Causality in Time Series

Challenges in Machine Learning, Volume 5

Causality in Time Series

Challenges in Machine Learning, Volume 5

Florin Popescu and Isabelle Guyon, editors

Nicola Talbot, production editor

Microtome Publishing
Brookline, Massachusetts
www.mtome.com

Causality in Time Series
Challenges in Machine Learning, Volume 5

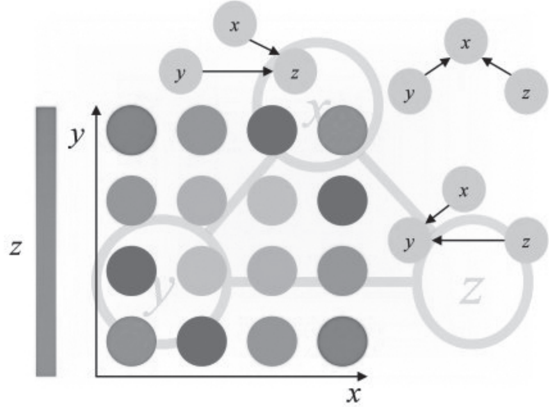
Florin Popescu and Isabelle Guyon, editors

Nicola Talbot, production editor

ISBN-13: 978-0-9719777-5-4

Causality Workbench

⟨http://clopinet.com/causality⟩



Foreword

The topic of causality has been subject to a lengthy academic debate in Western science and philosophy, as it forms the linchpin of systematic scientific explanations of nature and the basis of rational economic policy. The former dates back to Aristotle's momentous separation of inductive and deductive reasoning – as the inductive reasoning has lacked tools (statistics) to support its conclusions on a formal, objective basis, it has long taken a backseat to the rigor of logical deductive reasoning. Despite the 20th century rise to prominence of statistics, initially intended to resolve causal quandaries in agricultural and industrial process refinement, the field of statistical causal inference is relatively young. Although its pioneers have received wide praise (Clive Granger receiving the Nobel Prize and Judea Pearl receiving the ACM Turing Award) the methods they have developed are not yet widely known and are still subject to refinement. Although one of the least controversial necessary criterion of establishing a cause-effect is temporal precedence, this type of inference does not require time information – rather, it aims to establish possible causal relations among observations on other (logical) grounds based on conditional independence testing. The work of Clive Granger, built upon the 20th century development of time series modeling in engineering and economics, with some input from physiology, leads to a framework which admittedly does not allow us to identify causality unequivocally.

At the time of the Neural Information Processing Systems (NIPS 2009) Mini-Symposium on Time Series Causality (upon which this volume is based), there had been scant interaction among the Machine Learning researchers who undertake the annual pilgrimage to NIPS and the economists, engineers and neuro-physiologists who not only require causal inference methods, but also help develop them. Following the highly successful 2008 NIPS Causality Workshop (organized by Isabelle Guyon and featuring, among others, Judea Pearl), it was decided to follow-up with a symposium the following year aiming precisely to present related work by non- 'machine learners' to this community . The symposium presented current state-of-the-art and helped suggest future means of cross-disciplinary collaboration, while also featuring a tribute to the work of the late Clive Granger by his former friend and colleague Hal White. This work therefore presents an interdisciplinary exposition of both methodological challenges and recent innovations.

The chapter of White and Chalak presents a detailed formal exposition of causal inference in econometrics and, very importantly, provides the long awaited link between the time-series causality work of Clive Granger (based on relative information of the present/past states of a time series pair) and the Pearl-type inference based on conditional information (focused on triads or three-way dependence among variables), as well as a practical exposition of a testing procedure that expands the classical errors of type I and II in traditional statistics (causality is directional and therefore is not a simply a question of accepting or rejecting the null hypothesis).

In the chapter of Popescu, we attempted to introduce a re-formulation of Granger causality int terms of algorithmic information theory and also to formulate causal hypothesis testing subject to three (rather than two) error types. Building upon prior work in electro-physiology, notably the Phase Slope Index, PSI, introduced by Guido Nolte, which is robust to sub-sampling time interactions among variables, i.e. aliasing, we attempted to express a conditional Grange Causality test in auto-regressive terms familiar to engineers and economists analogously to the Fourier domain familiar to electro-physiology and physics in which PSI has hitherto been expressed. Finally, we attempted to extend PSI and AR-based modeling to three-way interactions among variables subject to aliasing effects while presenting comparative numerical results of a simulated electro-physiology problem subject to the aforementioned inference errors.

The chapter of Roerbroeck and collaborators presents the particular causal inference challenges of a wide-ranging recording technique in brain science, namely functional magnetic resonance imaging (fMRI), which affords a unique opportunity to non-invasively record intact whole brain activity, but which is hampered by relatively weak time-information accuracy relative to a high spatial dimensionality of the recording, due to both the relatively slow physiological oxygenation process it records and the measurement process itself. Granger causality and statistical pre-processing techniques for meaningful inference are presented as well as directions for future research.

The chapter of Moneta and collaborators presents an exciting mix of traditional econometrics and machine-learning techniques in which the dynamic dependence among time-series variables is auto-regressively modeled such that Pearl-type causal inference may be performed on the presumably uncorrelated residual vector, using independent component analysis and/or kernel-based density estimates for the conditional independence test.

As the overall aim of this volume is to present state-of-the-art research in time-series causality to the Machine Learning community as well as unify methodological interests in the various communities that require such inference, it was important to provide some relevant sample datasets upon which novel methods may be tested. The chapter of Guyon and collaborators introduces these datasets and the repository and pre-processing and analysis software which is available on the Causality Workbench web-site to aid future work. The datasets come from a wide-ranging set of sources (manufacturing, marketing, physiology among them) and can be expanded with other datasets provided by users, therefore allowing for novel method development and testing as well as targeted course-work.

Florin C. Popescu 05/28/12
Fraunhofer Institute FIRST

Acknowledgments

We would like to thank Guido Nolte for his support and fruitful discussions. We would also like to thank Pascal2 EU Network of Excellence and the NIPS foundation for supporting the NIPS Mini-symposium on Time Series Causality.

Preface

This book reprints papers of the Mini Symposium on Causality in Time Series, which was part of the Neural Information Processing Systems 2009 (NIPS 2009), December 10, 2009, Vancouver, Canada. The papers were initially published on-line in JMLR Workshop and Conference proceedings (JMLR W&CP), Volume 12: http://jmlr.csail.mit.edu/proceedings/papers/v12/.

The Editorial Team:

Florin Popescu
Fraunhofer Institute FIRST, Berkin
Florin.popescu@first.fraunhofer.de

Isabelle Guyon
Clopinet, California, USA
guyon@clopinet.com

Table of Contents

Time Series Analysis with the Causality Workbench

Isabelle Guyon ISABELLE@CLOPINET.COM
ClopiNet, Berkeley, California

Alexander Statnikov ALEXANDER.STATNIKOV@NYUMC.ORG
NYU Langone Medical Center, New York city

Constantin Aliferis CONSTANTIN.ALIFERIS@NYUMC.ORG
NYU Center for Health Informatics and Bioinformatics, New York city

Editors: Florin Popescu and Isabelle Guyon

Abstract

The Causality Workbench project is an environment to test causal discovery algorithms. Via a web portal (`http://clopinet.com/causality`), it provides a number of resources, including a repository of datasets, models, and software packages, and a virtual laboratory allowing users to benchmark causal discovery algorithms by performing virtual experiments to study artificial causal systems. We regularly organize competitions. In this paper, we describe what the platform offers for the analysis of causality in time series analysis.

Keywords: Causality, Benchmark, Challenge, Competition, Time Series Prediction.

1. Introduction

Uncovering cause-effect relationships is central in many aspects of everyday life in both highly industrialized and developing countries: what affects our health, the economy, climate changes, world conflicts, and which actions have beneficial effects? Establishing causality is critical to guiding policy decisions in areas including medicine and pharmacology, epidemiology, climatology, agriculture, economy, sociology, law enforcement, and manufacturing.

One important goal of causal modeling is to predict the consequences of given *actions*, also called *interventions*, *manipulations* or *experiments*. This is fundamentally different from the classical machine learning, statistics, or data mining setting, which focuses on making predictions from observations. Observations imply no manipulation on the system under study whereas actions introduce a disruption in the natural functioning of the system. In the medical domain, this is the distinction made between "diagnosis" and "prognosis" (prediction from observations of diseases or disease evolution) and "treatment" (intervention). For instance, smoking and coughing might be both predictive of respiratory disease and helpful for diagnosis purposes. However, if smoking is a cause and coughing a consequence, acting on the cause (smoking) can change your health status, but not acting on the symptom or consequence (coughing). Thus it is extremely important to distinguish between causes and consequences to predict the result of actions like predicting the effect of forbidding smoking in public places.

The need for assisting policy making while reducing the cost of experimentation and the availability of massive amounts of “observational” data prompted the proliferation of proposed computational causal discovery techniques (Glymour and Cooper, 1999; Pearl, 2000; Spirtes et al., 2000; Neapolitan, 2003; Koller and Friedman, 2009), but it is fair to say that to this day, they have not been widely adopted by scientists and engineers. Part of the problem is the lack of appropriate evaluation and the demonstration of the usefulness of the methods on a range of pilot applications. To fill this need, we started a project called the ”Causality Workbench”, which offers the possibility of exposing the research community to challenging causal problems and disseminating newly developed causal discovery technology. In this paper, we outline our setup and methods and the possibilities offered by the Causality Workbench to solve problems of causal inference in time series analysis.

2. Causality in Time Series

Causal discovery is a multi-faceted problem. The definition of causality itself has eluded philosophers of science for centuries, even though the notion of causality is at the core of the scientific endeavor and also a universally accepted and intuitive notion of everyday life. But, the lack of broadly acceptable definitions of causality has not prevented the development of successful and mature mathematical and algorithmic frameworks for inducing causal relationships.

Causal relationships are frequently modeled by causal Bayesian networks or structural equation models (SEM) (Pearl, 2000; Spirtes et al., 2000; Neapolitan, 2003). In the graphical representation of such models, an arrow between two variables $A \rightarrow B$ indicates the direction of a causal relationship: A causes B. A node in the graph corresponding to a particular variable X, represents a “mechanism” to evaluate the value of X, given the “parent” node variable values (immediate antecedents in the graph). For Bayesian networks, such evaluation is carried out by a conditional probability distribution $P(X|Parents(X))$ while for structural equation models it is carried out by a function of the parent variables and a noise model.

Our everyday-life concept of causality is very much linked to time dependencies (causes precede their effects). Hence an intuitive interpretation of an arrow in a causal network representing A causes B is that A preceded B.[1] But, in reality, Bayesian networks are a graphical representation of a factorization of conditional probabilities, hence a pure mathematical construct. The arrows in a “regular” Bayesian network (not a “causal Bayesian network”) do not necessarily represent either causal relationships nor precedence, which often creates some confusion. In particular, many machine learning problems are concerned with stationary systems or “cross-sectional studies”, which are

[1] More precise semantics have been developed. Such semantics assume discrete time point or interval time models and allow for continuous or episodic “occurrences” of the values of a variable as well as overlapping or non-overlapping intervals (Aliferis, 1998). Such practical semantics in Bayesian networks allow for abstracted and explicit time.

studies where many samples are drawn at a given point in time. Thus, sometimes the reference to time in Bayesian networks is replaced by the notion of "causal ordering". Causal ordering can be understood as fixing a particular time scale and considering only causes happening at time t and effects happening at time $t+\delta t$, where δt can be made as small as we want. Within this framework, causal relationships may be inferred from data including no explicit reference to time. Causal clues in the absence of temporal information include *conditional independencies* between variables and loss of information due to *irreversible transformations* or the *corruption of signal by noise* (Sun et al., 2006; Zhang and Hyvärinen, 2009).

In seems reasonable to think that temporal information should resolve many causal relationship ambiguities. Yet, the addition of the time dimension simplifies the problem of inferring causal relationships only to a limited extend. For one, it reduces, but does not eliminate, the problem of confounding: A correlated event A happening in the past of event B cannot be a consequence of B; however it is not necessarily a cause because a previous event C might have been a "common cause" of A and B. Secondly, it opens the door to many subtle modeling questions, including problems arising with modeling the dynamic systems, which may or may not be stationary. One of the charters of our Causality Workbench project is to collect both problems of practical and academic interest to push the envelope of research in inferring causal relationships from time series analysis.

3. A Virtual Laboratory

Methods for learning cause-effect relationships without experimentation (learning from observational data) are attractive because observational data is often available in abundance and experimentation may be costly, unethical, impractical, or even plain impossible. Still, many causal relationships cannot be ascertained without the recourse to experimentation and the use of a mix of observational and experimental data might be more cost effective. We implemented a *Virtual Lab* allowing researchers to perform experiments on artificial systems to infer their causal structure. The design of the platform is such that researchers can submit new artificial systems for others to experiment, experimenters can place queries and get answers, the activity is logged, and registered users have their own virtual lab space. This environment allows researchers to test computational causal discovery algorithms and, in particular, to test whether modeling assumptions made hold in real and simulated data.

We have released a first version `http://www.causality.inf.ethz.ch/workbench.php`. We plan to attach to the virtual lab sizeable realistic simulators such as the Spatiotemporal Epidemiological Modeler (STEM), an epidemiology simulator developed at IBM, now publicly available: `http://www.eclipse.org/stem/`. The virtual lab was put to work in a recent challenge we organized on the problem of "Active Learning" (see `http://clopinet.com/al`). More details on the virtual lab are given in the appendix.

4. A Data Repository

Part of our benchmarking effort is dedicated to collecting problems from diverse application domains. Via the organization of competitions, we have successfully channeled the effort or dozens of researchers to solve new problems of scientific and practical interest and identified effective methods. However, competition without collaboration is sterile. Recently, we have started introducing new dimensions to our effort of research coordination: stimulating creativity, collaborations, and data exchange. We are organizing regular teleconference seminars. We have created a data repository for the Causality Workbench already populated by 15 datasets. All the resources, which are the product of our effort, are freely available on the Internet at `http://clopinet.com/causality`. The repository already includes several time series datasets, illustrating problems of practical and academic interest (see table 1):

- Learning the structure of a fairly complex dynamic system that disobeys equilibration-manipulation commutability, and predicting the effect of manipulations that do not cause instabilities (the MIDS task) (Voortman et al., 2010);
- Learning causal relationships using time series when noise is corrupting data in a way that classical "Granger causality" fails (the NOISE task) (Nolte et al., 2010);
- Uncovering which promotions affect most sales in a marketing database (the PROMO task) (Pellet, 2010);
- Identifying in a manufacturing process (wafer production) faulty steps affecting a performance metric (the SEFTI task) (Tuv, 2010);
- Modeling a biological signalling process (the SIGNET task) (Jenkins, 2010).

The donor of the dataset NOISE (Guido Nolte) received the best dataset award. The reviewers appreciated that the task includes both real data from EEG time series and artificial data modeling EEG. We want to encourage future data donors to move in this direction.

5. Benchmarks and Competitions

Our effort has been gaining momentum with the organization of two challenges, which each attracted over 50 participants. The first causality challenge we have organized (Causation and Prediction challenge, December 15 2007 – April 30 2008) allowed researchers both from the causal discovery community and the machine learning community to try their algorithms on sizable tasks of real practical interest in medicine, pharmacology, and sociology (see `http://www.causality.inf.ethz.ch/challenge.php`). The goal was to train models exclusively on observational data, then make predictions of a target variable on data collected after intervention on the system under study were performed. This first challenge reached a number of goals that we had set to ourselves: familiarizing many new researchers and practitioners with causal discovery problems and existing tools to address them, pointing out the limitations of current methods on some particular difficulties, and fostering the development

Table 1: **Time dependent datasets.**"TP" is the data type, "NP" the number of participants who returned results and "V" the number of views as of January 2011. The semi-artificial datasets are obtained from simulators of real tasks. N is the number of variables, T is the number of time samples (not necessarily evenly spaced) and R the number of simulations with different initial states or conditions.

Name (TP; NP; V)	Size	Description	Objective
MIDS (Artificial; NA; 794)	T=12 sampled values in time (unevenly spaced); R=10000 simulations. N=9 variables.	**Mixed Dynamic Systems.** Simulated time-series based on linear Gaussian models with no latent common causes, but with multiple dynamic processes.	Use the training data to build a model able to predict the effects of manipulations on the system in test data.
NOISE (Real + artificial; NA; 783)	**Artificial:** T=6000 time points; R=1000 simul.; N=2 var. **Real:** R=10 subjects. T≃200000 points sampled at 256Hz. N=19 channels.	**Real and simulated EEG data.** Learning causal relationships using time series when noise is corrupting data causing the classical Granger causality method to fail.	**Artificial task:** find the causal dir. in pairs of var. **Real task:** Find which brain region influence each other.
PROMO (Semi-artificial; 3; 1601)	T=365*3 days; R=1 simulation; N=1000 promotions + 100 products.	**Simulated marketing task.** Daily values of 1000 promotions and 100 product sales for three years incorporating seasonal effects.	Predict a 1000x100 boolean matrix of causal influences of promotions on product sales.
SEFTI (Semi-artificial; NA; 908)	R=4000 manufacturing lots; T=300 async. operations (pair of values {one of N=25 tool IDs, date of proc.}) + cont. target (circuit perf. for each lot).	**Semiconductor manufacturing.** Each wafer undergoes 300 steps each involving one of 25 tools. A regression problem for quality control of end-of-line circuit performance.	Find the tools that are guilty of performance degradation and eventual interactions and influence of time.
SIGNET (Semi-artif.; 2; 2663)	T=21 asynchronous state updates; R=300 pseudodynamic simulations; N=43 rules.	**Abscisic Acid Signaling Network.** Model inspired by a true biological signaling network.	Determine the set of 43 boolean rules that describe the network.

of new algorithms. The results indicated that causal discovery from observational data is not an impossible task, but a very hard one and pointed to the need for further research and benchmarks (Guyon et al., 2008). The Causal Explorer package (Aliferis et al., 2003), which we had made available to the participants and is downloadable as shareware, proved to be competitive and is a good starting point for researchers new to the field. It is a Matlab toolkit supporting "local" causal discovery algorithms, efficient to discover the causal structure around a target variable, even for a large number of variables. The algorithms are based on structure learning from tests of conditional independence, as all the top ranking methods in this first challenge.

The first challenge (Guyon et al., 2008) explored an important problem in causal modeling, but is only one of many possible problem statements. The second challenge (Guyon et al., 2010) called "competition pot-luck" aimed at enlarging the scope of causal discovery algorithm evaluation by inviting members of the community to submit their own problems and/or solve problems proposed by others. The challenge started September 15, 2008 and ended November 20, 2008, see `http://www.causality.inf.ethz.ch/pot-luck.php`. One task proposed by a participant drew a lot of attention: the cause-effect pair task. The problem was to try to determine in pairs of variables (of known causal relationships), which one was the cause of the other. This problem is hard for a lot of algorithms, which rely on the result of conditional independence tests of three or more variables. Yet the winners of the challenge succeeded in unraveling 8/8 correct causal directions (Zhang and Hyvärinen, 2009).

Our planned challenge ExpDeCo (Experimental Design in Causal Discovery) will benchmark methods of experimental design in application to causal modeling. The goal will be to identify effective methods to unravel causal models, requiring a minimum of experimentation, using the Virtual Lab. A budget of virtual cash will be allocated to participants to "buy" the right to observe or manipulate certain variables, manipulations being more expensive that observations. The participants will have to spend their budget optimally to make the best possible predictions on test data. This setup lends itself to incorporating problems of relevance to development projects, in particular in medicine and epidemiology where experimentation is difficult while developing new methodology.

We are planning another challenge called CoMSICo for "Causal Models for System Identification and Control", which is more ambitious in nature because it will perform a continuous evaluation of causal models rather than separating training and test phase. In contrast with ExpDeCo in which the organizers will provide test data with prescribed manipulations to test the ability of the participants to make predictions of the consequences of actions, in CoMSICo, the participants will be in charge of making their own plan of action (policy) to optimize an overall objective (e.g., improve the life expectancy of a population, improve the GNP, etc.) and they will be judged directly with this objective, on an on-going basis, with no distinction between "training" and "test" data. This challenge will also be via the Virtual Lab. The participants will be given an initial amount of virtual cash, and, as previously, both actions and observations will have a price. New in CoMSICo, virtual cash rewards will be given for achieving good intermediate performance, which the participants will be allowed to re-invest to conduct additional experiments and improve their plan of action (policy). The winner will be the participant ending up with the largest amount of virtual cash.

6. Conclusion

Our program of data exchange and benchmark proposes to challenge the research community with a wide variety of problems from many domains and focuses on realistic settings. Causal discovery is a problem of fundamental and practical interest in many areas of science and technology and there is a need for assisting policy making in all these areas while reducing the costs of data collection and experimentation. Hence, the identification of efficient techniques to solve causal problems will have a widespread impact. By choosing applications from a variety of domains and making connections between disciplines as varied as machine learning, causal discovery, experimental design, decision making, optimization, system identification, and control, we anticipate that there will be a lot of cross-fertilization between different domains.

Acknowledgments

This project is an activity of the Causality Workbench supported by the Pascal network of excellence funded by the European Commission and by the U.S. National Science Foundation under Grant N0. ECCS-0725746. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We are very grateful to all the members of the causality workbench team for their contribution and in particular to our co-founders Constantin Aliferis, Greg Cooper, André Elisseeff, Jean-Philippe Pellet, Peter Spirtes, and Alexander Statnikov.

References

C. F. Aliferis, I. Tsamardinos, A. Statnikov, and L.E. Brown. Causal explorer: A probabilistic network learning toolkit for biomedical discovery. In *2003 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS)*, Las Vegas, Nevada, USA, June 23-26 2003. CSREA Press.

Constantin Aliferis. *A Temporal Representation and Reasoning Model for Medical Decision-Support Systems*. PhD thesis, University of Pittsburgh, 1998.

C. Glymour and G.F. Cooper, editors. *Computation, Causation, and Discovery*. AAAI Press/The MIT Press, Menlo Park, California, Cambridge, Massachusetts, London, England, 1999.

I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov. Design and analysis of the causation and prediction challenge. In *JMLR W&CP*,

volume 3, pages 1–33, WCCI2008 workshop on causality, Hong Kong, June 3-4 2008.

I. Guyon, D. Janzing, and B. Schölkopf. Causality: Objectives and assessment. *JMLR W&CP*, 6:1–38, 2010.

Jerry Jenkins. Signet: Boolean rile determination for abscisic acid signaling. In *Causality: objectives and assessment (NIPS 2008)*, volume 6, pages 215–224. JMLR W&CP, 2010.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

R. E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall series in Artificial Intelligence. Prentice Hall, 2003.

G. Nolte, A. Ziehe, N. Krämer, F. Popescu, and K.-R. Müller. Comparison of granger causality and phase slope index. In *Causality: objectives and assessment (NIPS 2008)*, volume 6, pages 267–276. JMLR W&CP, 2010.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

J.-P. Pellet. Detecting simple causal effects in time series. In *Causality: objectives and assessment (NIPS 2008)*. JMLR W&CP volume 6, supplemental material, 2010.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, Cambridge, Massachusetts, London, England, 2000.

X. Sun, D. Janzing, and B. Schölkopf. Causal inference by choosing graphs with most plausible Markov kernels. In *Ninth International Symposium on Artificial Intelligence and Mathematics*, 2006.

E. Tuv. Pot-luck challenge: Tied. In *Causality: objectives and assessment (NIPS 2008)*. JMLR W&CP volume 6, supplemental material, 2010.

M. Voortman, D. Dash, and M. J. Druzdzel. Learning causal models that make correct manipulation predictions. In *Causality: objectives and assessment (NIPS 2008)*, volume 6, pages 257–266. JMLR W&CP, 2010.

K. Zhang and A. Hyvärinen. Distinguishing causes from effects using nonlinear acyclic causal models. In *Causality: objectives and assessment (NIPS 2008)*, volume 6, pages 157–164. JMLR W&CP, 2009.

Causal Search in Structural Vector Autoregressive Models

Alessio Moneta MONETA@ECON.MPG.DE
Max Planck Institute of Economics
Jena, Germany

Nadine Chlaß NADINE.CHLASS@UNI-JENA.DE
Friedrich Schiller University of Jena, Germany

Doris Entner DORIS.ENTNER@CS.HELSINKI.FI
Helsinki Institute for Information Technology, Finland

Patrik Hoyer PATRK.HOYER@HELSINKI.FI
Helsinki Institute for Information Technology, Finland

Editors: Florin Popescu and Isabelle Guyon

Abstract

This paper reviews a class of methods to perform causal inference in the framework of a structural vector autoregressive model. We consider three different settings. In the first setting the underlying system is linear with normal disturbances and the structural model is identified by exploiting the information incorporated in the partial correlations of the estimated residuals. Zero partial correlations are used as input of a search algorithm formalized via graphical causal models. In the second, semi-parametric, setting the underlying system is linear with non-Gaussian disturbances. In this case the structural vector autoregressive model is identified through a search procedure based on independent component analysis. Finally, we explore the possibility of causal search in a nonparametric setting by studying the performance of conditional independence tests based on kernel density estimations.

Keywords: Causal inference, econometric time series, SVAR, graphical causal models, independent component analysis, conditional independence tests

1. Introduction

1.1. Causal inference in econometrics

Applied economic research is pervaded by questions about causes and effects. For example, what is the effect of a monetary policy intervention? Is energy consumption causing growth or the other way around? Or does causality run in both directions? Are economic fluctuations mainly caused by monetary, productivity, or demand shocks? Does foreign aid improve living standards in poor countries? Does firms' expenditure in R&D causally influence their profits? Are recent rises in oil prices in part caused by speculation? These are seemingly heterogeneous questions, but they all require some knowledge of the causal process by which variables came to take the values we observe.

A traditional approach to address such questions hinges on the explicit use of *a priori* economic theory. The gist of this approach is to partition a causal process in a deterministic, and a random part and to articulate the deterministic part such as to reflect the causal dependencies dictated by economic theory. If the formulation of the deterministic part is accurate and reliable enough, the random part is expected to display properties that can easily be analyzed by standard statistical tools. The touchstone of this approach is represented by the work of Haavelmo (1944), which inspired the research program subsequently pursued by the Cowles Commission (Koopmans, 1950; Hood and Koopmans, 1953). Therein, the causal process is formalized by means of a structural equation model, that is, a system of equations with endogenous variables, exogenous variables, and error terms, first developed by Wright (1921). Its coefficients were given a causal interpretation (Pearl, 2000).

This approach has been strongly criticized in the 1970s for being ineffective in both policy evaluation and forecasting. Lucas (1976) pointed out that the economic theory included in the SEM fails to take economic agents' (rational) motivations and expectations into consideration. Agents, according to Lucas, are able to anticipate policy intervention and act contrary to the prediction derived from the structural equation model, since the model usually ignores such anticipations. Sims (1980) puts forth another critique which runs parallel to Lucas' one. It explicitly addresses the status of exogeneity which the Cowles Commission approach attributes (arbitrarily, according to Sims) to some variables such that the structural model can be identified. Sims argues that theory is not a reliable source for deeming a variable as exogenous. More generally, the Cowles Commission approach with its strong *a priori* commitment to theory, risks falling into a vicious circle: if causal information (even if only about direction) can exclusively be derived from background theory, how do we obtain an empirically justified theory? (Cfr. Hoover, 2006, p.75).

An alternative approach has been pursued since Wiener (1956) and Granger's (1969) work. It aims at inferring causal relations directly from the statistical properties of the data relying only to a minimal extent on background knowledge. Granger (1980) proposes a probabilistic concept of causality, similar to Suppes (1970). Granger defines causality in terms of the incremental predictability (at horizon one) of a time series variable $\{Y_t\}$ (given the present and past values of $\{Y_t\}$ and of a set $\{Z_t\}$ of possible relevant variables) when another time series variable $\{X_t\}$ (in its present and past values) is not omitted. More formally:

$$\{X_t\} \text{ Granger-causes } \{Y_t\} \text{ if } P(Y_{t+1}|X_t, X_{t-1}, \ldots, Y_t, Y_{t-1}, \ldots, Z_t, Z_{t-1}, \ldots) \neq P(Y_{t+1}|Y_t, Y_{t-1}, \ldots, Z_t, Z_{t-1}, \ldots) \tag{1}$$

As pointed out by Florens and Mouchart (1982), testing the hypothesis of Granger noncausality corresponds to testing conditional independence. Given lags p, $\{X_t\}$ does not Granger cause $\{Y_t\}$, if

$$Y_{t+1} \perp\!\!\!\perp (X_t, X_{t-1}, \ldots, X_{t-p}) \mid (Y_t, Y_{t-1}, \ldots, Y_{t-p}, Z_t, Z_{t-1}, \ldots, Z_{t-p}) \tag{2}$$

To test Granger noncausality, researchers often specify linear vector autoregressive (VAR) models:

$$\mathbf{Y}_t = \mathbf{A}_1\mathbf{Y}_{t-1} + \ldots + \mathbf{A}_p\mathbf{Y}_{t-p} + \mathbf{u}_t, \tag{3}$$

in which $\mathbf{Y}_t$ is a $k \times 1$ vector of time series variables $(Y_{1,t}, \ldots, Y_{k,t})'$, where $()'$ is the transpose, the $\mathbf{A}_j$ $(j = 1, \ldots, p)$ are $k \times k$ coefficient matrices, and $\mathbf{u}_t$ is the $k \times 1$ vector of random disturbances. In this framework, testing the hypothesis that $\{Y_{i,t}\}$ does not Granger-cause $\{Y_{j,t}\}$, reduces to test whether the (j,i) entries of the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_p$ are vanishing simultaneously. Granger noncausality tests have been extended to nonlinear settings by Baek and Brock (1992), Hiemstra and Jones (1994), and Su and White (2008), using nonparametric tests of conditional independence (more on this topic in section 4).

The concept of Granger causality has been criticized for failing to capture 'structural causality' (Hoover, 2008). Suppose one finds that a variable A Granger-causes another variable B. This does not necessarily imply that an economic mechanism exists by which A can be manipulated to affect B. The existence of such a mechanism in turn does not necessarily imply Granger causality either (for a discussion see Hoover 2001, pp. 150-155). Indeed, the analysis of Granger causality is based on coefficients of reduced-form models, like those incorporated in equation (3), which are unlikely to reliably represent actual economic mechanisms. For instance, in equation (3) the simultaneous causal structure is not modeled in order to facilitate estimation. (However, note that Eichler (2007) and White and Lu (2010) have recently developed and formalized richer structural frameworks in which Granger causality can be fruitfully analyzed.)

1.2. The SVAR framework

Structural vector autoregressive (SVAR) models constitute a middle way between the Cowles Commission approach and the Granger-causality approach. SVAR models aim at recovering the concept of structural causality, but eschew at the same time the strong 'apriorism' of the Cowles Commission approach. The idea is, like in the Cowles Commission approach, to articulate an unobserved structural model, formalized as a dynamic generative model: at each time unit the system is affected by unobserved innovation terms, by which, once filtered by the model, the variables come to take the values we observe. But, differently from the Cowles Commission approach, and similarly to the Granger-VAR model, the data generating process is generally enough articulated so that time series variables are *not* distinguished *a priori* between exogenous and endogenous. A linear SVAR model is in principle a VAR model 'augmented' by the contemporaneous structure:

$$\mathbf{\Gamma}_0\mathbf{Y}_t = \mathbf{\Gamma}_1\mathbf{Y}_{t-1} + \ldots + \mathbf{\Gamma}_p\mathbf{Y}_{t-p} + \boldsymbol{\varepsilon}_t. \tag{4}$$

This is easily obtained by pre-multiplying each side of the VAR model

$$\mathbf{Y}_t = \mathbf{A}_1\mathbf{Y}_{t-1} + \ldots + \mathbf{A}_p\mathbf{Y}_{t-p} + \mathbf{u}_t, \tag{5}$$

by a matrix $\mathbf{\Gamma}_0$ so that $\mathbf{\Gamma}_i = \mathbf{\Gamma}_0\mathbf{A}_i$, for $i = 1, \ldots, k$ and $\boldsymbol{\varepsilon}_t = \mathbf{\Gamma}_0\mathbf{u}_t$. Note, however, that not *any* matrix $\mathbf{\Gamma}_0$ will be suitable. The appropriate $\mathbf{\Gamma}_0$ will be that matrix corresponding to the 'right' rotation of the VAR model, that is the rotation compatible both with the contemporaneous causal structure of the variable and the structure of the innovation term. Let us consider a matrix $\mathbf{B}_0 = \mathbf{I} - \mathbf{\Gamma}_0$. If the system is normalized such that the matrix $\mathbf{\Gamma}_0$ has all the elements of the principal diagonal equal to one (which can be done straightforwardly), the diagonal elements of $\mathbf{B}_0$ will be equal to zero. We can write:

$$\mathbf{Y}_t = \mathbf{B}_0 \mathbf{Y}_t + \mathbf{\Gamma}_1 \mathbf{Y}_{t-1} + \ldots + \mathbf{\Gamma}_p \mathbf{Y}_{t-p} + \boldsymbol{\varepsilon}_t \tag{6}$$

from which we see that $\mathbf{B}_0$ (and thus $\mathbf{\Gamma}_0$) determines in which form the values of a variable $Y_{i,t}$ will be dependent on the contemporaneous value of another variable $Y_{j,t}$. The 'right' rotation will also be the one which makes $\boldsymbol{\varepsilon}_t$ a vector of authentic innovation terms, which are expected to be independent (not only over time, but also contemporaneously) sources or shocks.

In the literature, different methods have been proposed to identify the SVAR model (4) on the basis of the estimation of the VAR model (5). Notice that there are more unobserved parameters in (4), whose number amounts to $k^2(p+1)$, than parameters that can be estimated from (5), which are $k^2 p + k(k+1)/2$, so one has to impose at least $k(k-1)/2$ restrictions on the system. One solution to this problem is to get a rotation of (5) such that the covariance matrix of the SVAR residuals $\mathbf{\Sigma}_\varepsilon$ is diagonal, using the Cholesky factorization of the estimated residuals $\mathbf{\Sigma}_\mathbf{u}$. That is, let $\mathbf{P}$ be the lower-triangular Cholesky factorization of $\mathbf{\Sigma}_\mathbf{u}$ (i.e. $\mathbf{\Sigma}_\mathbf{u} = \mathbf{PP}'$), let $\mathbf{D}$ be a $k \times k$ diagonal matrix with the same diagonal as $\mathbf{P}$, and let $\mathbf{\Gamma}_0 = \mathbf{DP}^{-1}$. By pre-multiplying (5) by $\mathbf{\Gamma}_0$, it turns out that $\mathbf{\Sigma}_\varepsilon = E[\mathbf{\Gamma}_0 \mathbf{u}_t \mathbf{u}_t' \mathbf{\Gamma}_0'] = \mathbf{DD}'$, which is diagonal. A problem with this method is that $\mathbf{P}$ changes if the ordering of the variables $(Y_{1t}, \ldots, Y_{kt})'$ in $\mathbf{Y}_t$ and, consequently, the order of residuals in $\mathbf{\Sigma}_\mathbf{u}$, changes. Since researchers who estimate a SVAR are often exclusively interested on tracking down the effect of a structural shock ε_{it} on the variables $Y_{1,t}, \ldots, Y_{k,t}$ over time (*impulse response functions*), Sims (1981) suggested investigating to what extent the impulse response functions remain robust under changes of the order of variables.

Popular alternatives to the Cholesky identification scheme are based either on the use of *a priori*, theory-based, restrictions or on the use of long-run restrictions. The former solution consists in imposing economically plausible constraints on the contemporaneous interactions among variables (Blanchard and Watson, 1986; Bernanke, 1986) and has the drawback of ultimately depending on the *a priori* reliability of economic theory, similarly to the Cowles Commission approach. The second solution is based on the assumptions that certain economic shocks have long-run effect to other variables, but do not influence in the long-run the level of other variables (see Shapiro and Watson, 1988; Blanchard and Quah, 1989; King et al., 1991). This approach has been criticized as not being very reliable unless strong *a priori* restrictions are imposed (see Faust and Leeper, 1997).

In the rest of the paper, we first present a method, based on the graphical causal model framework, to identify the SVAR (section 2). This method is based on conditional independence tests among the estimated residuals of the VAR estimated model. Such tests rely on the assumption that the shocks affecting the model are Gaussian.

We then relax the Gaussianity assumption and present a method to identify the SVAR model based on independent component analysis (section 3). Here the main assumption is that shocks are non-Gaussian and independent. Finally (section 4), we explore the possibility of extending the framework for causal inference to a nonparametric setting. In section 5 we wrap up the discussion and conclude by formulating some open questions.

2. SVAR identification via graphical causal models

2.1. Background

A data-driven approach to identify the structural VAR is based on the analysis of the estimated residuals $\hat{\mathbf{u}}_t$. Notice that when a basic VAR model is estimated (equation 3), the information about contemporaneous causal dependence is incorporated exclusively in the residuals (being not modeled among the variables). Graphical causal models, as originally developed by Pearl (2000) and Spirtes et al. (2000), represent an efficient method to recover, at least in part, the contemporaneous causal structure moving from the analysis of the conditional independencies among the estimated residuals. Once the contemporaneous causal structure is recovered, the estimation of the lagged autoregressive coefficients permits us to identify the complete SVAR model.

This approach was initiated by Swanson and Granger (1997), who proposed to test whether a particular causal order of the VAR is in accord with the data by testing all the partial correlations of order one among error terms and checking whether some partial correlations are vanishing. Reale and Wilson (2001), Bessler and Lee (2002), Demiralp and Hoover (2003), and Moneta (2008) extended the approach by using the partial correlations of the VAR residuals as input to graphical causal model search algorithms.

In graphical causal models, the structural model is represented as a causal graph (a *Directed Acyclic Graph* if the presence of causal loops is excluded), in which each node represents a random variable and each edge a causal dependence. Furthermore, a set of assumptions or 'rules of inference' are formulated, which regulate the relationship between causal and probabilistic dependencies: the *causal Markov* and the *faithfulness* conditions (Spirtes et al., 2000). The former restricts the joint probability distribution of modeled variables: each variable is independent of its graphical non-descendants conditional on its graphical parents. The latter makes causal discovery possible: all of the conditional independence relations among the modeled variables follow from the causal Markov condition. Thus, for example, if the causal structure is represented as $Y_{1t} \rightarrow Y_{2t} \rightarrow Y_{t,3}$, it follows from the Markov condition that $Y_{1,t} \perp\!\!\!\perp Y_{3,t}|Y_{2,t}$. If, on the other hand, the only (conditional) independence relation among $Y_{1,t}, Y_{2,t}, Y_{3,t}$ is $Y_{1,t} \perp\!\!\!\perp Y_{3,t}$, it follows from the faithfulness condition that $Y_{1,t} \rightarrow Y_{3,t} \longleftarrow Y_{2,t}$.

Constraint-based algorithms for causal discovery, like for instance, PC, SGS, FCI (Spirtes et al., 2000), or CCD (Richardson and Spirtes, 1999), use tests of conditional independence to constrain the possible causal relationships among the model variables. The first step of the algorithm typically involves the formation of a complete undirected graph among the variables so that they are all connected by an undirected edge. In a second step, conditional independence relations (or d-separations, which are the graphical characterization of conditional independence) are merely used to erase edges and, in further steps, to direct edges. The output of such algorithms are not necessarily one single graph, but a class of *Markov equivalent* graphs.

There is nothing neither in the Markov or faithfulness condition, nor in the constraint-based algorithms that limits them to linear and Gaussian settings. Graphical causal models do not require *per se* any a priori specification of the functional dependence between variables. However, in applications of graphical models to SVAR, conditional independence is ascertained by testing vanishing partial correlations (Swanson and Granger, 1997; Bessler and Lee, 2002; Demiralp and Hoover, 2003; Moneta, 2008). Since normal distribution guarantees the equivalence between zero partial correlation and conditional independence, these applications deal *de facto* with linear and Gaussian processes.

2.2. Testing residuals zero partial correlations

There are alternative methods to test zero partial correlations among the error terms $\hat{\mathbf{u}}_t = (u_{1t}, \ldots, u_{kt})'$. Swanson and Granger (1997) use the partial correlation coefficient. That is, in order to test, for instance, $\rho(u_{it}, u_{kt}|u_{jt}) = 0$, they use the standard t statistics from a least square regression of the model:

$$u_{it} = \alpha_j u_{jt} + \alpha_k u_{kt} + \varepsilon_{it}, \tag{7}$$

on the basis that $\alpha_k = 0 \Leftrightarrow \rho(u_{it}, u_{kt}|u_{jt}) = 0$. Since Swanson and Granger (1997) impose the partial correlation constraints looking only at the set of partial correlations of order one (that is conditioned on only one variable), in order to run their tests they consider regression equations with only two regressors, as in equation (7).

