeholder as we shall introduce probability measures of interest shortly.

If $\mathbf{x}$ was considered as a univariate sequence free from exogenous 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 $P^{\mathbf{x}}$ is defined over 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 $\omega \in \Omega$. 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,[2]

$$\begin{array}{l} \mathbf{x} := \{\mathbf{x}_t = f^{\mathbf{xx}}(\mathbf{x}_{t-1}), t \in \mathbb{Z}\} \\ \mathbf{y} := \{\mathbf{y}_t = f^{\mathbf{yy}}(\mathbf{y}_{t-1}), t \in \mathbb{Z}\} \end{array} . \tag{7.1}$$

The structure reveals that $P^{\mathbf{x}}$ is simply induced by the function $f^{\mathbf{xx}}$ on $\mathfrak{B}(\mathcal{X})$ according to $P^{\mathbf{x}}(B_{\mathbf{x}}) = P^{\mathbf{x}} \circ (f^{\mathbf{xx}})^{-1}(B_{\mathbf{x}}) \; \forall \; B_{\mathbf{x}} \in \mathfrak{B}(\mathcal{X}_\infty)$ and $P^{\mathbf{y}}$ is induced by the function $f^{\mathbf{yy}}$ on $\mathfrak{B}(\mathcal{Y})$ in a similar way, see Dudley (2002) p.118 and Davidson (1994) p.115. The notion is important to the extent that it has been argued (see Hendry (2017) 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 it's 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 that in turn would be guided by the research question of interest and the availability of data. Particularly, dynamical systems in economics are often modeled using aggregate macro-economic data that does not have the same granularity as micro-economic data that contains information about behavior of

[2]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 identify function.

individual economic agents.

If interrelated stochastic sequences are at the center of inference, the building blocks required for describing the processes are more complicated. This increases the potential complexity of $P^{\mathbf{x}}$ and $P^{\mathbf{y}}$ tremendously, but it also allows to conclude decisively between causality, non-causality and feedback. Consider a simple stochastic system:

$$\begin{aligned} \mathbf{x} &:= \{\mathbf{x}_t = f^{\mathbf{xx}}(\mathbf{x}_{t-1}) + f^{\mathbf{xy}}(\mathbf{y}_{t-1}), t \in \mathbb{Z}\} \\ \mathbf{y} &:= \{\mathbf{y}_t = f^{\mathbf{yx}}(\mathbf{x}_{t-1}) + f^{\mathbf{yy}}(\mathbf{y}_{t-1}), t \in \mathbb{Z}\} \end{aligned} . \tag{7.2}$$

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, and verifying their properties is central to causality studies, while $f^{\mathbf{xx}}$ 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 if that sequence turns inactive immediately after impact. The functions that control memory properties within channels in some sense determine the reflex of the future onto the past, and specifying correct empirical equivalents to $f^{\mathbf{xx}}$ and $f^{\mathbf{yy}}$ is just as crucial to the inference about the causal interdependencies as specifying mechanisms for the action of interest is. To understand directional cause,

and the role that $f^{\mathbf{xx}}$ and $f^{\mathbf{yy}}$ play, it is useful to consider the following:

$$\begin{aligned}\mathbf{x}^0 &:= \{\mathbf{x}_t^0 = f^{\mathbf{xy}}(\mathbf{y}_{t-1}), t \in \mathbb{Z}\} \\ \mathbf{y}^0 &:= \{\mathbf{y}_t^0 = f^{\mathbf{yx}}(\mathbf{x}_{t-1}), t \in \mathbb{Z}\}\end{aligned}, \tag{7.3}$$

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 eq. (7.2), the sequential system eq. (7.3) 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 out 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 out 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 eq. (7.3) produces any values.