Bessler and Lee (2002) and Demiralp and Hoover (2003) use Fisher's z that is incorporated in the software TETRAD (Scheines et al., 1998):

$$z(\rho_{XY.\mathbf{K}}, T) = \frac{1}{2}\sqrt{T - |\mathbf{K}| - 3}\,\log\left(\frac{|1 + \rho_{XY.\mathbf{K}}|}{|1 - \rho_{XY.\mathbf{K}}|},\right) \tag{8}$$

where $|\mathbf{K}|$ equals the number of variables in $\mathbf{K}$ and T the sample size. If the variables (for instance $X = u_{it}$, $Y = u_{kt}$, $\mathbf{K} = (u_{jt}, u_{ht})$) are normally distributed, we have that

$$z(\rho_{XY.\mathbf{K}}, T) - z(\hat{\rho}_{XY.\mathbf{K}}, T) \sim N(0,1) \tag{9}$$

(see Spirtes et al., 2000, p.94).

A different approach, which takes into account the fact that correlations are obtained from residuals of a regression, is proposed by Moneta (2008). In this case it is useful to write the VAR model of equation (3) in a more compact form:

$$\mathbf{Y}_t = \mathbf{\Pi}'\mathbf{X}_t + \mathbf{u}_t, \tag{10}$$

where $\mathbf{X}_t' = [\mathbf{Y}_{t-1}', ..., \mathbf{Y}_{t-p}']$, which has dimension $(1 \times kp)$ and $\mathbf{\Pi}' = [\mathbf{A}_1, \ldots, \mathbf{A}_p]$, which has dimension $(k \times kp)$. In case of stable VAR process (see next subsection), the conditional maximum likelihood estimate of Π for a sample of size T is given by

$$\hat{\mathbf{\Pi}}' = \left[\sum_{t=1}^{T} \mathbf{Y}_t \mathbf{X}_t'\right]\left[\sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t'\right]^{-1}.$$

Moreover, the ith row of $\hat{\Pi}'$ is

$$\hat{\pi}_i' = \left[\sum_{t=1}^{T} Y_{it} \mathbf{X}_t'\right]\left[\sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t'\right]^{-1},$$

which coincides with the estimated coefficient vector from an OLS regression of Y_{it} on $\mathbf{X}_t$ (Hamilton 1994: 293). The maximum likelihood estimate of the matrix of variance and covariance among the error terms $\mathbf{\Sigma}_u$ turns out to be $\hat{\mathbf{\Sigma}}_u = (1/T)\sum_{t=1}^{T} \hat{\mathbf{u}}_t \hat{\mathbf{u}}_t'$, where $\hat{\mathbf{u}}_t = \mathbf{Y}_t - \hat{\mathbf{\Pi}}'\mathbf{X}_t$. Therefore, the maximum likelihood estimate of the covariance between u_{it} and u_{jt} is given by the (i, j) element of $\hat{\mathbf{\Sigma}}_u$: $\hat{\sigma}_{ij} = (1/T)\sum_{t=1}^{T} \hat{u}_{it}\hat{u}_{jt}$. Denoting by σ_{ij} the (i, j) element of $\mathbf{\Sigma}_u$, let us first define the following matrix transform operators: *vec*, which stacks the columns of a $k \times k$ matrix into a vector of length k^2 and *vech*, which vertically stacks the elements of a $k \times k$ matrix on or below the principal diagonal into a vector of length $k(k+1)/2$. For example:

$$\text{vec}\begin{bmatrix} \sigma_{11}\sigma_{12} \\ \sigma_{21}\sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_{11} \\ \sigma_{21} \\ \sigma_{12} \\ \sigma_{22} \end{bmatrix}, \quad \text{vech}\begin{bmatrix} \sigma_{11}\sigma_{12} \\ \sigma_{21}\sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_{11} \\ \sigma_{21} \\ \sigma_{22} \end{bmatrix}.$$

The process being stationary and the error terms Gaussian, it turns out that:

$$\sqrt{T}\,[\text{vech}(\hat{\mathbf{\Sigma}}_u) - \text{vech}(\mathbf{\Sigma}_u)] \xrightarrow{d} N(\mathbf{0}, \mathbf{\Omega}), \tag{11}$$

where $\mathbf{\Omega} = 2\mathbf{D}_k^+(\mathbf{\Sigma}_u \otimes \mathbf{\Sigma}_u)(\mathbf{D}_k^+)'$, $\mathbf{D}_k^+ \equiv (\mathbf{D}_k'\mathbf{D}_k)^{-1}\mathbf{D}_k'$, $\mathbf{D}_k$ is the unique $(k^2 \times k(k+1)/2)$ matrix satisfying $\mathbf{D}_k\text{vech}(\mathbf{\Omega}) = \text{vec}(\mathbf{\Omega})$, and $\otimes$ denotes the Kronecker product (see Hamilton 1994: 301). For example, for $k = 2$, we have,

$$\sqrt{T}\begin{bmatrix} \hat{\sigma}_{11} - \sigma_{11} \\ \hat{\sigma}_{12} - \sigma_{12} \\ \hat{\sigma}_{22} - \sigma_{22} \end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2\sigma_{11}^2 & 2\sigma_{11}\sigma_{12} & 2\sigma_{12}^2 \\ 2\sigma_{11}\sigma_{12} & \sigma_{11}\sigma_{22} + \sigma_{12}^2 & 2\sigma_{12}\sigma_{22} \\ 2\sigma_{12}^2 & 2\sigma_{12}\sigma_{22} & 2\sigma_{22}^2 \end{bmatrix}\right)$$

Therefore, to test the null hypothesis that $\rho(u_{it}, u_{jt}) = 0$ from the VAR estimated residuals, it is possible to use the Wald statistic:

$$\frac{T\,(\hat{\sigma}_{ij})^2}{\hat{\sigma}_{ii}\hat{\sigma}_{jj} + \hat{\sigma}_{ij}^2} \approx \chi^2(1).$$

The Wald statistic for testing vanishing partial correlations of any order is obtained by applying the delta method, which suggests that if X_T is a $(r \times 1)$ sequence of vector-valued random-variables and if $[\sqrt{T}(X_{1T} - \theta_1), \ldots, \sqrt{T}(X_{rT} - \theta_r)] \xrightarrow{d} N(\mathbf{0}, \Sigma)$ and $h_1, \ldots, h_r$ are r real-valued functions of $\theta = (\theta_1, \ldots, \theta_r)$, $h_i : \mathbf{R}^r \rightarrow \mathbf{R}$, defined and continuously differentiable in a neighborhood ω of the parameter point θ and such that the matrix $B = \|\partial h_i / \partial \theta_j\|$ of partial derivatives is nonsingular in ω, then:

$$[\sqrt{T}[h_1(X_T) - h_1(\theta)], \ldots, \sqrt{T}[h_r(X_T) - h_r(\theta)]] \xrightarrow{d} N(\mathbf{0}, B\Sigma B')$$

(see Lehmann and Casella 1998: 61).

Thus, for $k = 4$, suppose one wants to test $corr(u_{1t}, u_{3t}|u_{2t}) = 0$. First, notice that $\rho(u_1, u_3|u_2) = 0$ if and only if $\sigma_{22}\sigma_{13} - \sigma_{12}\sigma_{23} = 0$ (by definition of partial correlation). One can define a function $g : \mathbf{R}^{k(k+1)/2} \rightarrow \mathbf{R}$, such that $g(vech(\Sigma_u)) = \sigma_{22}\sigma_{13} - \sigma_{12}\sigma_{23}$. Thus,

$$\nabla g' = (0,\ -\sigma_{23},\ \sigma_{22},\ 0,\ \sigma_{13},\ -\sigma_{12},\ 0,\ 0,\ 0,\ 0).$$

Applying the delta method:

$$\sqrt{T}[(\hat{\sigma}_{22}\hat{\sigma}_{13} - \hat{\sigma}_{12}\hat{\sigma}_{23})\ -\ (\sigma_{22}\sigma_{13} - \sigma_{12}\sigma_{23})] \xrightarrow{d} N(0, \nabla g' \mathbf{\Omega} \nabla g).$$

The Wald test of the null hypothesis $corr(u_{1t}, u_{3t}|u_{2t}) = 0$ is given by:

$$\frac{T(\hat{\sigma}_{22}\hat{\sigma}_{13} - \hat{\sigma}_{12}\hat{\sigma}_{23})^2}{\nabla g' \Omega \nabla g} \approx \chi^2(1).$$

Tests for higher order correlations and for $k > 4$ follow analogously (see also Moneta, 2003). This testing procedure has the advantage, with respect to the alternative methods, to be straightforwardly applied to the case of cointegrated data, as will be explained in the next subsection.

2.3. Cointegration case

A typical feature of economic time series data in which there is some form of causal dependence is cointegration. This term denotes the phenomenon that nonstationary processes can have linear combinations that are stationary. That is, suppose that each component Y_{it} of $\mathbf{Y}_t = (Y_{1t}, \ldots, Y_{kt})'$, which follows the VAR process

$$\mathbf{Y}_t = \mathbf{A}_1 \mathbf{Y}_{t-1} + \ldots + \mathbf{A}_p \mathbf{Y}_{t-p} + \mathbf{u}_t,$$

is nonstationary and integrated of order one ($\sim I(1)$). This means that the VAR process $\mathbf{Y}_t$ is not *stable*, i.e. $\det(I_k - A_1 z - A_p z^p)$ is equal to zero for some $|z| \leq 1$ (Lütkepohl, 2006), and that each component ΔY_{it} of $\Delta \mathbf{Y}_t = (\mathbf{Y}_t - \mathbf{Y}_{t-1})$ is stationary ($I(0)$), that is it has time-invariant means, variances and covariance structure. A linear combination between between the elements of $\mathbf{Y}_t$ is called a *cointegrating relationship* if there is a linear combination $c_1 Y_{1t} + \ldots + c_k Y_{kt}$ which is stationary ($I(0)$).

If it is the case that the VAR process is unstable with the presence of cointegrating relationships, it is more appropriate (Lütkepohl, 2006; Johansen, 2006) to estimate the following re-parametrization of the VAR model, called Vector Error Correction Model (VECM):

$$\Delta\mathbf{Y}_t = \mathbf{F}_1\Delta\mathbf{Y}_{t-1} + \ldots + \mathbf{F}_{p-1}\Delta\mathbf{Y}_{t-p+1} - \mathbf{G}\mathbf{Y}_{t-p} + \mathbf{u}_t, \tag{12}$$

where $\mathbf{F}_i = -(\mathbf{I}_k - \mathbf{A}_1 - \ldots - \mathbf{A}_i)$, for $i = 1,\ldots,p-1$ and $\mathbf{G} = \mathbf{I}_k - \mathbf{A}_1 - \ldots - \mathbf{A}_p$. The $(k\times k)$ matrix $\mathbf{G}$ has rank r and thus $\mathbf{G}$ can be written as $\mathbf{HC}$ with $\mathbf{H}$ and $\mathbf{C}'$ of dimension $(k\times r)$ and of rank r. $\mathbf{C} \equiv [c_1,\ldots,c_r]'$ is called the *cointegrating matrix*.

Let $\tilde{\mathbf{C}}, \tilde{\mathbf{H}}$, and $\tilde{\mathbf{F}}_i$ be the maximum likelihood estimator of $\mathbf{C}$, $\mathbf{H}$, $\mathbf{F}$ according to Johansen's (1988, 1991) approach. Then the asymptotic distribution of $\tilde{\mathbf{\Sigma}}_u$, that is the maximum likelihood estimator of the covariance matrix of u_t, is:

$$\sqrt{T}\,[\mathrm{vech}(\tilde{\mathbf{\Sigma}}_u) - \mathrm{vech}(\mathbf{\Sigma}_u)] \xrightarrow{d} N(\mathbf{0},\, 2\mathbf{D}_k^{+}(\mathbf{\Sigma}_u \otimes \mathbf{\Sigma}_u)\mathbf{D}_k^{+'}), \tag{13}$$

which is equivalent to equation (11) (see it again for the definition of the various operators). Thus, it turns out that the asymptotic distribution of the maximum likelihood estimator $\tilde{\mathbf{\Sigma}}_u$ is the same as the OLS estimation $\hat{\mathbf{\Sigma}}_u$ for the case of stable VAR.

Thus, the application of the method described for testing residuals zero partial correlations can be applied straightforwardly to cointegrated data. The model is estimated as a VECM error correction model using Johansen's (1988, 1991) approach, correlations are tested exploiting the asymptotic distribution of $\tilde{\Sigma}_u$ and finally can be parameterized back in its VAR form of equation (3).

2.4. Summary of the search procedure

The graphical causal models approach to SVAR identification, which we suggest in case of Gaussian and linear processes, can be summarized as follows.

Step 1 Estimate the VAR model $\mathbf{Y}_t = \mathbf{A}_1\mathbf{Y}_{t-1} + \ldots + \mathbf{A}_p\mathbf{Y}_{t-p} + \mathbf{u}_t$ with the usual specification tests about normality, zero autocorrelation of residuals, lags, and unit roots (see Lütkepohl, 2006). If the hypothesis of nonstationarity is rejected, estimate the VAR model via OLS (equivalent to MLE under the assumption of normality of the errors). If unit root tests do not reject $I(1)$ nonstationarity in the data, specify the model as VECM testing the presence of cointegrating relationships. If tests suggest the presence of cointegrating relationships, estimate the model as VECM. If cointegration is rejected estimate the VAR models taking first difference.

Step 2 Run tests for zero partial correlations between the elements $u_{1t},\ldots,u_{kt}$ of $\mathbf{u}_t$ using the Wald statistics on the basis of the asymptotic distribution of the covariance matrix of $\mathbf{u}_t$. Note that not all possible partial correlations $\rho(u_{it}, u_{jt}|u_{ht},\ldots)$ need to be tested, but only those necessary for step 3.

Step 3 Apply a causal search algorithm to recover the causal structure among $u_{1t}, \ldots, u_{kt}$, which is equivalent to the causal structure among $Y_{1t}, \ldots, Y_{kt}$ (cfr. section 1.2 and see Moneta 2003). In case of acyclic (no feedback loops) and causally sufficient (no latent variables) structure, the suggested algorithm is the PC algorithm of Spirtes et al. (2000, pp. 84-85). Moneta (2008) suggested few modifications to the PC algorithm in order to make the orientation of edges compatible with as many conditional independence tests as possible. This increases the computational time of the search algorithm, but considering the fact that VAR models deal with a few number of time series variables (rarely more than six to eight; see Bernanke et al. 2005), this slowing down does not create a serious concern in this context. Table 1 reports the modified PC algorithm. In case of acyclic structure without causal sufficiency (i.e. possibly including latent variables), the suggested algorithm is FCI (Spirtes et al. 2000, pp. 144-145). In the case of no latent variables and in the presence of feedback loops, the suggested algorithm is CCD (Richardson and Spirtes, 1999). There is no algorithm in the literature which is consistent for search when both latent variables and feedback loops may be present. If the goal of the study is only impulse response analysis (i.e. tracing out the effects of structural shocks $\varepsilon_{1t}, \ldots, \varepsilon_{kt}$ on $\mathbf{Y}_t, \mathbf{Y}_{t-1}, \ldots$) and neither contemporaneous feedbacks nor latent variables can be excluded *a priori*, a possible solution is to apply only steps (A) and (B) of the PC algorithm. If the resulting set of possible causal structures (represented by an undirected graph) contains a manageable number of elements, one can study the characteristics of the impulse response functions which are robust across all the possible causal structures, where the presence of both feedbacks and latent variables is allowed (Moneta, 2004).

Step 4 Calculate structural coefficients and impulse response functions. If the output of Step 3 is a set of causal structures, run sensitivity analysis to investigate the robustness of the conclusions under the different possible causal structures. Bootstrap procedures may also be applied to determine which is the most reliable causal order (see simulations and applications in Demiralp et al., 2008).

3. Identification via independent component analysis

The methods considered in the previous section use tests for zero partial correlation on the VAR-residuals to obtain (partial) information about the contemporaneous structure in an SVAR model with Gaussian shocks. In this section we show how non-Gaussian and independent shocks can be exploited for model identification by using the statistical method of 'Independent Component Analysis' (ICA, see Comon (1994); Hyvärinen et al. (2001)). The method is again based on the VAR-residuals $\mathbf{u}_t$ which can be obtained as in the Gaussian case by estimating the VAR model using for example ordinary least squares or least absolute deviations, and can be tested for non-Gaussianity using any normality test (such as the Shapiro-Wilk or Jarque-Bera test).

To motivate, we note that, from equations (3) and (4) (with matrix $\mathbf{\Gamma}_0$) or the Cholesky factorization in section 1.2 (with matrix $\mathbf{PD}^{-1}$), the VAR-disturbances $\mathbf{u}_t$ and

Table 1: Search algorithm (adapted from the PC Algorithm of Spirtes et al. (2000: 84-85); in bold character the modifications). Under the assumption of Gaussianity conditional independence is tested by zero partial correlation tests.

```
(A): (connect everything):
Form the complete undirected graph $\mathcal{G}$ on the vertex set $u_{1t}, \ldots, u_{kt}$ so that each vertex is connected to any other vertex by an undirected edge.
(B)(cut some edges):
$n = 0$
repeat :
        repeat :
                select an ordered pair of variables $u_{ht}$ and $u_{it}$ that are adjacent in $\mathcal{G}$ such that the number of variables adjacent to $u_{ht}$ is equal or greater than $n + 1$. Select a set $S$ of $n$ variables adjacent to $u_{ht}$ such that $u_{ti} \notin S$. If $u_{ht} \perp\!\!\!\perp u_{it} | S$ delete edge $u_{ht}$ — $u_{it}$ from $\mathcal{G}$;
        until all ordered pairs of adjacent variables $u_{ht}$ and $u_{it}$ such that the number of variables adjacent to $u_{ht}$ is equal or greater than $n + 1$ and all sets $S$ of $n$ variables adjacent to $u_{ht}$ such that $u_{it} \notin S$ have been checked to see if $u_{ht} \perp\!\!\!\perp u_{it} | S$;
        $n = n + 1$;
until for each ordered pair of adjacent variables $u_{ht}$, $u_{it}$, the number of adjacent variables to $u_{ht}$ is less than $n + 1$;
(C)(build colliders):
for each triple of vertices $u_{ht}, u_{it}, u_{jt}$ such that the pair $u_{ht}, u_{it}$ and the pair $u_{it}, u_{jt}$ are each adjacent in $\mathcal{G}$ but the pair $u_{ht}, u_{jt}$ is not adjacent in $\mathcal{G}$, orient $u_{ht}$ — $u_{it}$ — $u_{jt}$ as $u_{ht} \longrightarrow u_{it} \longleftarrow u_{jt}$ if and only if $u_{it}$ does not belong to **any set** of variables $S$ such that $u_{ht} \perp\!\!\!\perp u_{jt} | S$;
(D)(direct some other edges):
repeat :
        if $u_{at} \longrightarrow u_{bt}$, $u_{bt}$ and $u_{ct}$ are adjacent, $u_{at}$ and $u_{ct}$ are not adjacent and $u_{bt}$ belongs to **every set** $S$ such that $u_{at} \perp\!\!\!\perp u_{ct} | S$, then orient $u_{bt}$ — $u_{ct}$ as $u_{bt} \longrightarrow u_{ct}$;
        if there is a directed path from $u_{at}$ to $u_{bt}$, and an edge between $u_{at}$ and $u_{bt}$, then orient $u_{at}$ — $u_{bt}$ as $u_{at} \longrightarrow u_{bt}$;
until no more edges can be oriented.
```

the structural shocks $\boldsymbol{\varepsilon}_t$ are connected by

$$\mathbf{u}_t = \boldsymbol{\Gamma}_0^{-1}\boldsymbol{\varepsilon}_t = \mathbf{P}\mathbf{D}^{-1}\boldsymbol{\varepsilon}_t \tag{14}$$

with square matrices $\boldsymbol{\Gamma}_0$ and $\mathbf{P}\mathbf{D}^{-1}$, respectively. Equation (14) has two important properties: First, the vectors $\mathbf{u}_t$ and $\boldsymbol{\varepsilon}_t$ are of the same length, meaning that there are as many residuals as structural shocks. Second, the residuals $\mathbf{u}_t$ are linear mixtures of the shocks $\boldsymbol{\varepsilon}_t$, connected by the 'mixing matrix' $\boldsymbol{\Gamma}_0^{-1}$. This resembles the ICA model, when placing certain assumptions on the shocks $\boldsymbol{\varepsilon}_t$.

In short, the ICA model is given by $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{x}$ are the mixed components, $\mathbf{s}$ the independent, non-Gaussian sources, and $\mathbf{A}$ a square invertible mixing matrix (meaning that there are as many mixtures as independent components). Given samples from the mixtures $\mathbf{x}$, ICA estimates the mixing matrix $\mathbf{A}$ and the independent components $\mathbf{s}$, by linearly transforming $\mathbf{x}$ in such a way that the dependencies among the independent components $\mathbf{s}$ are minimized. The solution is unique up to ordering, sign and scaling (Comon, 1994; Hyvärinen et al., 2001).

By comparing the ICA model $\mathbf{x} = \mathbf{A}\mathbf{s}$ and equation (14), we see a one-to-one correspondence of the mixtures $\mathbf{x}$ to the residuals $\mathbf{u}_t$ and the independent components $\mathbf{s}$ to the shocks $\boldsymbol{\varepsilon}_t$. Thus, to be able to apply ICA, we need to assume that the shocks are non-Gaussian and mutually independent. We want to emphasize that no specific non-Gaussian distribution is assumed for the shocks, but only that they cannot be Gaussian.[1] For the shocks to be mutually independent their joint distribution has to factorize into the product of the marginal distributions. In the non-Gaussian setting, this implies zero partial correlation, but the converse is not true (as opposed to the Gaussian case where the two statements are equivalent). Thus, for non-Gaussian distributions conditional independence is a much stronger requirement than uncorrelatedness.

Under the assumption that the shocks $\boldsymbol{\varepsilon}_t$ are non-Gaussian and independent, equation (14) follows exactly the ICA-model and applying ICA to the VAR residuals $\mathbf{u}_t$ yields a unique solution (up to ordering, sign, and scaling) for the mixing matrix $\boldsymbol{\Gamma}_0^{-1}$ and the independent components $\boldsymbol{\varepsilon}_t$ (i.e. the structural shocks in our case). However, the ambiguities of ICA make it hard to directly interpret the shocks found by ICA since without further analysis we cannot relate the shocks directly to the measured variables.

Hence, we assume that the residuals $\mathbf{u}_t$ follow a linear non-Gaussian acyclic model (Shimizu et al., 2006), which means that the contemporaneous structure is represented by a DAG (directed acyclic graph). In particular, the model is given by

$$\mathbf{u}_t = \mathbf{B}_0\mathbf{u}_t + \boldsymbol{\varepsilon}_t \tag{15}$$

with a matrix $\mathbf{B}_0$, whose diagonal elements are all zero and, if permuted according to the causal order, is strictly lower triangular. By rewriting equation (15) we see that

$$\boldsymbol{\Gamma}_0 = \mathbf{I} - \mathbf{B}_0. \tag{16}$$

From this equation it follows that the matrix $\mathbf{B}_0$ describes the contemporaneous structure of the variables $\mathbf{Y}_t$ in the SVAR model as shown in equation (6). Thus, if we can

[1] Actually, the requirement is that *at most one* of the residuals can be Gaussian.

identify the matrix $\mathbf{\Gamma}_0$, we also obtain the matrix $\mathbf{B}_0$ for the contemporaneous effects. As pointed out above, the matrix $\mathbf{\Gamma}_0^{-1}$ (and hence $\mathbf{\Gamma}_0$) can be estimated using ICA up to ordering, scaling, and sign. With the restriction of $\mathbf{B}_0$ representing an acyclic system, we can resolve these ambiguities and are able to fully identify the model. For simplicity, let us assume that the variables are arranged according to a causal ordering, so that the matrix $\mathbf{B}_0$ is strictly lower triangular. From equation (16) then follows that the matrix $\mathbf{\Gamma}_0$ is lower triangular with all ones on the diagonal. Using this information, the ambiguities of ICA can be resolved in the following way.

The lower triangularity of $\mathbf{B}_0$ allows us to find the unique permutation of the rows of $\mathbf{\Gamma}_0$, which yields all non-zero elements on the diagonal of $\mathbf{\Gamma}_0$, meaning that we replace the matrix $\mathbf{\Gamma}_0$ with $\mathbf{Q}_1\mathbf{\Gamma}_0$ where $\mathbf{Q}_1$ is the uniquely determined permutation matrix. Finding this permutation resolves the ordering-ambiguity of ICA and links the shocks $\boldsymbol{\varepsilon}_t$ to the components of the residuals $\mathbf{u}_t$ in a one-to-one manner. The sign- and scaling-ambiguity is now easy to fix by simply dividing each row of $\mathbf{\Gamma}_0$ (the row-permuted version from above) by the corresponding diagonal element yielding all ones on the diagonal, as implied by Equation (16). This ensures that the connection strength of the shock $\boldsymbol{\varepsilon}_t$ on the residual $\mathbf{u}_t$ is fixed to one in our model (Equation (15)).

For the general case where $\mathbf{B}_0$ is not arranged in the causal order, the above arguments for solving the ambiguities still apply. Furthermore, we can find the causal order of the contemporaneous variables by performing simultaneous row- and column-permutations on $\mathbf{\Gamma}_0$ yielding the matrix closest to lower triangular, in particular $\tilde{\mathbf{\Gamma}}_0 = \mathbf{Q}_2\mathbf{\Gamma}_0\mathbf{Q}_2'$ with an appropriate permutation matrix $\mathbf{Q}_2$. In case non of these permutations leads to a close to lower triangular matrix a warning is issued.

Essentially, the assumption of acyclicity allows us to uniquely connect the structural shocks $\boldsymbol{\varepsilon}_t$ to the components of $\mathbf{u}_t$ and fully identify the contemporaneous structure. Details of the procedure can be found in Shimizu et al. (2006); Hyvärinen et al. (2010). In the sense of the Cholesky factorization of the covariance matrix explained in Section 1 (with $\mathbf{PD}^{-1} = \mathbf{\Gamma}_0^{-1}$), full identifiability means that a causal order among the contemporaneous variables can be determined.

In addition to yielding full identification, an additional benefit of using the ICA-based procedure when shocks are non-Gaussian is that it does not rely on the faithfulness assumption, which was necessary in the Gaussian case.

We note that there are many ways of exploiting non-Gaussian shocks for model identification as alternatives to directly using ICA. One such approach was introduced by Shimizu et al. (2009). Their method relies on iteratively finding an exogenous variable and regressing out their influence on the remaining variables. An exogenous variable is characterized by being independent of the residuals when regressing any other variable in the model on it. Starting from the model in equation (15), this procedure returns a causal ordering of the variables $\mathbf{u}_t$ and then the matrix $\mathbf{B}_0$ can be estimated using the Cholesky approach.

One relatively strong assumption of the above methods is the *acyclicity* of the contemporaneous structure. In Lacerda et al. (2008) an extension was proposed where feedback loops were allowed. In terms of the matrix $\mathbf{B}_0$ this means that it is not restricted to being lower triangular (in an appropriate ordering of the variables). While in general this model is not identifiable because we cannot uniquely match the shocks to the residuals, Lacerda et al. (2008) showed that the model is identifiable when assuming stability of the generating model in (15) (the absolute value of the biggest eigenvalue in $\mathbf{B}_0$ is smaller than one) and disjoint cycles.

Another restriction of the above model is that all relevant variables must be included in the model (causal sufficiency). Hoyer et al. (2008b) extended the above model by allowing for hidden variables. This leads to an overcomplete basis ICA model, meaning that there are more independent non-Gaussian sources than observed mixtures. While there exist methods for estimating overcomplete basis ICA models, those methods which achieve the required accuracy do not scale well. Additionally, the solution is again only unique up to ordering, scaling, and sign, and when including hidden variables the ordering-ambiguity cannot be resolved and in some cases leads to several observationally equivalent models, just as in the cyclic case above.

We note that it is also possible to combine the approach of section 2 with that described here. That is, if some of the shocks are Gaussian or close to Gaussian, it may be advantageous to use a combination of constraint-based search and non-Gaussianity-based search. Such an approach was proposed in Hoyer et al. (2008a). In particular, the proposed method does not make any assumptions on the distributions of the VAR-residuals $\mathbf{u}_t$. Basically, the PC algorithm (see Section 2) is run first, followed by utilization of whatever non-Gaussianity there is to further direct edges. Note that there is no need to know in advance which shocks are non-Gaussian since finding such shocks is part of the algorithm.

Finally, we need to point out that while the basic ICA-based approach does not require the faithfulness assumption, the extensions discussed at the end of this section do.

4. Nonparametric setting

4.1. Theory

Linear systems dominate VAR, SVAR, and more generally, multivariate time series models in econometrics. However, it is not always the case that we know how a variable X may cause another variable Y. It may be the case that we have little or no *a priori* knowledge about the way how Y depends on X. In its most general form we want to know whether X is independent of Y conditional on the set of potential graphical parents Z, i.e.

$$H_0 : Y \perp\!\!\!\perp X \,|\, Z, \tag{17}$$

where Y, X, Z is a set of time series variables. Thereby, we do not per se require an *a priori* specification of how Y possibly depends on X. However, constraint based algorithms typically specify conditional independence in a very restrictive way. In continuous settings, they simply test for nonzero partial correlations, or in other words, for linear (in)dependencies. Hence, these algorithms will fail whenever the data generation process (DGP) includes nonlinear causal relations.

In search for a more general specification of conditional independency, Chlaß and Moneta (2010) suggest a procedure based on nonparametric density estimation. Therein, neither the type of dependency between Y and X, nor the probability distributions of the variables need to be specified. The procedure exploits the fact that if two random variables are independent of a third, one obtains their joint density by the product of the joint density of the first two, and the marginal density of the third. Hence, hypothesis test (17) translates into:

$$H_0 : \frac{f(Y,X,Z)}{f(XZ)} = \frac{f(YZ)}{f(Z)}. \tag{18}$$

If we define $h_1(\cdot) := f(Y,X,Z)f(Z)$, and $h_2(\cdot) := f(YZ)f(XZ)$, we have:

$$H_0 : h_1(\cdot) = h_2(\cdot). \tag{19}$$

We estimate h_1 and h_2 using a kernel smoothing approach (see Wand and Jones, 1995, ch.4). Kernel smoothing has the outstanding property that it is insensitive to autocorrelation phenomena and, therefore, immediately applicable to longitudinal or time series settings (Welsh et al., 2002).

In particular, we use a so-called *product kernel* estimator:

$$\begin{aligned} \hat{h}_1(x,y,z;b) &= \tfrac{1}{N^2 b^{m+d}} \left\{\textstyle\sum_{i=1}^n K\left(\tfrac{X_i-x}{b}\right) K\left(\tfrac{Y_i-y}{b}\right) K\left(\tfrac{Z_i-z}{b}\right)\right\}\left\{\textstyle\sum_{i=1}^n K_p\left(\tfrac{Z_i-z}{b}\right)\right\} \\ \hat{h}_2(x,y,z;b) &= \tfrac{1}{N^2 b^{m+d}} \left\{\textstyle\sum_{i=1}^n K\left(\tfrac{X_i-x}{b}\right) K_Z\left(\tfrac{Z_i-z}{b}\right)\right\}\left\{\textstyle\sum_{i=1}^n K\left(\tfrac{Y_i-y}{b}\right) K_p\left(\tfrac{Z_i-z}{b}\right)\right\}, \end{aligned} \tag{20}$$

where X_i, Y_i, and Z_i are the i^{th} realization of the respective time series, K denotes the *kernel* function, b indicates a scalar bandwidth parameter, and K_p represents a product kernel[2].

So far, we have shown how we can estimate h_1 and h_2. To see whether these are different, we require some similarity measure between both conditional densities. There are different ways to measure the distance between a product of densities:

(i) The weighted Hellinger distance proposed by Su and White (2008):

$$d_H = \frac{1}{n} \sum_{i=1}^{n} \left\{ 1 - \sqrt{\frac{h_2(X_i,Y_i,Z_i)}{h_1(X_i,Y_i,Z_i)}} \right\}^2 a(X_i,Y_i,Z_i), \tag{21}$$

where $a(\cdot)$ is a nonnegative weighting function. Both the weighting function $a(\cdot)$, and the resulting test statistic are specified in Su and White (2008).

(ii) The Euclidean distance proposed by Szekely and Rizzo (2004) in their 'energy test':

$$d_E = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} ||h_{1_i} - h_{2_j}|| - \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} ||h_{1_i} - h_{1_j}|| - \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} ||h_{2_i} - h_{2_j}||, \tag{22}$$

[2] I.e. $K_p((Z_i - z)/b) = \prod_{j=1}^d K((Z_{j_i} - z_j)/b)$. For our simulations (see next section) we choose the kernel: $K(u) = (3 - u^2)\phi(u)/2$, with $\phi(u)$ the standard normal probability density function. We use a "rule-of-thumb" bandwidth: $b = n^{-1/8.5}$.

where $h_{1_i} = h_1(X_i, Y_i, Z_i)$, $h_{2_i} = h_2(X_i, Y_i, Z_i)$, and $||\cdot||$ is the Euclidean norm.[3]

Given these test statistics and their distributions, we compute the type-I error, or *p-value* of our test problem (19). If $Z = \emptyset$, the tests are available in R-packages `energy` and `cramer`. The Hellinger distance is not suitable here, since one can only test for $Z \neq \emptyset$.

For $Z \neq \emptyset$, our test problem (19) requires higher dimensional kernel density estimation. The more dimensions, i.e. the more elements in Z, the scarcer the data, and the greater the distance between two subsequent data points. This so-called *Curse of dimensionality* strongly reduces the accuracy of a nonparametric estimation (Yatchew, 1998). To circumvent this problem, we calculate the type-I errors for $Z \neq \emptyset$ by a local bootstrap procedure, as described in Su and White (2008, pp. 840-841) and Paparoditis and Politis (2000, pp. 144-145). Local bootstrap draws repeatedly with replacement from the sample and counts how many times the bootstrap statistic is larger than the test statistic of the entire sample. Details on the local bootstrap procedure ca be found in appendix A.

Now, let us see how this procedure fares in those time series settings, where other testing procedures failed - the case of nonlinear time series.

4.2. Simulation Design

Our simulation design should allow us to see how the search procedures of 4.1 perform in terms of *size* and *power*. To identify size properties (type-I error), H_0 (19) must hold everywhere. We call data generation processes for which H_0 holds everywhere, *size-DGPs*. We induce a system of time series $\{V_{1,t}, V_{2,t}, V_{3,t}\}_{t=1}^{n}$ whereby each time series follows an autoregressive process AR(1) with $a_1 = 0.5$ and error term $e_t \sim N(0,1)$, for instance, $V_{1,t} = a_1 V_{1,t-1} + e_{V_1,t}$. These time series may cause each other as illustrated in Fig. 1.

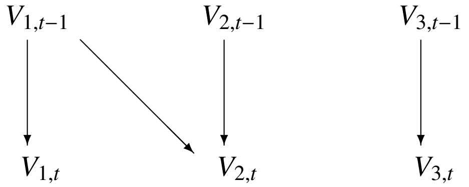


Figure 1: Time series DAG.

Therein, $V_{1,t} \perp\!\!\!\perp V_{2,t}|V_{1,t-1}$, since $V_{1,t-1}$ *d*-separates $V_{1,t}$ from $V_{2,t}$, while $V_{2,t} \perp\!\!\!\perp V_{3,s}$, for any t and s. Hence, the set of variables Z, conditional on which two sets of variables X and Y are independent of each other, contains zero elements, i.e. $V_{2,t} \perp\!\!\!\perp V_{3,t-1}$, contains one element, i.e. $V_{1,t} \perp\!\!\!\perp V_{2,t}|V_{1,t-1}$, or contains two elements, i.e. $V_{1,t} \perp\!\!\!\perp$

[3] An alternative Euclidean distance is proposed by Baringhaus and Franz (2004) in their Cramer test. This distance turns out to be $d_E/2$. The only substantial difference from the distance proposed in (ii) lies in the method to obtain the critical values (see Baringhaus and Franz 2004).

$V_{2,t}|V_{1,t-1}, V_{3,t-1}$.

In our simulations, we vary two aspects. The first aspect is the *functional form of the causal dependency*. To systematically vary nonlinearity and its impact, we characterize the causal relation between, say, $V_{1,t-1}$ and $V_{2,t}$, in a polynomial form, i.e. via $V_{2,t} = f(V_{1,t-1}) + e$, where $f = \sum_{j=0}^{p} b_j V_{1,t-1}^{j}$. Herein, j reflects the degree of nonlinearity, while b_j would capture the impact nonlinearity exerts. For polynomials of any degree, we set only $b_p \neq 0$. An additive error term e completes the specification.

The second aspect is the number of variables in Z conditional on which X and Y can be independent. Either zero, one, but maximally two variables may form the set $Z = \{Z_1, \ldots, Z_d\}$ of conditioned variables; hence Z has cardinality $\#Z = \{0, 1, 2\}$.

To identify power properties, H_0 must not hold anywhere, i.e. $X \not\!\perp\!\!\!\perp Y|Z$. We call data generation processes where H_0 does not hold anywhere, *power-DGPs*. Such DGPs can be induced by (i) a direct path between X and Y which does not include Z, (ii) a common cause for X and Y which is not an element of Z, or (iii) a "collider" between X and Y belonging to Z.[4] As before, we vary the functional form f of these causal paths polynomially where again, only $b_p \neq 0$. Third, we investigate different cardinalities $\#Z = \{0, 1, 2\}$ of set Z conditional on which X and Y become dependent.

4.3. Simulation Results

Let us start with $\#Z = 0$, that is, $H_0 := X \perp\!\!\!\perp Y$. Table 2 reports our simulation results for both size and power DGPs. Rejection frequencies are reported for three different tests, for a theoretical level of significance of 0.05, and 0.1.

Table 2: Proportion of rejection of H_0 (no conditioned variables)

	Energy	Cramer	Fisher	Energy	Cramer	Fisher
	level of significance 5%			*level of significance* 10%		
Size DGPs						
S0.1 (ind. time series)	0.065	0.000	0.151	0.122	0.000	0.213
Power DGPs						
P0.1 (time series linear)	0.959	0.308	0.999	0.981	0.462	1
P0.2 (time series quadratic)	0.986	0.255	0.432	0.997	0.452	0.521
P0.3 (time series cubic)	1	0.905	1	1	0.975	1
P0.3 (time series quartic)	1	0.781	0.647	1	0.901	0.709

Note: length series (n) = 100; number of iterations = 1000

Take the first line of Table 2. For *size DGPs*, H_0 holds everywhere. A test performs accurately if it rejects H_0 in accordance with the respective theoretical significance level. We see that the energy test rejects H_0 slightly more often than it should ($0.065 > 0.05; 0.122 > 0.1$), whereas the Cramer test does not reject H_0 often enough ($0.000 < 0.05, 0.000 < 0.1$). In comparison to the standard parametric Fisher's z, we see that the

[4] An example of collider is displayed in Figure 1: $V_{2,t}$ forms a collider between $V_{1,t-1}$ and $V_{2,t-1}$.

latter rejects H_0 much more often than it should. The energy test keeps the type-I error most accurately. Contrary to both nonparametric tests, the parametric procedure leads us to suspect a lot more causal relationships than there actually are, if $\#Z = 0$.

How well do these tests perform if H_0 does not hold anywhere? That is, how accurately do they reject H_0 if it is false *(power-DGPs)*? For linear time series, we see that the nonparametric energy test has nearly as much power as Fisher's z. For nonlinear time series, the energy test clearly outperforms Fisher's z[5]. As it did for *size*, Cramer's test generally underperforms in terms of power. Interestingly, its power appears to be higher for higher degrees of nonlinearity. In summary, if one wishes to test for marginal independence without any information on the type of a potential dependence, one would opt for the energy test. It has a size close to the theoretical significance level, and has power similar to a parametric specification.

Let us turn to $\#Z = 1$, where $H_0 := X \perp\!\!\!\perp Y|Z$, for which the results are shown in Table 3. Starting with *size DGPs*, tests based on Hellinger and Euclidian distance slightly underreject H_0 whereas for the highest polynomial degree, the Hellinger test strongly overrejects H_0. The parametric Fisher's z slightly overrejects H_0 in case of linearity, and for higher degrees, starts to underreject H_0.

Table 3: Proportion of rejection of H_0 (one conditioned variable)

	Hellinger	Euclid	Fisher	Hellinger	Euclid	Fisher
	level of significance 5%			*level of significance* 10%		
Size DGPs						
S1.1 (time series linear)	0.035	0.035	0.062	0.090	0.060	0.103
S1.2 (time series quadratic)	0.040	0.020	0.048	0.065	0.035	0.104
S1.3 (time series cubic)	0.010	0.010	0.050	0.020	0.015	0.093
S1.4 (time series quartic)	0.13	0	0.023	0.2	0.1	0.054
Power DGPs						
P1.1 (time series linear)	0.875	0.910	0.999	0.925	0.950	1
P1.2 (time series quadratic)	0.905	0.895	0.416	0.940	0.950	0.504
P1.3 (time series cubic)	0.990	1	1	1	1	1
P1.4 (time series quartic)	0.84	0.995	0.618	0.91	0.995	0.679

Note: n = 100; number of iterations = 200; number of bootstrap iterations (I) = 200

Turning to *power DGPs*, Fisher's z suffers a dramatic loss in power for those polynomial degrees which depart most from linearity, i.e. quadratic, and quartic relations. Nonparametric tests which do not require linearity have high power in absolute terms, and nearly twice as much as compared to Fisher's z. The power properties of the nonparametric procedures indicate that our local bootstrap succeeds in mitigating the Curse of dimensionality. In sum, nonparametric tests exhibit good power properties for $\#Z = 1$ whereas Fisher's z would fail to discover underlying quadratic or quartic relationships in some 60%, and 40% of the cases, respectively.

[5] For cubic time series, Fisher's z performs as well as the energy test does. This may be due to the fact that a cubic relation resembles more to a line than other polynomial specifications do.

The results for #Z = 2 are presented in Table 4. We find that both nonparametric tests have a size which is notably smaller than the theoretical significance level we induce. Hence, both have a strong tendency to underreject H_0. Turning to *power DGPs*, we find that the Euclidean test still has over 90% power to correctly reject H_0. For those polynomial degrees which depart most from linearity, i.e. quadratic and quartic, the Euclidean test has three times as much power as Fisher's z. However, the Hellinger test performs even worse than Fisher's z. Here, it may be the Curse of dimensionality which starts to show an impact.

Table 4: Proportion of rejection of H_0 (two conditioned variables)

	Hellinger	Euclid	Fisher	Hellinger	Euclid	Fisher
	level of significance 5%			*level of significance* 10%		
Size DGPs						
S2.1 (time series linear)	0.006	0.020	0.050	0.033	0.046	0.102
S2.2 (time series quadratic)	0.000	0.010	0.035	0.000	0.040	0.087
S2.3 (time series cubic)	0	0.007	0.056	0	0.007	0.109
S2.4 (time series quartic)	0.006	0	0.031	0.013	0	0.067
Power DGPs						
P2.1 (time series linear)	0.28	0.92	1	0.4	0.973	1
P2.2 (time series quadratic)	0.170	0.960	0.338	0.250	0.980	0.411
P2.3 (time series cubic)	0.667	1	1	0.754	1	1
P2.4 (time series quartic)	0.086	0.946	0.597	0.133	0.966	0.665

Note: n = 100; number of iterations = 150; number of bootstrap iterations (I) = 100

To sum up, we can say that both marginal independencies, and higher dimensional conditional independencies, i.e. ($\#Z = 1,2$) are best tested for using Euclidean tests. The Hellinger test seems to be more affected by the Curse of dimensionality. We see that our local bootstrap procedure mitigates the latter, but we admit that the number of variables our nonparametric procedure can deal with is very small. Here, it might be promising to opt for semiparametric (Chu and Glymour, 2008), rather than nonparametric procedures which combine parametric and nonparametric approaches.

5. Conclusions

The difficulty of learning causal relations from passive, that is non-experimental, observations is one of the central challenges of econometrics. Traditional solutions involve the distinction between structural and reduced form model. The former is meant to formalize the unobserved data generating process, whereas the latter aims to describe a simpler transformation of that process. The structural model is articulated hinging on *a priori* economic theory. The reduced form model is formalized in such a way that it can be estimated directly from the data. In this paper, we have presented an approach to identify the structural model which minimizes the role of *a priori* economic theory and emphasizes the need of an appropriate and rich statistical model of the data. Graphical causal models, independent component analysis, and tests of conditional independence are the tools we propose for structural identification in vector autoregressive models. We conclude with an overview of some important issues which are left open in this domain.

1. *Specification of the statistical model.* Data driven procedures for SVAR identification depend upon the specification of the (reduced form) VAR model. Therefore, it is important to make sure that the estimated VAR model is an accurate description of the dynamics of the included variables (whereas the contemporaneous structure is intentionally left out, as seen in section 1.2). The usual criterion for accuracy is to check that the model estimates residuals conform to white noise processes (although serial independence of residuals is not a sufficient criterion for model validation). This implies stable dependencies captured by the relationships among the modeled variables, and an unsystematic noise. It may be the case, as in many empirical applications, that different VAR specifications pass the model checking tests equally well. For example, a VAR with Gaussian errors and p lags may fit the data equally well as a VAR with non-Gaussian errors and q lags and these two specifications justify two different causal search procedures. So far, we do not know how to adjudicate among alternative and seemingly equally accurate specifications.

2. *Background knowledge and assumptions.* Search algorithms are based on different assumptions, such as, for example, causal sufficiency, acyclicity, the Causal Markov Condition, Faithfulness, and/or the existence of independent components. Maybe, background knowledge could justify some of these assumptions and reject others. For example, institutional or theoretical knowledge about an economic process might inform us that Faithfulness is a plausible assumption in some contexts rather than in others, or instead, that one should expect feedback loops if data are collected at certain levels of temporal aggregation. Yet, if background information could inform us here, this might again provoke a problem of circularity mentioned at the outset of the paper.

3. *Search algorithms in nonparametric settings.* We have provided some information on which nonparametric test procedures might be more appropriate in certain circumstances. However, it is not clear which causal search algorithms are most efficient in exploiting the nonparametric conditional independence tests proposed in Section 4. The more variables the search algorithm needs to be informed about at the same point of the search, the higher the number of conditioned variables, and hence, the slower, or the more inaccurate, the test.

4. *Number of shocks and number of variables.* To conserve degrees of freedom, SVARs rarely model more than six to eight time series variables (Bernanke et al., 2005, p.388). It is an open question how the procedures for causal inference we reviewed can be applied to large scale systems such as dynamic factor models. (Forni et al., 2000)

5. *Simulations and empirical applications.* Graphical causal models for identifying SVARs, equivalent or similar to the search procedures described in section 2, have been applied to several sets of macroeconomic data (Swanson and Granger, 1997; Bessler and Lee, 2002; Demiralp and Hoover, 2003; Moneta, 2004; Demiralp et al., 2008; Moneta, 2008; Hoover et al., 2009). Demiralp and Hoover (2003) present Monte Carlo

simulations to evaluate the performance of the PC algorithm for such an identification. There are no simulation results so far about the performance of the alternative tests on residual partial correlations presented in section 2.2. Moneta et al. (2010) applied an independent component analysis as described in section 3, to microeconomic US data about firms' expenditures on R&D and performance, as well as to macroeconomic US data about monetary policy and its effects on the aggregate economy. Hyvärinen et al. (2010) assess the performance of independent component analysis for identifying SVAR models. It is yet to be established how independent component analysis applied to SVARs fares compared to graphical causal models (based on the appropriate conditional independence tests) in non-Gaussian settings. Nonparametric tests of conditional independence, as those proposed in section 4, have been applied to test for Granger non-causality (Su and White, 2008), but there are not yet any applications where these test results inform a graphical causal search algorithm. Overall, there is a need for more empirical applications of the procedures described in this paper. Such applications will be useful to test, compare, and improve different search procedures, to suggest new problems, and obtain new causal knowledge.

6. Appendix

6.1. Appendix 1 - Details of the bootstrap procedure from 4.1.

(1) Draw a bootstrap sampling Z_t^* (for $t = 1,\dots,n$) from the estimated kernel density $\hat{f}(z) = n^{-1}b^{-d}\sum_{t=1}^{n} K_p((Z_t - z)/b)$.

(2) For $t = 1,\dots,n$, given Z_t^*, draw X_t^* and Y_t^* *independently* from the estimated kernel density $\hat{f}(x|Z_t^*)$ and $\hat{f}(y|Z_t^*)$ respectively.

(3) Using X_t^*, Y_t^*, and Z_t^*, compute the bootstrap statistic S_n^* using one of the distances defined above.

(4) Repeat steps (1) and (2) I times to obtain I statistics $\{S_{ni}^*\}_{i=1}^{I}$.

(5) The p-value is then obtained by:

$$p \equiv \frac{\sum_{i=1}^{I} 1\{S_{ni}^* > S_n\}}{I},$$

where S_n is the statistic obtained from the original data using one of the distances defined above, and $1\{\cdot\}$ denotes an indicator function taking value one if the expression between brackets is true and zero otherwise.

References

E. Baek and W. Brock. A general test for nonlinear Granger causality: Bivariate model. *Discussin paper, Iowa State University and University of Wisconsin, Madison*, 1992.

L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004.

B. S. Bernanke. Alternative explanations of the money-income correlation. In *Carnegie-Rochester Conference Series on Public Policy*, volume 25, pages 49–99. Elsevier, 1986.

B.S. Bernanke, J. Boivin, and P. Eliasz. Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach. *Quarterly Journal of Economics*, 120(1):387–422, 2005.

D. A. Bessler and S. Lee. Money and prices: US data 1869-1914 (a study with directed graphs). *Empirical Economics*, 27:427–446, 2002.

O. J. Blanchard and D. Quah. The dynamic effects of aggregate demand and supply disturbances. *The American Economic Review*, 79(4):655–673, 1989.

O. J. Blanchard and M. W. Watson. Are business cycles all alike? *The American business cycle: Continuity and change*, 25:123–182, 1986.

N. Chlaß and A. Moneta. Can Graphical Causal Inference Be Extended to Nonlinear Settings? *EPSA Epistemology and Methodology of Science*, pages 63–72, 2010.

T. Chu and C. Glymour. Search for additive nonlinear time series causal models. *The Journal of Machine Learning Research*, 9:967–991, 2008.

P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3): 287–314, 1994.

S. Demiralp and K. D. Hoover. Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and Statistics*, 65:745–767, 2003.

S. Demiralp, K. D. Hoover, and D. J. Perez. A Bootstrap method for identifying and evaluating a structural vector autoregression. *Oxford Bulletin of Economics and Statistics, 65, 745-767*, 2008.

M. Eichler. Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, 137(2):334–353, 2007.

J. Faust and E. M. Leeper. When do long-run identifying restrictions give reliable results? *Journal of Business & Economic Statistics*, 15(3):345–353, 1997.

J. P. Florens and M. Mouchart. A note on noncausality. *Econometrica*, 50(3):583–591, 1982.

M. Forni, M. Hallin, M. Lippi, and L. Reichlin. The generalized dynamic-factor model: Identification and estimation. *Review of Economics and Statistics*, 82(4):540–554, 2000.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 37(3):424–438, 1969.

C. W. J. Granger. Testing for causality:: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352, 1980.

T. Haavelmo. The probability approach in econometrics. *Econometrica*, 12:1–115, 1944.

C. Hiemstra and J. D. Jones. Testing for linear and nonlinear Granger causality in the stock price-volume relation. *Journal of Finance*, 49(5):1639–1664, 1994.

W. C. Hood and T. C. Koopmans. *Studies in econometric method, Cowles Commission Monograph, No. 14*. New York: John Wiley & Sons, 1953.

K. D. Hoover. *Causality in macroeconomics*. Cambridge University Press, 2001.

K. D. Hoover. The methodology of econometrics. *New Palgrave Handbook of Econometrics*, 1:61–87, 2006.

K. D. Hoover. Causality in economics and econometrics. In *The New Palgrave Dictionary of Economics*. London: Palgrave Macmillan, 2008.

K.D. Hoover, S. Demiralp, and S.J. Perez. Empirical Identification of the Vector Autoregression: The Causes and Effects of US M2. In *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry*, pages 37–58. Oxford University Press, 2009.

P. O. Hoyer, A. Hyvärinen, R. Scheines, P. Spirtes, J. Ramsey, G. Lacerda, and S. Shimizu. Causal discovery of linear acyclic models with arbitrary distributions. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, 2008a.

P. O. Hoyer, S. Shimizu, A. J. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49:362–378, 2008b.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.

A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer. Estimation of a Structural Vector Autoregression model using non-Gaussianity. *Journal of Machine Learning Research*, 11:1709–1731, 2010.

S. Johansen. Statistical analysis of cointegrating vectors. *Journal of Economic Dynamics and Control*, 12:231–254, 1988.

S. Johansen. Estimation and hypothesis testing of cointegrating vectors in Gaussian vector autoregressive models. *Econometrica*, 59:1551–1580, 1991.

S. Johansen. Cointegration: An Overview. In *Palgrave Handbook of Econometrics. Volume 1. Econometric Theory*, pages 540–577. Palgrave Macmillan, 2006.

R. G. King, C. I. Plosser, J. H. Stock, and M. W. Watson. Stochastic trends and economic fluctuations. *American Economic Review*, 81:819–840, 1991.

T. C. Koopmans. *Statistical Inference in Dynamic Economic Models, Cowles Commission Monograph, No. 10*. New York: John Wiley & Sons, 1950.

G. Lacerda, P. Spirtes, J. Ramsey, and P. O. Hoyer. Discovering cyclic causal models by Independent Components Analysis. In *Proc. 24th Conference on Uncertainty in Artificial Intelligence (UAI-2008)*, Helsinki, Finland, 2008.

R. E. Lucas. Econometric policy evaluation: A critique. In *Carnegie-Rochester Conference Series on Public Policy*, volume 1, pages 19–46. Elsevier, 1976.

H. Lütkepohl. Vector Autoregressive Models. In *Palgrave Handbook of Econometrics. Volume 1. Econometric Theory*, pages 477–510. Palgrave Macmillan, 2006.

A. Moneta. Graphical Models for Structural Vector Autoregressions. *LEM Papers Series, Sant'Anna School of Advanced Studies, Pisa*, 2003.

A. Moneta. Identification of monetary policy shocks: a graphical causal approach. *Notas Económicas, 20, 39-62*, 2004.

A. Moneta. Graphical causal models and VARs: an empirical assessment of the real business cycles hypothesis. *Empirical Economics*, 35(2):275–300, 2008.

A. Moneta, D. Entner, P.O. Hoyer, and A. Coad. Causal inference by independent component analysis with applications to micro-and macroeconomic data. *Jena Economic Research Papers*, 2010:031, 2010.

E. Paparoditis and D. N. Politis. The local bootstrap for kernel estimators under general dependence conditions. *Annals of the Institute of Statistical Mathematics*, 52(1):139–159, 2000.

J. Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, Cambridge, 2000.

M. Reale and G. T. Wilson. Identification of vector AR models with recursive structural errors using conditional independence graphs. *Statistical Methods and Applications, 10, 49-65*, 2001.

T. Richardson and P. Spirtes. Automated discovery of linear feedback models. In *Computation, causation and discovery*. AAAI Press and MIT Press, Menlo Park, 1999.

R. Scheines, P. Spirtes, C. Glymour, C. Meek, and T. Richardson. The TETRAD project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1):65–117, 1998.

M. D. Shapiro and M. W. Watson. Sources of business cycle fluctuations. *NBER Macroeconomics annual*, 3:111–148, 1988.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

S. Shimizu, A. Hyvärinen, Y. Kawahara, and T. Washio. A direct method for estimating a causal ordering in a linear non-Gaussian acyclic model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 2009.

C. A. Sims. Macroeconomics and Reality. *Econometrica, 48, 1-47*, 1980.

C. A. Sims. An autoregressive index model for the u.s. 1948-1975. In J. Kmenta and J.B. Ramsey, editors, *Large-scale macro-econometric models: theory and practice*, pages 283–327. North-Holland, 1981.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search.* MIT Press, Cambridge MA, 2nd edition, 2000.

L. Su and H. White. A nonparametric Hellinger metric test for conditional independence. *Econometric Theory*, 24(04):829–864, 2008.

P. Suppes. A probabilistic theory of causation. *Acta Philosophica Fennica, XXIV*, 1970.

N. R. Swanson and C. W. J. Granger. Impulse response function based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association*, 92:357–367, 1997.

G. J. Szekely and M. L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5, 2004.

M. P. Wand and M. C. Jones. Kernel smoothing. *Chapman&Hall Ltd., London*, 1995.

A. H. Welsh, X. Lin, and R. J. Carroll. Marginal Longitudinal Nonparametric Regression. *Journal of the American Statistical Association*, 97(458):482–493, 2002.

H. White and X. Lu. Granger Causality and Dynamic Structural Systems. *Journal of Financial Econometrics*, 8(2):193, 2010.

N. Wiener. The theory of prediction. *Modern mathematics for engineers, Series*, 1: 125–139, 1956.

S. Wright. Correlation and causation. *Journal of agricultural research*, 20(7):557–585, 1921.

A. Yatchew. Nonparametric regression techniques in economics. *Journal of Economic Literature*, 36(2):669–721, 1998.

Causal Time Series Analysis of functional Magnetic Resonance Imaging Data

Alard Roebroeck A.ROEBROECK@MAASTRICHTUNIVERSITY.NL
Faculty of Psychology & Neuroscience
Maastricht University, the Netherlands

Anil K. Seth A.K.SETH@SUSSEX.AC.UK
Sackler Centre for Consciousness Science
University of Sussex, UK

Pedro Valdes-Sosa PETER@CNEURO.EDU.CU
Cuban Neuroscience Centre, Playa, Cuba

Editors: Florin Popescu and Isabelle Guyon

Abstract

This review focuses on dynamic causal analysis of functional magnetic resonance (fMRI) data to infer brain connectivity from a time series analysis and dynamical systems perspective. Causal influence is expressed in the Wiener-Akaike-Granger-Schweder (WAGS) tradition and dynamical systems are treated in a state space modeling framework. The nature of the fMRI signal is reviewed with emphasis on the involved neuronal, physiological and physical processes and their modeling as dynamical systems. In this context, two streams of development in modeling causal brain connectivity using fMRI are discussed: time series approaches to causality in a discrete time tradition and dynamic systems and control theory approaches in a continuous time tradition. This review closes with discussion of ongoing work and future perspectives on the integration of the two approaches.

Keywords: fMRI, hemodynamics, state space model, Granger causality, WAGS influence

1. Introduction

Understanding how interactions between brain structures support the performance of specific cognitive tasks or perceptual and motor processes is a prominent goal in cognitive neuroscience. Neuroimaging methods, such as Electroencephalography (EEG), Magnetoencephalography (MEG) and functional Magnetic Resonance Imaging (fMRI) are employed more and more to address questions of functional connectivity, inter-region coupling and networked computation that go beyond the 'where' and 'when' of task-related activity (Friston, 2002; Horwitz et al., 2000; McIntosh, 2004; Salmelin and Kujala, 2006; Valdes-Sosa et al., 2005a). A network perspective onto the parallel and distributed processing in the brain - even on the large scale accessible by neuroimaging methods - is a promising approach to enlarge our understanding of perceptual, cognitive and motor functions. Functional Magnetic Resonance Imaging (fMRI) in particular is increasingly used not only to localize structures involved in cognitive and perceptual processes but also to study the connectivity in large-scale brain networks that support these functions.

Generally a distinction is made between three types of brain connectivity. *Anatomical connectivity* refers to the physical presence of an axonal projection from one brain area to another. Identification of large axon bundles connecting remote regions in the brain has recently become possible non-invasively in vivo by diffusion weighted Magnetic resonance imaging (DWMRI) and fiber tractography analysis (Johansen-Berg and Behrens, 2009; Jones, 2010). *Functional connectivity* refers to the correlation structure (or more generally: any order of statistical dependency) in the data such that brain areas can be grouped into interacting networks. Finally, *effective connectivity* modeling moves beyond statistical dependency to measures of directed influence and causality within the networks constrained by further assumptions (Friston, 1994).

Recently, effective connectivity techniques that make use of the temporal dynamics in the fMRI signal and employ time series analysis and systems identification theory have become popular. Within this class of techniques two separate developments have been most used: Granger causality analysis (GCA; Goebel et al., 2003; Roebroeck et al., 2005; Valdes-Sosa, 2004) and Dynamic Causal Modeling (DCM; Friston et al., 2003). Despite the common goal, there seem to be differences between the two methods. Whereas GCA explicitly models temporal precedence and uses the concept of Granger causality (or G-causality) mostly formulated in a discrete time-series analysis framework, DCM employs a biophysically motivated generative model formulated in a continuous time dynamic system framework. In this chapter we will give a general causal time-series analysis perspective onto both developments from what we have called the Wiener-Akaike-Granger-Schweder (WAGS) influence formalism (Valdes-Sosa et al., in press).

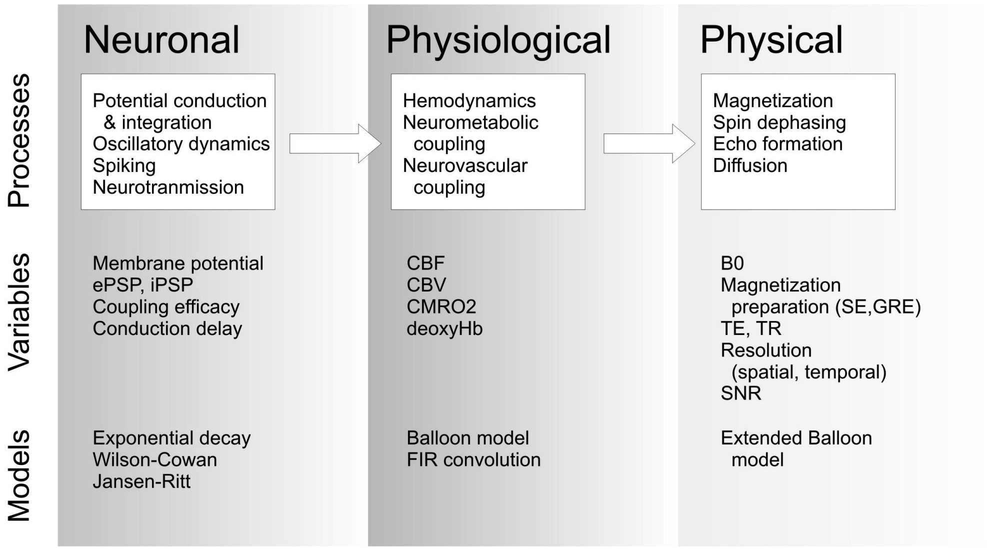
Effective connectivity modeling of neuroimaging data entails the estimation of multivariate mathematical models that benefits from a state space formulation, as we will discuss below. Statistical inference on estimated parameters that quantify the directed influence between brain structures, either individually or in groups (model comparison) then provides information on directed connectivity. In such models, brain structures are defined from at least two viewpoints. From a *structural* viewpoint they correspond to a set of “nodes" that comprise a graph, the purpose of causal discovery being the identification of active links in the graph. The structural model contains i) a selection of the structures in the brain that are assumed to be of importance in the cognitive process or task under investigation, ii) the possible interactions between those structures and iii) the possible effects of exogenous inputs onto the network. The exogenous inputs may be under control of the experimenter and often have the form of a simple indicator function that can represent, for instance, the presence or absence of a visual stimulus in the subject’s view. From a *dynamical* viewpoint brain structures are represented by states or variables that describe time varying neural activity within a time-series model of the measured fMRI time-series data. The functional form of the model equations can embed assumptions on signal dynamics, temporal precedence or physiological processes from which signals originate.

We start this review by focusing on the nature of the fMRI signal in some detail in section 2, separating the treatment into neuronal, physiological and physical processes. In section 3 we review two important formal concepts: causal influence in the Wiener-Akaike-Granger-Schweder tradition and the state space modeling framework, with some emphasis on the relations between discrete and continuous time series models. Building on this discussion, section 4 reviews time series modeling of causality in fMRI data. The review proceeds somewhat chronologically, discussing and comparing the two separate streams of development (GCA and DCM) that have recently begun to be integrated. Finally, section 5 summarizes and discusses the main topics in general dynamic state space models of brain connectivity and provides an outlook on future developments.

2. The fMRI Signal

The fMRI signal reflects the activity within neuronal populations non-invasively with excellent spatial resolution (millimeters down to hundreds of micrometers at high field strength), good temporal resolution (seconds down to hundreds of milliseconds) and whole-brain coverage of the human or animal brain (Logothetis, 2008). Although fMRI is possible with a few different techniques, the Blood Oxygenation Level Dependent (BOLD) contrast mechanism is employed in the great majority of cases. In short, the BOLD fMRI signal is sensitive to changes in blood oxygenation, blood flow and blood volume that result from oxidative glucose metabolism which, in turn, is needed to fuel local neuronal activity (Buxton et al., 2004). This is why fMRI is usually classified as a 'metabolic' or 'hemodynamic' neuroimaging modality. Its superior spatial resolution, in particular, distinguishes it from other functional brain imaging modalities used in humans, such as EEG, MEG and Positron Emission Tomography (PET). Although its temporal resolution is far superior to PET (another 'metabolic' neuroimaging modality) it is still an order of magnitude below that of EEG and MEG, resulting in a relatively sparse sampling of fast neuronal processes, as we will discuss below. The final fMRI signal arises from a complex chain of processes that we can classify into neuronal, physiological and physical processes (Uludag et al., 2005), each of which contain some crucial parameters and variables and have been modeled in various ways as illustrated in Figure 1. We will discuss each of the three classes of processes to explain the intricacies involved in trying to model this causal chain of events with the ultimate goal of estimating neuronal activity and interactions from the measured fMRI signal.

On the neuronal level, it is important to realize that fMRI reflects certain aspects of neuronal functioning more than others. A wealth of processes are continuously in operation at the microscopic level (i.e. in any single neuron), including maintaining a resting potential, post-synaptic conduction and integration (spatial and temporal) of graded excitatory and inhibitory post synaptic potentials (EPSPs and IPSPs) arriving at the dendrites, subthreshold dynamic (possibly oscillatory) potential changes, spike generation at the axon hillock, propagation of spikes by continuous regeneration of



	Neuronal	Physiological	Physical
Processes	Potential conduction & integration Oscillatory dynamics Spiking Neurotranmission	Hemodynamics Neurometabolic coupling Neurovascular coupling	Magnetization Spin dephasing Echo formation Diffusion
Variables	Membrane potential ePSP, iPSP Coupling efficacy Conduction delay	CBF CBV CMRO2 deoxyHb	B0 Magnetization preparation (SE,GRE) TE, TR Resolution (spatial, temporal) SNR
Models	Exponential decay Wilson-Cowan Jansen-Ritt	Balloon model FIR convolution	Extended Balloon model

Figure 1: The neuronal, physiological and physical processes (top row) and variables and parameters involved (middle row) in the complex causal chain of events that leads to the formation of the fMRI signal. The bottom row lists some mathematical models of the sub-processes that play a role in the analysis and modeling of fMRI signals. See main text for further explanation.

the action potential along the axon, and release of neurotransmitter substances into the synaptic cleft at arrival of an action potential at the synaptic terminal. There are many different types of neurons in the mammalian brain that express these processes in different degrees and ways. In addition, there are other cells, such as glia cells, that perform important processes, some of them possibly directly relevant to computation or signaling. As explained below, the fMRI signal is sensitive to the local oxidative metabolism in the brain. This means that, indirectly, it mainly reflects the most energy consuming of the neuronal processes. In primates, post-synaptic processes account for the great majority (about 75%) of the metabolic costs of neuronal signaling events (Attwell and Iadecola, 2002). Indeed, the greater sensitivity of fMRI to post-synaptic activity, rather than axon generation and propagation ('spiking'), has been experimentally verified. For instance, in a simultaneous invasive electrophysiology and fMRI measurement in the primate, Logothetis and colleagues (Logothetis et al., 2001) found the fMRI signal to be more correlated to the mean Local Field Potential (LFP) of the electrophysiological signal, known to reflect post-synaptic graded potentials, than to high-frequency and multi-unit activity, known to reflect spiking. In another study it was shown that, by suppressing action potentials while keeping LFP responses intact by injecting a serotonin agonist, the fMRI response remained intact, again suggesting that LFP is a better predictor for fMRI activity (Rauch et al., 2008). These results confirmed earlier results obtained on the cerebellum of rats (Thomsen et al., 2004).

Neuronal activity, dynamics and computation can be modeled at a different levels of abstraction, including the macroscopic (whole brain areas), mesoscopic (sub-areas to cortical columns) and microscopic level (individual neurons or groups of these). The levels most relevant to modeling fMRI signals are at the macro- and mesoscopic levels. Macroscopic models used to represent considerable expanses of gray matter tissue or sub-cortical structures as Regions Of Interest (ROIs) prominently include single variable deterministic (Friston et al., 2003) or stochastic (autoregressive; Penny et al., 2005; Roebroeck et al., 2005; Valdes-Sosa et al., 2005b) exponential activity decay models. Although the simplicity of such models entail a large degree of abstraction in representing neuronal activity dynamics, their modest complexity is generally well matched to the limited temporal resolution available in fMRI. Nonetheless, more complex multi-state neuronal dynamics models have been investigated in the context of fMRI signal generation. These include the 2 state variable Wilson-Cowan model (Marreiros et al., 2008), with one excitatory and one inhibitory sub-population per ROI and the 3 state variable Jansen-Rit model with a pyramidal excitatory output population and an inhibitory and excitatory interneuron population, particularly in the modeling of simultaneously acquired fMRI and EEG (Valdes-Sosa et al., 2009).

The physiology and physics of the fMRI signal is most easily explained by starting with the physics. We will give a brief overview here and refer to more dedicated overviews (Haacke et al., 1999; Uludag et al., 2005) for extended treatment. The hallmark of Magnetic Resonance (MR) spectroscopy and imaging is the use of the resonance frequency of magnetized nuclei possessing a magnetic moment, mostly protons (hydrogen nuclei, 1H), called 'spins'. Radiofrequency antennas (RF coils) can measure signal from ensembles of spins that resonate in phase at the moment of measurement. The first important physical factor in MR is the main magnetic field strength (B_0), which determines both the resonance frequency (directly proportional to field-strength) and the baseline signal-to-noise ratio of the signal, since higher fields make a larger proportion of spins in the tissue available for measurement. The most used field strengths for fMRI research in humans range from 1,5T (Tesla) to 7T. The second important physical factor – containing several crucial parameters – is the MR pulse-sequence that determines the magnetization preparation of the sample and the way the signal is subsequently acquired. The pulse sequence is essentially a series of radiofrequency pulses, linear magnetic gradient pulses and signal acquisition (readout) events (Bernstein et al., 2004; Haacke et al., 1999). An important variable in a BOLD fMRI pulse sequence is whether it is a gradient-echo (GRE) sequence or a spin-echo (SE) sequence, which determines the granularity of the vascular processes that are reflected in the signal, as explained later this section. These effects are further modulated by the echo-time (time to echo; TE) and repetition time (time to repeat; TR) that are usually set by the end-user of the pulse sequence. Finally, an important variable within the pulse sequence is the type of spatial encoding that is employed. Spatial encoding can primarily be achieved with gradient pulses and it embodies the essence of 'Imaging' in MRI. It is only with spatial encoding that signal can be localized to certain 'voxels' (volume elements) in the tissue. A strength of fMRI as a neuroimaging technique is that an adjustable trade-off is available to the user between spatial resolution, spatial coverage, temporal resolution and signal-to-noise ratio (SNR) of the acquired data. For instance, although fMRI can achieve excellent spatial resolution at good SNR and reasonable temporal resolution, one can choose to sacrifice some spatial resolution to gain a better temporal resolution for any given study. Note, however, that this concerns the resolution and SNR of the data *acquisition*. As explained below, the physiology of fMRI can put fundamental limitations on the nominal resolution and SNR that is achieved in relation to the neuronal processes of interest.

On the physiological level, the main variables that mediate the BOLD contrast in fMRI are cerebral blood flow (CBF), cerebral blood volume (CBV) and the cerebral metabolic rate of oxygen (CMRO2) which all change the oxygen saturation of the blood (as usefully quantified by the concentration of deoxygenated hemoglobin). The BOLD contrast is made possible by the fact that oxygenation of the blood changes its magnetic susceptibility, which has an effect on the MR signal as measured in GRE and SE sequences. More precisely, oxygenated and deoxygenated hemoglobin (oxy-Hb and deoxy-Hb) have different magnetic properties, the former being diamagnetic and the latter paramagnetic. As a consequence, deoxygenated blood creates local microscopic magnetic field gradients, such that local spin ensembles dephase, which is reflected in a lower MR signal. Conversely oxygenation of blood above baseline lowers the concentration of deoxy-Hb, which decreases local spin dephasing and results in a higher MR signal. This means that fMRI is directly sensitive to the relative amount of oxy- and deoxy Hb and to the fraction of cerebral tissue that is occupied by blood (the CBV), which are controlled by local neurovascular coupling processes. Neurovascular processes, in turn, are tightly coupled to neurometabolic processes controlling the rate of oxidative glucose metabolism (the CMRO2) that is needed to fuel neural activity.

Naively one might expect local neuronal activity to quickly increase CMRO2 and increase the local concentration of deoxy-Hb, leading to a lowering of the MR signal. However, this transient increase in deoxy-Hb or the initial dip in the fMRI signal is not consistently observed and, thus, there is a debate whether this signal is robust, elusive or simply not existent (Buxton, 2001; Ugurbil et al., 2003; Uludag, 2010). Instead, early experiments showed that the dynamics of blood flow and blood volume, the hemodynamics, lead to a robust BOLD signal *increase*. Neuronal activity is quickly followed by a large CBF increase that serves the continued functioning of neurons by clearing metabolic by-products (such as CO2) and supplying glucose and oxy-Hb. This CBF response is an overcompensating response, supplying much more oxy-Hb to the local blood system than has been metabolized. As a consequence, within 1-2 seconds, the oxygenation of the blood increases and the MR signal increases. The increased flow also induces a 'ballooning' of the blood vessels, increasing CBV, the proportion of volume taken up by blood, further increasing the signal.

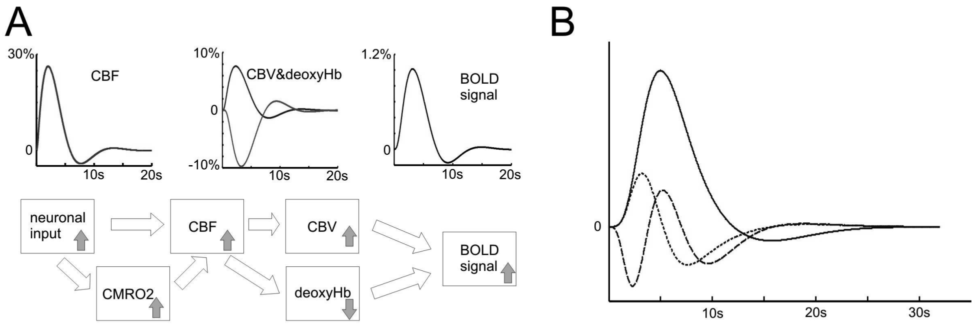


Figure 2: A: Simplified causal chain of hemodynamic events as modeled by the balloon model. Grey arrows show how variable increases (decreases) tend to relate to each other. The dynamic changes after a brief pulse of neuronal activity are plotted for CBF (in red), CBV (in purple), deoxyHb (in green) and BOLD signal (in blue). B: A more abstract representation of the hemodynamic response function as a set of linear basis functions acting as convolution kernels (arbitrary amplitude scaling). Solid line: canonical two-gamma HRF; Dotted line: time derivative; Dashed line: dispersion derivative.

A mathematical characterization of the hemodynamic processes in BOLD fMRI at 1.5-3T has been given in the biophysical balloon model (Buxton et al., 2004, 1998), schematized in Figure 2A. A simplification of the full balloon model has become important in causal models of brain connectivity (Friston et al., 2000). In this simplified model, the dependence of fractional fMRI signal change $\frac{\Delta S}{S}$, on normalized cerebral blood flow f, normalized cerebral blood volume v and normalized deoxyhemoglobin content q is modeled as:

$$\frac{\Delta S}{S} = V_0 \cdot \left[k_1 \cdot (1-q) + k_2 \cdot (1-\frac{q}{v}) + k_3 \cdot (1-v) \right] \tag{1}$$

$$\dot{v}_t = \frac{1}{\tau_0} \left(f_t - v_t^{1/\alpha} \right) \tag{2}$$

$$\dot{q}_t = \frac{1}{\tau_0} \left(\frac{f_t \left(1 - (1-E_0)^{1/f_t} \right)}{E_0} - \frac{q_t}{v_t^{1-1/\alpha}} \right) \tag{3}$$

The term E_0 is the resting oxygen extraction fraction, V_0 is the resting blood volume fraction, τ_0 is the mean transit time of the venous compartment, α is the stiffness component of the model balloon and $\{k_1, k_2, k_3\}$ are calibration parameters. The main simplifications of this model with respect to a more complete balloon model (Buxton et al., 2004) are a one-to-one coupling of flow and volume in (2), thus neglecting the actual balloon effect, and a perfect coupling between flow and metabolism in (3). Friston et al. (2000) augment this model with a putative relation between the a neuronal activity variable z, a flow-inducing signal s, and the normalized cerebral blood flow f. They propose the following relations in which neuronal activity z causes an increase in a vasodilatory signal that is subject to autoregulatory feedback:

$$\dot{s}_t = z_t - \frac{1}{\tau_s} s_t - \frac{1}{\tau_f{}^2} (f_t - 1) \tag{4}$$

$$\dot{f}_t = s_t \tag{5}$$

Here τ_s is the signal decay time constant, τ_f is the time-constant of the feedback autoregulatory mechanism[1], and f is the flow normalized to baseline flow. The physiological interpretation of the autoregulatory mechanism is unspecified, leaving us with a neuronal activity variable z that is measured in units of s^{-2}. The physiology of the hemodynamics contained in differential equations (2) to (5), on the other hand, is more readily interpretable, and when integrated for a brief neuronal input pulse shows the behavior as described above (Figure 2A, upper panel). This simulation highlights a few crucial features. First, the hemodynamic response to a brief neural activity event is sluggish and delayed, entailing that the fMRI BOLD signal is a delayed and low-pass filtered version of underlying neuronal activity. More than the distorting effects of hemodynamic processes on the temporal structure of fMRI signals per se, it is the *difference* in hemodynamics in different parts of the brain that forms a severe confound for dynamic brain connectivity models. Particularly, the delay imposed upon fMRI signals with respect to the underlying neural activity is known to vary between subjects and between different brain regions of the same subject (Aguirre et al., 1998; Saad et al., 2001). Second, although CBF, CBV and deoxyHb changes range in the tens of percents, the BOLD signal change at 1.5T or 3T is in the range of 0.5-2%. Nevertheless,

[1] Note that we have reparametrized the equation here in terms of $\tau_f{}^2$ to make τ_f a proper time constant in units of seconds

the SNR of BOLD fMRI in general is very good in comparison to electrophysiological techniques like EEG and MEG.

Although the balloon model and its variations have played an important role in describing the transient features of the fMRI response and inferring neuronal activity, simplified ways of representing the BOLD signal responses are very often used. Most prominent among these is a linear finite impulse response (FIR) convolution with a suitable kernel. The most used single convolution kernel characterizing the 'canonical' hemodynamic reponse is formed by a superposition of two gamma functions (Glover, 1999), the first characterizing the initial signal increase, the second the later negative undershoot (Figure 2B, solid line):

$$\begin{gathered} h(t) = m_1 t^{\tau_1} e^{(-l_1 t)} - c m_2 t^{\tau_2} e^{(-l_2 t)} \\ m_i = \max\left(t^{\tau_i} e^{(-l_i t)}\right) \end{gathered} \tag{6}$$

With times-to-peak in seconds $\tau_1 = 6$, $\tau_2 = 16$, scale parameters l_i (typically equal to 1) and a relative amplitude of undershoot to peak of $c = 1/6$.

Often, the canonical two-gamma HRF kernel is augmented with one or two additional orthogonalized convolution kernels: a temporal derivative and a dispersion derivative. Together the convolution kernels form a flexible basis function expansion of possible HRF shapes, with the temporal derivative of the canonical accounting for variation in the response delay and the dispersion derivative accounting for variations in temporal response width (Henson et al., 2002; Liao et al., 2002). Thus, the linear basis function representation is a more abstract characterization of the HRF (i.e. further away from the physiology) that still captures the possible variations in responses.

It is an interesting property of hemodynamic processes that, although they are characterized by a large overcompensating reaction to neuronal activity, their effects are highly local. The locality of the hemodynamic reponse to neuronal activity limits the actual spatial resolution of fMRI. The path blood inflow in the brain is from large arteries through arterioles into capillaries where exchange with neuronal tissue takes place at a microscopic level. Blood outflow takes place via venules into the larger veins. The main regulators of blood flow are the arterioles that are surrounded by smooth muscle, although arteries and capillaries are also thought to be involved in blood flow regulation (Attwell et al., 2010). Different hemodynamic parameters have different spatial resolutions. While CBV and CBF changes in all compartments but mostly venules, oxygenation changes mostly in the venules and veins. Thus, the achievable spatial resolution with fMRI is limited by its specificity to the smaller arterioles and venules and microscopic capillaries supplying the tissue, rather than the larger supplying arteries draining veins. The larger vessels have a larger domain of supply or extraction and, as a consequence, their signal is blurred and mislocalized with respect to active tissue. Here, physiology and physics interact in an important way. It can be shown theoretically – by the effects of thermal motion of spin diffusion over time and the distance of the spins to deoxy-Hb – that the origin of the BOLD signal in SE sequences at high main field strengths (larger than 3T) is much more specific to the microscopic vasculature than to the larger arteries and veins. This does not hold for GRE sequences or SE sequences at lower field strengths. The cost of this greater specificity and higher effective spatial resolution is that SE-BOLD has a lower intrinsic SNR than GRE-BOLD. The balloon model equations above are specific to GRE-BOLD at 1.5T and 3T and have been extended to reflect diffusion effects for higher field strengths (Uludag et al., 2009).