Importantly, the system also 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 eq. (7.3) 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 $f_{\mathbf{y}}^0(\mathbf{y}_t^0, \mathbf{y}_{t-1}) = f^{\mathbf{xy}}(\mathbf{y}_t^0 + f^{\mathbf{yy}}(\mathbf{y}_{t-1}))$ and $f_{\mathbf{x}}^0(\mathbf{x}_t^0, \mathbf{x}_{t-1}) = f^{\mathbf{yx}}(\mathbf{x}_t^0 + f^{\mathbf{xx}}(\mathbf{x}_{t-1}))$, one can write

$$\begin{aligned}\mathbf{x}^0 &:= \{\mathbf{x}_t^0 = f_{\mathbf{y}}^0(\mathbf{y}_{t-1}^0, \mathbf{y}_{t-2}), t \in \mathbb{Z}\} \\ \mathbf{y}^0 &:= \{\mathbf{y}_t^0 = f_{\mathbf{x}}^0(\mathbf{x}_{t-1}^0, \mathbf{x}_{t-2}), t \in \mathbb{Z}\}\end{aligned}. \tag{7.4}$$

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

$$\begin{aligned}\mathbf{x}^0 &:= \{\mathbf{x}_\infty^0 = (f_{\mathbf{y}}^0)^\infty(\mathbf{y}_1^0, \mathbf{y}_1), t \in \mathbb{Z}\} \\ \mathbf{y}^0 &:= \{\mathbf{y}_\infty^0 = (f_{\mathbf{x}}^0)^\infty(\mathbf{x}_1^0, \mathbf{x}_1), t \in \mathbb{Z}\}\end{aligned}. \tag{7.5}$$

$(f_{\mathbf{y}}^0)^\infty$ and $(f_{\mathbf{x}}^0)^\infty$ 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 random functions that are all defined on $\mathbf{y}$, and $\mathbf{y}^0$ defined on $\mathbf{x}$ in a similar way.[3] For ease of notation, let us write

$$\begin{aligned} \mathbf{x}^0 &:= \{\mathbf{x}_t^0 = \mathbf{f}_{\mathbf{y}}^0(\mathbf{y}_{-\infty:t}), t \in \mathbb{Z}\} \\ \mathbf{y}^0 &:= \{\mathbf{y}_t^0 = \mathbf{f}_{\mathbf{x}}^0(\mathbf{x}_{-\infty:t}), t \in \mathbb{Z}\} \end{aligned} . \tag{7.6}$$

where bold-faced $\mathbf{f}^0$ is used to refer to the entire sequence of functions $f^0$ 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.

Note first that $\mathbf{f}_{\mathbf{y}}^0 : \mathcal{Y} \to \mathcal{X} \subseteq \mathbb{R}$ is a $\mathfrak{B}(\mathcal{Y})/\mathfrak{B}(\mathcal{X})$-measurable mapping, and $\mathbf{f}_{\mathbf{x}}^0 : \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}$ is a $\mathfrak{B}(\mathcal{X})/\mathfrak{B}(\mathcal{Y})$-measurable mapping. The sequence $\mathbf{x}^0$ thus lives on $(\mathcal{X}_\infty, \mathfrak{B}(\mathcal{X}_\infty), P_0^{\mathbf{x}})$, where $P_0^{\mathbf{x}}$ is induced according to $P_0^{\mathbf{x}}(B_{\mathbf{x}}) = P_0^{\mathbf{y}} \circ (\mathbf{f}_{\mathbf{y}}^0)^{-1}(B_{\mathbf{x}}) \ \forall \ B_{\mathbf{x}} \in \mathfrak{B}(\mathcal{X}_\infty)$, and $\mathbf{y}^0$ lives on $(\mathcal{Y}_\infty, \mathfrak{B}(\mathcal{Y}_\infty), P_0^{\mathbf{y}})$, where $P_0^{\mathbf{y}}$ is induced according to $P_0^{\mathbf{y}}(B_{\mathbf{y}}) = P_0^{\mathbf{x}} \circ (\mathbf{f}_{\mathbf{x}}^0)^{-1}(B_{\mathbf{y}}) \ \forall \ B_{\mathbf{y}} \in \mathfrak{B}(\mathcal{Y}_\infty)$. 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, and the probability measures underlying the causal sequences are induced by the functional relationships that describe these dynamical dependencies.