In summary, fMRI is an indirect measure of neuronal and synaptic activity. The physiological quantities directly determining signal contrast in BOLD fMRI are hemodynamic quantities such as cerebral blood flow and volume and oxygen metabolism. fMRI can achieve a excellent spatial resolution (millimeters down to hundreds of micrometers at high field strength) with good temporal resolution (seconds down to hundreds of milliseconds). The potential to resolve neuronal population interactions at a high spatial resolution is what drives attempts at causal time series modeling of fMRI data. However, the significant aspects of fMRI that pose challenges for such attempts are i) the enormous dimensionality of the data that contains hundreds of thousands of channels (voxels) ii) the temporal convolution of neuronal events by sluggish hemodynamics that can differ between remote parts of the brain and iii) the relatively sparse temporal sampling of the signal.

3. Causality and state-space models

The inference of causal influence relations from statistical analysis of observed data has two dominant approaches. The first approach is in the tradition of *Granger causality* or *G-causality*, which has its signature in improved predictability of one time series by another. The second approach is based on graphical models and the notion of intervention (Glymour, 2003), which has been formalized using a Bayesian probabilistic framework termed *causal calculus* or *do-calculus* (Pearl, 2009). Interestingly, recent work has combined of the two approaches in a third line of work, termed *Dynamic Structural Systems* (White and Lu, 2010). The focus here will be on the first approach, initially firmly rooted in econometrics and time-series analysis. We will discuss this tradition in a very general form, acknowledging early contributions from Wiener, Akaike, Granger and Schweder and will follow (Valdes-Sosa et al., in press) in refering to the crucial concept as *WAGS influence*.

3.1. Wiener-Akaike-Granger-Schweder (WAGS) influence

The crucial premise of the WAGS statistical causal modeling tradition is that a cause must precede and increase the predictability of its effect. In other words: a variable X_2 influences another variable X_1 if the prediction of X_1 improves when we use past values of X_2, given that all other relevant information (importantly: the past of X1 itself) is taken into account. This type of reasoning can be traced back at least to Hume and is particularly popular in analyzing dynamical data measured as time series. In a formal framework it was originally proposed (in an abstract form) by Wiener (Wiener, 1956), and then introduced into practical data analysis and popularized by Granger (Granger, 1969). A point stressed by Granger is that increased predictability is a necessary but not sufficient condition for a causal relation between time series.

In fact, Granger distinguished true causal relations – only to be inferred in the presence of knowledge of the state of the whole universe – from “prima facie" causal relations that we refer to as “influence" in agreement with other authors (Commenges and Gegout-Petit, 2009). Almost simultaneous with Grangers work, Akaike (Akaike, 1968), and Schweder (Schweder, 1970) introduced similar concepts of influence, prompting (Valdes-Sosa et al., in press) to coin the term WAGS influence (for Wiener-Akaike-Granger-Schweder). This is a generalization of a proposal by placeAalen (Aalen, 1987; Aalen and Frigessi, 2007) who was among the first to point out the connections between Granger's and Schweder's influence concepts. Within this framework we can define several general types of WAGS influence, which are applicable to both Markovian and non-Markovian processes, in discrete or continuous time.

For three vector time series $X_1(t), X_2(t), X_3(t)$ we wish to know if time series $X_1(t)$ is influenced by time series $X_2(t)$ conditional on $X_3(t)$. Here $X_3(t)$ can be considered any set of relevant time series to be controlled for. Let $X[a,b] = \{X(t), t \in [a,b]\}$ denote the history of a time series in the discrete or continuous time interval $[a,b]$ The first categorical distinction is based on what part of the present or future of $X_1(t)$ can be predicted by the past or present of $X_2(\tau_2)$ $\tau_2 \le t$. This leads to the following classification (Florens, 2003; Florens and Fougere, 1996):

1. If $X_2(\tau_2) : \tau_2 < t$, can influence any future value of $X_1(t)$ it is a *global* influence.
2. If $X_2(\tau_2) : \tau_2 < t$, can influence $X_1(t)$ at time t it is a *local* influence.
3. If $X_2(\tau_2) : \tau_2 = t$, can influence $X_1(t)$ it is a *contemporaneous* influence.

A second distinction is based on predicting the whole probability distribution (*strong* influence) or only given moments (*weak* influence). Since the most natural formal definition is one of independence, every influence type amounts to the negation of an independence statement. The two classifications give rise to six types of independence and corresponding influence as set out in Table 1.

To illustrate, $X_1(t)$ is **strongly, conditionally, and globally independent** of $X_2(t)$ given $X_3(t)$, if

$$P(X_1(\infty,t]|X_1(t,-\infty],X_2(t,-\infty],X_3(t,-\infty]) = P(X_1(\infty,t]|X_1(t,-\infty],X_3(t,-\infty])$$

That is: the probability distribution of the future values of X_1 does not depend on the past of X_2, given that the influence of the past of both X_1 and X_3 has been taken into account. When this condition does not hold we say $X_2(t)$ **strongly, conditionally, and globally influences** (**SCGi**) $X_1(t)$ given $X_3(t)$. Here we use a convention for intervals [a,b) which indicates that the left endpoint is included but not the right and that b precedes a. Note that the whole future of X_t is included (hence the term “global"). And the whole past of all time series is considered. This means these definitions accommodate non-Markovian processes (for Markovian processes, we only consider the previous time point). Furthermore, these definitions do not depend on an assumption of linearity or any given functional relationship between time series. Note also that

Table 1: Types of Influence defined by absence of the corresponding independence relations. See text for acronym definitions.

	Strong (Probability Distribution)	Weak (Expectation)
Global (All horizons)	By absence of **strong, conditional, global** independence: $X_2\,(t)\mathbf{SCGi}\;X_1\,(t)\|X_3\,(t)$	By absence of **weak, conditional, global** independence: $X_2\,(t)\mathbf{WCGi}\;X_1\,(t)\|X_3\,(t)$
Local (Immediate future)	By absence of **strong, conditional, local** independence: $X_2\,(t)\mathbf{SCLi}\;X_1\,(t)\|X_3\,(t)$	By absence of **weak, conditional, local** independence: $X_2\,(t)\mathbf{WCLi}\;X_1\,(t)\|X_3\,(t)$
Contemporaneous	By absence of **strong, conditional, contemporaneous** independence: $X_2\,(t)\mathbf{SCCi}\;X_1\,(t)\|X_3\,(t)$	By absence of **weak, conditional, contemporaneous** independence: $X_2\,(t)\mathbf{WCCi}\;X_1\,(t)\|X_3\,(t)$

this definition is appropriate for point processes, discrete and continuous time series, even for categorical (qualitative valued) time series. The only problem with this formulation is that it calls on the whole probability distribution and therefore its practical assessment requires the use of measures such as mutual information that estimate the probability densities nonparametrically.

As an alternative, weak concepts of influence can be defined based on expectations. Consider **weak conditional local independence** *in discrete time*, which is defined:

$$E\left[X_1\left[t+\Delta t\right]\|X_1[t,-\infty],X_2[t,-\infty],X_3[t,-\infty]\right]=E\left[X_1\left[t+\Delta t\right]\|X_1[t,-\infty],X_3[t,-\infty]\right] \tag{7}$$

When this condition does not hold we say X_2 **weakly, conditionally and locally influences** (**WCLi**) X_1 given X_3. To make the implementation this definition insightful, consider a discrete first-order vector auto-regressive (VAR) model for $X = [X_1X_2X_3]$:

$$X\left[t+\Delta t\right]=AX\left[t\right]+e\left[t+\Delta t\right] \tag{8}$$

For this case $E\left[X[t+\Delta t]\|X[t,-\infty]\right] = AX\left[t\right]$, and analyzing influence reduces to finding which of the autoregressive coefficients are zero. Thus, many proposed operational tests of WAGS influence, particularly in fMRI analysis, have been formulated as tests of discrete autoregressive coefficients, although not always of order 1. Within the same model one can operationalize **weak conditional instantaneous independence** *in dis-*

crete time as zero off-diagonal entries in the co-variance matrix of the innovations $e[t]$:

$$\Sigma_e = cov\left[X\left[t+\Delta t\right]\left|X\left[t,-\infty\right]\right]\right. = E\left[X\left[t+\Delta t\right]X'\left[t+\Delta t\right]\left|X\left[t,-\infty\right]\right]\right.$$

In comparison **weak conditional local independence** in *continuous time* is defined:

$$E\left[\,Y_1[t]\right|Y_1(t,-\infty],Y_2(t,-\infty],Y_3(t,-\infty]] = E\left[\,Y_1[t]\right|Y_1(t,-\infty],Y_3(t,-\infty]] \tag{9}$$

Now consider a first-order stochastic differential equation (SDE) model for $Y = [Y_1 Y_2 Y_3]$:

$$dY = BYdt + d\omega \tag{10}$$

Then, since ω is a Wiener process with zero-mean white Gaussian noise as a derivative, $E\left[\,Y[t]\right|Y(t,-\infty]] = B\,Y(t)$and analysing influence amounts to estimating the parameters B of the SDE. However, if one were to observe a discretely sampled version$X[k] = Y(k\Delta t)$ at sampling interval $\Delta t$and model this with the discrete autoregressive model above, this would be inadequate to estimate the SDE parameters for large Δt, since the exact relations between continuous and discrete system matrices are known to be:

$$\begin{array}{l} \mathbf{A} = e^{\mathbf{B}\Delta t} = \mathbf{I} + \sum\limits_{i=1}^{\infty} \frac{\Delta t^i}{i!}\mathbf{B}^i \\ \Sigma_e = \int_t^{t+\Delta t} e^{\mathbf{B}s} \sum_\omega e^{\mathbf{B}s} ds \end{array} \tag{11}$$

The power series expansion of the matrix exponential in the first line shows A to be a weighted sum of successive matrix powers $B^i$of the continuous time system matrix. Thus, the Awill contain contributions from direct (in B) and indirect (in i steps inB^i) causal links between the modeled areas. The contribution of the more indirect links is progressively down-weighted with the number of causal steps from one area to another and is smaller when the sampling interval Δt is smaller. This makes clear that multivariate *discrete* signal models have some undesirable properties for coarsely sampled signals (i.e. a large Δt with respect to the system dynamics), such as fMRI data. Critically, entirely ruling out *indirect* influences is not actually achieved merely by employing a multivariate discrete model. Furthermore, estimated WAGS influence (particularly the relative contribution of indirect links) is dependent on the employed sampling interval. However, the discrete system matrix still represents the presence and direction of influence, possibly mediated through other regions.

When the goal is to estimate WAGS influence for discrete data starting from a continuous time model, one has to model explicitly the mapping to discrete time. Mapping continuous time predictions to discrete samples is a well known topic in engineering and can be solved by explicit integration over discrete time steps as performed in (11) above. Although this defines the mapping from continuous to discrete parameters, it does not solve the reverse assignment of estimating continuous model parameters from discrete data. Doing so requires a solution to the aliasing problem (Mccrorie, 2003) in continuous stochastic system identification by setting sufficient conditions on the matrix logarithm function to make Babove identifiable (uniquely defined) in terms of A. Interesting in this regard is a line of work initiated by Bergstrom (Bergstrom, 1966,

1984) and Phillips (Phillips, 1973, 1974) studying the estimation of continuous time Autoregressive models (McCrorie, 2002), and continuous time Autoregressive Moving Average Models (Chambers and Thornton, 2009) from discrete data. This work rests on the observation that the lag zero covariance matrix $\Sigma_e$will show contemporaneous covariance even if the continuous covariance matrix Σ_ω is diagonal. In other words, the discrete noise becomes correlated over the discrete time-series because the random fluctuations are aggregated over time. Rather than considering this a disadvantage, this approach tries to use both lag information (the AR part) and zero-lag covariance information to identify the underlying continuous linear model.

Notwithstanding the desirability of a continuous time model for consistent inference on WAGS influence, there are a few invariances of discrete VAR models, or more generally discrete Vector Autoregressive Moving Average (VARMA) models that allow their carefully qualified usage in estimating causal influence. The VAR formulation of WAGS influence has the property of invariance under invertible linear filtering. More precisely, a general measure of influence remains unchanged if channels are each premultiplied with different invertible lag operators (Geweke, 1982). However, in practice the order of the estimated VAR model would need to be sufficient to accommodate these operators. Beyond invertible linear filtering, a VARMA formulation has further invariances. Solo (2006) showed that causality in a VARMA model is preserved under sampling and additive noise. More precisely, if both local and contemporaneous influence is considered (as defined above) the VARMA measure is preserved under sampling and under the addition of independent but colored noise to the different channels. Finally, Amendola et al. (2010) shows the class of VARMA models to be closed under aggregation operations, which include both sampling and time-window averaging.

3.2. State-space models

A general state-space model for a continuous vector time-series $y(t)$can be formulated with the set of equations:

$$\begin{aligned} \dot{x}(t) &= f(x(t), v(t), \Theta) + \omega(t) \\ y(t) &= g(x(t), v(t), \Theta) + \varepsilon(t) \end{aligned} \tag{12}$$

This expresses the observed time-series $y(t)$as a function of the state variables $x(t)$, which are possibly hidden (i.e. unobserved) and observed exogenous inputs $v(t)$, which are possibly under control. All parameters in the model are grouped into Θ. Note that some generality is sacrificed from the start since $f$and $g$do not depend on t (The model is autonomous and generates stationary processes) or on $\omega(t)$ or $\varepsilon(t)$, that is: noise enters only additively. The first set of equations, the *transition equations* or *state equations*, describe the evolution of the dynamic system over time in terms of stochastic differential equations (SDEs, though technically only when $\omega(t) = \Sigma\dot{w}(t)$ with $w(t)$ a Wiener process), capturing relations among the hidden state variables $x(t)$ themselves and the influence of exogenous inputs $v(t)$. The second set of equations, the *observation equations* or *measurement equations*, describe how the measurement variables $y(t)$ are obtained from the instantaneous values of the hidden state variables $x(t)$ and the inputs

$v(t)$. In fMRI experiments the exogenous inputs $v(t)$ mostly reflect experimental control and often have the form of a simple indicator function that can represent, for instance, the presence or absence of a visual stimulus. The vector-functions $f$and g can generally be non-linear.

The state-space formalism allows representation of a very large class of stochastic processes. Specifically, it allows representation of both so-called '*black-box*' models, in which parameters are treated as means to adjust the fit to the data without reflecting physically meaningful quantities, and '*grey-box*' models, in which the adjustable parameters *do* have a physical or physiological (in the case of the brain) interpretation. A prominent example of a black-box model in econometric time-series analysis and systems identification is the (discrete) Vector Autoregressive Moving Average model with exogenous inputs (VARMAX model) defined as (Ljung, 1999; Reinsel, 1997):

$$\begin{array}{l} F(B)y_t = G(B)v_t + L(B)e_t \Leftrightarrow \\ \sum_{i=0}^{p} F_i y_{t-i} = \sum_{j=0}^{s} G_j v_{t-j} + \sum_{k=0}^{q} L_k e_{t-k} \end{array} \tag{13}$$

Here, the backshift operator B is defined, for any η_t as $B^i\eta_t = \eta_{t-i}$ and F, G and L are polynomials in the backshift operator, such that e.g. $F(B) = \sum_{i=0}^{p} F_i B^i$ and p, s and q are the dynamic orders of the VARMAX(p,s,q) model. The minimal constraints on (13) to make it identifiable are $F_0 = L_0 = I$, which yields the *standard* VARMAX representation. The VARMAX model and its various reductions (by use of only one or two of the polynomials, e.g. VAR, VARX or VARMA models) have played a large role in time-series prediction and WAGS influence modeling. Thus, in the context of state space models it is important to consider that the VARMAX model form can be equivalently formulated in a discrete linear state space form:

$$\begin{array}{rcl} x_{k+1} & = & \mathbf{A}x_k + \mathbf{B}v_k + \mathbf{K}e_k \\ y_k & = & \mathbf{C}x_k + \mathbf{D}v_k + e_k \end{array} \tag{14}$$

In turn the *discrete* linear state space form can be explicitly accommodated by the *continuous* general state-space framework in (12) when we define:

$$\begin{array}{ll} f(x(t),v(t),\Theta) \simeq Fx(t) + Gv(t) & \omega(t) = \tilde{K}\varepsilon(t) \\ g(x(t),v(t),\Theta) \simeq Hx(t) + Dv(t) & \Theta = \left\{F,G,H,D,\tilde{K},\Sigma_e\right\} \end{array} \tag{15}$$

Again, the exact relations between the discrete and continuous state space parameter matrices can be derived analytically by explicit integration over time (Ljung, 1999). And, as discussed above, wherever discrete data is used to model continuous influence relations the problems of temporal aggregation and aliasing have to be taken into account.

Although analytic solutions for the discretely sampled continuous linear systems exist, the discretization of the *nonlinear* stochastic model (12) does not have a unique global solution. However, physiological models of neuronal population dynamics and hemodynamics are formulated in continuous time and are mostly nonlinear while fMRI data is inherently discrete with low sampling frequencies. Therefore, it is the discretization of the *nonlinear* dynamical stochastic models that is especially relevant to causal analysis of fMRI data. A local linearization approach was proposed by (Ozaki, 1992) as bridge between discrete time series models and nonlinear continuous dynamical systems model. Considering the nonlinear state equation without exogenous input:

$$\dot{x}(t) = f(x(t)) + \omega(t). \tag{16}$$

The essential assumption in local linearization (LL) of this nonlinear system is to consider the Jacobian matrix $J(l,m) = \frac{\partial f_l(X)}{\partial X_m}$ as constant over the time period $[t+\Delta t, t]$. This Jacobian plays the same role as the autoregressive matrix in the linear systems above. Integration over this interval gives the solution:

$$x_{k+1} = x_k + J^{-1}(e^{J\Delta t} - I)f(x_k) + e_{k+1} \tag{17}$$

where I is the identity matrix. Note integration should not be computed this way since it is numerically unstable, especially when the Jacobian is poorly conditioned. A list of robust and fast procedures is reviewed in (Valdes-Sosa et al., 2009). This solution is locally linear but crucially it changes with the state at the beginning of each integration interval; this is how it accommodates nonlinearity (i.e., a state-dependent autoregression matrix). As above, the discretized noise shows instantaneous correlations due to the aggregation of ongoing dynamics within the span of a sampling period. Once again, this highlights the underlying mechanism for problems with temporal sub-sampling and aggregation for some discrete time models of WAGS influence.

4. Dynamic causality in fMRI connectivity analysis

Two streams of developments have recently emerged that make use of the temporal dynamics in the fMRI signal to analyse directed influence (effective connectivity): Granger causality analysis (GCA; Goebel et al., 2003; Roebroeck et al., 2005; Valdes-Sosa, 2004) in the tradition of time series analysis and WAGS influence on the one hand, and Dynamic Causal Modeling (DCM; Friston et al., 2003) in the tradition of system control on the other hand. As we will discuss in the final section, these approaches have recently started developing into an integrated single direction. However, initially each was focused on separate issues that pose challenges for the estimation of causal influence from fMRI data. Whereas DCM is formulated as an explicit grey box state space model to account for the temporal convolution of neuronal events by sluggish hemodynamics, GCA analysis has been mostly aimed at solving the problem of region selection in the enormous dimensionality of fMRI data.

4.1. Hemodynamic deconvolution in a state space approach

While having a long history in engineering, state space modeling was only introduced recently for the inference of neural states from neuroimaging signals. The earliest attempts targeted estimating hidden neuronal population dynamics from scalp-level EEG data (Hernandez et al., 1996; Valdes-Sosa et al., 1999). This work first advanced the idea that state space models and appropriate filtering algorithms are an important tool to estimate the trajectories of hidden neuronal processes from observed neuroimaging data if one can formulate an accurate model of the processes leading from neuronal activity to data records. A few years later, this idea was robustly transferred to fMRI data in the form of DCM (Friston et al., 2003). DCM combines three ideas about causal influence analysis in fMRI data (or neuroimaging data in general), which can be understood in terms of the discussion of the fMRI signal and state space models above (Daunizeau et al., 2009a).

First, neuronal interactions are best modeled at the level of unobserved (latent) signals, instead of at the level of observed BOLD signals. This requires a state space model with a dynamic model of neuronal population dynamics and interactions. The original model that was formulated for the dynamics of neuronal states $x = \{x_1, \ldots, x_N\}$ is a bilinear ODE model:

$$\dot{x} = \mathbf{A}x + \sum v_j \mathbf{B}^j x + \mathbf{C}v \tag{18}$$

That is, the noiseless neuronal dynamics are characterized by a linear term (with entries in **A** representing intrinsic coupling between populations), an exogenous term (with **C** representing driving influence of experimental variables) and a bilinear term (with $\mathbf{B}^j$ representing the modulatory influence of experimental variables on coupling between populations). More recent work has extended this model, e.g. by adding a quadratic term (Stephan et al., 2008), stochastic dynamics (Daunizeau et al., 2009b) or multiple state variables per region (Marreiros et al., 2008).

Second, the latent neuronal dynamics are related to observed data by a generative (forward) model that accounts for the temporal convolution of neuronal events by slow and variably delayed hemodynamics. This generative forward model in DCM for fMRI is exactly the (simplified) balloon model set out in section 2. Thus, for every selected region a single state variable represents the neuronal or synaptic activity of a local population of neurons and (in DCM for BOLD fMRI) four or five more represent hemodynamic quantities such as capillary blood volume, blood flow and deoxy-hemoglobin content. All state variables (and the equations governing their dynamics) that serve the mapping of neuronal activity to the fMRI measurements (including the observation equation) can be called the *observation model*. Most of the physiologically motivated generative model in DCM for fMRI is therefore concerned with an observation model encapsulating hemodynamics. The parameters in this model are estimated conjointly with the parameters quantifying neuronal connectivity. Thus, the forward biophysical model of hemodynamics is ‘inverted’ in the estimation procedure to achieve a deconvolution of fMRI time series and obtain estimates of the underlying neuronal states. DCM has also been applied to EEG/MEG, in which case the observation model encapsulates the lead-field matrix from neuronal sources to EEG electrodes or MEG sensors (Kiebel et al., 2009).

Third, the approach to estimating the hidden state trajectories (i.e. filtering and smoothing) and parameter values in DCM is cast in a Bayesian framework. In short, Bayes’ theorem is used to combine priors $p(\Theta|M)$and likelihood $p(y|\Theta,M)$into the *marginal likelihood* or *evidence*:

$$p(y|M) = \int p(y|\Theta, M)\, p(\Theta|M)\, d\Theta \tag{19}$$

Here, the model M is understood to define the priors on all parameters and the likelihood through the generative models for neuronal dynamics and hemodynamics. A posterior for the parameters $p(\Theta|y, M)$ can be obtained as the distribution over parameters which maximizes the evidence (19). Since this optimization problem has no analytic solution and is intractable with numerical sampling schemes for complex models, such as DCM, approximations must be used. The inference approach for DCM relies on variational Bayes methods (Beal, 2003) that optimize an approximation density $q(\Theta)$ to the posterior. The approximation density is taken to have a Gaussian form, which is often referred to as the "Laplace approximation" (Friston et al., 2007). In addition to the approximate posterior on the parameters, the variational inference will also result into a lower bound on the evidence, sometimes referred to as the "free energy". This lower bound (or other approximations to the evidence, such as the Akaike Information Criterion or the Bayesian Information Criterion) are used for model comparison (Penny et al., 2004). Importantly, these quantities explicitly balance goodness-of-it against model complexity as a means of avoiding overfitting.

An important limiting aspect of DCM for fMRI is that the models M that are compared also (implicitly) contain an *anatomical model* or *structural model* that contains i) a selection of the ROIs in the brain that are assumed to be of importance in the cognitive process or task under investigation, ii) the possible interactions between those structures and iii) the possible effects of exogenous inputs onto the network. In other words, each model M specifies the nodes and edges in a directed (possibly cyclic) structural graph model. Since the anatomical model also determines the selected part y of the total dataset (all voxels) one cannot use the evidence to compare different anatomical models. This is because the evidence of different anatomical models is defined over different data. Applications of DCM to date invariably use very simple anatomical models (typically employing 3-6 ROIs) in combination with its complex parameter-rich dynamical model discussed above. The clear danger with overly simple anatomical models is that of spurious influence: an erroneous influence found between two selected regions that in reality is due to interactions with additional regions which have been ignored. Prototypical examples of spurious influence, of relevance in brain connectivity, are those between unconnected structures A and B that receive common input from, or are intervened by, an unmodeled region C.

4.2. Exploratory approaches for model selection

Early applications of WAGS influence to fMRI data were aimed at counteracting the problems with overly restrictive anatomical models by employing more permissive anatomical models in combination with a simple dynamical model (Goebel et al., 2003; Roebroeck et al., 2005; Valdes-Sosa, 2004). These applications reflect the observation that estimation of mathematical models from time-series data generally has two important aspects: model selection and model identification (Ljung, 1999). In the *model selection* stage a class of models is chosen by the researcher that is deemed suitable for the problem at hand. In the *model identification* stage the parameters in the chosen model class are estimated from the observed data record. In practice, model selection and identification often occur in a somewhat interactive fashion where, for instance, model selection can be informed by the fit of different models to the data achieved in an identification step. The important point is that model selection involves a mixture of choices and assumptions on the part of the researcher and the information gained from the data-record itself. These considerations indicate that an important distinction must be made between exploratory and confirmatory approaches, especially in structural model selection procedures for brain connectivity. Exploratory techniques use information in the data to investigate the relative applicability of many models. As such, they have the potential to detect 'missing' regions in structural models. Confirmatory approaches, such as DCM, test hypotheses about connectivity within a set of models assumed to be applicable. Sources of common input or intervening causes are taken into account in a multivariate confirmatory model, but only if the employed structural model allows it (i.e. if the common input or intervening node is incorporated in the model).

The technique of Granger Causality Mapping (GCM) was developed to explore all regions in the brain that interact with a single selected reference region using autoregressive modeling of fMRI time-series (Roebroeck et al., 2005). By employing a simple bivariate model containing the reference region and, in turn, every other voxel in the brain, the sources and targets of influence for the reference region can be mapped. It was shown that such an 'exploratory' mapping approach can form an important tool in structural model selection. Although a bivariate model does not discern direct from indirect influences, the mapping approach locates potential sources of common input and areas that could act as intervening network nodes. In addition, by settling for a bivariate model one trivially avoids the conflation of direct and indirect influences that can arise in discrete AR model due to temporal aggregation, as discussed above. Other applications of autoregressive modeling to fMRI data have considered full multivariate models on large sets of selected brain regions, illustrating the possibility to estimate high-dimensional dynamical models. For instance, Valdes-Sosa (2004) and Valdes-Sosa et al. (2005b) applied these models to parcellations of the entire cortex in conjunction with sparse regression approaches that enforce an implicit structural model selection within the set of parcels. In another more recent example (Deshpande et al., 2008) a full multivariate model was estimated over 25 ROIs (that were found to be activated in the investigated task) together with an explicit reduction procedure to prune regions from the full model as a structural model selection procedure. Additional variants of VAR model based causal inference that has been applied to fMRI include time varying influence (Havlicek et al., 2010), blockwise (or 'cluster-wise') influence from one group of variables to another (Barrett et al., 2010; Sato et al., 2010) and frequency-decomposed influence (Sato et al., 2009).

The initial developments in autoregressive modeling of fMRI data led to a number of interesting applications studying human mental states and cognitive processes, such as gestural communication (Schippers et al., 2010), top-down control of visual spatial attention (Bressler et al., 2008), switching between executive control and default-mode networks (Sridharan et al., 2008), fatigue (Deshpande et al., 2009) and the resting state (Uddin et al., 2009). Nonetheless, the lack of AR models to account for the varying hemodynamics convolving the signals of interest and aggregation of dynamics between time samples has prompted a set of validation studies evaluating the conditions under which discrete AR models can provide reliable connectivity estimates. In (Roebroeck et al., 2005) simulations were performed to validate the use of bivariate AR models in the face of hemodynamic convolution and sampling. They showed that under these conditions (even without variability in hemodynamics) AR estimates for a unidirectional influence are biased towards inferring bidirectional causality, a well known problem when dealing with aggregated time series (Wei, 1990). They then went on to show that instead unbiased non-parametric inference for bivariate AR models can be based on a difference of influence terms ($X \to Y - Y \to X$). In addition, they posited that inference on such influence estimates should always include experimental modulation of influence, in order to rule out hemodynamic variation as an underlying reason for spurious causality. In Deshpande et al. (2010) the authors simulated fMRI data by manipulating the causal influence and neuronal delays between local field potentials (LFPs) acquired from the macaque cortex and varying the hemodynamic delays of a convolving hemodynamic response function and the signal-to-noise ratio (SNR) and the sampling period of the final simulated fMRI data. They found that in multivariate (4 dimensional) simulations with hemodynamic and neuronal delays drawn from a uniform random distribution correct network detection from fMRI was well above chance and was up to 90% under conditions of fast sampling and low measurement noise. Other studies confirmed the observation that techniques with intermediate temporal resolution, such as fMRI, can yield good estimates of the causal connections based on AR models (Stevenson and Kording, 2010), even in the face of variable hemodynamics (Ryali et al., 2010). However, another recent simulation study, investigating a host of connectivity methods concluded low detection performance of directed influence by AR models under general conditions (Smith et al., 2010).

4.3. Toward integrated models

David et al. (2008) aimed at direct comparison of autoregressive modeling and DCM for fMRI time series and explicitly pointed at deconvolution of variable hemodynamics for causality inferences. The authors created a controlled animal experiment where gold standard validation of neuronal connectivity estimation was provided by intracranial EEG (iEEG) measurements. As discussed extensively in Friston (2009b) and Roebroeck et al. (2009a) such a validation experiment can provide important information on best practices in fMRI based brain connectivity modeling that, however, need to be carefully discussed and weighed. In David et al.'s study, simultaneous fMRI, EEG, and iEEG were measured in 6 rats during epileptic episodes in which spike-and-wave discharges (SWDs) spread through the brain. fMRI was used to map the hemodynamic response throughout the brain to seizure activity, where ictal and interictal states were quantified by the simultaneously recorded EEG. Three structures were selected by the authors as the crucial nodes in the network that generates and sustains seizure activity and further analysed with i) DCM, ii) simple AR modeling of the fMRI signal and iii) AR modeling applied to neuronal state-variable estimates obtained with a hemodynamic deconvolution step. By applying G-causality analysis to deconvolved fMRI time-series, the stochastic dynamics of the linear state-space model are augmented with the complex biophysically motivated observation model in DCM. This step is crucial if the goal is to compare the dynamic connectivity models and draw conclusions on the relative merits of linear stochastic models (explicitly estimating WAGS influence) and bilinear deterministic models. The results showed both AR analysis after deconvolution and DCM analysis to be in accordance with the gold-standard iEEG analyses, identifying the most pertinent influence relations undisturbed by variations in HRF latencies. In contrast, the final result of simple AR modeling of the fMRI signal showed less correspondence with the gold standard, due to the confounding effects of different hemodynamic latencies which are not accounted for in the model.

Two important lessons can be drawn from David et al.'s study and the ensuing discussions (Bressler and Seth, 2010; Daunizeau et al., 2009a; David, 2009; Friston, 2009b,a; Roebroeck et al., 2009a,b). First, it confirms again the distorting effects of hemodynamic processes on the temporal structure of fMRI signals and, more importantly, that the difference in hemodynamics in different parts of the brain can form a confound for dynamic brain connectivity models (Roebroeck et al., 2005). Second, state-space models that embody observation models that connect latent neuronal dynamics to observed fMRI signals have a potential to identify causal influence unbiased by this confound. As a consequence, substantial recent methodological work has aimed at combining different models of latent neuronal dynamics with a form of a hemodynamic observation model in order to provide an inversion or filtering algorithm for estimation of parameters and hidden state trajectories. Following the original formulation of DCM that provides a bilinear ODE form for the hidden neuronal dynamics, attempts have been made at explicit integration of hemodynamics convolution with stochastic dynamic models that are interpretable in the framework of WAGS influence.

For instance in (Ryali et al., 2010), following earlier work (Penny et al., 2005; Smith et al., 2009), a discrete state space model with a bi-linear vector autoregressive model to quantify dynamic neuronal state evolution and both intrinsic and modulatory interactions is proposed:

$$\begin{aligned} x_k &= \mathbf{A}x_{k-1} + \textstyle\sum_{j=1} v_k^j \mathbf{B}^j x_{k-1} + \mathbf{C}v_k^j + \varepsilon_k \\ \mathbf{x}_k^m &= \left[x_k^m, x_{k-1}^m, \cdots, x_{k-L+1}^m\right] \\ y_k^m &= \beta^m \Phi \mathbf{x}_k^m + e_k^m \end{aligned} \tag{20}$$

Here, we index exogenous inputs with j and ROIs with $m$in superscripts. The entries in the autoregressive matrix $\mathbf{A}$, exogenous influence matrix $\mathbf{C}$ and bi-linear matrices $\mathbf{B}^j$have the same interpretation as in deterministic DCM. The relation between observed BOLD-fMRI data $y$and latent neuronal sources x is modeled by a temporal embedding of into $\mathbf{x}^m$for each region or ROI m. This allows convolution with a flexible basis function expansion of possible HRF shapes to be represented by a simple matrix multiplication $\beta^m \Phi x_k^m$in the observation equation. Here Φ contains the temporal basis functions in Figure 2B and β^m the basis function parameters to be estimated. By estimating basis function parameters individually per region, variations in the HRF shape between region can be accounted for and the confounding effects of these on WAGS influence estimate can be avoided. Ryali et al. found that robust estimates of parameters $\Theta = \left\{\mathbf{A}, \mathbf{B}^j, \mathbf{C}, \beta^m, \Sigma_\varepsilon, \Sigma_e\right\}$ and states x_k can be obtained from a variational Bayesian approach. In their simulations, they show that a state-space model with interactions modeled at the latent level can compensate well for the effects of HRF variability, even when relative HRF delays are opposed to delayed interactions. Note, however, that subsampling of the BOLD signal is not explicitly characterized in their state-space model.

A few interesting variations on this discrete state-space modeling have recently been proposed. For instance in (Smith et al., 2009) a switching linear systems model for latent neuronal state evolution, rather than a bi-linear model was used. This model represents experimental modulation of connections as a random variable, to be learned from the data. This variable switches between different linear system instantiations that each characterize connectivity in a single experimental condition. Such a scheme has the important advantage that an n-fold cross validation approach can be used to obtain a measure of absolute model-evidence (rather than relative between a selected set of models). Specifically, one could learn parameters for each context-specific linear system with knowledge of the timing of changing experimental conditions in a training data set. Then the classification accuracy of experimental condition periods in a test data set based on connectivity will provide a absolute model-fit measure, controlled for model complexity, which can be used to validate overall usefulness of the fitted model. In particular, this can point to important brain regions missing from the model incase of poor classification accuracy.

Another related line of developments instead has involved generalizing the ODE models in DCM for fMRI to stochastic dynamic models formulated in continuous time (Daunizeau et al., 2009b; Friston et al., 2008). An early exponent of this approach used local linearization in a (generalized) Kalman filter to estimate states and parameters in a non-linear SDE models of hemodynamics (Riera et al., 2004). Interestingly, the inclusion of stochastics in the state equations makes inference on coupling parameters of such models usefully interpretable in the framework of WAGS influence. This hints at the ongoing convergence, in modeling of brain connectivity, of time series approaches to causality in a discrete time tradition and dynamic systems and control theory approaches in a continuous time tradition.

5. Discussion and Outlook

The modeling of an enormously complex biological system such as the brain has many challenges. The abstractions and choices to be made in useful models of brain connectivity are therefore unlikely to be accommodated by one single ‘master’ model that does better than all other models on all counts. Nonetheless, the ongoing development efforts towards improved approaches are continually extending and generalizing the contexts in which dynamic time series models can be applied. It is clear that state space modeling and inference on WAGS influence are fundamental concepts within this endeavor. We end here with some considerations of dynamic brain connectivity models that summarize some important points and anticipate future developments.

We have emphasized that WAGS influence models of brain connectivity have largely been aimed at data driven exploratory analysis, whereas biophysically motivated state space models are mostly used for hypothesis-led confirmatory analysis. This is especially relevant in the interaction between model selection and model identification. Exploratory techniques use information in the data to investigate the relative applicability of many models. As such, they have the potential to detect 'missing' regions in anatomical models. Confirmatory approaches test hypotheses about connectivity within a set of models assumed to be applicable.

As mentioned above, the WAGS influence approach to statistical analysis of causal influence that we focused on here is complemented by the interventional approach rooted in the theory of graphical models and causal calculus. Graphical causal models have been recently applied to brain connectivity analysis of fMRI data (Ramsey et al., 2009). Recent work combining the two approaches (White and Lu, 2010) possibly leads the way to a combined causal treatment of brain imaging data incorporating dynamic models and interventions. Such a combination could enable incorporation of direct manipulation of brain activity by (for example) transcranial magnetic stimulation (Pascual-Leone et al., 2000; Paus, 1999; Walsh and Cowey, 2000) into the current state space modeling framework.

Causal models of brain connectivity are increasingly inspired by biophysical theories. For fMRI this is primarily applicable in modeling the complex chain of events separating neuronal population activity from the BOLD signal. Inversion of such a model (in state space form) by a suitable filtering algorithm amounts to a model-based deconvolution of the fMRI signal resulting in an estimate of latent neuronal population activity. If the biophysical model is appropriately formulated to be identifiable (possibly including priors on relevant parameters), it can take variation in the hemodynamics between brain regions into account that can otherwise confound time series causality analyses of fMRI signals. Although models of hemodynamics for causal fMRI analysis have reached a reasonable level of complexity, the models of neuronal dynamics used to date have remained simple, comprising one or two state variables for an entire cortical region or subcortical structure. Realistic dynamic models of neuronal activity have a long history and have reached a high level of sophistication (Deco et al., 2008; Markram, 2006). It remains an open issue to what degree complex realistic equation systems can be embedded in analysis of fMRI – or in fact: any brain imaging modality – and result in identifiable models of neuronal connectivity and computation.

Two recent developments create opportunities to increase complexity and realism of neuronal dynamics models and move the level of modeling from the macroscopic (whole brain areas) towards the mesoscopic level comprising sub-populations of areas and cortical columns. First, the fusion of multiple imaging modalities, possibly simultaneously recorded, has received a great deal of attention. Particularly, several attempts to model-driven fusion of simultaneousy recorded fMRI and EEG data, by inverting a separate observation model for each modality while using the same underlying neuronal model, have been reported (Deneux and Faugeras, 2010; Riera et al., 2007; Valdes-Sosa et al., 2009). This approach holds great potential to fruitfully combine the superior spatial resolution of fMRI with the superior temporal resolution of EEG. In (Valdes-Sosa et al., 2009) anatomical connectivity information obtained from diffusion tensor imaging and fiber tractography is also incorporated. Second, advances in MRI technology, particularly increases of main field strength to 7T (and beyond) and advances in parallel imaging (de Zwart et al., 2006; Heidemann et al., 2006; Pruessmann, 2004; Wiesinger et al., 2006), greatly increase the level spatial detail that are accessible with fMRI. For instance, fMRI at 7T with sufficient spatial resolution to resolve orientation columns in human visual cortex has been reported (Yacoub et al., 2008).

The development of state space models for causal analysis of fMRI data has moved from discrete to continuous and from deterministic to stochastic models. Continuous models with stochastic dynamics have desirable properties, chief among them a robust inference on causal influence interpretable in the WAGS framework, as discussed above. However, dealing with continuous stochastic models leads to technical issues such as the properties and interpretation of Wiener processes and Ito calculus (Friston, 2008). A number of inversion or filtering methods for continuous stochastic models have been recently proposed, particularly for the goal of causal analysis of brain imaging data, including the local linearization and innovations approach (Hernandez et al., 1996; Riera et al., 2004), dynamic expectation maximization (Friston et al., 2008) and generalized filtering (Friston et al., 2010). The ongoing development of these filtering methods, their validation and their scalability towards large numbers of state variables will be a topic of continuing research.

Acknowledgments

The authors thank Kamil Uludag for comments and discussion.

References

Odd O. Aalen. Dynamic modeling and causality. *Scandinavian Actuarial journal*, pages 177–190, 1987.

O.O. Aalen and A. Frigessi. What can statistics contribute to a causal understanding? *Board of the Foundation of the Scandinavian journal of Statistics*, 34:155–168, 2007.