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 $f^{\mathbf{z}}(\mathbf{z})$,

[3] Equation (7.5) reveals an important implication for causality studies. 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 assurances is required that the initialization effect on the causal pathway must eventually not matter given sufficient observations. This is central to contraction studies.

is considered, the researcher is confronted with the situation that

$$\begin{aligned}\mathbf{x} &:= \{\mathbf{x}_t = f^{\mathbf{xx}}(\mathbf{x}_{t-1}) + f^{\mathbf{xy}}(\mathbf{y}_{t-1}) + f^{\mathbf{xz}}(\mathbf{z}_{t-1}), t \in \mathbb{Z}\} \\ \mathbf{y} &:= \{\mathbf{y}_t = f^{\mathbf{yx}}(\mathbf{x}_{t-1}) + f^{\mathbf{yy}}(\mathbf{y}_{t-1}) + f^{\mathbf{yz}}(\mathbf{z}_{t-1}), t \in \mathbb{Z}\}\end{aligned} . \tag{7.7}$$

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:

$$\begin{aligned}\mathbf{z} &:= \{\mathbf{z}_t = (f^{\mathbf{xz}})^{-1}(\mathbf{x}_{t+1} - (f^{\mathbf{xx}}(\mathbf{x}_t) + f^{\mathbf{xy}}(\mathbf{y}_t))), t \in \mathbb{Z}\} \\ \mathbf{z} &:= \{\mathbf{z}_t = (f^{\mathbf{yz}})^{-1}(\mathbf{y}_{t+1} - (f^{\mathbf{yx}}(\mathbf{x}_t) + f^{\mathbf{yy}}(\mathbf{y}_t))), t \in \mathbb{Z}\}\end{aligned} . \tag{7.8}$$

Equation (7.8) suggests to write eq. (7.7) in terms of $\mathbf{y}$ and $\mathbf{x}$ only by defining $\mathbf{z}$ as a difference combination of $\mathbf{x}$ and $\mathbf{y}$.[4] This allows us to define the spaces and measures on which the multivariate process operates in terms of $\mathbf{x}$ and $\mathbf{y}$ only even in the presence of $\mathbf{z}$. If the process is invertible, one can simply write:[5]

$$\begin{aligned}\mathbf{x} &:= \{\mathbf{x}_t = f^{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}), t \in \mathbb{Z}\} \\ \mathbf{y} &:= \{\mathbf{y}_t = f^{\mathbf{y}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}), t \in \mathbb{Z}\}\end{aligned} . \tag{7.9}$$

For every $t \in \mathbb{Z}$, the map $f^{\mathbf{x}} \circ (\mathbf{y}_{t-1}, \mathbf{x}_{t-1}) : \Omega \to \mathcal{Y}$ is $\mathcal{F}/\mathfrak{B}(\mathcal{Y} \times \mathcal{X})$-measurable and $\mathbf{y}(\omega)$ lives on the space $(\mathcal{Y}_\infty, \mathfrak{B}(\mathcal{Y}_\infty \times \mathcal{X}_\infty), P^{\mathbf{y}})$ where the probability measure $P^{\mathbf{y}}$ is induced by $f^{\mathbf{x}}$ on $\mathfrak{B}(\mathcal{Y}_\infty \times \mathcal{X}_\infty)$ according to the point-wise application of $P^{\mathbf{x}}$ and the inverse of $f^{\mathbf{x}}$.[6] Similar arguments follow for $f^{\mathbf{y}}$. This tells us that in the multivariate case with possibly unobserved variables, the probability measures underlying the individual

[4] Apart from stability conditions on the endogenous process, one requires also that the exogenous impacts enter the system in some suitable manner such that $(f^{\mathbf{y}z})^{-1}$ and $(f^{\mathbf{x}z})^{-1}$ are absolute summable. Following the same arguments that resulted in eq. (7.5), 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.

[5] By aggregating the functions

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

[6] $P^{\mathbf{y}}(B_{\mathbf{y}} \times B_{\mathbf{x}}) = P^{\mathbf{x}} \circ (f^{\mathbf{x}})^{-1}(B_{\mathbf{y}} \times B_{\mathbf{x}}) \; \forall \; (B_{\mathbf{y}} \times B_{\mathbf{x}}) \in \mathfrak{B}(\mathcal{Y}_\infty \times \mathcal{X}_\infty)$.

sequences are possibly intertwined with those of the other sequences. This strongly complicates candidates and studies for the probability measure $P^{\mathbf{w}}$ that underlies the joint process $\mathbf{w} := \{\mathbf{w}_t = (\mathbf{y}_t, \mathbf{x}_t), t \in \mathbb{Z}\}$ operating on $(\mathcal{W}_\infty, \mathfrak{B}(\mathcal{W}_\infty), P^{\mathbf{w}})$.[7]

Nevertheless, when the correct invertible filters for all the time dynamics of the observed part of the system are specified, one can still rewrite general systems of the form eq. (7.7) into a representation that follows eq. (7.6). One can thus always state causality conditions relevant for correct inference, based on the subsystems that produce the directly caused effects eq. (7.6). 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 components of $P^{\mathbf{w}}$ that defines the complete estimation objective.