G. K. Aguirre, E. Zarahn, and M. D'Esposito. The variability of human, bold hemodynamic responses. *Neuroimage*, 8(4):360–9, 1998.

H Akaike. On the use of a linear model for the identification of feedback systems. *Annals of the Institute of statistical mathematics*, 20(1):425–439, 1968.

A Amendola, M Niglio, and C Vitale. Temporal aggregation and closure of VARMA models: Some new results. In F. Palumbo et al., editors, *Data Analysis and Classification: Studies in Classification, Data Analysis, and Knowledge Organization*, pages 435–443. Springer Berlin/Heidelberg, 2010.

D. Attwell and C. Iadecola. The neural basis of functional brain imaging signals. *Trends Neurosci*, 25(12):621–5, 2002.

D. Attwell, A. M. Buchan, S. Charpak, M. Lauritzen, B. A. Macvicar, and E. A. Newman. Glial and neuronal control of brain blood flow. *Nature*, 468(7321):232–43, 2010.

A. B. Barrett, L. Barnett, and A. K. Seth. Multivariate granger causality and generalized variance. *Phys Rev E Stat Nonlin Soft Matter Phys*, 81(4 Pt 1):041907, 2010.

M.J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, University College London, 2003.

A R Bergstrom. Nonrecursive models as discrete approximations to systems of stochastic differential equations. *Econometrica*, 34:173–182, 1966.

A R Bergstrom. Continuous time stochastic models and issues of aggregation. In Z. Griliches and M.D. Intriligator, editors, *Handbook of econometrics*, volume II. Elsevier, 1984.

M.A. Bernstein, K.F. King, and X.J. Zhou. *Handbook of MRI Pulse Sequences*. Elsevier Academic Press, urlington, 2004.

S. L. Bressler and A. K. Seth. Wiener-granger causality: A well established methodology. *Neuroimage*, 2010.

S. L. Bressler, W. Tang, C. M. Sylvester, G. L. Shulman, and M. Corbetta. Top-down control of human visual cortex by frontal and parietal cortex in anticipatory visual spatial attention. *J Neurosci*, 28(40):10056–61, 2008.

R. B. Buxton. The elusive initial dip. *Neuroimage*, 13(6 Pt 1):953–8, 2001.

R. B. Buxton, E. C. Wong, and L. R. Frank. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magn Reson Med*, 39(6):855–64, 1998.

R. B. Buxton, K. Uludag, D. J. Dubowitz, and T. T. Liu. Modeling the hemodynamic response to brain activation. *Neuroimage*, 23 Suppl 1:S220–33, 2004.

Marcus J Chambers and Michael A Thornton. Discrete time representation of continuous time arma processes, 2009.

Daniel Commenges and Anne Gegout-Petit. A general dynamical statistical model with possible causal interpretation. *journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):1–43, 2009.

J. Daunizeau, O. David, and K. E. Stephan. Dynamic causal modelling: A critical review of the biophysical and statistical foundations. *Neuroimage*, 2009a.

J. Daunizeau, K. J. Friston, and S. J. Kiebel. Variational bayesian identification and prediction of stochastic nonlinear dynamic causal models. *Physica D*, 238(21):2089–2118, 2009b.

O. David. fmri connectivity, meaning and empiricism comments on: Roebroeck et al. the identification of interacting networks in the brain using fmri: Model selection, causality and deconvolution. *Neuroimage*, 2009.

O. David, I. Guillemain, S. Saillet, S. Reyt, C. Deransart, C. Segebarth, and A. Depaulis. Identifying neural drivers with functional mri: an electrophysiological validation. *PLoS Biol*, 6(12):2683–97, 2008.

J. A. de Zwart, P. van Gelderen, X. Golay, V. N. Ikonomidou, and J. H. Duyn. Accelerated parallel imaging for functional imaging of the human brain. *NMR Biomed*, 19 (3):342–51, 2006.

G. Deco, V. K. Jirsa, P. A. Robinson, M. Breakspear, and K. Friston. The dynamic brain: from spiking neurons to neural masses and cortical fields. *PLoS Comput Biol*, 4(8):e1000092, 2008.

T. Deneux and O. Faugeras. Eeg-fmri fusion of paradigm-free activity using kalman filtering. *Neural Comput*, 22(4):906–48, 2010.

G. Deshpande, X. Hu, R. Stilla, and K. Sathian. Effective connectivity during haptic perception: a study using granger causality analysis of functional magnetic resonance imaging data. *Neuroimage*, 40(4):1807–14, 2008.

G. Deshpande, S. LaConte, G. A. James, S. Peltier, and X. Hu. Multivariate granger causality analysis of fmri data. *Hum Brain Mapp*, 30(4):1361–73, 2009.

G. Deshpande, K. Sathian, and X. Hu. Effect of hemodynamic variability on granger causality analysis of fmri. *Neuroimage*, 52(3):884–96, 2010.

J Florens. Some technical issues in defining causality. *journal of Econometrics*, 112: 127–128, 2003.

J.P. Florens and D. Fougere. Noncausality in continuous time. *Econometrica*, 64(5): 1195–1212, 1996.

K. Friston. Functional and effective connectivity in neuroimaging: A synthesis. *Hum Brain Mapp*, 2:56–78, 1994.

K. Friston. Beyond phrenology: what can neuroimaging tell us about distributed circuitry? *Annu Rev Neurosci*, 25:221–50, 2002.

K. Friston. Dynamic causal modeling and granger causality comments on: The identification of interacting networks in the brain using fmri: Model selection, causality and deconvolution. *Neuroimage*, 2009a.

K. Friston, J. Mattout, N. Trujillo-Barreto, J. Ashburner, and W. Penny. Variational free energy and the laplace approximation. *Neuroimage*, 34(1):220–34, 2007.

K. J. Friston, A. Mechelli, R. Turner, and C. J. Price. Nonlinear responses in fmri: the balloon model, volterra kernels, and other hemodynamics. *Neuroimage*, 12(4): 466–77, 2000.

K. J. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *Neuroimage*, 19 (4):1273–302, 2003.

K. J. Friston, N. Trujillo-Barreto, and J. Daunizeau. Dem: a variational treatment of dynamic systems. *Neuroimage*, 41(3):849–85, 2008.

Karl Friston. Hierarchical models in the brain. *PLoS Computational Biology*, 4, 2008.

Karl Friston. Causal modelling and brain connectivity in functional magnetic resonance imaging. *PLoS biology*, 7:e33, 2009b.

Karl Friston, Klaas Stephan, Baojuan Li, and Jean Daunizeau. Generalised filtering. *Mathematical Problems in Engineering*, 2010:1–35, 2010.

J. F. Geweke. Measurement of linear dependence and feedback between multiple time series. *journal of the American Statistical Association*, 77(378):304–324, 1982.

G. H. Glover. Deconvolution of impulse response in event-related bold fmri. *Neuroimage*, 9(4):416–29, 1999.

C. Glymour. Learning, prediction and causal bayes nets. *Trends Cogn Sci*, 7(1):43–48, 2003.

R. Goebel, A. Roebroeck, D. S. Kim, and E. Formisano. Investigating directed cortical interactions in time-resolved fmri data using vector autoregressive modeling and granger causality mapping. *Magn Reson Imaging*, 21(10):1251–61, 2003.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

E.M. Haacke, R.W. Brown, M.R. Thompson, and R. Venkatesan. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. John Wiley and Sons, Inc, New York, 1999.

M. Havlicek, J. Jan, M. Brazdil, and V. D. Calhoun. Dynamic granger causality based on kalman filter for evaluation of functional network connectivity in fmri data. *Neuroimage*, 53(1):65–77, 2010.

R. M. Heidemann, N. Seiberlich, M. A. Griswold, K. Wohlfarth, G. Krueger, and P. M. Jakob. Perspectives and limitations of parallel mr imaging at high field strengths. *Neuroimaging Clin N Am*, 16(2):311–20, 2006.

R. N. Henson, C. J. Price, M. D. Rugg, R. Turner, and K. J. Friston. Detecting latency differences in event-related bold responses: application to words versus nonwords and initial versus repeated face presentations. *Neuroimage*, 15(1):83–97, 2002.

J. L. Hernandez, P. A. Valdés, and P. Vila. Eeg spike and wave modelled by a stochastic limit cycle. *NeuroReport*, 1996.

B. Horwitz, K. J. Friston, and J. G. Taylor. Neural modeling and functional brain imaging: an overview. *Neural Netw*, 13(8-9):829–46, 2000.

H. Johansen-Berg and T.E.J Behrens, editors. *Diffusion MRI: From quantitative measurement to in-vivo neuroanatomy*. Academic Press, London, 2009.

D.K. Jones, editor. *Diffusion MRI: Theory, Methods, and Applications*. Oxford University Press, Oxford, 2010.

S. J. Kiebel, M. I. Garrido, R. Moran, C. C. Chen, and K. J. Friston. Dynamic causal modeling for eeg and meg. *Hum Brain Mapp*, 30(6):1866–76, 2009.

C. H. Liao, K. J. Worsley, J. B. Poline, J. A. Aston, G. H. Duncan, and A. C. Evans. Estimating the delay of the fmri response. *Neuroimage*, 16(3 Pt 1):593–606, 2002.

L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, New Jersey, 2nd edition, 1999.

N. K. Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453 (7197):869–78, 2008.

N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann. Neurophysiological investigation of the basis of the fmri signal. *Nature*, 412(6843):150–7, 2001.

H. Markram. The blue brain project. *Nat Rev Neurosci*, 7(2):153–60, 2006.

A. C. Marreiros, S. J. Kiebel, and K. J. Friston. Dynamic causal modelling for fmri: a two-state model. *Neuroimage*, 39(1):269–78, 2008.

J. Roderick McCrorie. The likelihood of the parameters of a continuous time vector autoregressive model. *Statistical Inference for Stochastic Processes*, 5:273–286, 2002.

J. Roderick Mccrorie. The problem of aliasing in identifying finite parameter continuous time stochastic models. *Acta Applicandae Mathematicae*, 79:9–16, 2003.

A. R. McIntosh. Contexts and catalysts: a resolution of the localization and integration of function in the brain. *Neuroinformatics*, 2(2):175–82, 2004.

T Ozaki. A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: A local linearization approach. *Statistica Sinica*, 2:113–135, 1992.

A. Pascual-Leone, V. Walsh, and J. Rothwell. Transcranial magnetic stimulation in cognitive neuroscience–virtual lesion, chronometry, and functional connectivity. *Curr Opin Neurobiol*, 10(2):232–7, 2000.

T. Paus. Imaging the brain before, during, and after transcranial magnetic stimulation. *Neuropsychologia*, 37(2):219–24, 1999.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, 2nd edition, 2009.

W. Penny, Z. Ghahramani, and K. Friston. Bilinear dynamical systems. *Philos Trans R Soc Lond B Biol Sci*, 360(1457):983–93, 2005.

W. D. Penny, K. E. Stephan, A. Mechelli, and K. J. Friston. Comparing dynamic causal models. *Neuroimage*, 22(3):1157–72, 2004.

Peter C.B. Phillips. The problem of identification in finite parameter continuous time models. *journal of Econometrics*, 1:351–362, 1973.

Peter C.B. Phillips. The estimation of some continuous time models. *Econometrica*, 42:803–823, 1974.

K. P. Pruessmann. Parallel imaging at high field strength: synergies and joint potential. *Top Magn Reson Imaging*, 15(4):237–44, 2004.

J. D. Ramsey, S. J. Hanson, C. Hanson, Y. O. Halchenko, R. A. Poldrack, and C. Glymour. Six problems for causal inference from fmri. *Neuroimage*, 49(2):1545–58, 2009.

A. Rauch, G. Rainer, and N. K. Logothetis. The effect of a serotonin-induced dissociation between spiking and perisynaptic activity on bold functional mri. *Proc Natl Acad Sci U S A*, 105(18):6759–64, 2008.

G.C. Reinsel. *Elements of Multivariate Time Series Analysis*. Springer-Verlag, New York, 2nd edition, 1997.

J. J. Riera, J. Watanabe, I. Kazuki, M. Naoki, E. Aubert, T. Ozaki, and R. Kawashima. A state-space model of the hemodynamic approach: nonlinear filtering of bold signals. *Neuroimage*, 21(2):547–67, 2004.

J. J. Riera, J. C. Jimenez, X. Wan, R. Kawashima, and T. Ozaki. Nonlinear local electrovascular coupling. ii: From data to neuronal masses. *Hum Brain Mapp*, 28(4): 335–54, 2007.

A. Roebroeck, E. Formisano, and R. Goebel. Mapping directed influence over the brain using granger causality and fmri. *Neuroimage*, 25(1):230–42, 2005.

A. Roebroeck, E. Formisano, and R. Goebel. The identification of interacting networks in the brain using fmri: Model selection, causality and deconvolution. *Neuroimage*, 2009a.

A. Roebroeck, E. Formisano, and R. Goebel. Reply to friston and david after comments on: The identification of interacting networks in the brain using fmri: Model selection, causality and deconvolution. *Neuroimage*, 2009b.

S. Ryali, K. Supekar, T. Chen, and V. Menon. Multivariate dynamical systems models for estimating causal interactions in fmri. *Neuroimage*, 2010.

Z. S. Saad, K. M. Ropella, R. W. Cox, and E. A. DeYoe. Analysis and use of fmri response delays. *Hum Brain Mapp*, 13(2):74–93, 2001.

R. Salmelin and J. Kujala. Neural representation of language: activation versus long-range connectivity. *Trends Cogn Sci*, 10(11):519–25, 2006.

J. R. Sato, D. Y. Takahashi, S. M. Arcuri, K. Sameshima, P. A. Morettin, and L. A. Baccala. Frequency domain connectivity identification: an application of partial directed coherence in fmri. *Hum Brain Mapp*, 30(2):452–61, 2009.

J. R. Sato, A. Fujita, E. F. Cardoso, C. E. Thomaz, M. J. Brammer, and Jr. E. Amaro. Analyzing the connectivity between regions of interest: an approach based on cluster granger causality for fmri data analysis. *Neuroimage*, 52(4):1444–55, 2010.

M. B. Schippers, A. Roebroeck, R. Renken, L. Nanetti, and C. Keysers. Mapping the information flow from one brain to another during gestural communication. *Proc Natl Acad Sci U S A*, 107(20):9388–93, 2010.

T Schweder. Composable markov processes. *journal of Applied Probability*, 7(2): 400–410, 1970.

J. F. Smith, A. Pillai, K. Chen, and B. Horwitz. Identification and validation of effective connectivity networks in functional magnetic resonance imaging using switching linear dynamic systems. *Neuroimage*, 52(3):1027–40, 2009.

S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich. Network modelling methods for fmri. *Neuroimage*, 2010.

V Solo. On causality i: Sampling and noise. *Proceedings of the 46th IEEE Conference on Decision and Control*, pages 3634–3639, 2006.

D. Sridharan, D. J. Levitin, and V. Menon. A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proc Natl Acad Sci U S A*, 105(34):12569–74, 2008.

K. E. Stephan, L. Kasper, L. M. Harrison, J. Daunizeau, H. E. den Ouden, M. Breakspear, and K. J. Friston. Nonlinear dynamic causal models for fmri. *Neuroimage*, 42 (2):649–62, 2008.

I. H. Stevenson and K. P. Kording. On the similarity of functional connectivity between neurons estimated across timescales. *PLoS One*, 5(2):e9206, 2010.

K. Thomsen, N. Offenhauser, and M. Lauritzen. Principal neuron spiking: neither necessary nor sufficient for cerebral blood flow in rat cerebellum. *J Physiol*, 560(Pt 1):181–9, 2004.

L. Q. Uddin, A. M. Kelly, B. B. Biswal, F. Xavier Castellanos, and M. P. Milham. Functional connectivity of default mode network components: correlation, anticorrelation, and causality. *Hum Brain Mapp*, 30(2):625–37, 2009.

K. Ugurbil, L. Toth, and D. S. Kim. How accurate is magnetic resonance imaging of brain function? *Trends Neurosci*, 26(2):108–14, 2003.

K. Uludag. To dip or not to dip: reconciling optical imaging and fmri data. *Proc Natl Acad Sci U S A*, 107(6):E23; author reply E24, 2010.

K. Uludag, D. J. Dubowitz, and R. B. Buxton. Basic principles of functional mri. In R. Edelman, J. Hesselink, and M. Zlatkin, editors, *Clinical MRI*. Elsevier, San Diego, 2005.

K. Uludag, B. Muller-Bierl, and K. Ugurbil. An integrative model for neuronal activity-induced signal changes for gradient and spin echo functional imaging. *Neuroimage*, 48(1):150–65, 2009.

P Valdes-Sosa, J C Jimenez, J Riera, R Biscay, and T Ozaki. Nonlinear eeg analysis based on a neural mass model. *Biological cybernetics*, 81:415–24, 1999.

P. Valdes-Sosa, A. Roebroeck, J. Daunizeau, and K. Friston. Effective connectivity: Influence, causality and biophysical modeling. *Neuroimage*, in press.

P. A. Valdes-Sosa. Spatio-temporal autoregressive models defined over brain manifolds. *Neuroinformatics*, 2(2):239–50, 2004.

P. A. Valdes-Sosa, R. Kotter, and K. J. Friston. Introduction: multimodal neuroimaging of brain connectivity. *Philos Trans R Soc Lond B Biol Sci*, 360(1457):865–7, 2005a.

P. A. Valdes-Sosa, J. M. Sanchez-Bornot, A. Lage-Castellanos, M. Vega-Hernandez, J. Bosch-Bayard, L. Melie-Garcia, and E. Canales-Rodriguez. Estimating brain functional connectivity with sparse multivariate autoregression. *Philos Trans R Soc Lond B Biol Sci*, 360(1457):969–81, 2005b.

P. A. Valdes-Sosa, J. M. Sanchez-Bornot, R. C. Sotero, Y. Iturria-Medina, Y. Aleman-Gomez, J. Bosch-Bayard, F. Carbonell, and T. Ozaki. Model driven eeg/fmri fusion of brain oscillations. *Hum Brain Mapp*, 30(9):2701–21, 2009.

V. Walsh and A. Cowey. Transcranial magnetic stimulation and cognitive neuroscience. *Nat Rev Neurosci*, 1(1):73–9, 2000.

W. W. S. Wei. *Time Series Analysis: Univariate and Multivariate Methods*. Addison-Wesley, Redwood City, 1990.

Halbert White and Xun Lu. Granger causality and dynamic structural systems. *Journal of Financial Econometrics*, 8(2):193–243, 2010.

N. Wiener. The theory of prediction. In E.F. Berkenbach, editor, *Modern Mathematics for Engineers*. McGraw-Hill, New York, 1956.

F. Wiesinger, P. F. Van de Moortele, G. Adriany, N. De Zanche, K. Ugurbil, and K. P. Pruessmann. Potential and feasibility of parallel mri at high field. *NMR Biomed*, 19 (3):368–78, 2006.

E. Yacoub, N. Harel, and K. Ugurbil. High-field fmri unveils orientation columns in humans. *Proc Natl Acad Sci U S A*, 105(30):10607–12, 2008.

Robust statistics for describing causality in multivariate time-series.

Florin Popescu FLORIN.POPESCU@FIRST.FRAUNHOFER.DE
Fraunhofer Institute FIRST
Kekulestr. 7, Berlin 12489 Germany

Editors: Florin Popescu and Isabelle Guyon

Abstract

A widely agreed upon definition of time series causality inference, established in the seminal 1969 article of Clive Granger (1969), is based on the relative ability of the history of one time series to predict the current state of another, conditional on all other past information. While the Granger Causality (GC) principle remains uncontested, its literal application is challenged by practical and physical limitations of the process of discretely sampling continuous dynamic systems. Advances in methodology for time-series causality subsequently evolved mainly in econometrics and brain imaging: while each domain has specific data and noise characteristics the basic aims and challenges are similar. Dynamic interactions may occur at higher temporal or spatial resolution than our ability to measure them, which leads to the potentially false inference of causation where only correlation is present. Causality assignment can be seen as the principled partition of spectral coherence among interacting signals using both auto-regressive (AR) modelling and spectral decomposition. While both approaches are theoretically equivalent, interchangeably describing linear dynamic processes, the purely spectral approach currently differs in its somewhat higher ability to accurately deal with mixed additive noise.

Two new methods are introduced 1) a purely auto-regressive method named Causal Structural Information is introduced which unlike current AR-based methods is robust to mixed additive noise and 2) a novel means of calculating multivariate spectra for unevenly sampled data based on cardinal trigonometric functions is incorporated into the recently introduced phase slope index (PSI) spectral causal inference method (Nolte et al., 2008). In addition to these, PSI, partial coherence-based PSI and existing AR-based causality measures were tested on a specially constructed data-set simulating possible confounding effects of mixed noise and another additionally testing the influence of common, background driving signals. Tabulated statistics are provided, in which true causality influence is subjected to an acceptable level of false inference probability. The new methods as well as PSI are shown to allow reliable inference for signals as short as 100 points and to be robust to additive colored mixed noise and to the influence commonly coupled driving signals, as well as provide for a useful measure of strength of causal influence.

Keywords: Causality, spectral decomposition, cross-correlation, auto regressive models.

1. Introduction

Causality is the *sine qua non* of scientific inference methodology, allowing us, among other things to advocate effective policy, diagnose and cure disease and explain brain function. While it has recently attracted much interest within Machine Learning, it bears reminding that a lot of this recent effort has been directed toward *static* data rather than time series. The 'classical' statisticians of the early 20th century, such as Fisher, Gosset and Karl Pearson, aimed at a rational and general recipe for causal inference and discovery (Gigerenzer et al., 1990) but the tools they developed applied to simple types of inference which required the pres-selection, through consensus or by design, of a handful of candidate causes (or 'treatments') and a handful of subsequently occurring candidate effects. Numerical experiments yielded tables which were intended to serve as a technician's almanac (Pearson, 1930; Fisher, 1925), and are today an essential part of the vocabulary of scientific discourse, although tables have been replaced by precise formulae and specialized software. These methods rely on *removing* possible causal links at a certain 'significance level', on the basic premise that a twin experiment on data of similar size generated by a hypothetical non-causal mechanism would yield a result of similar strength only with a known (small) probability. While it may have been hoped that a generalization of the statistical test of difference among population means (e.g. the t-test) to the case of time series causal structure may be possible using a similar almanac or recipe book approach, in reality causality has proven to be a much more contentious - and difficult - issue.

Time series theory and analysis immediately followed the development of classical statistics (Yule, 1926; Wold, 1938) and was spurred thereafter by exigence (a severe economic boom/bust cycle, an intense high-tech global conflict) as well as opportunity (the post-war advent of a machine able to perform large linear algebra calculations). From a wide historical perspective, Fisher's 'almanac' has rendered the industrial age more orderly and understandable. It can be argued, however, that the 'scientific method', at least in its accounting/statistical aspects, has not kept up with the explosive growth of data tabulated in history, geology, neuroscience, medicine, population dynamics, economics, finance and other fields in which causal structure is at best partially known and understood, but is needed in order to cure or to advocate policy. While it may have been hoped that the advent of the computer might give rise to an automatic inference machine able to 'sort out' the ever-expanding data sphere, the potential of a computer of any conceivable power to condense the world to predictable patterns has long been proven to be shockingly limited by mathematicians such as Turing (Turing, 1936) and Kolmogorov (Kolmogorov and Shiryayev, 1992) - even before the ENIAC was built. The basic problem reduces itself to the curse of dimensionality: being forced to choose among combinations of members of a large set of hypotheses (Lanterman, 2001). Scientists as a whole took a more positive outlook, in line with post-war boom optimism, and focused on accessible automatic inference problems. One of these was scientists was Norbert Wiener, who, besides founding the field of cybernetics (the precursor of ML), introduced some of the basic tools of modern time-series analysis, a line of research he began during wartime and focused on feedback control in ballistics. The time-series causality definition of Granger (1969) owes inspiration to earlier discussion of causality by Wiener (1956). Granger's approach blended spectral analysis with vector auto-regression, which had long been basic tools of economics (Wold, 1938; Koopmans, 1950), and appeared nearly at the same time as similar work by Akaike (1968) and Gersch and Goddard (1970).

It is useful to highlight the differences in methodological principle and in motivation for static *vs.* time series data causality inference, starting with the former as it comprises a large part of the pertinent corpus in Machine Learning and in data mining. Static causal inference is important in the sense that any classification or regression presumes some kind of causality, for the resulting relation to be useful in identifying elements or features of the data which 'cause' or predict target labels or variables and are to be selected at the exclusion of other confounding 'features'. In learning and generalization of static data, sample ordering is either uninformative or unknown. Yet order is implicitly relevant to learning both in the sense that some calculation occurs in the physical world in some finite number of steps which transform independent inputs (*stimuli*) to dependent output (*responses*), and in the sense that generalization should occur on expected *future stimuli*. To ably generalize from a limited set of samples implies making accurate causal inference. With this priority in mind prior NIPS workshops have concentrated on feature selection and on graphical model-type causal inference (Guyon and Elisseeff, 2003; Guyon et al., 2008, 2010) inspired by the work of Pearl (2000) and Spirtes et al. (2000). The basic technique or underlying principle of this type of inference is vanishing partial correlation or the inference of *static* conditional independence among 3 or more random variables. While it may seem limiting that no unambiguous, generally applicable causality assignment procedure exists among single *pairs* of random variables, for large ensembles the ambiguity may be partially resolved. Statistical tests exist which assign, with a controlled probability of false inference, random variable X_1 as dependent on X_2 given no other information, but as independent on X_2 given X_3, a conceptual framework proposed for time-series causality soon after Granger's 1969 paper using partial *coherence* rather than static correlation (Gersch and Goddard, 1970). Applied to an ensemble of observations $X_1..X_N$, efficient polynomial time algorithms have been devised which combine information about pairs, triples and other sub-ensembles of random variables into a complete dependency graph including, but not limited to, a directed acyclical graph (DAG). Such inference algorithms operate in a nearly deductive manner but are not guaranteed to have unique, optimal solution. Underlying predictive models upon which this type of inference can operate includes linear regression (or structural equation modeling) (Richardson and Spirtes, 1999; Lacerda et al., 2008; Pearl, 2000) and Markov chain probabilistic models (Scheines et al., 1998; Spirtes et al., 2000). Importantly, a previously unclear conceptual link between the notions of time series causality and static causal inference has been formally described: see White and Lu (2010) in this volume.

Likewise, algorithmic and functional relation constraints, or at least likelihoods thereof, have been proposed as to assign causality for co-observed random variable pairs (i.e. simply by analyzing the scatter plot of X_1 vs. X_2) (Hoyer et al., 2009). In general terms, if we are presented a scatter plot X_1 vs. X_2 which looks like a noisy sine wave, we may reasonably infer that X_2 causes X_1, since a given value of X_2 'determines' X_1 and not vice versa. We may even make some mild assumptions about the noise process which superimposes on a functional relation ($X_2 = X_1 +$ additive noise which is independent of X_1) and by this means turn our intuition into a proper *asymmetric* statistic, i.e. a controlled probability that X_1 does *not* determine X_2, an approach that has proven remarkably successful in some cases where the presence of a causal relation was known but the direction was not (Hoyer et al., 2009). The challenge here is that, unlike in traditional statistics, there is not simply the case of the null hypothesis and its converse, but one of 4 mutually exclusive cases. A) X_1 is independent of X_2 B) X_1 causes X_2 C) X_2 causes X_1 and D) X_1 and X_2 are observations of dependent and non-causally related random variables (bidirectional information flow or feedback). The appearance of a symmetric bijection (with additive noise) between X_1 and X_2 does not mean absence of causal relation, as asymmetry in the apparent relations is merely a clue and not a determinant of causality. Inference over static data is not without ambiguities without additional assumptions and requires observations of interacting triples (or more) of variables as to allow somewhat reliable descriptions of causal relations or lack thereof (see Guyon et al. (2010) for a more comprehensive overview). Statistical evaluation requires estimation of relative likelihood of various candidate models or causal structures, including a null hypothesis of non-causality. In the case of complex multidimensional data theoretical derivation of such probabilities is quite difficult, since it is hard to analytically describe the class of dynamic systems we may be expected to encounter. Instead, common ML practice consists in running toy experiments in which the 'ground truth' (in our case, causal structure) is only known to those who run the experiment, while other scientists aim to test their discovery algorithms on such data, and methodological validity (including error rate) of any candidate method rests on its ability to predict responses to a set of 'stimuli' (test data samples) available only to the scientists organizing the challenge. This is the underlying paradigm of the Causality Workbench (Guyon, 2011). In time series causality, we fortunately have far more information at our disposal relevant to causality than in the static case. Any type of reasonable interpretation of causality implies a physical mechanism which accepts a modifiable input and performs some operations in some finite time which then produce an output and includes a source of randomness which gives it a stochastic nature, be it inherent to the mechanism itself or in the observation process. Intuitively, the structure or connectivity among input-output blocks that govern a data generating process are related to causality no matter (within limits) what the exact input-output relationships are: this is what we mean by structural causality. However, not all structures of data generating processes are obviously causal, nor is it self evident how structure corresponds to Granger (*non*) causality (GC), as shown in further detail by White and Lu (2010). Granger causality is a measure of relative predictive information among variables and not evidence of a direct physical mechanism linking the two processes: no amount of analysis can exclude a latent unobserved cause. Strictly speaking the GC statistic is not a measure of causal relation: it is the possible non-rejection of a null hypothesis of time-ordered independence.

Although time information helps solve many of the ambiguities of static data several problems, and despite the large body of literature on time-series modeling, several problems in time-series causality remain vexing. Knowledge of the structure of the overall multivariate data generating process is an indispensable aid to inferring causal relationships: but how to infer the structure using weak *a priori* assumptions is an open research question. Sections 3, 4 and 5 will address this issue. Even in the simplest case (the bivariate case) the observation process can introduce errors in time-series causal inference by means of co-variate observation noise (Nolte et al., 2010). The bivariate dataset NOISE in the Causality Workbench addresses this case, and is extended in this study to the evaluation datasets PAIRS and TRIPLES. Two new methods are introduced: an autoregressive method named Causal Structural Information (Section 7) and a method for estimating spectral coherence in the case of unevenly sampled data (Section 8.1). A principled comparison of different methods as well as their performance in terms of type I, II and III errors is necessary, which addresses both the presence/absence of causal interaction and directionality. In discussing causal influence in real-world processes, we may reasonably expect that not inferring a potentially weak causal link may be acceptable but positing one where none is missing may be problematic. Sections 2, 6, 7 and 8 address robustness of bivariate causal inference, introducing a pair of novel methods and evaluating them along with existing ones. Another common source of argument in discussions of causal structure is the case of false inference by neglecting to condition the proposed causal information on other background variables which may explain the proposed effect equally well. While the description of a general deductive method of causal connectivity in multivariate time series is beyond the scope of this article, Section 9 evaluates numerical and statistical performance in the tri-variate case, using methods such as CSI and partial coherence based PSI which can apply to bivariate interactions conditioned by an arbitrary number of background variables.

2. Causality statistic

Causality inference is subject to a wider class of errors than classical statistics, which tests independence among variables. A general hypothesis evaluation framework can be:

$$\text{Null Hypothesis} = \text{No causal interaction} \;\; H_0 = A \perp_C B \,|C$$

$$\text{Hypothesis 1a} = \text{A } \textit{drives} \text{ B} \;\; H_a = A \rightarrow B \,|C$$

$$\text{Hypothesis 1b} = \text{B } \textit{drives} \text{ A} \;\; H_b = B \rightarrow A \,|C$$

$$\text{Type I error prob. } \alpha = P\Big(\hat{H}_a \text{ or } \hat{H}_b |\, H_0\Big) \tag{1}$$

$$\text{Type II error prob. } \beta = P\Big(\hat{H}_0 |\, H_a \text{ or } H_b\Big)$$

$$\text{Type III error prob. } \gamma = P\Big(\hat{H}_a \,|H_b \text{ or } \hat{H}_b \,|H_a\Big)$$

The notation $\hat{H}$ means that our statistical estimate of the estimated likelihood of H exceeds the threshold needed for our decision to confirm it. This formulation carries some caveats the justification for which is pragmatic and will be expounded upon in later sections. The main one is the use of the term '*drives*' in place of '*causes*'. The null hypothesis can be viewed as equivalent to *strong* Granger non-causality (as it will be argued is necessary), but it does not mean that the signals **A** and **B** are independent: they may well be correlated to one another. Furthermore, we cannot realistically aim at statistically supporting *strict* Granger causality, i.e. strictly one-sided causal interaction, since asymmetry in bidirectional interaction may be more likely in real-world observations and is equally meaningful. By '*driving*' we mean instead that the history of one time series element **A** is more useful to predicting the current state of **B** than vice-versa, and not that the history of **B** is irrelevant to predicting **A**. In the latter case we would specify '*G-causes*' instead of '*drives*' and for H_0 we would employ non-parametric independence tests of Granger non causality (GNC) which have already been developed as in Su and White (2008) and Moneta et al. (2010). Note that the definition in (1) is different from that recently proposed in White and Lu (2010), which goes further than GNC testing to make the point that structural causality inference must also involve a further conditional independence test: Conditional Exogeneity (CE). In simple terms, CE tests whether the innovations process of the potential effect is conditionally independent of the cause (or, by practical consequence, whether the innovations processes are uncorrelated). White and Lu argue that if both GNC and CE fail we ought not make any decision regarding causality, and combine the power of both tests in a principled manner such that the probability of false causal inference, or non-decision, is controlled. The difference in this study is that the concurrent failure of GNC and CE is *precisely* the difficult situation requiring additional focus and it will be argued that methods that can cope with this situation can also perform well for the case of CE, although they require stronger assumptions. In effect, it is assumed that real-world signals feature a high degree of non-causal correlation, due to aliasing effects as described in the following section, and that strong evidence to the contrary is required, i.e. that non-decision is equivalent to inference of non-causality. The precise meaning of 'driving' will also be made explicit in the description of Causal Structural Information, which is implicitly a proposed definition of H_0. Also different in Definition (1) than in White and Lu is the accounting of potential error in causal *direction* assignment under a framework which forces the practitioner to make such a choice if GNC is rejected.

One of the difficulties of causality inference methodology is that it is difficult to ascertain what true causality in the real world ('ground truth') is for a sufficiently comprehensive class of problems (such that we can reliably gage error probabilities): hence the need for extensive simulation. A clear means of validating a causal hypothesis would be *intervention* Pearl (2000), i.e. modification of the presumed cause, but in instances such as historic and geological data this is not feasible. The basic approach will be to assume a non-informative probability distribution of the degree degree of mixing, or non-causal dynamic interactions, as well as over individual spectra and compile inference error probabilities over a wide class of coupled dynamic systems. In constructing a 'robust causality' statistic there is more than simply null-hypothesis rejection and accurate directionality to consider, however. In scientific practice we are not only interested to know that **A** and **B** are causally related or not, but which is the *main* driver in case of bidirectional coupling, and among a time series vector **A**, **B**, **C**, **D**... it is important to determine which of these factors are the main causes of the target variable, say **A**. The relative effect size and relative causal influence strength, lest the analysis be misused (Ziliak and McCloskey, 2008). The rhetorical and scientific value of effect size in no way devalues the underlying principle of robust statistics and controlled inference error probabilities used to quantify it.

3. Auto-regression and aliasing

A simple multivariate time series model is the multivariate auto-regressive model (abbreviated as MVAR or VAR). It assumes that the data generating process (DGP) that created the observations is a linear dynamic model and, as such, it contains poles only i.e. the numerator of the transfer function between innovations process and observation is a scalar. The more complex auto-regressive moving average model (ARMA) includes zeros as well. Despite the rather stringent assumptions of VAR, a time-series extension of ordinary least squares linear regression, it has been hugely successful in applications from neuroscience to engineering to sociology and economics. Its familiar VAR (or VARX) formulation is:

$$y_i = \sum_{k=1}^{K} A_k y_{i-k} + Bu + w_i \tag{2}$$

Where $\{y_{i,d=1..D}\}$ is a real valued vector of dimension D. Notice the absence of a subscript in the exogenous input term u. This is because a general treatment of exogenous inputs requires a lagged sum, i.e. $\sum_{k=1}^{K} B_k u_{i-k}$. Since exogenous inputs are not explicitly addressed in the following derivations the general linear operator placeholder Bu is used instead and can be re-substituted for subsequent use.

Granger non-causality for this system, expressed in terms of conditional independence, would place a relation among elements of y subject to knowledge of u. If $D = 2$, for all i

$$y_{1,i} \perp y_{2,i-1..i-K} \mid y_{1,i-1..i-K} \tag{3}$$

If the above is true, we would say that y_2 does not finite-order G cause y_1. If the world was made exclusively of linear VARs, it would not be terribly difficult to devise a reliable statistic for G causality. We would, given a sequence of N data points, identify the maximum-likelihood parameters A and B via ordinary least squares (OLS) linear regression after having, via some model selection criterion, determined the order K. Furthermore we would choose another criterion (e.g. test and p-value) which tells us whether any particular coefficient is likely to be statistically indistinguishable from 0, which would correspond to a vanishing partial correlation. If all A's are lower triangular G non-causality is satisfied (in one direction but not the converse). It is however very rare that the physical mechanism we are observing is indeed the embodiment of a VAR, and therefore even in the case in which G non-causality can be safely rejected, it is not likely that the best VAR approximation of the data observed is strictly lower/upper triangular. The necessity of a distinction between strict causality, which has a structural interpretation, and a causality statistic, which does not measure independence in the sense of Granger-non causality, but rather *relative* degree of dependence in both directions among two signals (driving) is most evident in this case. If the VAR in question had very small (and statistically observable) upper triangular elements would a discussion of causality of the observed time series be rendered moot?

One of the most common physical mechanisms which is incompatible with VAR is aliasing, i.e. dynamics which are faster than the (shortest) sampling interval. The standard interpretation of aliasing is the false representation of frequency components of a signal due to sub-Nyquist frequency sampling: in the multivariate time-series case this can also lead to spurious correlations in the observed innovations process (Phillips, 1973). Consider a continuous bivariate VAR of order 1 with Gaussian innovations in which the sampling frequency is several orders of magnitude smaller than the Nyquist frequency. In this case we would observe a covariate time independent Gaussian process since for all practical purposes the information travels 'instantaneously'. In economics, this effect could be due to social interactions or market reactions to news which happen faster than the sampling interval (be it daily, hourly or monthly). In fMRI analysis sub- sampling interval brain dynamics are observed over a relatively slow time convolution process of hemodynamic response of neural activity (for a detailed exposition of causality inference in fMRI see Roebroeck et al. (2011) in this volume). Although 'aliasing' normally refers to temporal aliasing, the same process can occur *spatially*. In neuroscience and in economics the observed variables are summations (dimensionality reductions) of a far larger set of interacting agents, be they individuals or neurons. In electroencephalography (EEG) the propagation of electrical potential from cortical axons arrives via multiple pathways to the same recording location on the scalp: the summation of micrometer scale electric potentials on the scalp at centimeter scale. Once again there are spurious observable correlations: this is known as the *mixing* problem. Such effects can be modeled, albeit with significant information loss, by the same DGP class which is a superset of VAR and known in econometrics as SVAR (structural vector auto-regression, the time series equivalent of structural equation modeling (SEM), often used in static causality inference (Pearl, 2000)). Another basic problem in dynamic system identification is that we not only discard much information from the world in sampling it, but that our observations are susceptible to additive noise, and that the randomness we see in the data is not entirely the randomness of the mechanism we intend to study. One of the most problematic of additive noise models is *mixed colored noise,* in which there are structured correlations both in time and across elements of the time-series, but not in any causal way: there is only a linear transformation of colored noise, sometimes called mixing, due to spatial aliasing.

Mixing may occur due to temporal aliasing in sampling a coupled continuous-variable VAR system. In EEG analysis mixed colored noise models the background electrical activity of the brain. In other domains such as economics, one can imagine the influence of unpredictable events such as natural cataclysms or macroenomic cycles which are not white noise and which are reflect nearly 'instantaneously' but to varying degree in all our measurements. In this case, since each additive noise component is colored (it has temporal auto- correlation), its past helps predict its current value. Since the observation is a linear mixture of noise components, all current observations are correlated, and the past of any component can help predict the current state of any other. In this case, the strict definition of Granger causality would not make practical sense, since this cross-predictability is not meaningful.

It should be noted on this point that the literature contains (sometimes inconsistent) sub-classifications of Granger Causality, such as *weak* and *strong* Granger causality. One definition which is particularly pertinent to this work is that given in Caines (1976) and Solo (2006) and is that strong Granger causality allows instantaneous dependence and that weak Granger causality does not (i.e. it is strictly time ordered). We are aiming in this work at strong Granger causality inference, i.e. one which is robust to aliasing effects such as colored noise. While we should *account* for instantaneous interactions, we do not have to assign causal interpretations to them, since they are symmetric (the cross-correlation of independent mixed signals is symmetric).

4. Auto-regression, learning and Granger Causality

Learning is the process of discovering predictable patterns in the real world, where a 'pattern' is described by an algorithm or an automaton. Besides the object of learning, i.e. the algorithm which we infer and which maps stimuli to responses, we need to consider the algorithm which performs the learning process and outputs the former. The third algorithm we should consider is the algorithm embodied in the real world, which we do not know, which generates the data we observe, and which we hope to be able to recover, or at least approximate. How can we formally describe it? A Data Generating Process (DGP) can be a machine or automaton: an algorithm that performs every operation deterministically in a *finite* number of steps, but which contains an oracle that generates perfectly random numbers. It is sufficient that this oracle generate **1**'s and **0**'s only: all other computable probability distributions can be calculated from it. A DGP contains rational valued parameters (rational as to comply with finite computability), in this case the integer K and all elements of the matrices A. Last but not least a DGP specification may limit the set of admissible parameter values and probability distributions of the oracle-generated values. The set of all possible outputs of a DGP corresponds to the set of all probability distributions generated by it over all admissible parameter values, which we shall call the DGP class.

Definition 1 *Let $i \in \mathbb{N}$ and let s_a, s_w, p_w be finite length prefix-free binary strings. Furthermore let $\boldsymbol{y}$ and $\boldsymbol{u}$ be rational valued matrices of size $N \times i$ and $M \times i$, and $\boldsymbol{t}$ be rational valued vector with distinct elements, of length $\boldsymbol{i}$. Let $\boldsymbol{a}$ also be a finite rational* *valued vector. A Data Generating Process is a quintuple {s_a,p_w,T_a,T_w} where T_a, T_w are finite time Turing machines which perform the following operations: Given an input of the incompressible string p_w the machine T_w calculates a rational valued matrix* $\mathbf{w}$*. The machine T_a when given matrices* $\mathbf{y}$, $\mathbf{a}$, $\mathbf{u}$, $\mathbf{t}$, $\mathbf{w}$ *and a positive rational Δt outputs a vector y_{i+1} which is assigned for future operations to the time $t_{i+1} = max(t) + \Delta t$*

The definition is somewhat unusual in terms of the definition of stochastic systems as embodiments of Turing machines, but it is quite standard in terms of defining an innovations term w, a probability distribution thereof p_w, a state y, a generating function p_a with parameters a and an exogenous input u. The motivation for using the terminology of algorithmic information theory is to analyse causality assignment as a computational problem. For reasons of finite description and computability our variables are rational, rather than real valued. Notice that there is no real restriction on how the time series is to be generated, recursively or otherwise. The initial condition in case of recursion is implicit, and time is specified as distinct and increasing but otherwise arbitrarily distributed - it does not necessarily grow in constant increments (it is asynchronous). The slight paradox about describing stochastic dynamical systems in algorithmic terms is the necessity of postulating a random number generator (an oracle) which in some ways is our main tool for abstracting the complexity of the real world, but yet is a physical impossibility (since such an oracle would require infinite computational time see Li and Vitanyi (1997) for overview). Also, the Turing machines we consider have finite memory and are time restricted (they implement a predefined maximum number of operations before yielding a default output). Otherwise the rules of algebra (since they perform algebraic operations) apply normally. The *cover* of a DGP can be defined as:

Definition 2 *The cover of a Data Generating Process (DGP) class is the cover of the set of all outputs* $\mathbf{y}$ *that a DGP calculates for each member of the set of admissible parameters* ***a,u,t,w*** *and for each initial condition y_1. Two DGPs are stochastically equivalent if the cover of the set of their possible outputs (for fixed parameters) is the same.*

Let us now attempt to define a Granger Causality statistic in algorithmic terms. Allowing for the notation $j..k = \{j-1, j-2.., k+1, k\}$ if $j > k$ and in reverse order if $j < k$

$$\frac{1}{i}\sum_{j=1}^{i} K(y_{1,j} \mid y_{1,j-1..1}, u_{j-1..1}) - K(y_{1,j} \mid y_{2,j-1..1}, y_{1,j-1..1}, u_{j-1..1}) \tag{4}$$

This differs from Equation (3) in two elemental ways: it is not a statement of independence but a number (statistic), namely the average difference (rate) of conditional (or prefix) Kolmogorov complexity of each point in the presumed effect vector when given both vector histories or just one, and given the exogenous input history. It is a generalized conditional entropy rate, and may be reasonably be normalized as such:

$$\mathcal{F}^K_{2\to 1|u} = \frac{1}{i}\sum_{j=1}^{i}\left(1 - \frac{K(y_{1,j} \mid y_{2,j-1..1}, y_{1,j-1..1}, u_{j-1..1})}{K(y_{1,j} \mid y_{1,j-1..1}, u_{j-1..1})}\right) \tag{5}$$

which is a fraction ranging from 0 - meaning no influence of y_1 by y_2 - to 1, corresponding to complete determination of y_1 by y_2 and can be transformed into a statistic comparing different data sets and processes, and which gives probabilities of spurious results. Another difference with Equation (3) is that we do not refer to finite-order G causality but simply G causality (in the general case we do not know the maximum lag order but must infer it). For a more in depth look at DGPs, structure and G-causality, see White and Lu (2010). The larger the value $\mathcal{F}^K_{2\to 1|u}$, the more likely that y_2 G-causes y_1. The definition is one of conditional information and it is one of an averaged process rather than a single instance (time point). However, Kolmogorov complexity is incomputable, and as such Granger (non) causality must also be, in general, incomputable. A detailed look at this issue is beyond the scope of this article, but in essence, we can never test all possible models that could tell us wether the history of a time series helps or does not help predict (compress) another, and the set of finite running time Turing machines is not enumerable. We've partially circumvented the halting problem since we've specified finite-state, finite-operation machines as the basis of DGPs but have not specified a search procedure over all DGPs that enumerates them. Even if we limit ourselves to DGPs which are MVAR, the necessary computational time to calculate the description length (instead of $K(.)$) is NP-complete, i.e. it requires an enumeration of all possible parameters of a DGP class, barring any special properties thereof: finding the optimal model order requires such a search (keep in mind VAR estimation is convex only once we know the model order and AR structure).

In practice, we should limit the class of DGPs we consider within out statistic to one which allows the possibility of polynomial time computation. Let us take Equation (2), and further make the common assumption that the input vector w is an *i.i.d.* normally distributed sequence independent along dimension d, we've specified the linear VAR Gaussian DGP class (which we shall shorten as VAR class). This DGP class, again, has proven remarkably useful in cases where nothing else except the time series vector y is known. Re-writing (2):

$$y_i = \sum_{k=1}^{K} A_k y_{i-k} + \mathcal{D}\, w_{i-1}\,, \mathcal{D}_{ii} > 0, \mathcal{D}_{ij} = 0 \tag{6}$$

The matrix $\mathcal{D}$ is a positive diagonal matrix containing the scaling, or effective standard deviations of the innovation terms. The standard deviation of each element of the innovations term w **is assumed hereafter to be equal to 1**.

5. Equivalence of auto-regressive data generation processes.

In econometrics the following formulation is familiar (SVAR):

$$y_i = \sum_{k=0}^{K} A_k y_{i-k} + Bu + \mathcal{D}w_i \tag{7}$$

The difference between this and Equation (6) is the presence of a 0-lag matrix A_0 which, for easy tractability has zero diagonal entries and is sometimes present on the LHS. This 0-lag matrix is meant to model the sub-sampling interval dynamic interactions among observations, which appear instantaneous, see Moneta et al. (2011) in this volume. Let us call this form *zero lag SVAR*. In electric- and magneto- encephalography (EEG/MEG) we often encounter the following form:

$$x_i = \sum_{k=1}^{K} {}_\mu A_k x_{i-k} +_\mu Bu + \mathcal{D}w_i,$$

$$y_i = Cx_i \tag{8}$$

Where C represents the observation matrix, or *mixing matrix* and is determined by the conductivity/permeability of tissue, and accounts for the superposition of the electromagnetic fields created by neural activity, which happens at nearly the speed of light and therefore appears instantaneous. Let us call this *mixed output SVAR*. Finally, in certain engineering applications we may see structured disturbances:

$$y_i = \sum_{k=1}^{K} {}_\theta A_k y_{i-k} +_\theta Bu + D_w w_i \tag{9}$$

Which we shall call *covariate innovations SVAR* (D_w is a general nonsingular matrix unlike $\mathcal{D}$ which is diagonal). Another final SVAR form to consider would be one in which the 0-lag matrix ${}_\triangleleft A_0$ is strictly upper triangular (*upper triangular zero lag* SVAR):

$$y_i =_\triangleleft A_0 y_i + \sum_{k=1}^{K} A_k y_{i-k} +_\triangleleft Bu + \mathcal{D}w_i \tag{10}$$

Finally, we may consider a upper or lower triangular co-variate innovations SVAR:

$$y_i = \sum_{k=0}^{K} A_k y_{i-k} + Bu +_\triangleleft Dw_i \tag{11}$$

Where ${}_\triangleleft D$ is upper/lower triangular. The SVAR forms (6)-(10) may look different, and in fact each of them may uniquely represent physical processes and allow for direct interpretation of parameters. From a statistical point of view, however, all four SVAR DGPs introduced above are equivalent since they have identical cover.

Lemma 3 *The Gaussian covariate innovations SVAR DGP has the same cover as the Gaussian mixed output SVAR DGP. Each of these sets has a redundancy of* $2^N N!$ *for* *instances in which the matrices D_w is the product of and unitary and diagonal matrices, the matrix C is a unitary matrix and the matrix A_0 is a permutation of an upper triangular matrix.*

Proof Staring with the definition of covariate innovations SVAR in Equation (9) we use the variable transformation $y = D_w x$ and obtain the mixed-output form (trivial). The set of Guassian random variables is closed under scalar multiplication (and hence sign change) and addition. This means that the variance if the innovations term in Equation (9) can be written as:

$$\Sigma_w = D_w^T D_w = D_w^T U^T U D_w$$

Where U is a unitary (orthogonal, unit 2-norm) matrix. Since all innovations term elements are zero mean, the covariance matrix is the sole descriptor of the Gaussian innovations term. This in turn means that any other matrix $D_w^{'} = D_w^T U^T$ substituted into the DGP described in Equation (9) amounts to a stochastically equivalent DGP. The matrix $D_w^{'}$ can belong to a number of general sets of matrices, one of which is the set of nonsingular upper triangular matrices (the transformation is achievable through the QR decomposition of Σ_w). Another such set is lower triangular matrix set. Both are subsets of the set of matrices sometimes named 'psychologically upper triangular', meaning a permutation of an upper triangular matrix.

If we constrain D_w to be of the form $D_w = U\mathcal{D}$, i.e. such that (by polar decomposition) it is the product of a unitary and a diagonal positive definite matrix, the only stochastically equivalent transformations of D_w are a symmetry preserving permutation of its rows/columns and a sign change in one of the columns (this is a property of orthogonal matrices such as U). There are $N!$ such permutations and 2^N possible sign changes. For the general case, in which the input u has no special properties, there are no other redundancies in the SVAR model (since changing any parameter in A and B will otherwise change the output). Without loss of generality then, we can write the transformation from covariate innovations to mixed output SVAR form as:

$$y_i = \sum_{k=1}^{K} {}_{\theta}A_k y_{i-k} +_{\theta} Bu + U\mathcal{D}_w w_i$$

$$x_i = \sum_{k=1}^{K} U^T ({}_{\theta}A_k) U x_{i-k} + U^T ({}_{\theta}B) u + \mathcal{D}_w w_i$$

$$y_i = U^T x_i$$

Since the transformation U is one to one and invertible, and since this transformation is what allows a (restricted) a covariate noise SVAR to map, one to one, onto a mixed output SVAR, the cardinality of both covers is the same.

Now consider the zero-lag SVAR form:

$$y_i = \sum_{k=0}^{K} A_k y_{i-k} + Bu + \mathcal{D} w_i$$

$$\mathcal{D}^{-1}\left(1 - A_0\right) y_i = \sum_{k=1}^{K} \mathcal{D}^{-1} A_k y_{i-k} + \mathcal{D}^{-1} Bu + w$$

Taking the singular value decomposition of the (nonsingular) matrix coefficient on the LHS:

$$U_0 S V_0^T y_i = \sum_{k=1}^{K} \mathcal{D}^{-1} A_k y_{i-k} + \mathcal{D}^{-1} Bu + w_i$$

$$V_0^T y_i = S^{-1} U_0^T \sum_{k=1}^{K} \mathcal{D}^{-1} A_k y_{i-k} + S^{-1} U_0^T \mathcal{D}^{-1} Bu + S^{-1} U_0^T w_i$$

Using the coordinate transformation $z = V_0^T y$. The unitary transformation U_0^T can be ignored due closure properties of the Gaussian. This leaves us with the mixed-output form:

$$z_i = \sum_{k=1}^{K} S^{-1} U_0^T \mathcal{D}^{-1} A_k V_0 z_{i-k} + S^{-1} U_0^T \mathcal{D}^{-1} Bu + S^{-1} w_i'$$

$$y = V_0 z$$

So far we've shown that for every zero-lag SVAR there is at least one mixed-output VAR. Let us for a moment consider the covariate noise SVAR (after pre-multiplication)

$$D_w^{-1} y_i = \sum_{k=1}^{K} D_w^{-1}{}_{\theta} A_k y_{i-k} + D_w^{-1}{}_{\theta} Bu + w_i$$

We can easily then write it in terms of zero lag:

$$y_i = \left(I - D_w^{-1}\right) y_i + \sum_{k=1}^{K} D_w^{-1}{}_{\theta} A_k y_{i-k} + D_w^{-1}{}_{\theta} Bu + w_i$$

However, the entries of $I - D_w^{-1}$ are not zero (as required by definition). This can be done by scaling by the diagonal:

$$diag(D_w^{-1}) y_i = (diag(D_w^{-1}) - D_w^{-1}) y_i + \sum_{k=1}^{K} D_w^{-1}{}_{\theta} A_k y_{i-k} + D_w^{-1}{}_{\theta} Bu + w_i$$

$$\mathcal{D}_0 \triangleq diag(\mathcal{D}_w^{-1})$$

$$y_i = (I - \mathcal{D}_0^{-1} D_w^{-1}) y_i + \sum_{k=1}^{K} \mathcal{D}_0^{-1} D_w^{-1}{}_{\theta} A_k y_{i-k} + \mathcal{D}_0^{-1} D_w^{-1}{}_{\theta} B u + \mathcal{D}_0^{-1} w_i$$

$$A_0 = (I - \mathcal{D}_0^{-1} D_w^{-1})$$

$$D_w^{-1} = diag(D_w^{-1})(I - A_0)$$

While the following constant relation preserves DGP equivalence:

$$(D_w^T D_w)^{-1} = \Sigma_w^{-1} = \mathcal{D}_w^{-1} \mathcal{D}_w^{-T} = \mathcal{D}_0 (I - A_0)(I - A_0)^T \mathcal{D}_0$$

$$A_0 = (I - \mathcal{D}_0^{-1} D_w^{-1})^T (I - \mathcal{D}_0^{-1} D_w^{-1})$$

The zero lag matrix is a function of D_w^{-1}, the inverse of which is an eigenvalue problem. However, as long as the covariance matrix or its inverse is constant, the DGP is unchanged and this allows $N(N-1)/2$ degrees of freedom. Let us consider only mixed input systems for which the innovations terms are of unit variance. There is no real loss of generality since a simple row-division by each element of $\mathcal{D}_0$ normalized the covariate noise form (to be regained by scaling the output). In this case the equivalence constraint on is one of in which:

$$(I -_{\triangleleft} A_0)^T (I -_{\triangleleft} A_0) = (I - A_0)^T (I - A_0)$$

If $(I - A_0)$ is full rank, a strictly upper triangular matrix ${}_{\triangleleft}A_0$ may be found that is equivalent (this would be the Cholesky decomposition of the inverse covariance matrix in reverse order). As D_w is equivalent to a unitary transformation $U\mathcal{D}_w$ this will include permutations and orthogonal rotations. Any permutation of D_w will imply a corresponding permutation of A_0, which (along with rotations) has $2^N N!$ solutions.

■

The non-uniqueness of SVAR and the problematic interpretation of AR coefficients with respect to variable permutation is a known problem Sims (1981), as is the fact that modeling zero-lag matrices is equivalent to covariance estimation for the Gaussian case in the other lag coefficients are zero. In fact, statistically vanishing elements of the covariance matrix are used in Structural Equation Modeling and are given causality interpretations Pearl (2000). It is not clear how robust such inferences are with respect to equivalent permutations. The point of the lemma above is to illustrate the ambiguity of interpretation if the structure of (sparse or full) AR systems in the case of covariate innovations, zero-lag, or mixed output, which are equivalent to each other. In the case of SVAR, one approach is to perform standard AR followed by a Cholesky decomposition of the covariance of the residuals and then pre-multiplying. In Popescu (2008), the upper triangular SVAR estimation is done directly by singular value decomposition after regression and the innovations covariance estimated from the zero-lag matrix.

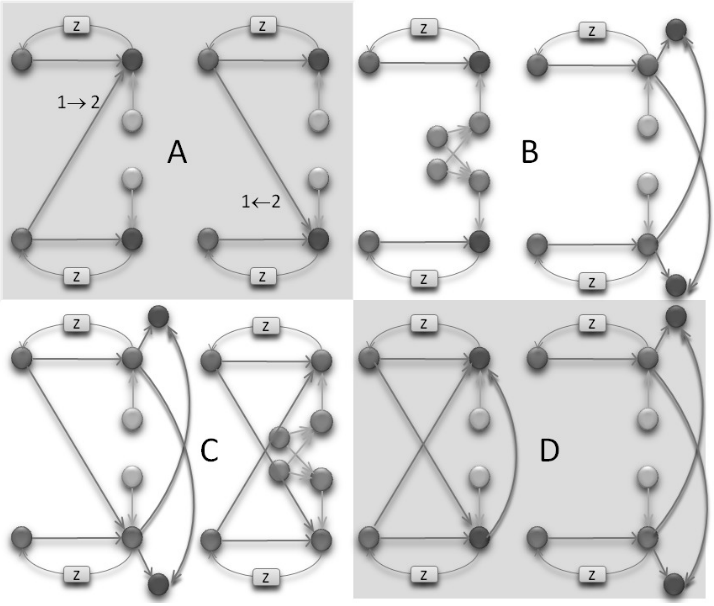


Figure 1: SVAR causality and equivalence. Structural VAR equivalence and causality. A) direct structural Granger causality (both directions shown). z stands for the delay operator. B) equivalent covariate innovations (left) and mixed output systems. Neither representation shows dynamic interaction C) sparse, one sided covariate innovations DAG is non sparse in the mixed output case (and vice-versa). D) upper triangular structure of the zero-lag matrix is not informative in the 2 variable Gaussian case, and is equivalent to a full mixed output system.

Granger, in his 1969 paper, suggests that 'instantaneous' (i.e. covariate) effects be ignored and only the temporal structure be used. Whether or not we accept instantaneous causality depends on prior knowledge: in the case of EEG, the mixing matrix cannot have any physical 'causal' explanation even if it is sparse. Without additional *a priori* assumptions, either we infer causality on unseen and presumably interacting hidden variables (mixed output form, the case of EEG/MEG) or we assume a a non-causal mixed innovations input. Note also that the zero-lag system appears to be causal but can be written in a form which suggest the opposite difference causal influence (hence it is sometimes termed 'spurious causality'). In short, since instantaneous interaction in the Gaussian case cannot be resolved causally purely in terms of prediction and conditional information (as intended by Wiener and Granger), it is proposed that such interactions be accounted for but not given causal interpretation (as in 'strong' Granger non-causality) .

There are at least four distinct overall approaches to dealing with aliasing effects in time series causality. 1) is to make prior assumptions about covariance matrices and limit inference to domain relevant and interpretable posteriors, as in Bernanke et al. (2005) in economics and Valdes-Sosa et al. (2005) in neuroscience. 2) to allow for unconstrained graphical causal model type inference among covariate innovations, by either assuming Gaussianity or non-Gaussianity, the latter allowing for stronger causal inferences (see Moneta et al. (2011) in this volume). One possible drawback of this approach is that DAG-type inference, at least in the Gaussian case in which there is so-called 'Markov equivalence' among candidate graphs, is non-unique. 3) a physically interpretable mixed output or co-variate innovations is assumed and the inferred sparsity structure (or the intersection thereof over the nonzero lag coefficient matrices) as the connection graph. Popescu (2008) implemented such an approach by using the minimum description length principle to provide a universal prior over rational-valued coefficients, and was able to recover structure in the majority of simulated co-variate innovations processes of arbitrary sparsity. This approach is computationally laborious, as it is NP and non-convex, and moreover a system that is sparse in one form (covariate innovations or mixed-ouput) is not necessarily sparse in another equivalent SVAR form. Moreover completely dense SVAR systems may be non-causal (in the strong GC sense). 4) Causality is not interpreted as a binary value, but rather direction of interaction is determined as a continuous valued statistic, and one which is theoretically robust to covariate innovations or mixtures. This is the principle of the recently introduced phase slope index (PSI), which belongs to a class of methods based on spectral decomposition and partition of coherency. Although auto-regressive, spectral and impulse response convolution are theoretically equivalent representation of linear dynamics, they do differ numerically and spectral representations afford direct access to phase estimates which are crucial to the interpretation of lead and lag as it relates to causal influence. These methods are reviewed in the next section.

6. Spectral methods and phase estimation

Cross- and auto spectral densities of a time series, assuming zero-mean or de-trended values, are defined as:

$$\rho_{Lij}(\tau) = E\Big(y_i(t) y_j(t-\tau)\Big)$$
$$S_{ij}(\omega) = \mathcal{F}(\rho_{Lij}(\tau)) \tag{12}$$

Note that continuous, linear, raw correlation values are used in the above definition as well as the continuous Fourier transform. Bivariate *coherency* is defined as:

$$C_{ij}(\omega) = \frac{S_{ij}(\omega)}{\sqrt{S_{ii}(\omega) S_{jj}(\omega)}} \tag{13}$$

Which consists of a complex numerator and a real-valued denominator. The coherence is the squared magnitude of the coherency:

$$c_{ij}(\omega) = C_{ij}(\omega)^* C_{ij}(\omega) \tag{14}$$

Besides various histogram and discrete (fast) Fourier transform methods available for the computation of coherence, AR methods may be also used, since they are also linear transforms, the Fourier transform of the delay operator being simply $z^k = e^{-j2\pi\omega\tau_S}$ where τ_S is the sampling time and $k = \omega\tau_S$. Plugging this into Equation (9) we obtain:

$$X(j\omega) = \left(\sum_{k=1}^{K} A_k e^{-j2\pi\omega\tau_S k}\right) X(j\omega) + BU(j\omega) + \mathcal{D}$$
$$Y(j\omega) = CX(j\omega) \tag{15}$$

$$Y(j\omega) = C\left(I - \sum_{k=1}^{K} A_k e^{-j2\pi\omega\tau_S k}\right)^{-1} (BU(j\omega) + \mathcal{D}W(j\omega)) \tag{16}$$

In terms of a SVAR therefore (as opposed to VAR) the mixing matrix C does not affect stability, nor the dynamic response (i.e. the poles). The transfer functions from ith innovations to jth output are entries of the following matrix of functions:

$$H(j\omega) = C\left(I - \sum_{k=1}^{K} A_k e^{-j2\pi\omega\tau_S k}\right)^{-1} D \tag{17}$$

The spectral matrix is simply (having already assumed independent unit Gaussian noise):

$$S(j\omega) = H(j\omega)^* H(j\omega) \tag{18}$$

The coherency as the coherence following definitions above. The partial coherence considers the pair (i, j) of signals conditioned on all *other* signals, the (ordered) set of which we denote $\overline{(i, j)}$:

$$S_{i,j|\overline{(i,j)}}(j\omega) = S_{(i,j),(i,j)} + S_{(i,j),\overline{(i,j)}} S^{-1}_{\overline{(i,j)},\overline{(i,j)}} S_{\overline{(i,j)},(i,j)} \tag{19}$$

Where the subscripts refer to row/column subsets of the matrix $S(j\omega)$. The partial spectrum, substituted into Equation (13) gives us partial coherency $C_{i,j|\overline{(i,j)}}(j\omega)$ and correspondingly, partial coherence $c_{i,j|\overline{(i,j)}}(j\omega)$. These functions are symmetric and therefore cannot indicate direction of interaction in the pair (i, j). Several alternatives have been proposed to account for this limitation. Kaminski and Blinowska (1991); Blinowska et al. (2004) proposed the following normalization of $H(j\omega)$ which attempts to measure the relative magnitude of the transfer function from any innovations process to any output (which is equivalent to measuring the normalized strength of Granger causality) and is called the directed transfer function (DTF):

$$\gamma_{ij}(j\omega) = \frac{H_{ij}(j\omega)}{\sqrt{\sum_k |H_{ik}(j\omega)|^2}}$$

$$\gamma^2_{ij}(j\omega) = \frac{\left|H_{ij}(j\omega)\right|^2}{\sum_k |H_{ik}(j\omega)|^2} \tag{20}$$

A similar measure is called directed coherence Baccalá et al. (Feb 1991), later elaborated into a method complimentary to DTF, called partial directed coherence (PDC) Baccalá and Sameshima (2001); Sameshima and Baccalá (1999), based on the inverse of H:

$$\pi_{ij}(j\omega) = \frac{H^{-1}_{ij}(j\omega)}{\sqrt{\sum_k \left|H^{-1}_{ik}(j\omega)\right|^2}}$$

The objective of these coherency-like measures is to place a measure of directionality on the otherwise information-symmetric coherency. While SVAR is not generally used as a basis of the autoregressive means of spectral and coherence estimation, or of DTF/PDC is is done so in this paper for completeness (otherwise it is assumed $C = I$). Granger's 1969 paper did consider a mixing matrix (indirectly, by adding non-diagonal terms to the zero-lag matrix), and suggested ignoring the role of that part of coherency which depends on mixing terms as non-informative 'instantaneous causality'. Note that the ambiguity of the role and identifiability of the full zero lag matrix, as described herein, was fully known at the time and was one of the justifications given for separating sub-sampling time dynamics. Another measure of directionality, proposed by Schreiber (2000) is a Shannon-entropy interpretation of Granger Causality, and therefore will be referred to as GC herein. The Shannon entropy, and conditional Shannon entropy of a random process is related to its spectrum. The conditional entropy formulation of Granger Causality for AR models in the multivariate case is (where $\overline{(i)}$ denotes, as above, all other elements of the vector except i):

$$\mathcal{H}^{GC}_{j\to i|u} = \mathcal{H}(y_{i,t+1}|y_{:,t:t-K}, u_{:,t:t-K}) - \mathcal{H}(y_{i,t+1}|y_{\overline{(j)},t:t-K}, u_{:,t:t-K})$$

$$\mathcal{H}^{GC}_{j\to i|u} = \log \mathcal{D}_i - \log \mathcal{D}_i^{\overline{(j)}} \tag{21}$$

The Shannon entropy of a Gaussian random variable is the logarithm of its standard deviation plus a constant. Notice than in this paper the definition of Granger Causality is slightly different than the literature in that it relates to the innovations process of a mixed output SVAR system of closest rotation and not a regular MVAR. The second term $\mathcal{D}_i^{\overline{(j)}}$ is formed by computing a reduced SVAR system which omits the jth variable. Recently Barrett et al. have proposed an extension of GC, based on prior work by Geweke (1982) from interaction among pairs of variables to groups of variables, termed multivariate Granger Causality (MVGC) Barrett et al. (2010). The above definition is straightforwardly extensible to the group case, where I ad J are subsets of $1..D$, since total entropy of independent variables is the sum of individual entropies.

$$\mathcal{H}^{GC}_{J\to I|u} = \sum_{i\in I}\left(\log \mathcal{D}_i - \log \mathcal{D}_i^{\overline{(J)}}\right) \tag{22}$$

The Granger entropy can be calculated directly from the transfer function, using the Shannon-Hartley theorem:

$$\mathcal{H}^{GCH}_{j\to i} = -\sum_{\omega} \Delta\omega \ln\left(1 - \frac{\left|H_{ij}(\omega)\right|^2}{S_{ii}(\omega)}\right) \tag{23}$$

Finally Nolte (Nolte et al., 2008) introduced a method called Phase Slope Index which evaluates bilateral causal interaction and is robust to mixing effects (i.e. zero lag, observation or innovations covariance matrices that depart from MVAR):

$$PSI_{ij\to i} = Im\left(\sum_{\omega} C^*_{ij}(\omega)\; C_{ij}(\omega + d\omega)\right) \tag{24}$$

PSI, as a method is based on the observation that pure mixing (that is to say, all effects stochastically equivalent to output mixing as outlined above) does not affect the imaginary part of the coherency C_{ij} just as (equivalently) it does not affect the antisymmetric part of the auto-correlation of a signal. It does not place a measure the phase relationship *per se*, but rather the slope of the coherency phase weighted by the magnitude of the coherency.

7. Causal Structural Information

Currently, Granger causality estimation based on linear VAR modeling has been shown to be susceptible to mixed noise, in the presence of which it may produce false causality assignment Nolte et al. (2010). In order to allow for accurate causality assignment in the presence of instantaneous interaction and aliasing the Causal Structural Information (CSI) method and statistic for causality assignment is introduced below.

Consider the SVAR lower triangular form in (10) for a set of observations y. The information transfer from i to j may be obtained by first defining the index re-orderings:

$$ij* \triangleq \{i, j, \overline{ij}\}$$

$$i* \triangleq \{i, \overline{ij}\}$$

This means that we reorder the (identified) mixed-innovations system by placing the target time series first and the driver series second, followed by all the rest. The same ordering, minus the driver is also useful. We define CSI as

$$CSI(j \to i|\overline{ij}) \triangleq \log({}_{\triangleleft ij*}D_{11}) - \log({}_{\triangleleft i*}D_{11}) \tag{25}$$

$$CSI(i, j|\overline{ij}) \triangleq CSI(j \to i|\overline{ij}) - CSI(i \to j|\overline{ij}) \tag{26}$$

Where the D is *upper-triangular* form in each instance. This Granger Causality formulation requires the identification of 3 different SVAR models, one for the entire time series vector, and one each for all elements except i and all elements except j. Via Cholesky decomposition, the logarithm of the top vertex of the triangle is proportional to the entropy rate (conditional information) of the innovations process for the target series given all other (past and present) information including the innovations process. While this definition is clearly an interpretation of the core idea of Granger causality, it is, like DTF and PDC, not an independence statistic but a measure of (causal) information flow among elements of a time-series vector. Note the anti-symmetry (by definition) of this information measure $CSI(i, j|\overline{ij}) = -CSI(j, i|\overline{ij})$. Note also that $CSI(j \to i|\overline{ij})$ and $CSI(i \to j|\overline{ij})$ may very conceivably have the same sign: the various triangular forms used to derive this measure are purely for calculation purposes, and do not carry intrinsic meaning. As a matter of fact other re-orderings and SVAR forms may be employed for convenient calculation as well. In order to improve the explanatory power of the CSI statistic the following normalization is proposed, mirroring that defined in Equation (5) :

$$\mathcal{F}^{CSI}_{j \to i|\overline{ij}} \triangleq \frac{CSI(i, j|\overline{ij})}{\log({}_{\triangleleft i*}D_{11}) + \log({}_{\triangleleft j*}D_{11}) + \zeta} \tag{27}$$

This normalization effectively measures the ratio of causal to non-causal information, where ζ is a constant which depends on the number of dimensions and quantization width and is necessary to transform continuous entropy to discrete entropy.

8. Estimation of multivariate spectra and causality assignment

In Section 6 and a series of causality assignment methods based on spectral decomposition of a multivariate signal were described. In this section spectral decomposition itself will be discussed, and a novel means of doing so for unevenly sampled data will be introduced and evaluated along with the other methods for a bivariate benchmark data set.

8.1. The cardinal transform of the autocorrelation

Currently there are few commonly used methods for cross- *power* spectrum estimation (i.e. multivariate spectral power estimation) as opposed to univariate power spectrum estimation, and these methods average over repeated, or shifting, time windows and therefore require a lot of data points. Furthermore all commonly used multivariate spectral power estimation methods rely on synchronous, evenly spaced sampling, despite the fact that much of available data is unevenly sampled, has missing values, and can be composed of relatively short sequences. Therefore a novel method is presented below for multivariate spectral power estimation which can be estimated on asynchronous data.

Returning to the definition of coherency as the Fourier transform of the autocorrelation, which are both continuous transforms, we may extend the conceptual means of its estimation in the discrete sense as a regression problem (as a discrete Fourier transform, DFT) in the evenly sampled case as:

$$\Omega_n \triangleq \frac{n}{2\tau_0(N-1)}, \quad n = -\lfloor N/2 \rfloor \ldots \lfloor N/2 \rfloor \tag{28}$$

$$\hat{C}_{ij}(\omega)|_{\,\omega=\Omega} = a_{ij,n} + \mathrm{j} b_{ij,n} \tag{29}$$

$$\rho_{ji}(-k\tau) = \rho_{ij}(k\tau) = E(x_i(t)x_j(t+k\tau)) \tag{30}$$

$$\rho_{ij}(k\tau_0) \cong \frac{1}{N-k} \sum_{q=1:N-k} x_i(q)x_j(q+k) \tag{31}$$

$$\left\{a_{ij}, b_{ij}\right\} = \arg\min \sum_{k=-N/2}^{N/2} \left(\rho_{ij}(k\tau_0) - a_{ij,n} cos(2\pi\Omega_n\tau_0 k) - b_{ij,n}\sin(2\pi\Omega_n\tau_0 k)\ \right)^2 \tag{32}$$

where τ_0 is the sampling interval. Note that for an odd number of points the regression above is actually a well determined set of equations, corresponding to the 2-sided DFT. Note also that by replacing the expectation with the geometric mean, the above equation can also be written (with a slight change in weighting at individual lags) as:

$$\left\{a_{ij}, b_{ij}\right\} = \arg\min \sum_{p,q\in 1..N} \left(x_{i,p}x_{j,q} - a_{ij,n}\cos(2\pi\Omega_k(t_{i,p} - t_{j,q})) - b_{ij,n}\sin(2\pi\Omega_n(t_{i,p} - t_{j,q}))\ \right)^2 \tag{33}$$

The above equation holds even for time series sampled at unequal (but overlapping) times (x_i, t_i) and (x_j, t_j) as long as the frequency basis definition is adjusted (for example $\tau_0 = 1$). It represents a discrete, finite approximation of the continuous, infinite autoregression function of an infinitely long random process. It is a regression on the outer product of the vectors x_i and x_j. Since autocorrelations for finite memory systems tend to fall off to zero with increasing lag magnitude, a novel coherency estimate is proposed based on the cardinal sine and cosine functions, which also decay, as a compact basis:

$$\hat{C}_{ij}(\omega) = a_{ij,n} \sum C(\Omega_n) + \mathrm{j}\, b_{ij,n} \mathcal{S}(\Omega_n) \tag{34}$$

$$\left\{a_{ij}, b_{ij}\right\} = \arg\min$$

$$\sum_{p,q \in 1..N} \left(x_{i,p} x_{j,q} - a_{ij,n} cosc(2\pi\Omega_k(t_{i,p} - t_{j,q})) - b_{ij,n} sinc(2\pi\Omega_n(t_{i,p} - t_{j,q})) \right)^2 \tag{35}$$

Where the sine cardinal is defined as $sinc(x) = sin(\pi x)/x$, and its Fourier transform is $\mathcal{S}(j\omega) = 1, |j\omega| < 1$ and $\mathcal{S}(j\omega) = 0$ otherwise. Also the Fourier transform of the cosine cardinal can be written as $C(j\omega) = j\omega\, S(j\omega)$. Although in principle we could choose any complete basis as a means of Fourier transform estimation, the cardinal transform preserves the odd-even function structure of the standard trigonometric pair. Computationally this means that for autocorrelations, which are real valued and even, only *sinc* needs to be calculated and used, while for cross-correlation both functions are needed. As linear mixtures of independent signals only have symmetric cross-correlations, any nonzero values of the *cosc* coefficients would indicate the presence of dynamic interaction. Note that the Fast Fourier Transform earns its moniker thanks to the orthogonality of *sin* and *cos* which allows us to avoid a matrix inversion. However their orthogonality holds true only for infinite support, and slight correlations are found for finite windows - in practice this effect requires further computation (windowing) to counteract. The cardinal basis is not orthogonal, requires full regression and may have demanding memory requirements. For moderate size data this not problematic and implementation details will be discussed elsewhere.

8.2. Robustness evaluation based on the NOISE dataset

A dataset named NOISE, intended as a benchmark for the bivariate case, has been introduced in the preceding NIPS workshop on causality Nolte et al. (2010) and can be found online at `www.causality.inf.ethz.ch/repository.php`, along with the code that generated the data. It was awarded best dataset prize in the previous NIPS causality workshop and challenge Guyon et al. (2010). For further discussion of Causality Workbench and current dataset usage see Guyon (2011). NOISE is created by the summation of the output of a strictly causal VAR DGP and a non-causal SVAR DGP which consists of mixed colored noise:

$$y_{C,i} = \sum_{k=1}^{K} \begin{bmatrix} a_{11} & a_{12} \\ 0 & a_{22} \end{bmatrix}_{C,k} y_{C,i-k} + w_{C,i} \tag{36}$$

$$x_{N,i} = \sum_{k=1}^{K} \begin{bmatrix} a_{11} & 0 \\ 0 & a_{22} \end{bmatrix}_{N,k} x_{N,i-k} + w_{N,i}$$

$$y_{N,i} = Bx_{N,i} \tag{37}$$

$$y = (1 - |\beta|)y_N + |\beta|y_C \frac{||y_N||_F}{||y_C||_F} \tag{38}$$

The two sub-systems are pictured graphically as systems A and B in Figure 1. If $\beta < 0$ the AR matrices that create y_C are transposed (meaning that y_{1C} causes y_{2C} instead of the opposite). The coefficient β is represented in Nolte et al. (2010) by 'γ ' where $\beta = sgn(\gamma)(1 - |\gamma|)$. All coefficients are generated as independent Gaussian random variables of unit variance, and unstable systems are discarded. While both the causal and noise generating systems have the same order, note that the system that would generate the sum thereof requires an infinite order SVAR DGP to generate (it is not stochastically equivalent to any SVAR DGP but instead is a SVARMA DGP, having both poles and zeros). Nevertheless it is an interesting benchmark since the exact parameters are not fully recoverable via the commonly used VAR modeling procedure and because the causality interpretation is fairly clear: the sum of a strictly causal DGP and a stochastic noncausal DGP should retain the causality of the former.

In this study, the same DGPs were used as in NOISE but as one of the current aims is to study the influence of sample size on the reliability of causality assignment, signals of 100, 500, 1000 and 5000 points were generated (as opposed to the original 6000). This is the dataset referred to as PAIRS below, which only differs in numbers of samples per time series. For each evaluation 500 time series were simulated, with the order for each system of being uniformly distributed from 1 to 10. The following methods were evaluated:

- PSI (Ψ) using Welch's method, and segment and epoch lengths being equal and set to $\lceil \sqrt{N} \rceil$ and otherwise is the same as Nolte et al. (2010).
- Directed transfer function DTF. estimation using an automatic model order selection criterion (BIC, Bayesian Information Criterion) using a maximum model order of 20. DTF has been shown to be equivalent to GC for linear AR models (Kaminski et al., 2001) and therefore GC itself is not shown . The covariance matrix of the residuals is also included in the estimate of the transfer function. The same holds for all methods described below.
- Partial directed coherence PDC. As described in the previous section, it is similar to DTF except it operates on the signal-to-innovations (i.e. inverse) transfer function.
- Causal Structural Information. As a described above this is based on the triangular innovations equivalent to the estimated SVAR (of which there are 2 possible forms in the bivariate case) and which takes into account instantaneous interaction / innovations process covariance.

All methods were statistically evaluated for robustness and generality by performing a 5-fold jackknife, which gave both a mean and standard deviation estimate for each method and each simulation. All statistics reported below are mean normalized by standard deviation (from jackknife). For all methods the final output could be -1, 0, or 1, corresponding to causality assignment 1→2, no assignment, and causality 2 → 1. A true positive (TP) was the rate of correct causality assignment, while a false positive (FP) was the rate of incorrect causality assignment (type III error), such that TP+FP+NA=1, where NA stands for rate of no assignment (or neutral assignment). The TP and FP rates are be co-modulated by increasing/decreasing the threshold of the absolute value of the *mean/std* statistic, under which no causality assignment is made:

$$STAT = rawSTAT/\mathrm{std}(rawSTAT),\ \ \mathit{rawSTAT{=}PSI,\ DTF,\ PDC\ ..}$$

$$c = sign(STAT) \text{ if } STAT > TRESH,\ 0 \text{ otherwise}$$

In Table 1 we see results of overall accuracy and controlled True Positive rate for the ***non-mixed*** colored noise case (meaning the matrix *B* above is diagonal). In Table 1 and Table 2 methods are ordered according to the mean TP rate over time series length (highest on top).