DEFINITION. 5 (Non-causality). *The stochastic sequences* $\mathbf{x}(\omega)$ *and* $\mathbf{y}(\omega)$ *are not causality 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. 6 (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. 7 (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* $\mathbf{y}^0(\omega) \in \mathcal{Y} \ \forall \ (\omega, t) \in \Omega \times \mathbb{Z}$.

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 emptyset with probability 1. This is because the functions that induce the probability measure cancel out, hence they can be removed from the

[7] The sequence is more complicated, and 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 $\sigma$-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, Dudley (2002) p119.).

equations resulting in a probability measure that is not induced by any remaining rule or relationship. Respectively, conditioning on impacts in $\mathbf{x}$, these probabilistic causality definitions can thus be understood as:

1. Whenever an intervention in $\mathbf{x}$ occurs, there is no chance that $\mathbf{y}$ reacts as a result of that.

2. Whenever an intervention in $\mathbf{x}$ occurs, there is positive chance that $\mathbf{y}$ reacts as a result of that.

3. Whenever an intervention in $\mathbf{x}$ occurs, there is positive chance that $\mathbf{y}$ 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.

## 7.3 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 nonzero 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 parametrization 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

$$\hat{\mathbf{w}} := \{\hat{\mathbf{w}}_t = f(\mathbf{w}_{t-1}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta, t \in \mathbb{Z}\}, \tag{7.10}$$

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 $\Theta$ is trivially 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 $\Theta$ indexes families of parametric models each inducing a space of parametric functions generated under $\Theta$. In this discussion the focus remains limited to parametric inference, hence a compact set of potential hypotheses is considered. The arguments are trivially extended to the nonparametric case, by focusing on a compact subset $\Theta_s \subset \Theta$ of solutions.[8] 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. 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 $\Theta$ 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

[8] For example, by letting $\Theta_s$ grow as $T \to \infty$, hence focusing on the case $\Theta_{s1} \subset \Theta_{s2} ... \subset \Theta_{s\infty} \subseteq \Theta$, see for example Geman, Stuart; Hwang (1982).

the parameter space $\boldsymbol{\theta} \in \Theta$, there is a triplet $(\mathcal{W}_\infty, \mathfrak{B}(\mathcal{W}_\infty), P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$ that describes the probability space of modeled data under $\boldsymbol{\theta}$. The triplet $(\mathcal{W}_\infty, \mathfrak{B}(\mathcal{W}_\infty), P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$ is thus itself an element of the measure spaces indexed by $\boldsymbol{\theta}$ across all $\Theta$. Given the *true* probability measure $P^{\mathbf{w}}$ on $\mathfrak{B}(\mathcal{W})$, this process is summarized by a functional $\mathfrak{P} : F_\Theta(\mathcal{W}) \to \mathcal{P}_\Theta^{\hat{\mathbf{w}}}$, that maps elements from the space of parametric functions generated by the entire parameter space $F_\Theta(\mathcal{W})$, onto the space $\mathcal{P}_\Theta^{\hat{\mathbf{w}}}$ of probability measures defined on the sets of $\mathfrak{B}(\mathcal{W}_\infty)$ generated by $\Theta$ through $f(\cdot; \boldsymbol{\theta})$.