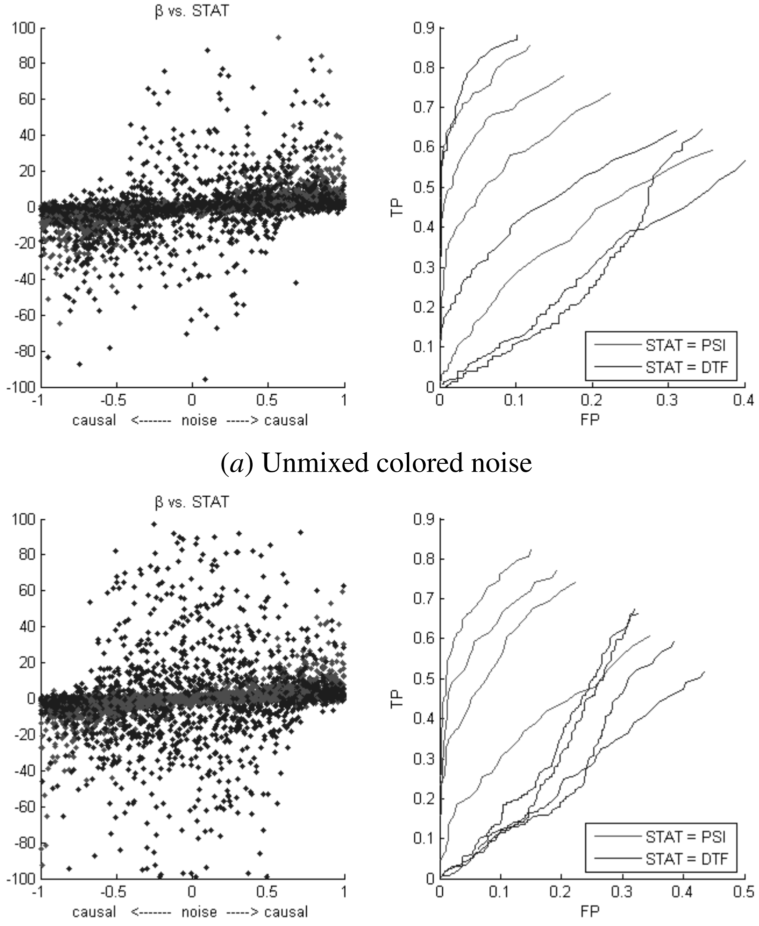
Table 1: Unmixed colored noise PAIRS

	Max. Accuracy				TP , FP < 0.10			
	100	**500**	**1000**	**5000**	**100**	**500**	**1000**	**5000**
Ψ	0.62	0.73	0.83	0.88	0.25	0.56	0.75	0.85
DTF	0.58	0.79	0.82	0.88	0.18	0.58	0.72	0.86
CSI	0.62	0.72	0.79	0.89	0.23	0.53	0.66	0.88
Ψ_C	0.57	0.68	0.81	0.88	0.19	0.29	0.70	0.87
PDC	0.64	0.67	0.75	0.78	0.23	0.33	0.48	0.57

In Table 2 we can see results for a PAIRS, in which the noise mixing matrix B is not strictly diagonal.

Table 2: Mixed colored noise PAIRS

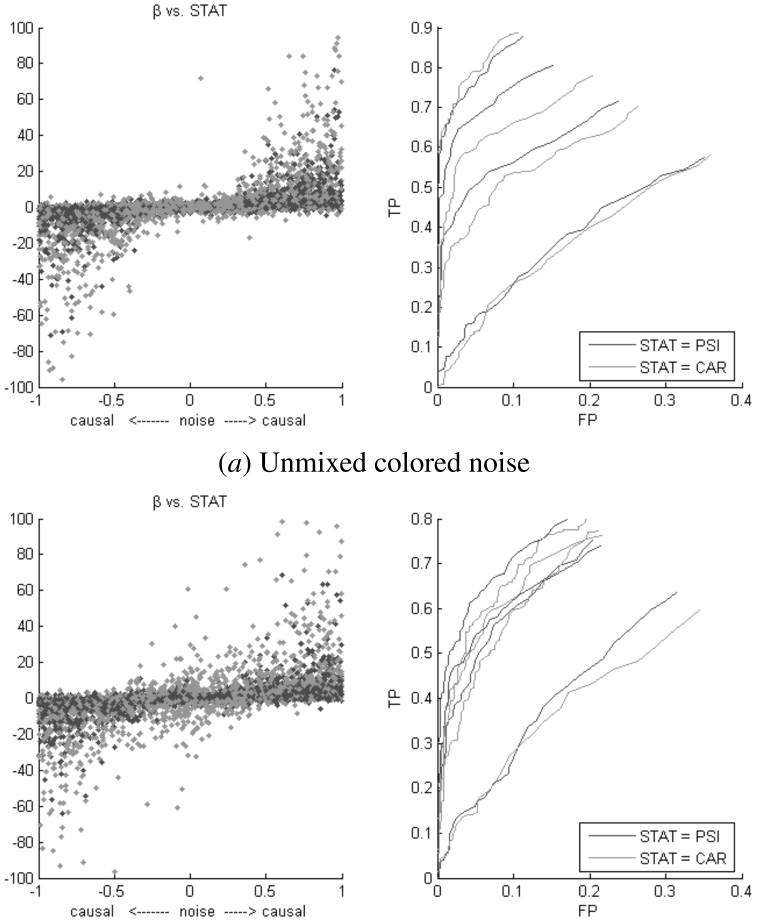
	Max. Accuracy				TP , FP< 0.10			
	N=100	**500**	**1000**	**5000**	**N=100**	**500**	**1000**	**5000**
Ψ_C	0.64	0.74	0.81	0.83	0.31	0.49	0.64	0.73
Ψ	0.66	0.76	0.78	0.81	0.25	0.59	0.61	0.71
CSI	0.63	0.77	0.79	0.80	0.27	0.62	0.59	0.66
PDC	0.64	0.71	0.69	0.66	0.24	0.30	0.29	0.24
DTF	0.55	0.61	0.66	0.66	0.11	0.10	0.09	0.12



(*a*) Unmixed colored noise

(*b*) Mixed colored noise

Figure 2: PSI vs. DTF Scatter plots of β vs. STAT (to the left of each panel), TP vs. FP curves for different time series lengths (100, 500,1000 and 500) (right). a) colored unmixed additive noise. b) colored mixed additive noise. DTF is equivalent to Granger Causality for linear systems. All STAT values are jackknife mean normalized by standard deviation.



(*a*) Unmixed colored noise

(*b*) Mixed colored noise

Figure 3: PSI vs. CSI Scatter plots of β vs. STAT (to the left of each panel), TP vs. FP curves for different time series lengths (right). a) unmixed additive noise. b) mixed additive noise

As we can see in both Figure 2 and Table 1, all methods are almost equally robust to unmixed colored additive noise (except PDC). However, while *addition of mixed colored noise* induces a mild gap in maximum accuracy, it creates a large gap in terms of TP/FP rates. Note the dramatic drop-off of the TP rate of VAR/SVAR based methods PDC and DTF. Figure 3 shows this most clearly, by a wide scatter of STAT outputs for DTF around $\beta = 0$ that is to say with no actual causality in the time series and a corresponding fall-off of TP vs. FP rates. Note also that PSI methods still allow a fairly reasonable TP rate determination at low FP rates of 10% even at 100 points per time-series, while the CSI method was also robust to the addition of colored mixed noise, not showing any significant difference with respect to PSI except a higher FP rate for longer time series (N=5000). The advantage of PSIcardinal was near PSI in overall accuracy. In conclusion, DTF (or weak Granger causality) and PDC are not robust with respect to additive mixed colored noise, although they perform similarly to PSI and CSI for independent colored noise.[1]

9. Conditional causality assignment

In multivariate time series analysis we are often concerned with inference of causal relationship among more than 2 variables, in which the role of a potential common cause must be accounted for, analogously to vanishing partial correlation in the static data case. For this reason the PAIRS data set was extended into a set called TRIPLES in which the degree of common driver influence versus direct coupling was controlled.

In effect, the TRIPLES DGP is similar to PAIRS, in that additive noise is mixed colored noise (in 3 dimensions) but in this case another variable x_3 may drive the pair x_1, x_2 independently of each other, also with random coefficients (but either one set to 1/10 of the other randomly). That is to say, the signal is itself a mixture of one where there is a direct one sided causal link among x_1, x_2 as in PAIRS and one where they are actually independent but commonly driven, according to a parameter χ which at 0 is commonly driven and at 1 is causal.

$$\beta < 0 \quad y_{C,i} = \sum_{k=1}^{K} \begin{bmatrix} a_{11} & a_{12} & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}_{C,k} y_{C,i-k} + w_{C,i}$$

[1] Correlation and rank correlation analysis was performed (for N=5000) to shed light on the reason for the discrepancy between PSI and CSI. The linear correlation between rawSTAT and STAT was .87 and .89 for PSI and CSI. No influence of model order K of the simulated system was seen in the error of either PSI or CSI, where error is estimated as the difference in rank of $rankerr(STAT) = |rank(\beta) - rank(STAT)|$. There were however significant correlations between $rank(|\beta|)$ and $rankerr(STAT)$, -.13 for PSI and -.27 for CSI. Note that as expected, standard Granger causality (GC) performed the same as DTF (TP=0.116 for FP<.10). Using Akaike's Information Criterion (AIC) instead of BIC for VAR model order estimation did not significantly affect AR-based STAT values.

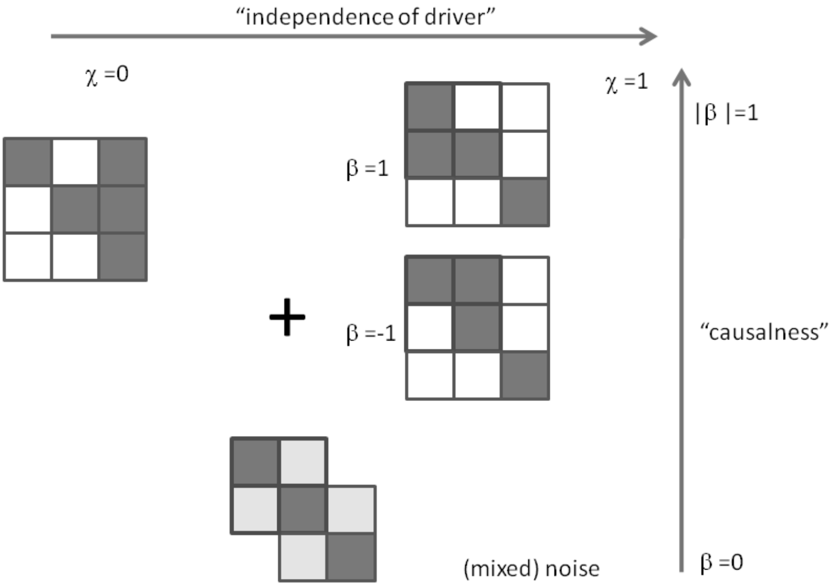


Figure 4: Diagram of TRIPLES dataset with common driver

$$\beta > 0 \quad y_{C,i} = \sum_{k=1}^{K} \begin{bmatrix} a_{11} & 0 & 0 \\ a_{12} & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}_{C,k} y_{C,i-k} + w_{C,i} \tag{39}$$

$$x_{N,i} = \sum_{k=1}^{K} \begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{22} \end{bmatrix}_{N,k} x_{N,i-k} + w_{N,i}$$

$$y_{N,i} = B x_{N,i} \tag{40}$$

$$x_{D,i} = \sum_{k=1}^{K} \begin{bmatrix} a_{11} & 0 & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{22} \end{bmatrix}_{D,k} x_{D,i-k} + w_{D,i}$$

$$y_{MC} = (1 - |\beta|) y_N + |\beta| y_C \frac{||y_N||_F}{||y_C||_F} \tag{41}$$

$$y_{DC} = (1 - \chi) y_{MC} + \chi \, y_D \frac{||y_{MC}||_F}{||y_D||_F} \tag{42}$$

The Table 3, similar to the tables in the preceding section, shows results for all usual methods, except for PSIpartial which is PSI calculated on the partial coherence as defined above and calculated from Welch (cross-spectral) estimators in the case of mixed noise and a common driver.

Table 3: TRIPLES: Commonly driven, additive mixed colored noise

	Max. Accuracy				TP , FP< 0.10			
	100	**500**	**1000**	**5000**	**100**	**500**	**1000**	**5000**
Ψ_p	0.53	0.61	0.71	0.75	0.12	0.31	0.49	0.56
Ψ	0.54	0.60	0.70	0.72	0.10	0.25	0.40	0.52
CSI	0.51	0.60	0.69	0.76	0.09	0.27	0.38	0.45
PDC	0.55	0.54	0.60	0.58	0.13	0.12	0.16	0.13
DTF	0.51	0.56	0.59	0.61	0.12	0.09	0.09	0.11

Notice that the TP rates are lower for all methods with respect to Table 2 which represents the mixed noise situation without any common driver.

10. Discussion

In a recent talk, Emanuel Parzen (Parzen, 2004) proposed, both in hindsight and for future consideration, that aim of statistics consist in an ‘answer machine’, i.e. a more intelligent, automatic and comprehensive version of Fisher’s almanac, which currently consists in a plenitude of chapters and sections related to different types of hypotheses and assumption sets meant to model, insofar as possible, the ever expanding variety of data available. These categories and sub-categories are not always distinct, and furthermore there are competing general approaches to the same problems (e.g. Bayesian vs. frequentist). Is an ‘answer machine’ realistic in terms of time-series causality, prerequisites for which are found throughout this almanac, and which has developed in parallel in different disciplines?

This work began by discussing Granger causality in abstract terms, pointing out the implausibility of finding a general method of causal discovery, since that depends on the general learning and time-series prediction problem, which are incomputable. However, if any consistent patterns that can be found mapping the history of one time series variable to the current state of another (using non-parametric tests), there is sufficient evidence of causal interaction and the null hypothesis is rejected. Such a determination still does not address direction of interaction and relative strength of causal influence, which may require a complete model of the DGP. This study - like many others - relied on the rather strong assumption of stationary linear Gaussian DGPs but otherwise made weak assumptions on model order, sampling and observation noise. Are there, instead, more general assumptions we can use? The following is a list of competing approaches in increasing order of (subjectively judged) strength of underlying assumption(s):

- Non-parametric tests of conditional probability for Granger non-causality rejection. These directly compare the probability distributions $P(y_{1,j} \,|\, y_{1,j-1..1}, u_{j-1..1})$ $P(y_{1,j} \,|\, y_{1,j-1..1}, u_{j-1..1})$ to detect a possible statistically significant difference. Proposed approaches (see chapter in this volume by (Moneta et al., 2011) for a detailed overview and tabulated robustness comparison) include product kernel density with kernel smoothing (Chlaß and Moneta, 2010), made robust by bootstrapping and with density distances such as the Hellinger (Su and White, 2008), Euclidean (Szekely and Rizzo, 2004), or completely nonparametric difference tests such Cramer-Von Mises or Kolmogorov-Smirnov. A potential pitfall of nonparametric approaches is their loss of power for higher dimensionality of the space over which the probabilities are estimated - *aka* the curse of dimensionality (Yatchew, 1998). This can occur if the lag order K needed to be considered is high, if the system memory is long, or the number of other variables over which GC must be conditioned ($u_{j-1..1}$) is high. In the case of mixed noise, strong GC estimation would require accounting for all observed variables (which in neuroscience can number in the hundreds). While non-parametric non-causality rejection is a very useful tool (and could be valid even if the lag considered in analysis is much smaller than the true lag K), in practice we would require robust estimated of causal direction and relative strength of different factors, which implies a complete accounting of all relevant factors. As was already discussed, in many cases Granger non-causality is likely to be rejected in both directions: it is useful to find the dominant one.

- General parametric or semi-parametric (black-box) predictive modeling subject to GC interpretation which can provide directionality, factor analysis and interpretation of information flow. A large body of literature exists on neural network time series modeling (in this context see White (2006)), complemented in recent years by support vector regression and Bayesian processes. The major concern with black-box predictive approaches is model validation: does the fact that a given model features a high cross-validation score automatically imply the implausibility of another predictive model with equal CV-score that would lead to different conclusions about causal structure? A reasonable compromise between nonlinearity and DGP class restriction can be seen in (Chu and Glymour, 2008) and Ryali et al. (2010), in which the VAR model is augmented by additive nonlinear functions of the regressed variable and exogenous input. Robustness to noise, sample size influence and accuracy of effect strength and direction determination are open questions.

- Linear dynamic models which incorporate (and often require) non-Gaussianity in the innovations process such as ICA and TDSEP (Ziehe and Mueller, 1998). See Moneta et al. (2011) in this volume for a description of causality inference using ICA and causal modeling of innovation processes (i.e. independent components). Robustness under permutation is necessary for a principled accounting of dynamic interaction and partition of innovations process entropy. Note that many ICA variants assume that at most one of the innovations processes is Gaussian, a strong assumption which requires a posteriori checks. To be elucidated is the robustness to filtering and additive noise.

- Non-stationary Gaussian linear models. In neuroscience non-stationarity is important (the brain may change state abruptly, respond to stimuli, have transient pathological episodes etc). Furthermore accounting for non-stochastic exogenous inputs needs further consideration. Encouragingly, the current study shows that even in the presence of complex confounders such as common driving signals and co-variate noise, segments as small as 100 points can yield accurate causality estimation, such that changes in longer time series can be adaptively tracked. Note that in establishing statistical significance we must take into account signal bandwidth: up-sampling the same process would arbitrarily increase the number of samples but not the information contained in the signal. See Appendix A for a proposal on non-parametric bandwidth estimation.

- Linear Gaussian dynamic models: in this work we have considered SVAR but not wider classes of linear DGPs such as VARMA and heteroskedastic (GARCH) models. In comparing PSI and CSI note that overall accuracy of directionality assignment was virtually identical, but PSI correlated slightly better with effect size. While CSI made slightly more errors at low strengths of ‘causality’, PSI made slightly more errors at high strengths. Nevertheless, PSI was most robust to (colored, mixed) noise and hidden driving/conditioning signal (tabulated significance results are provided in Appendix A). Jackknifed, normalized estimates can help establish causality at low strength levels, although a large raw PSI statistic value may also suffice. A potential problem with the jackknife (or bootstrap) procedure is the strong stationarity assumption which allows segmentation and rearrangement of the data.

Although AR modeling was commonly used to model interaction in time series and served as a basis for (linear) Granger causality modeling (Blinowska et al., 2004; Baccalá and Sameshima, 2001), robustness to mixed noise remained a problem, which the spectral method PSI was meant to resolve (Nolte et al., 2008). While 'phase', if structured, already implies prediction, precedence and mutual information among time series elements, it was not clear how SVAR methods would reconcile with PSI performance, until now. This prompted the introduction in this article of the causal AR method (CSI) which takes into account 'instantaneous' causality . A prior study had shown that strong Granger causality is preserved under addition of colored noise, as opposed to weak (i.e. strictly time ordered) causality Solo (2006). This is consistent with the results obtained herein. The CSI method, measuring strong Granger Causality, was in fact robust with respect to a type of noise not studied in (Solo, 2006), which is *mixed* colored noise; other VAR based methods and (weak) Granger causality measures were not. While and SVAR DGP observed under additive colored noise is a VARMA process (the case of the PAIRS and TRIPLES datasets), SVAR modeling did not result in a severe loss of accuracy. AR processes of longer lags can approximate VARMA processes by using higher orders and more parameters, even if doing so increases exposure to over-fit and may have resulted in a small number of outliers. Future work must be undertaken to ascertain what robustness and specificity advantages result from VARMA modeling, and if it is worth doing so considering the increased computational overload. One of the common 'defects' of real-life data are missing/outlier samples, or uneven sampling in time, or that the time stamps of two time series to be compared are unrelated though overlapping: it is for these case that the method PSIcardinal was developed and shown to be practically equal in numerical performance to the Welch estimate-based PSI method (though it is slower computationally). Both PSI estimates were robust to common driver influence even when not based on partial but direct coherency because it is the *asymmetry* in influence of the driver on phase that is measured rather than its overall strength. While 2-way interaction with conditioning was considered, future work must organize multivariate signals using directed graphs, as in DAG-type static causal inference. Although only 1 conditioning signal was analysed in this paper, the methods apply to higher numbers of background variables. Directed Transfer Function and Partial Directed Coherence did not perform as well under additive colored noise, but their formulation does address a theoretically important question, namely the partition of strength of influence among various candidate causes of an observation; CSI also proposes an index for this important purpose. Whether the assumptions about stationarity or any other data properties discussed are warranted may be checked by performing appropriate *a posteriori* tests. If these tests justify prior assumptions and a correspondingly significant causal effect is observed, we can assign statistical confidence intervals to the causality structure of the system under study. The 'almanac' chapter on time series causality is rich and new alternatives are emerging. For the entire corpus of time-series causality statistics to become an 'answer machine', however, it is suggested that a principled bottom-up investigation be undertaken, beginning with the simple SVAR form studied in this paper and all proposed criteria be quantified: type I, II and III errors, accurate determination of causality strength and direction and robustness in the presence of conditioning variables and colored mixed noise.

Acknowledgments

I would like to thank Guido Nolte for his insightful feedback and informative discussions.

References

H. Akaike. On the use of a linear model for the identification of feedback systems. *Annals of the Institute of statistical mathematics*, 20(1):425–439, 1968.

L. A Baccalá and K. Sameshima. Partial directed coherence: a new concept in neural structure determination. *Biological cybernetics*, 84(6):463–474, 2001.

L. A. Baccalá, M. A. Nicolelis, C. H. Yu, and M. Oshiro. Structural analysis of neural circuits using the theory of directed graphs. *Computers and Biomedical Research, an International Journal*, 24:7–28, Feb 1991. URL `http://www.ncbi.nlm.nih.gov/pubmed/2004525`.

A. B. Barrett, L. Barnett, and A. K. Seth. Multivariate granger causality and generalized variance. *Physical Review E*, 81(4):041907, April 2010. doi: 10.1103/PhysRevE.81.041907. URL `http://link.aps.org/doi/10.1103/PhysRevE.81.041907`.

B. S. Bernanke, J. Boivin, and P. Eliasz. Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach. *Quarterly Journal of Economics*, 120(1):387–422, 2005.

K. J. Blinowska, R. Kuś, and M. Kamiński. Granger causality and information flow in multivariate processes. *Physical Review E*, 70(5):050902, November 2004. doi: 10.1103/PhysRevE.70.050902. URL `http://link.aps.org/doi/10.1103/PhysRevE.70.050902`.

P. E. Caines. Weak and strong feedback free processes. *IEEE. Trans. Autom . Contr*, 21:737–739, 1976.

N. Chlaß and A. Moneta. Can Graphical Causal Inference Be Extended to Nonlinear Settings? *EPSA Epistemology and Methodology of Science*, pages 63–72, 2010.

T. Chu and C. Glymour. Search for additive nonlinear time series causal models. *The Journal of Machine Learning Research*, 9:967–991, 2008.

R. A. Fisher. *Statistical Methods for Research Workers.* Macmillan Pub Co, 1925. ISBN 0028447301.

W. Gersch and G. V. Goddard. Epileptic focus location: spectral analysis method. *Science (New York, N.Y.)*, 169(946):701–702, August 1970. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/5429908. PMID: 5429908.

J. Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77:304–313, 1982.

G. Gigerenzer, Z. Swijtink, T. Porter, L. Daston, J. Beatty, and L. Kruger. *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge University Press, October 1990. ISBN 052139838X.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, August 1969. ISSN 00129682. URL http://www.jstor.org/stable/1912791.

I. Guyon. Time series analysis with the causality workbench. *Journal of Machine Learning Research, Workshop and Conference Proceedings*, XX. Time Series Causality:XX–XX, 2011.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *JMLR*, 3: 1157–1182, March 2003.

I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov. Design and analysis of the causation and prediction challenge. wcci2008 workshop on causality, hong kong, june 3-4 2008. *Journal of Machine Learning Research Workshop and Conference Proceedings*, 3:1–33, 2008.

I. Guyon, D. Janzing, and B. Schölkopf. Causality: Objectives and assessment. *JMLR W&CP*, 6:1–38, 2010.

P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 689–696, 2009.

M Kaminski, M Ding, W A Truccolo, and S L Bressler. Evaluating causal relations in neural systems: granger causality, directed transfer function and statistical assessment of significance. *Biological Cybernetics*, 85(2):145–157, August 2001. ISSN 0340-1200. URL http://www.ncbi.nlm.nih.gov/pubmed/11508777. PMID: 11508777.

M. J. Kaminski and K. J. Blinowska. A new method of the description of the information flow in the brain structures. *Biological Cybernetics*, 65(3):203–210, 1991. ISSN 0340-1200. doi: 10.1007/BF00198091. URL http://dblp.uni-trier.de/rec/bibtex/journals/bc/KaminskiB91.

A. N. Kolmogorov and A. N. Shiryayev. *Selected Works of A.N. Kolmogorov: Probability theory and mathematical statistics*. Springer, 1992. ISBN 9789027727978.

T. C. Koopmans. *Statistical Inference in Dynamic Economic Models, Cowles Commission Monograph, No. 10.* New York: John Wiley & Sons, 1950.

G. Lacerda, P. Spirtes, J. Ramsey, and P. O. Hoyer. Discovering Cyclic Causal Models by Independent Component Analysis. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI-2008), Helsinki, Finland*, 2008.

A. D. Lanterman. Schwarz, wallace and rissanen: Intertwining themes in theories of model selection. *International Statistical Review*, 69(2):185–212, 2001.

M. Li and P. M. B. Vitanyi. *An introduction to Kolmogorov complexity and its applications, 2nd edition*. Springer-Verlag, 1997.

A. Moneta, D. Entner, P.O. Hoyer, and A. Coad. Causal inference by independent component analysis with applications to micro-and macroeconomic data. *Jena Economic Research Papers*, 2010:031, 2010.

A. Moneta, N. Chlaß, D. Entner, and P.Hoyer. Causal search in structural vector autoregression models. *Journal of Machine Learning Research, Workshop and Conference Proceedings*, XX. Time Series Causality:XX–XX, 2011.

G. Nolte, A. Ziehe, V.V. Nikulin, A. Schlögl, N. Krämer, T. Brismar, and K.-R. Müller. Robustly estimating the flow direction of information in complex physical systems. *Physical Review Letters*, 00(23):234101, 2008.

G. Nolte, A. Ziehe, N. Kraemer, F. Popescu, and K.-R. Müller. Comparison of granger causality and phase slope index. *Journal of Machine Learning Research Workshop & Conference Proceedings.*, Causality: Objectives and Assessment:267:276, 2010.

E. Parzen. Long memory of statistical time series modeling. presented at the 2004 nber/nsf time series conference at smu, dallas, usa. Technical report, Texas A&M University, `http://www.stat.tamu.edu/~eparzen/Long%20Memory%20of%20Statistical%20Time%20Series%20Modeling.pdf`, 2004.

J. Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, Cambridge, 2000.

K. Pearson. *Tables for statisticians and biometricians,*. University Press, University College, London, [Cambridge Eng., 1930.

Peter C.B. Phillips. The problem of identification in finite parameter continous time models. *Journal of Econometrics*, 1:351–362, 1973.

F. Popescu. Identification of sparse multivariate autoregressive models. *Proceedings of the European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland*, 2008.

T. Richardson and P. Spirtes. Automated discovery of linear feedback models. In *Computation, causation and discovery*. AAAI Press and MIT Press, Menlo Park, 1999.

A. Roebroeck, A. K. Seth, and P. Valdes-Sosa. Causal time series analysis of functional magnetic resonance imaging data. *Journal of Machine Learning Research, Workshop and Conference Proceedings*, XX. Time Series Causality:XX–XX, 2011.

S. Ryali, K. Supekar, T. Chen, and V. Menon. Multivariate dynamical systems models for estimating causal interactions in fmri. *Neuroimage*, 2010.

K. Sameshima and L. A Baccalá. Using partial directed coherence to describe neuronal ensemble interactions. *Journal of Neuroscience Methods*, 94(1):93:103, 1999.

R. Scheines, P. Spirtes, C. Glymour, C. Meek, and T. Richardson. The TETRAD project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1):65–117, 1998.

T. Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461, July 2000. doi: 10.1103/PhysRevLett.85.461. URL `http://link.aps.org/doi/10.1103/PhysRevLett.85.461`.

C. A. Sims. An autoregressive index model for the u.s. 1948-1975. In J. Kmenta and J.B. Ramsey, editors, *Large-scale macro-econometric models: theory and practice*, pages 283–327. North-Holland, 1981.

V. Solo. On causality i: Sampling and noise. *Proceedings of the 46th IEEE Conference on Decision and Control*, pages 3634–3639, 2006.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT Press, Cambridge MA, 2nd edition, 2000.

L. Su and H. White. A nonparametric Hellinger metric test for conditional independence. *Econometric Theory*, 24(04):829–864, 2008.

G. J. Szekely and M. L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5, 2004.

A. M. Turing. On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42:230–65, 1936.

J. M Valdes-Sosa, P. A aand Sanchez-Bornot, A. Lage-Castellanos, M. Vega-Hernandez, J. Bosch-Bayard, L. Melie-Garca, and E. Canales-Rodriguez. Estimating brain functional connectivity with sparse multivariate autoregression. *Neuroinformatics*, 360(1457):969, 2005.

H White. Approximate nonlinear forecasting methods. In G. Elliott, C. W. J. Granger, and A. Timmermann, editors, *Handbook of Economic Forecasting*, chapter 9, pages 460–512. Elsevier, New York, 2006.

H. White and X. Lu. Granger Causality and Dynamic Structural Systems. *Journal of Financial Econometrics*, 8(2):193, 2010.

N. Wiener. The theory of prediction. *Modern mathematics for engineers, Series*, 1: 125–139, 1956.

H. O. Wold. *A Study in the Analysis of Stationary Time Series.* Stockholm: Almqvist and Wiksell., 1938.

A. Yatchew. Nonparametric regression techniques in economics. *Journal of Economic Literature*, 36(2):669–721, 1998.

G. U. Yule. Why do we sometimes get nonsense correlations between time series? a study in sampling and the nature of time series. *Journal of the Royal Statistical Society*, 89:1–64, 1926.

A. Ziehe and K.-R. Mueller. Tdsep- an efficient algorithm for blind separation using time structure. *ICANN Proceedings*, pages 675–680, 1998.

S. T. Ziliak and D. N. McCloskey. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives.* University of Michigan Press, February 2008. ISBN 0472050079.

Appendix A. Statistical significance tables for Type I and Type III errors

In order to assist practitioners in evaluating the statistical significance of bivariate causality testing, tables were prepared for type I and type III error probabilities as defined in (1) for different values of the base statistic. Below tables are provided for both the jacknifed statistic $\Psi/std(\Psi)$ and for the raw statistic Ψ, which is needed in case the number of points is too low to allow a jackknife/cross-validation/bootstrap or computational speed is at a premium. The spectral evaluation method is Welch's method as described in Section 6. There were 2000 simulations for each condition. The tables in this Appendix differ in one important aspect with respect to those in the main text. In order to avoid non-informative comparison of datasets which are, for example, analyses of the same physical process sampled at different sampling rates, the number of points is scaled by the 'effective' number of points which is essentially the number of samples relative to a simple estimate of the observed signal bandwidth:

$$N^* = N\tau_S / \widehat{BW}$$

$$\widehat{BW} = \frac{||X||_F}{||\Delta X/\Delta T||_F}$$

The values marked with an asterisk have values of both α and γ which are less than 5%. Note also that Ψ is non-dimensional index.

Table 4: α vs. $\Psi/std(\Psi)$

$N^* \rightarrow$	**50**	**100**	**200**	**500**	**750**	**1000**	**1500**	**2000**	**5000**
0.125	0.82	0.83	0.86	0.87	0.88	0.90	0.89	0.89	0.89
0.25	0.67	0.69	0.73	0.77	0.76	0.78	0.78	0.78	0.77
0.5	0.41	0.44	0.50	0.54	0.55	0.56	0.56	0.59	0.57
0.75	0.26	0.26	0.32	0.36	0.36	0.37	0.38	0.40	0.39
1	0.15	0.15	0.20	0.23	0.23	0.25	0.25	0.26	0.26
1.25	0.09	0.09	0.11	0.13	0.14	0.16	0.14	0.15	0.16
1.5	0.06	0.05	0.06	0.07	0.08	0.09	0.08	0.10	0.10
1.75	0.04	0.03	0.04	0.05	0.05 *	0.05 *	0.04 *	0.06	0.06
2	0.03	0.02	0.03	0.02 *	0.02 *	0.03 *	0.02 *	0.03 *	0.03 *
2.5	0.01	0.01	0.01	0.01 *	0.01 *	0.01 *	0.00 *	0.01 *	0.01 *
3	0.01	0.00	0.00	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *
4	0.00	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *
8	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *
16	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *

Table 5: γ vs. $\Psi/std(\Psi)$

$N^* \rightarrow$	**50**	**100**	**200**	**500**	**750**	**1000**	**1500**	**2000**	**5000**
0.125	0.46	0.41	0.35	0.26	0.23	0.21	0.19	0.18	0.16
0.25	0.46	0.39	0.33	0.24	0.21	0.19	0.18	0.16	0.14
0.5	0.44	0.36	0.29	0.21	0.18	0.15	0.14	0.13	0.12
0.75	0.43	0.35	0.28	0.17	0.14	0.12	0.11	0.10	0.09
1	0.43	0.31	0.23	0.13	0.12	0.09	0.08	0.07	0.06
1.25	0.42	0.31	0.20	0.09	0.08	0.06	0.05	0.05	0.04
1.5	0.40	0.26	0.20	0.06	0.05	0.04	0.04	0.04	0.02
1.75	0.42	0.26	0.16	0.05	0.04 *	0.02 *	0.02 *	0.03	0.01
2	0.41	0.23	0.11	0.03 *	0.03 *	0.02 *	0.02 *	0.02 *	0.01 *
2.5	0.41	0.20	0.06	0.02 *	0.02 *	0.01 *	0.01 *	0.01 *	0.01 *
3	0.33	0.09	0.06	0.03 *	0.01 *	0.00 *	0.00 *	0.00 *	0.00 *
4	0.33	0.00 *	0.00 *	0.02 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *
8	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *
16	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *

Table 6: α vs. Ψ

$N^* \rightarrow$	**50**	**100**	**200**	**500**	**750**	**1000**	**1500**	**2000**	**5000**
0.125	0.25	0.22	0.25	0.24	0.24	0.24	0.21	0.21	0.15
0.25	0.09	0.09	0.09	0.09	0.10	0.09	0.09	0.08	0.05 *
0.5	0.02 *	0.01 *	0.01 *	0.02 *	0.02 *	0.02 *	0.01 *	0.01 *	0.01 *
0.75	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *
1	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *

Table 7: γ vs. Ψ

$N^* \rightarrow$	**50**	**100**	**200**	**500**	**750**	**1000**	**1500**	**2000**	**5000**
0.125	0.21	0.17	0.14	0.10	0.08	0.07	0.06	0.05	0.03
0.25	0.11	0.08	0.06	0.04	0.03	0.03	0.02	0.02	0.01 *
0.5	0.03 *	0.01 *	0.01 *	0.01 *	0.01 *	0.00 *	0.00 *	0.01 *	0.00 *
0.75	0.02 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *
1	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *

JMLR: Workshop and Conference Proceedings 12:1–29

Linking Granger Causality and the Pearl Causal Model with Settable Systems

Halbert White HWHITE@UCSD.EDU
Department of Economics
University of California, San Diego
La Jolla, CA 92093

Karim Chalak CHALAK@BC.EDU
Department of Economics
Boston College
140 Commonwealth Avenue
Chestnut Hill, MA 02467

Xun Lu XUNLU@UST.HK
Department of Economics
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong

Editors: Florin Popescu and Isabelle Guyon

Abstract

The causal notions embodied in the concept of Granger causality have been argued to belong to a different category than those of Judea Pearl's Causal Model, and so far their relation has remained obscure. Here, we demonstrate that these concepts are in fact closely linked by showing how each relates to straightforward notions of direct causality embodied in settable systems, an extension and refinement of the Pearl Causal Model designed to accommodate optimization, equilibrium, and learning. We then provide straightforward practical methods to test for direct causality using tests for Granger causality.

Keywords: Causal Models, Conditional Exogeneity, Conditional Independence, Granger Non-causality

1. Introduction

The causal notions embodied in the concept of Granger causality ("$G-$causality") (e.g., Granger, C.W.J., 1969; Granger, C.W.J. and P. Newbold, 1986) are probabilistic, relating to the ability of one time series to predict another, conditional on a given information set. On the other hand, the causal notions of the Pearl Causal Model ("PCM") (e.g., Pearl, J., 2000) involve specific notions of interventions and of functional rather than probabilistic dependence. The relation between these causal concepts has so far remained obscure. For his part, Granger, C.W.J. (1969) acknowledged that $G-$causality was not "true" causality, whatever that might be, but that it seemed likely to be an important part of the full story. On the other hand, Pearl, J. (2000, p. 39) states that "econometric concepts such as 'Granger causality' (Granger, C.W.J., 1969) and 'strong exogeneity' (Engle, R., D. Hendry, and J.-F. Richard, 1983) will be classified as statistical rather than causal." In practice, especially in economics, numerous studies have used $G-$causality either explicitly or implicitly to draw structural or policy conclusions, but without any firm foundation.

Recently, White, H. and X. Lu (2010a, "WL") have provided conditions under which $G-$causality is equivalent to a form of direct causality arising naturally in dynamic structural systems, defined in the context of *settable systems*. The settable systems framework, introduced by White, H. and K. Chalak (2009, "WC"), extends and refines the PCM to accommodate optimization, equilibrium, and learning. In this paper, we explore the relations between direct structural causality in the settable systems framework and notions of direct causality in the PCM for both recursive and non-recursive systems. The close correspondence between these concepts in the recursive systems relevant to $G-$causality then enables us to show that there is in fact a close linkage between $G-$causality and PCM notions of direct causality. This enables us to provide straightforward practical methods to test for direct causality using tests for Granger causality.

In a related paper, Eichler, M. and V. Didelez (2009) also study the relation between $G-$causality and interventional notions of causality. They give conditions under which Granger non-causality implies that an intervention has no effect. In particular, Eichler, M. and V. Didelez (2009) use graphical representations as in Eichler, M. (2007) of given $G-$causality relations satisfying the "global Granger causal Markov property" to provide graphical conditions for the identification of effects of interventions in "stable" systems. Here, we pursue a different route for studying the interrelations between $G-$causality and interventional notions of causality. Specifically, we see that $G-$causality and certain settable systems notions of direct causality based on functional dependence are *equivalent* under a conditional form of exogeneity. Our conditions are alternative to "stability" and the "global Granger causal Markov property," although particular aspects of our conditions have a similar flavor.

As a referee notes, the present work also provides a rigorous complement, in discrete time, to work by other authors in this volume (for example Roebroeck, A., Seth, A.K., and Valdes-Sosa, P., 2011) on combining structural and dynamic concepts of causality.

The plan of the paper is as follows. In Section 2, we briefly review the PCM. In Section 3, we motivate settable systems by discussing certain limitations of the PCM using a series of examples involving optimization, equilibrium, and learning. We then specify a formal version of settable systems that readily accommodates the challenges to causal discourse presented by the examples of Section 3. In Section 4, we define direct structural causality for settable systems and relate this to corresponding notions in the PCM. The close correspondence between these concepts in recursive systems establishes the first step in linking $G-$causality and the PCM. In Section 5, we discuss how the results of WL complete the chain by linking direct structural causality and $G-$causality. This also involves a conditional form of exogeneity. Section 6 constructs convenient practical tests for structural causality based on proposals of WL, using tests for $G-$causality and conditional exogeneity. Section 7 contains a summary and concluding remarks.

2. Pearl's Causal Model

Pearl's definition of a *causal model* (Pearl, J., 2000, def. 7.1.1, p. 203) provides a formal statement of elements supporting causal reasoning. The PCM is a triple $M := (u, v, f)$, where $u := \{u_1, \ldots, u_m\}$ contains "background" variables determined outside the model, $v := \{v_1, \ldots, v_n\}$ contains "endogenous" variables determined within the model, and $f := \{f_1, \ldots, f_n\}$ contains "structural" functions specifying how each endogenous variable is determined by the other variables of the model, so that $v_i = f_i(v_{(i)}, u)$, $i = 1, \ldots, n$. Here, $v_{(i)}$ is the vector containing every element of v but v_i. The integers m and n are finite. The elements of u and v are system "units."