Now, $f^{\mathbf{w}}$ is generally not only unknown, but for a finite $\Theta$ 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 $P^{\mathbf{w}} \notin \mathcal{P}_\Theta^{\hat{\mathbf{w}}}$. However, if $\exists P^{\mathbf{w}} \in \mathcal{P}_\Theta^{\hat{\mathbf{w}}}$ , one can learn all about $P^{\mathbf{w}}$, by uncovering the properties of $f$, given a sufficient amount of observations is available.[9] Let

$$\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta} \in \Theta} Q_T(\mathbf{w}_T; \boldsymbol{\theta}), \tag{7.11}$$

$\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$, 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 Kolmogorov and Fomin (1975) 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 $\mathcal{P}_\Theta^{\hat{\mathbf{w}}}$ 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. It is important to again note that $\mathcal{P}_\Theta^{\hat{\mathbf{w}}}$ is induced by the proposed candidates for $P^{\mathbf{w}}$.

[9] As discussed in literature on miss-specification, even when the axiom of correct specification is abandoned, $f$ may converge to a function that produces the optimal conditional a density which may reveal important properties of $f^{\mathbf{w}}$.

Studies on causality thus rely on flexible model design as the researcher determines which hypotheses are considered in a study by exerting control over $\Theta$. Naturally if $\Theta_1 \subset \Theta_2$, then $\Theta_2$ produces a larger $\mathcal{P}_{\Theta_2}^{\hat{\mathbf{w}}} \supset \mathcal{P}_{\Theta_1}^{\hat{\mathbf{w}}}$. This suggests that minimizing this divergence metric over a large as possible $\mathcal{P}_{\Theta}^{\hat{\mathbf{w}}}$ results in selecting $P^{\hat{\mathbf{w}}}$ at a point in $\mathcal{P}_{\Theta}^{\hat{\mathbf{w}}}$ that attains equivalence to $P^{\mathbf{w}}$ only when $\Theta$ 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 $\Theta$, under the injective $f_{\mathcal{W}} : \Theta \to F_\Theta(\mathcal{W})$ and the functional $\mathfrak{P} : F_\Theta(\mathcal{W}) \to \mathcal{P}_{\Theta}^{\hat{\mathbf{w}}}$ that induces the space of probability measures, is defined on the sample space $\mathcal{W}$. This highlights that the correct specification argument $P^{\mathbf{w}} \in \mathcal{P}_{\Theta}^{\hat{\mathbf{w}}}$, not only stresses flexible parametrization in the sense that parameterized dependencies can take on many values, but also in the sense of using correct data.[10] When little is known about $f$, one is thus not only concerned with flexibility in terms of the type of parametric functions generated under $\Theta$, 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 $Q_\infty(\boldsymbol{\theta}) = Q_\infty^{\mathcal{P}}(P^{\mathbf{w}}; P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}) \ \forall \ \boldsymbol{\theta} \in \Theta$ with the criterion $Q_\infty(\boldsymbol{\theta}) = Q_\infty^{\mathcal{P}}$ as a measure of divergence $d_{\mathcal{P}}$ on the

[10] 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 $\Theta$. The distinction is nevertheless useful because nonparametric models are often popularized as methods to reduce misspecification bias as $\Theta$ becomes infinite dimensional, but this does not imply that $P^{\mathbf{w}} \in \mathcal{P}_{\Theta}^{\hat{\mathbf{w}}}$ if important data is missing.

*true* probability measure and the modeled measure. More specifically, $d_{\mathcal{P}} \equiv Q_{\infty}^{\mathcal{P}} : \mathcal{P}_{\Theta}^{\hat{\mathbf{w}}} \times \mathcal{P}_{\Theta}^{\hat{\mathbf{w}}} \to \mathbb{R}_{\geq 0}$. By definition of $Q_{\infty}^{\mathcal{P}}$ as a divergence on the space that contains $P^{\mathbf{w}}$ and $P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}$ $\forall\ \boldsymbol{\theta} \in \Theta$, 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 $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta} \in \Theta} Q_{\infty}^{\mathcal{P}} \equiv \arg\min_{\boldsymbol{\theta} \in \Theta} Q_{\infty}^{F}(f^{\mathbf{w}}, f_{\mathcal{W}}(\boldsymbol{\theta}))$, with $Q_{\infty}^{F} : F(\mathcal{W}) \times F(\mathcal{W}) \to \mathbb{R}_{\geq 0}$, 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 converge to the *true* IRF at $\boldsymbol{\theta}_0$. This implies that deciding between DEFINITION. 5-DEFINITION. 7 can be read from the responses produced by the IRF that minimizes divergence w.r.t. the *true* IRF

Not necessarily, but convenient for a proof that holds easily in practical situations, is to assume existence of a strictly increasing function $r : \mathbb{R} \to \mathbb{R}_{\geq 0}$ that ensures 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 strictly increasing function. Under these assumptions a simple result follows. For convenience all assumptions are summarized in ASSUMPTION. 13.

ASSUMPTION. 13. *For a limit criterion $Q_{\infty} : \Theta \to \mathbb{R}$ of the form $Q_{\infty}(\boldsymbol{\theta}) \equiv Q_{\infty}^{\mathcal{P}}(P^{\mathbf{w}}, P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})\ \forall\ \boldsymbol{\theta} \in \Theta$, $d_{\mathcal{P}} \equiv Q_{\infty}^{\mathcal{P}} : \mathcal{P}^{\mathbf{w}} \times \mathcal{P}^{\mathbf{w}} \to \mathbb{R}_{\geq 0}$ 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. 6. *Assume ASSUMPTION. 13, then the following are equiv-*

*alent 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. $\arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty^{\mathcal{P}}(P^{\mathbf{w}}, P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$,
5. $\arg\min_{\boldsymbol{\theta}\in\Theta} d_{\mathcal{P}}^*(P^{\mathbf{w}}, P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$.

REMARK. 6. *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. 6 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 $P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}$ (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. 7. *PROPOSITION. 6 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 7.4 we also shall give an example 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. 6 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 uniqueness of $\boldsymbol{\theta}_0$ within the compact hypothesis space $\Theta$.

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 White and Pettenuzzo (2014). Summarized:

COROLLARY. 6. *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. *5.*
2. $\mathbf{x}$ *causes* $\mathbf{y}$ *if the probability measure* $P^{\hat{\mathbf{y}}}_0$ *adheres to* DEFINITION. *6.*
3. $\mathbf{y}$ *causes* $\mathbf{x}$ *if the probability measure* $P^{\hat{\mathbf{x}}}_0$ *adheres to* DEFINITION. *6.*
4. *There is bi-directional causality if* $P^{\hat{\mathbf{x}}}_0$ *and* $P^{\hat{\mathbf{y}}}_0$ *adhere to* DEFINITION. *7.*

Finally, in the case of a miss-specified model, REMARK. 6 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 $d^*_{P^{\mathbf{w}}} = r \circ d_{\mathcal{P}}(P^{\mathbf{w}}, \arg\min_{\boldsymbol{\theta} \in \Theta} Q^{\mathcal{P}}_{\infty}(P^{\mathbf{w}}, P^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}})) > 0$. 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 strong 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 $\Theta$ by allowing flexibility and investigating a wide range of data. Disregard the value of $d^*_{P^{\hat{\mathbf{w}}}}$, the following holds.

PROPOSITION. 7. *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* $\Theta$*. 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. *6 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. 7. *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. *5.*
2. $\mathbf{x}$ *causes* $\mathbf{y}$ *if the probability measure* $P^{\hat{\mathbf{y}}}_0$ *adheres to* DEFINITION. *6.*
3. $\mathbf{y}$ *causes* $\mathbf{x}$ *if the probability measure* $P^{\hat{\mathbf{x}}}_0$ *adheres to* DEFINITION. *6.*
4. *There is bi-directional causality if* $P^{\hat{\mathbf{x}}}_0$ *and* $P^{\hat{\mathbf{y}}}_0$ *adhere to* DEFINITION. *7.*

Respectively, conditioning on interventions in $\mathbf{x}$, the results can be understood as:

1. Whenever an intervention in $\mathbf{x}$ occurs, our best hypothesis is that there is no chance that $\mathbf{y}$ 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.

## 7.4 Limit Squared Hellinger distance

Both COROLLARY. 6 and COROLLARY. 7 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 (Cheney and Respess, 1982). While this retains simplicity of the argument, it also shows that a direct interpretation of COROLLARY. 6 and COROLLARY. 7

can be obtained within the framework if Maximum Likelihood. Let us first define the our criterion as the Maximum Likelihood Estimator:

$$\arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{w}_T;\boldsymbol{\theta}) := \arg\max_{\boldsymbol{\theta}\in\Theta} \sum_{t=1}^{T} \ln p_t(\mathbf{w}_t|\boldsymbol{\theta}). \tag{7.12}$$

Note that this is conform the form

$$Q_\infty(\boldsymbol{\theta}) := Q_T\left(f^{\mathbf{w}}(\mathbf{w}_T), f\left(\mathbf{w}_T; \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{w}_T;\boldsymbol{\theta})\right)\right)$$

with $T \to \infty$ and $Q_\infty : \Theta \to \mathbb{R}$. It can be shown that under this definition with $Q_\infty(\boldsymbol{\theta}) = Q_\infty^{\mathcal{P}}(P^{\mathbf{w}}; P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})\ \forall\ \boldsymbol{\theta} \in \Theta$ that 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 Lebesgue measure (such that they are simply probability density functions), then the following is a divergence from the true probability measure to the modeled probability measure (Kullback-Leibler divergence, see Kullback and Leibler (1951)):

$$KL\big(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})||P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})\big)$$
$$= \begin{cases} \int_{-\infty}^{\infty} p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) \ln \dfrac{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})}{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})} d\mathbf{w} & \forall\ p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) << p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}}) \\ \infty & \forall\ p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}}) >> p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) \end{cases}. \tag{7.13}$$

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 COROLLARY. 6 and COROLLARY. 7 to establish equivalences by partition-

ing into classes of zero-distance points. In particular, it is asymmetric

$$KL\big(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})||P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})\big) \neq KL\big(P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})||P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})\big), \quad (7.14)$$

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

$$KL(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})||P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})) = \sum_{t}^{T} \ln KL(p_t^{\mathbf{w}}(\mathbf{w}_t|\boldsymbol{\theta}_{\mathbf{w}})||p_t^{\hat{\mathbf{w}}}(\mathbf{w}_t|\boldsymbol{\theta}_{\hat{\mathbf{w}}})), \quad (7.15)$$

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:

$$\begin{aligned} \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{w}_T;\boldsymbol{\theta}) &:= \arg\max_{\boldsymbol{\theta}\in\Theta} \sum_{t=1}^{T} \ln \frac{p^{\mathbf{w}}(\mathbf{w}_t|\boldsymbol{\theta}_{\mathbf{w}})}{p^{\hat{\mathbf{w}}}(\mathbf{w}_t|\boldsymbol{\theta}_{\hat{\mathbf{w}}})} \\ &= \arg\min_{\boldsymbol{\theta}\in\Theta} KL\big(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})||P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})\big). \end{aligned} \quad (7.16)$$

Note that under standard assumptions, a Law of Large Numbers can be applied to obtain the convergence, hence by maximizing 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 $d^*_{P^{\hat{\mathbf{w}}}} = r \circ d_{\mathcal{P}}(P^{\mathbf{w}}, P_\theta^{\hat{\mathbf{w}}}) \sim 0$. 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 (Hellinger, 1909):

$$H\Big(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}), P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})\Big) = \sqrt{\frac{1}{2}\int \left(\sqrt{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})} - \sqrt{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}\right)^2 d\mathbf{w}}. \quad (7.17)$$

Specifically, the squared Hellinger distance provides a lower bound for

the Kullback-Leibler divergence. Therefore, maximizing likelihood implies minimizing Kullback-Leibler divergence which implies minimizing Hellinger distance. This is easily seen by the following:

PROPOSITION. 8. *Squared Hellinger distance provides a lower bound to Kullback-Leibler divergence:*

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

We end this sections 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. 6, 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 simple quadratic expansion

$$\begin{aligned}&\lim_{T\to\infty} \mathbb{E}\Big(L_T(\hat{\boldsymbol{\theta}}_T) - \mathbb{E}L_T(\boldsymbol{\theta}_0)\Big)\\ &= \lim_{T\to\infty} \mathbb{E}\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0)'\frac{1}{T}L_T''(\boldsymbol{\theta}_T)\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \neq 0. \qquad (7.18)\end{aligned}$$

Under considerably restrictive conditions original work by Akaike (1973, 1974) 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 Hurvich and Tsai (1989, 1991); Sin and White (1996). The AIC is also valid to decide between economic theories for which no test statistics can

be found Granger et al. (1995).

This means that while maximizing Log Likelihood over $\Theta$ 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 $\Theta$. Hence, in practice, a researcher can minimize the AIC as the practical objective to minimize Hellinger distance, and use correct specification tests to decide whether COROLLARY. 6 or COROLLARY. 7 is relevant.

## 7.5 Concluding remarks

During the 20th century, probability theory and economic theory have been closely developed together. While empirical studies in economics rely heavily on probabilistic concepts for inference, definitions for causality are often viewed through a deterministic lens. This paper discussed a probabilistic view on causality. In this view, a theory about causality is seen as a statement about the properties of the *true* measure that describes an observed process stochastically. The correct economic theory thus concerns the *true* frequencies in Markov chains of iterated processes of causes and effects, in which the transitions from one phase to another are regulated by the *true* probability law. This *true* probability law has been used to define causality in terms of stochastic sequences of caused effects.

Some argue that similar system theoretic definitions of causality, most notably the one from Granger, are not causal in the sense that they do not provide economic insight in the origin of the *true* probability law, but rather describe (correctly) the probabilistic behavior of the outcome of a causal origin. Clearly, these definitional discussions lie outside the scope of the statistical framework used in an empirical setting and relate to the

structure of the research question itself. In fact, we have seen that the relation between the functional behavior of a system and the probability measure that regulates its transitions from one phase to another, can be made explicit such that the direct relationship between theorized functional behavior and the stochastic properties of data produced under that functional behavior, is easily established. This thus suggests that the critiquing views rather relate to disagreements around whether the functional behavior that is looked at in an application, is critically of interest to policy.

Apart from definitional issues, the distinction between "good" predictors and causal effects is another central part of discussion. In many cases, researchers do not accept an empirical result to be causal, but settle by agreeing that the relationship that is found constitutes a good predictor. From the point of view currently presented, it is not acceptable that a suboptimal predictor could in fact be a better candidate for the causal description of the mechanisms that produced the data. An empirical model of reality found by a distance-minimization process, attains the status of the one closest to the *true* model. Proofs that sample averages approach their infinite counterparts, are among the most fundamental results in probability theory. In practice there may be various violations to the required regularity conditions for the convergence of a criterion function, and attention must be paid to ensure that empirical models are constructed in an appropriate manner. The *true* probability measure, however, is by definition the optimal description of observed data sequences when tested infinitely many times against other ones, and doing away the result that is closest to this description as merely predictive, and not as one that is close to the causal origin of the observed data, seems therefore a flawed attack.

Still, economics has been criticized to not deliver on a number of important prediction problems, even though economists, disregard their differences

in views on causality, have paid important attention to uncover causal relationships in their analyses of economic systems. Some examples include not being able to accurately predict a downturn in markets or find a definitive answer to the relationship between employment and government expenditures. The argument that not observing an outcome that was predicted by a supposedly causal model, invalidates the causal claim, is naturally flawed as well. A prediction made with the correct probability measure of a dice is only correct in the frequency domain – e.g., one out of six for an ordinary dice. In a similar manner, we would say that stress and bad lifestyle habits cause increased risk of a heart attack, which is similarly a probabilistic statement that provides accurate predictions only in the frequency domain. The optimal, causal, predictor must hence always be understood as the predictor that minimizes distance between predicted probability of occurrence and the *true* future probabilistic occurrence, and those laws will only ever be correct within the frequency domain.

## Proofs

### Proof for Proposition 1.

*Proof.* By construction of the criterion as stated in ASSUMPTION. 13, $\arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty(\boldsymbol{\theta})$ 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 $d^*_F(f^{\mathbf{w}}, f(\cdot, \boldsymbol{\theta}))\ \forall\ \boldsymbol{\theta} \in \Theta$ providing that item 3 is equivalent to item 2.

□

**Proof for Proposition 2.**

*Proof.* The result follows immediately by the arguments used in PROPOSITION. 6 dropping only the first equivalence. □

**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 $-\ln \int \left[ \sqrt{\frac{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})}} p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) \right]$ and take expectations to get

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

Note that

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

by Jensen's inequality.

Finally, $\mathbb{E} \left[ \ln \sqrt{\frac{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})}} p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) \right]$ can be written as

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

where the last expression is equivalent to Kullback-Leibler divergence by an elementary row operation

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

□