Finally, the PCM requires that for each u, f yields a unique fixed point. Thus, there must be a unique collection $g := \{g_1, \ldots, g_n\}$ such that for each u,

$$v_i = g_i(u) = f_i(g_{(i)}(u), u), \quad i = 1, \ldots, n. \tag{1}$$

The unique fixed point requirement is crucial to the PCM, as this is necessary for defining the *potential response function* (Pearl, J., 2000, def. 7.1.4). This provides the foundation for discourse about causal relations between endogenous variables; without the potential response function, causal discourse is not possible in the PCM. A variant of the PCM (Halpern, J., 2000) does not require a fixed point, but if any exist, there may be multiple collections of functions g yielding a fixed point. We call this a Generalized Pearl Causal Model (GPCM). As GPCMs also do not possess an analog of the potential response function in the absence of a unique fixed point, causal discourse in the GPCM is similarly restricted.

In presenting the PCM, we have adapted Pearl's notation somewhat to facilitate subsequent discussion, but all essential elements are present and complete.

Pearl, J. (2000) gives numerous examples for which the PCM is ideally suited for supporting causal discourse. As a simple game-theoretic example, consider a market in which there are exactly two firms producing similar but not identical products (e.g., Coke and Pepsi in the cola soft-drink market). Price determination in this market is a two-player game known as "Bertrand duopoly."

In deciding its price, each firm maximizes its profit, taking into account the prevailing cost and demand conditions it faces, as well as the price of its rival. A simple system representing price determination in this market is

$$\begin{aligned} p_1 &= a_1 + b_1 p_2 \\ p_2 &= a_2 + b_2 p_1. \end{aligned}$$

Here, p_1 and p_2 represent the prices chosen by firms 1 and 2 respectively, and a_1, b_1, a_2, and b_2 embody the prevailing cost and demand conditions.

We see that this maps directly to the PCM with $n = 2$, endogenous variables $v = (p_1, p_2)$, background variables $u = (a_1, b_1, a_2, b_2)$, and structural functions

$$\begin{aligned} f_1(v_2, u) &= a_1 + b_1\, p_2 \\ f_2(v_1, u) &= a_2 + b_2\, p_1. \end{aligned}$$

These functions are the Bertrand "best response" or "reaction" functions.

Further, provided $b_1 b_2 \neq 1$, this system has a unique fixed point,

$$\begin{aligned} p_1 &= g_1(u) = (a_1 + b_1 a_2)/(1 - b_1 b_2) \\ p_2 &= g_2(u) = (a_2 + b_2 a_1)/(1 - b_1 b_2). \end{aligned}$$

This fixed point represents the Nash equilibrium for this two-player game.

Clearly, the PCM applies perfectly, supporting causal discourse for this Bertrand duopoly game. Specifically, we see that p_1 causes p_2 and vice-versa, and that the effect of p_2 on p_1 is b_1, whereas that of p_1 on p_2 is b_2.

In fact, the PCM applies directly to a wide variety of games, provided that the game has a unique equilibrium. But there are many important cases where there may be no equilibrium or multiple equilibria. This limits the applicability of the PCM. We explore examples of this below, as well as other features of the PCM that limit its applicability.

3. Settable Systems

3.1. Why Settable Systems?

WC motivate the development of the settable system (SS) framework as an extension of the PCM that accommodates optimization, equilibrium, and learning, which are central features of the explanatory structures of interest in economics. But these features are of interest more broadly, especially in machine learning, as optimization corresponds to any intelligent or rational behavior, whether artificial or natural; equilibrium (e.g., Nash equilibrium) or transitions toward equilibrium characterize stable interactions between multiple interacting systems; and learning corresponds to adaptation and evolution within and between interacting systems. Given the prevalence of these features in natural and artificial systems, it is clearly desirable to provide means for explicit and rigorous causal discourse relating to systems with these features.

To see why an extension of the PCM is needed to handle optimization, equilibrium, and learning, we consider a series of examples that highlight certain limiting features of the PCM: (*i*) in the absence of a unique fixed point, causal discourse is undefined; (*ii*) background variables play no causal role; (*iii*) the role of attributes is restricted; and (*iv*) only a finite rather than a countable number of units is permitted. WC discuss further relevant aspects of the PCM, but these suffice for present purposes.

Example 3.1 (Equilibria in Game Theory) Our first example concerns general two-player games, extending the discussion that we began above in considering Bertrand duopoly.

Let two players $i = 1,2$ have strategy sets S_i and utility functions u_i, such that $\pi_i = u_i(z_1, z_2)$ gives player i's payoff when player 1 plays $z_1 \in S_1$ and player 2 plays $z_2 \in S_2$. Each player solves the optimization problem

$$\max_{z_i \in S_i} u_i(z_1, z_2).$$

The solution to this problem, when it exists, is player i's *best response*, denoted

$$y_i = r_i^e(z_{(i)}; \mathbf{a}),$$

where r_i^e is player i's *best response function* (the superscript "e" stands for "elementary," conforming to notation formally introduced below); $z_{(i)}$ denotes the strategy played by the player other than i; and $\mathbf{a} := (S_1, u_1, S_2, u_2)$ denotes given attributes defining the game. For simplicity here, we focus on "pure strategy" games; see Gibbons, R. (1992) for an accessible introduction to game theory.

Different configurations for **a** correspond to different games. For example, one of the most widely known games is *prisoner's dilemma*, where two suspects in a crime are separated and offered a deal: if one confesses and the other does not, the confessor is released and the other goes to jail. If both confess, both receive a mild punishment. If neither confesses, both are released. The strategies are whether to confess or not. Each player's utility is determined by both players' strategies and the punishment structure.

Another well known game is *hide and seek*. Here, player 1 wins by matching player 2's strategy and player 2 wins by mismatching player 1's strategy. A familiar example is a penalty kick in soccer: the goalie wins by matching the direction (right or left) of the kicker's kick; the kicker wins by mismatching the direction of the goalie's lunge. The same structure applies to baseball (hitter vs. pitcher) or troop deployment in battle (aggressor vs. defender).

A third famous game is *battle of the sexes*. In this game, Ralph and Alice are trying to decide how to spend their weekly night out. Alice prefers the opera, and Ralph prefers boxing; but both would rather be together than apart.

Now consider whether the PCM permits causal discourse in these games, e.g., about the effect of one player's action on that of the other. We begin by mapping the elements of the game to the elements of the PCM. First, we see that **a** corresponds to PCM background variables u, as these are specified outside the system. The variables determined within the system, i.e., the PCM endogenous variables are $z := (z_1, z_2)$ corresponding to v, provided that (for now) we drop the distinction between y_i and z_i. Finally, we see that the best response functions r_i^e correspond to the PCM structural functions f_i.

To determine whether the PCM permits causal discourse in these games, we can check whether there is a unique fixed point for the best responses. In prisoner's dilemma, there is indeed a unique fixed point (both confess), provided the punishments are suitably chosen. The PCM therefore applies to this game to support causal discourse. But there is no fixed point for hide and seek, so the PCM cannot support causal discourse there. On the other hand, there are two fixed points for battle of the sexes: both Ralph and Alice choose opera or both choose boxing. The PCM does not support causal discourse there either. Nor does the GPCM apply to the latter games, because even though it does not require a unique fixed point, the potential response functions required for causal discourse are not defined.

The importance of game theory generally in describing the outcomes of interactions of goal-seeking agents and the fact that the unique fixed point requirement prohibits the PCM from supporting causal discourse in important cases strongly motivates formulating a causal framework that drops this requirement. As we discuss below, the SS framework does not require a unique fixed point, and it applies readily to games generally. Moreover, recognizing and enforcing the distinction between y_i (i's best response strategy) and z_i (an arbitrary setting of i's strategy) turns out to be an important component to eliminating this requirement.

Another noteworthy aspect of this example is that $\mathbf{a}$ is a *fixed* list of elements that define the game. Although elements of $\mathbf{a}$ may differ across players, they do not vary for a given player. This distinction should be kept in mind when referring to the elements of $\mathbf{a}$ as background “variables."

Example 3.2 (Optimization in Consumer Demand) The neoclassical theory of consumer demand posits that consumers determine their optimal goods consumption by maximizing utility subject to a budget constraint (see, e.g., Varian, H., 2009). Suppose for simplicity that there are just two goods, say beer and pizza. Then a typical consumer solves the problem

$$\max_{z_1,z_2} \mathcal{U}(z_1,z_2) \quad s.t. \quad m = z_1 + pz_2,$$

where z_1 and z_2 represent quantities consumed of goods 1 (beer) and 2 (pizza) respectively and $\mathcal{U}$ is the utility function that embodies the consumer's preferences for the two goods. For simplicity, let the price of a beer be \$1, and let p represent the price of pizza; m represents funds available for expenditure, “income" for short[1]. The budget constraint $m = z_1 + pz_2$ ensures that total expenditure on beer and pizza does not exceed income (no borrowing) and also that total expenditure on beer and pizza is not less than m. (As long as utility is increasing in consumption of the goods, it is never optimal to expend less than the funds available.)

Solving the consumer's demand problem leads to the optimal consumer demands for beer and pizza, y_1 and y_2. It is easy to show that these can be represented as

$$y_1 = r_1^a(p,m;\mathbf{a}) \quad \text{and} \quad y_2 = r_2^a(p,m;\mathbf{a}),$$

where r_1^a and r_2^a are known as the consumer's *market demand functions* for beer and pizza. The “a" superscript stands for “agent," corresponding to notation formally introduced below. The attributes $\mathbf{a}$ include the consumer's utility function $\mathcal{U}$ (preferences) and the admissible values for z_1, z_2, p, and m, e.g., $\mathbb{R}^+ := [0,\infty)$.

Now consider how this problem maps to the PCM. First, we see that $\mathbf{a}$ and (p,m) correspond to the background variables u, as these are not determined within the system. Next, we see that $y := (y_1,y_2)$ corresponds to PCM endogenous variables v. Finally,

[1] Since a beer costs a dollar, it is the “numeraire," implying that income is measured in units of beer. This is a convenient convention ensuring that we only need to keep track of the price *ratio* between pizza and beer, p, rather than their two separate prices.

we see that the consumer demand functions r_i^a correspond to the PCM structural functions f_i. Also, because the demand for beer, y_1, does not enter the demand function for pizza, r_2^a, and vice versa, there is a unique fixed point for this system of equations. Thus, the PCM supports causal discourse in this system.

Nevertheless, this system is one where, in the PCM, the causal discourse natural to economists is unavailable. Specifically, economists find it natural to refer to “price effects" and “income effects" on demand, implicitly or explicitly viewing price p and income m as causal drivers of demand. For example, the pizza demand price effect is $(\partial/\partial p)r_2^a(p,m;\mathbf{a})$. This represents how much optimal pizza consumption (demand) will change as a result of a small (marginal) increase in the price of pizza. Similarly, the pizza demand income effect is $(\partial/\partial m)r_2^a(p,m;\mathbf{a})$, representing how much optimal pizza consumption will change as a result of a small increase in income. But in the PCM, causal discourse is reserved only for endogenous variables y_1 and y_2. The fact that background variables p and m do not have causal status prohibits speaking about their effects.

Observe that the “endogenous" status of y and “exogenous" status of p and m is determined in SS by utility maximization, the “governing principle" here. In contrast, there is no formal mechanism in the PCM that permits making these distinctions. Although causal discourse in the PCM can be rescued for such systems by “endogenizing" p and m, that is, by positing additional structure that explains the genesis of p and m in terms of further background variables, this is unduly cumbersome. It is much more natural simply to permit p and m to have causal status from the outset, so that price and income effects are immediately meaningful, without having to specify their determining processes. The SS framework embodies this direct approach. Those familiar with theories of price and income determination will appreciate the considerable complications avoided in this way. The same simplifications occur with respect to the primitive variables appearing in any responses determined by optimizing behavior.

Also noteworthy here is the important distinction between **a,** which represents fixed attributes of the system, and p and m, which are true variables that can each take a range of different possible values. As WC (p.1774) note, restricting the role of attributes by “lumping together" attributes and structurally exogenous variables as background objects without causal status creates difficulties for causal discourse in the PCM:

> [this] misses the opportunity to make an important distinction between invariant aspects of the system units on the one hand and counterfactual variation admissible for the system unit values on the other. Among other things, assigning attributes to u interferes with assigning natural causal roles to structurally exogenous variables.

By distinguishing between attributes and structurally exogenous variables, settable systems permit causal status for variables determined outside a given system, such as when price and income drive consumer demand.

Example 3.3 (Learning in Structural Vector Autoregressions) Structural vector autoregressions (VARs) are widely used to analyze time-series data. For example, consider the structural VAR

$$\begin{aligned} y_{1,t} &= a_{11}y_{1,t-1} + a_{12}y_{2,t-1} + u_{1,t} \\ y_{2,t} &= a_{21}y_{1,t-1} + a_{22}y_{2,t-1} + u_{2,t}, \quad t = 1,2,\ldots, \end{aligned}$$

where $y_{1,0}$ and $y_{2,0}$ are given scalars, $\mathbf{a} := (a_{11}, a_{12}, a_{21}, a_{22})'$ is a given real "coefficient" vector, and $\{u_t := (u_{1,t}, u_{2,t}) : t = 1,2,\ldots\}$ is a given sequence. This system describes the evolution of $\{y_t := (y_{1,t}, y_{2,t}) : t = 1,2,\ldots\}$ through time.

Now consider how this maps to the PCM. We see that $y_0 := (y_{1,0}, y_{2,0})$, $\{u_t\}$, and $\mathbf{a}$ correspond to the PCM background variables u, as these are not determined within the system. Further, we see that the sequence $\{y_t\}$ corresponds to the endogenous variables v, and that the PCM structural functions f_i correspond to

$$\begin{aligned} r_{1,t}(y^{t-1}, u^t; \mathbf{a}) &= a_{11}y_{1,t-1} + a_{12}y_{2,t-1} + u_{1,t} \\ r_{2,t}(y^{t-1}, u^t; \mathbf{a}) &= a_{21}y_{1,t-1} + a_{22}y_{2,t-1} + u_{2,t}, \quad t = 1,2,\ldots, \end{aligned}$$

where $y^{t-1} := (y_0, \ldots, y_{t-1})$ and $u^t := (u_1, \ldots, u_t)$ represent finite "histories" of the indicated variables. We also see that this system is recursive, and therefore has a unique fixed point.

The challenge to the PCM here is that it permits only a finite rather than a countable number of units: both the number of background variables (m) and endogenous variables (n) must be finite in the PCM, whereas the structural VAR requires a countable infinity of background and endogenous variables. In contrast, settable systems permit (but do not require) a countable infinity of units, readily accommodating structural VARs.

In line with our previous discussion, settable systems distinguish between system attributes $\mathbf{a}$ (a fixed vector) and structurally exogenous causal variables y_0 and $\{u_t\}$. The difference in the roles of y_0 and $\{u_t\}$ on the one hand and $\mathbf{a}$ on the other are particularly clear in this example. In the PCM, these are lumped together as background variables devoid of causal status. Since $\mathbf{a}$ is fixed, its lack of causal status is appropriate; indeed, $\mathbf{a}$ represents *effects* here[2], not causes. But the lack of causal status is problematic for the variables y_0 and $\{u_t\}$; for example, this prohibits discussing the effects of structural "shocks" u_t.

Observe that the structural VAR represents $u_{1,t}$ as a causal driver of $y_{1,t}$, as is standard. Nevertheless, settable systems do not admit "instantaneous" causation, so even though $u_{1,t}$ has the same time index as $y_{1,t}$, i.e. t, we adopt the convention that $u_{1,t}$ is realized prior to $y_{1,t}$. That is, there must be some positive time interval $\delta > 0$, no matter how small, separating these realizations. For example, δ can represent the amount of time it takes to compute $y_{1,t}$ once all its determinants are in place. Strictly speaking, then, we could write $u_{1,t-\delta}$ in place of $u_{1,t}$, but for notational convenience, we leave this implicit. We refer to this as "contemporaneous" causation to distinguish it from instantaneous causation.

[2] For example, $(\partial/\partial y_{1,t-1}) r_{1,t}(y^{t-1}, e^t; a) = a_{11}$ can be interpreted as the marginal effect of $y_{1,t-1}$ on $y_{1,t}$.

A common focus of interest when applying structural VARs is to learn the coefficient vector $\mathbf{a}$. In applications, it is typically assumed that the realizations $\{y_t\}$ are observed, whereas $\{u_t\}$ is unobserved. The least squares estimator for a sample of size T, say $\hat{\mathbf{a}}_T$, is commonly used to learn (estimate) $\mathbf{a}$ in such cases. This estimator is a straightforward function of y^T, say $\hat{\mathbf{a}}_T = r_{\mathbf{a},T}(y^T)$. If $\{u_t\}$ is generated as a realization of a sequence of mean zero finite variance independent identically distributed (IID) random variables, then $\hat{\mathbf{a}}_T$ generally converges to $\mathbf{a}$ with probability one as $T \to \infty$, implying that $\mathbf{a}$ can be fully learned in the limit. Viewing $\hat{\mathbf{a}}_T$ as causally determined by y^T, we see that we require a countable number of units to treat this learning problem.

As these examples demonstrate, the PCM exhibits a number of features that limit its applicability to systems involving optimization, equilibrium, and learning. These limitations motivate a variety of features of settable systems, extending the PCM in ways that permit straightforward treatment of such systems. We now turn to a more complete description of the SS framework.

3.2. Formal Settable Systems

We now provide a formal description of settable systems that readily accommodates causal discourse in the foregoing examples and that also suffices to establish the desired linkage between Granger causality and causal notions in the PCM. The material that follows is adapted from Chalak, K. and H. White (2010). For additional details, see WC.

A *stochastic settable system* is a mathematical framework in which a countable number of *units* i, $i = 1, \dots, n$, interact under uncertainty. Here, $n \in \bar{\mathbb{N}}^+ := \mathbb{N}^+ \cup \{\infty\}$, where $\mathbb{N}^+$ denotes the positive integers. When $n = \infty$, we interpret $i = 1, \dots, n$ as $i = 1, 2, \dots$. Units have *attributes* $a_i \in A$; these are fixed for each unit, but may vary across units. Each unit also has associated random variables, defined on a measurable space $(\Omega, \mathcal{F})$. It is convenient to define a *principal space* Ω_0 and let $\Omega := \times_{i=0}^n \Omega_i$, with each Ω_i a copy of Ω_0. Often, $\Omega_0 = \mathbb{R}$ is convenient. A probability measure $P_{\mathbf{a}}$ on $(\Omega, \mathcal{F})$ assigns probabilities to events involving random variables. As the notation suggests, $P_{\mathbf{a}}$ can depend on the attribute vector $\mathbf{a} := (a_1, \dots, a_n) \in \mathbf{A} := \times_{i=1}^n A$.

The random variables associated with unit i define a *settable variable* $\mathcal{X}_i$ for that unit. A settable variable $\mathcal{X}_i$ has a dual aspect. It can be *set* to a random variable denoted by Z_i (the *setting*), where $Z_i : \Omega_i \to \mathbb{S}_i$. $\mathbb{S}_i$ denotes the *admissible setting values* for Z_i, a multi-element subset of $\mathbb{R}$. Alternatively, the settable variable can be *free* to respond to settings of other settable variables. In the latter case, it is denoted by the *response* $Y_i : \Omega \to \mathbb{S}_i$. The response Y_i of a settable variable $\mathcal{X}_i$ to the settings of other settable variables is determined by a *response function*, r_i. For example, r_i can be determined by optimization, determining the response for unit i that is best in some sense, given the settings of other settable variables. The dual role of a settable variable $\mathcal{X}_i : \{0,1\} \times \Omega \to \mathbb{S}_i$, distinguishing responses $\mathcal{X}_i(0, \omega) := Y_i(\omega)$ and settings $\mathcal{X}_i(1, \omega) := Z_i(\omega_i)$, $\omega \in \Omega$, permits formalizing the directional nature of causal relations, whereby settings of some variables (causes) determine responses of others.

The *principal unit* $i = 0$ also plays a key role. We let the *principal setting* Z_0 and *principal response* Y_0 of the *principal settable variable* $\mathcal{X}_0$ be such that $Z_0 : \Omega_0 \to \Omega_0$ is the identity map, $Z_0(\omega_0) := \omega_0$, and we define $Y_0(\omega) := Z_0(\omega_0)$. The setting Z_0 of the principal settable variable may directly influence all other responses in the system, whereas its response Y_0 is unaffected by other settings. Thus, $\mathcal{X}_0$ supports introducing an aspect of "pure randomness" to responses of settable variables.

3.2.1. ELEMENTARY SETTABLE SYSTEMS

In *elementary* settable systems, Y_i is determined (actually or potentially) by the settings of *all* other system variables, denoted $Z_{(i)}$. Thus, in elementary settable systems, $Y_i = r_i(Z_{(i)}; \mathbf{a})$. The settings $Z_{(i)}$ take values in $\mathbb{S}_{(i)} \subseteq \Omega_0 \times_{j \neq i} \mathbb{S}_j$. We have that $\mathbb{S}_{(i)}$ is a strict subset of $\Omega_0 \times_{j \neq i} \mathbb{S}_j$ if there are joint restrictions on the admissible settings values, for example, when certain elements of $\mathbb{S}_{(i)}$ represent probabilities that sum to one.

We now give a formal definition of elementary settable systems.

Definition 3.1 (Elementary Settable System) *Let* $\mathbf{A}$ *be a set and let* **attributes** $\mathbf{a} \in \mathbf{A}$ *be given. Let* $n \in \bar{\mathbb{N}}^+$ *be given, and let* $(\Omega, \mathcal{F}, P_\mathbf{a})$ *be a complete probability space such that* $\Omega := \times_{i=0}^n \Omega_i$*, with each* Ω_i *a copy of the* **principal space** Ω_0*, containing at least two elements.*

Let the **principal setting** $Z_0 : \Omega_0 \to \Omega_0$ *be the identity mapping. For* $i = 1, 2, \ldots, n$*, let* $\mathbb{S}_i$ *be a multi-element Borel-measurable subset of* $\mathbb{R}$ *and let* **settings** $Z_i : \Omega_i \to \mathbb{S}_i$ *be surjective measurable functions. Let* $Z_{(i)}$ *be the vector including every setting except* Z_i *and taking values in* $\mathbb{S}_{(i)} \subseteq \Omega_0 \times_{j \neq i} \mathbb{S}_j$, $\mathbb{S}_{(i)} \neq \varnothing$. *Let* **response functions** $r_i(\,\cdot\,; \mathbf{a}) : \mathbb{S}_{(i)} \to \mathbb{S}_i$ *be measurable functions and define* **responses** $Y_i(\omega) := r_i(Z_{(i)}(\omega); \mathbf{a})$. *Define* **settable variables** $\mathcal{X}_i : \{0,1\} \times \Omega \to \mathbb{S}_i$ *as*

$$\mathcal{X}_i(0,\omega) := Y_i(\omega) \quad \textit{and} \quad \mathcal{X}_i(1,\omega) := Z_i(\omega_i), \quad \omega \in \Omega.$$

Define Y_0 *and* $\mathcal{X}_0$ *by* $Y_0(\omega) := \mathcal{X}_0(0,\omega) := \mathcal{X}_0(1,\omega) := Z_0(\omega_0)$, $\omega \in \Omega$.

Put $\mathcal{X} := \{\mathcal{X}_0, \mathcal{X}_1, \ldots\}$. *The triple* $\mathcal{S} := \{(\mathbf{A}, \mathbf{a}), (\Omega, \mathcal{F}, P_\mathbf{a}), \mathcal{X}\}$ *is an* **elementary settable system**.

An elementary settable system thus comprises an attribute component, $(\mathbf{A}, \mathbf{a})$, a stochastic component, $(\Omega, \mathcal{F}, P_\mathbf{a})$, and a structural or causal component $\mathcal{X}$, consisting of settable variables whose properties are crucially determined by response functions $r := \{r_i\}$. It is formally correct to write $\mathcal{X}_\mathbf{a}$ instead of $\mathcal{X}$; we write $\mathcal{X}$ for simplicity.

Note the absence of any fixed point requirement, the distinct roles played by fixed attributes $\mathbf{a}$ and setting variables Z_i (including principal settings Z_0), and the countable number of units allowed.

Example 3.1 is covered by this definition. There, $n = 2$. Attributes $\mathbf{a} := (S_1, u_1, S_2, u_2)$ belong to a suitably chosen set $\mathbf{A}$. Here, $\mathbb{S}_i = S_i$. We take $z_i = Z_i(\omega_i)$, $\omega_i \in \Omega_i$ and $y_i = Y_i(\omega) = r_i^e(Z_{(i)}(\omega); \mathbf{a}) = r_i^e(z_{(i)}; \mathbf{a})$, $i = 1, 2$. The "e" superscript in r_i^e emphasizes that the response function is for an elementary settable system. In the example games, the responses y_i only depend on settings (z_1, z_2). In more elaborate games, dependence on $z_0 = \omega_0$ can accommodate random responses.

3.2.2. PARTITIONED SETTABLE SYSTEMS

In elementary settable systems, each single response Y_i can freely respond to settings of all other system variables. We now consider systems where several settable variables jointly respond to settings of the remaining settable variables, as when responses represent the solution to a joint optimization problem. For this, *partitioned* settable systems group jointly responding variables into blocks. In elementary settable systems, every unit i forms a block by itself. We now define general partitioned settable systems.

Definition 3.2 (Partitioned Settable System) *Let* $(\mathbf{A},\mathbf{a}),(\Omega,\mathcal{F},P_{\mathbf{a}})$, $\mathcal{X}_0$, n, *and* $\mathbb{S}_i$, $i = 1,\ldots,n$, *be as in Definition 3.1. Let* $\Pi = \{\Pi_b\}$ *be a partition of* $\{1,\ldots,n\}$, *with cardinality* $B \in \bar{\mathbb{N}}^+$ $(B := \#\Pi)$.

For $i = 1,2,\ldots,n$, *let* Z_i^Π *be settings and let* $Z_{(b)}^\Pi$ *be the vector containing* Z_0 *and* $Z_i^\Pi, i \notin \Pi_b$, *and taking values in* $\mathbb{S}_{(b)}^\Pi \subseteq \Omega_0 \times_{i \notin \Pi_b} \mathbb{S}_i$, $\mathbb{S}_{(b)}^\Pi \neq \varnothing$, $b = 1,\ldots,B$. *For* $b = 1,\ldots,B$ *and* $i \in \Pi_b$, *suppose there exist measurable functions* $r_i^\Pi(\,\cdot\,;\mathbf{a}) : \mathbb{S}_{(b)}^\Pi \to \mathbb{S}_i$, *specific to* Π *such that responses* $Y_i^\Pi(\omega)$ *are jointly determined as*

$$Y_i^\Pi := r_i^\Pi(Z_{(b)}^\Pi;\mathbf{a}).$$

Define the settable variables $\mathcal{X}_i^\Pi : \{0,1\} \times \Omega \to \mathbb{S}_i$ as

$$\mathcal{X}_i^\Pi(0,\omega) := Y_i^\Pi(\omega) \quad \textit{and} \quad \mathcal{X}_i^\Pi(1,\omega) := Z_i^\Pi(\omega_i) \quad \omega \in \Omega.$$

Put $\mathcal{X}^\Pi := \{\mathcal{X}_0, \mathcal{X}_1^\Pi, \mathcal{X}_2^\Pi \ldots\}$. *The triple* $\mathcal{S} := \{(\mathbf{A},\mathbf{a}),(\Omega,\mathcal{F}),(\Pi,\mathcal{X}^\Pi)\}$ *is a* **partitioned settable system**.

The settings $Z_{(b)}^\Pi$ may be partition-specific; this is especially relevant when the admissible set $\mathbb{S}_{(b)}^\Pi$ imposes restrictions on the admissible values of $Z_{(b)}^\Pi$. Crucially, response functions and responses are partition-specific. In Definition 3.2, the joint response function $r_{[b]}^\Pi := (r_i^\Pi, i \in \Pi_b)$ specifies how the settings $Z_{(b)}^\Pi$ outside of block Π_b determine the joint response $Y_{[b]}^\Pi := (Y_i^\Pi, i \in \Pi_b)$, i.e., $Y_{[b]}^\Pi = r_{[b]}^\Pi(Z_{(b)}^\Pi;\mathbf{a})$. For convenience below, we let $\Pi_0 = \{0\}$ represent the block corresponding to $\mathcal{X}_0$.

Example 3.2 makes use of partitioning. Here, we have $n = 4$ settable variables with $B = 2$ blocks. Let settable variables 1 and 2 correspond to beer and pizza consumption, respectively, and let settable variables 3 and 4 correspond to price and income. The *agent* partition groups together all variables under the control of a given agent. Let the consumer be agent 2, so $\Pi_2 = \{1,2\}$. Let the rest of the economy, determining price and income, be agent 1, so $\Pi_1 = \{3,4\}$. The agent partition is $\Pi^a = \{\Pi_1,\Pi_2\}$. Then for block 2,

$$\begin{aligned} y_1 &= Y_1^a(\omega) = r_1^a(Z_0(\omega_0), Z_3^a(\omega_3), Z_4^a(\omega_4);\mathbf{a}) = r_1^a(p,m;\mathbf{a}) \\ y_2 &= Y_2^a(\omega) = r_2^a(Z_0(\omega_0), Z_3^a(\omega_3), Z_4^a(\omega_4);\mathbf{a}) = r_2^a(p,m;\mathbf{a}) \end{aligned}$$

represents the joint demand for beer and pizza (belonging to block 2) as a function of settings of price and income (belonging to block 1). This joint demand is unique under mild conditions. Observe that $z_0 = Z_0(\omega_0)$ formally appears as an allowed argument of r_i^a after the second equality, but when the consumer's optimization problem has a unique solution, there is no need for a random component to demand. We thus suppress this argument in writing $r_i^a(p,m;\mathbf{a})$, $i = 1,2$. Nevertheless, when the solution to the consumer's optimization problem is not unique, a random component can act to ensure a unique consumer demand. We do not pursue this here; WC provide related discussion.

We write the block 1 responses for the price and income settable variables as

$$\begin{aligned} y_3 &= Y_3^a(\omega) = r_3^a(Z_0(\omega_0), Z_1^a(\omega_1), Z_2^a(\omega_2); \mathbf{a}) = r_3^a(z_0; \mathbf{a}) \\ y_4 &= Y_4^a(\omega) = r_4^a(Z_0(\omega_0), Z_1^a(\omega_1), Z_2^a(\omega_2); \mathbf{a}) = r_4^a(z_0; \mathbf{a}). \end{aligned}$$

In this example, price and income are not determined by the individual consumer's demands, so although $Z_1^a(\omega_1)$ and $Z_2^a(\omega_2)$ formally appear as allowed arguments of r_i^a after the second equality, we suppress these in writing $r_i^a(z_0;\mathbf{a})$, $i = 3,4$. Here, price and income responses (belonging to block 1) are determined solely by block 0 settings $z_0 = Z_0(\omega_0) = \omega_0$. This permits price and income responses to be randomly distributed, under the control of $P_{\mathbf{a}}$.

It is especially instructive to consider the elementary partition for this example, $\Pi^e = \{\{1\},\{2\},\{3\},\{4\}\}$, so that $\Pi_i = \{i\}$, $i = 1,\dots,4$. The elementary partition specifies how each system variable freely responds to settings of all other system variables. In particular, it is easy to verify that when consumption of pizza is set to a given level, the consumer's optimal response is to spend whatever income is left on beer, and vice versa. Thus,

$$\begin{aligned} y_1 &= r_1^e(Z_0(\omega_0), Z_2^e(\omega_2), Z_3^e(\omega_3), Z_4^e(\omega_4); \mathbf{a}) = r_1^e(z_2, p, m; \mathbf{a}) = m - pz_2 \\ y_2 &= r_2^e(Z_0(\omega_0), Z_1^e(\omega_2), Z_3^e(\omega_3), Z_4^e(\omega_4); \mathbf{a}) = r_2^e(z_1, p, m; \mathbf{a}) = (m - z_1)/p. \end{aligned}$$

Replacing (y_1, y_2) with (z_1, z_2), we see that this system does not have a unique fixed point, as any (z_1, z_2) such that $m = z_1 + pz_2$ satisfies both

$$z_1 = m - pz_2 \quad \text{and} \quad z_2 = (m - z_1)/p.$$

Causal discourse in the PCM is ruled out by the lack of a fixed point. Nevertheless, the settable systems framework supports the natural economic causal discourse here about effects of prices, income, and, e.g., pizza consumption on beer demand. Further, in settable systems, the governing principle of optimization (embedded in $\mathbf{a}$) ensures that the response functions for both the agent partition and the elementary partition are mutually consistent.

3.2.3. Recursive and Canonical Settable Systems

The link between Granger causality and the causal notions of the PCM emerges from a particular class of *recursive* partitioned settable systems that we call *canonical* settable systems, where the system evolves naturally without intervention. This corresponds to what are also called "idle regimes" in the literature (Pearl, J., 2000; Eichler, M. and V. Didelez, 2009; Dawid, 2010).

To define recursive settable systems, for $b \geq 0$ define $\Pi_{[0:b]} := \Pi_0 \cup \ldots \cup \Pi_{b-1} \cup \Pi_b$.

Definition 3.3 (Recursive Partitioned Settable System) *Let $\mathcal{S}$ be a partitioned settable system. For $b = 0, 1, \ldots, B$, let $Z^{\Pi}_{[0:b]}$ denote the vector containing the settings Z^{Π}_i for $i \in \Pi_{[0:b]}$ and taking values in $\mathbb{S}_{[0:b]} \subseteq \Omega_0 \times_{i \in \Pi_{[1:b]}} \mathbb{S}_i$, $\mathbb{S}_{[0:b]} \neq \emptyset$. For $b = 1, \ldots, B$ and $i \in \Pi_b$, suppose that $r^{\Pi} := \{r^{\Pi}_i\}$ is such that the responses $Y^{\Pi}_i = \mathcal{X}^{\Pi}_i(1, \cdot)$ are determined as*

$$Y^{\Pi}_i := r^{\Pi}_i(Z^{\Pi}_{[0:b-1]}; \mathbf{a}).$$

Then we say that Π is a **recursive partition**, *that r^{Π} is* **recursive**, *and that $\mathcal{S} := \{(\mathbf{A}, \mathbf{a}), (\Omega, \mathcal{F}), (\Pi, \mathcal{X}^{\Pi})\}$ is a* **recursive partitioned settable system** *or simply that $\mathcal{S}$ is* **recursive**.

Example 3.2 is a recursive settable system, as the responses of block 1 depend on the settings of block 0, and the responses of block 2 depend on the settings of block 1.

Canonical settable systems are recursive settable systems in which the settings for a given block equal the responses for that block, i.e.,

$$Z^{\Pi}_{[b]} = Y^{\Pi}_{[b]} := r^{\Pi}_{[b]}(Z^{\Pi}_{[0:b-1]}; \mathbf{a}), \qquad b = 1, \ldots, B.$$

Without loss of generality, we can represent canonical responses and settings solely as a function of ω_0, so that

$$Z^{\Pi}_{[b]}(\omega_0) = Y^{\Pi}_{[b]}(\omega_0) := r^{\Pi}_{[b]}(Z^{\Pi}_{[0:b-1]}(\omega_0); \mathbf{a}), \qquad b = 1, \ldots, B.$$

The *canonical representation* drops the distinction between settings and responses; we write

$$Y^{\Pi}_{[b]} = r^{\Pi}_{[b]}(Y^{\Pi}_{[0:b-1]}; \mathbf{a}), \qquad b = 1, \ldots, B.$$

It is easy to see that the structural VAR of Example 3.3 corresponds to the canonical representation of a canonical settable system. The canonical responses y_0 and $\{u_t\}$ belong to the first block, and canonical responses $y_t = (y_{1,t}, y_{2,t})$ belong to block $t+1$, $t = 1, 2, \ldots$ Example 3.3 implements the *time* partition, where joint responses for a given time period depend on previous settings.

4. Causality in Settable Systems and in the PCM

In this section we examine the relations between concepts of direct causality in settable systems and in the PCM, specifically the PCM notions of direct cause and controlled direct effect (Pearl, J. (2000, p. 222); Pearl, J. (2001, definition 1)). The close correspondence between these notions for the recursive systems relevant to Granger causality enables us to take the first step in linking Granger causality and causal notions in the PCM. Section 5 completes the chain by linking direct structural causality and Granger causality.

4.1. Direct Structural Causality in Settable Systems

Direct structural causality is defined for both recursive and non-recursive partitioned settable systems. For notational simplicity in what follows, we may drop the explicit partition superscript Π when the specific partition is clearly understood. Thus, we may write Y, Z, and $\mathcal{X}$ in place of the more explicit Y^{Π}, Z^{Π}, and $\mathcal{X}^{\Pi}$ when there is no possibility of confusion.

Let $\mathcal{X}_j$ belong to block b ($j \in \Pi_b$). Heuristically, we say that a settable variable $\mathcal{X}_i$, outside of block b, *directly causes* $\mathcal{X}_j$ *in* $\mathcal{S}$ when the response for $\mathcal{X}_j$ differs for different settings of $\mathcal{X}_i$, while holding all other variables outside of block b to the same setting values. There are two main ingredients to this notion. The first ingredient is an admissible intervention. To define this, let $z^*_{(b);i}$ denote the vector otherwise identical to $z_{(b)}$, but replacing elements z_i with z^*_i. An *admissible intervention* $z_{(b)} \rightarrow z^*_{(b);i} :=$ $(z_{(b)}, z^*_{(b);i})$ is a pair of distinct elements of $\mathbb{S}_{(b)}$. The second ingredient is the behavior of the response under this intervention.

We formalize this notion of direct causality as follows.

Definition 4.1 (Direct Causality) *Let* $\mathcal{S}$ *be a partitioned settable system. For given positive integer* b, *let* $j \in \Pi_b$. (*i*) *For given* $i \notin \Pi_b$, $\mathcal{X}_i$ **directly causes** $\mathcal{X}_j$ **in** $\mathcal{S}$ *if there exists an admissible intervention* $z_{(b)} \rightarrow z^*_{(b);i}$ *such that*

$$r_j(z^*_{(b);i}; \mathbf{a}) - r_j(z_{(b)}; \mathbf{a}) \neq 0,$$

and we write $\mathcal{X}_i \overset{D}{\Rightarrow}_{\mathcal{S}} \mathcal{X}_j$. *Otherwise, we say* $\mathcal{X}_i$ **does not directly cause** $\mathcal{X}_j$ **in** $\mathcal{S}$ *and write* $\mathcal{X}_i \overset{D}{\not\Rightarrow}_{\mathcal{S}} \mathcal{X}_j$. (*ii*) *For* $i, j \in \Pi_b, \mathcal{X}_i \overset{D}{\not\Rightarrow}_{\mathcal{S}} \mathcal{X}_j$.

We emphasize that although we follow the literature in referring to "interventions," with their mechanistic or manipulative connotations, the formal concept only involves the properties of a response function on its domain.

By definition, variables within the same block do not directly cause each other. In particular $\mathcal{X}_i \overset{D}{\not\Rightarrow}_{\mathcal{S}} \mathcal{X}_i$. Also, Definition 4.1 permits *mutual causality*, so that $\mathcal{X}_i \overset{D}{\Rightarrow}_{\mathcal{S}} \mathcal{X}_j$ and $\mathcal{X}_j \overset{D}{\Rightarrow}_{\mathcal{S}} \mathcal{X}_i$ without contradiction for i and j in different blocks. Nevertheless, in recursive systems, mutual causality is ruled out: if $\mathcal{X}_i \overset{D}{\Rightarrow}_{\mathcal{S}} \mathcal{X}_j$ then $\mathcal{X}_j \overset{D}{\not\Rightarrow}_{\mathcal{S}} \mathcal{X}_i$.

We call the response value difference in Definition 4.1 the *direct effect of* $\mathcal{X}_i$ *on* $\mathcal{X}_j$ *in* $\mathcal{S}$ of the specified intervention. Chalak, K. and H. White (2010) also study various notions of indirect and total causality.

These notions of direct cause and direct effect are well defined regardless of whether or not the system possesses a unique fixed point. Further, all settable variables, including $\mathcal{X}_0$, can act as causes and have effects. On the other hand, attributes $\mathbf{a}$, being fixed, do not play a causal role. These definitions apply regardless of whether there is a finite or countable number of units. It is readily verified that this definition rigorously supports causal discourse in each of the examples of Section 3.

As we discuss next, in the recursive systems relevant for $G-$causality, these concepts correspond closely to notions of direct cause and "controlled" direct effect in

Pearl, J. (2000, 2001). To distinguish the settable system direct causality concept from Pearl's notion and later from Granger causality, we follow WL and refer to direct causality in settable systems as *direct structural causality*.

4.2. Direct Causes and Effects in the PCM

Pearl, J. (2000, p. 222), drawing on Galles, D. and J. Pearl (1997), gives a succinct statement of the notion of direct cause, coherent with the PCM as specified in Section 2:

> *X is a direct cause of Y* if there exist two values x and x' of X and a value u of U such that $Y_{xr}(u) \neq Y_{x'r}(u)$, where r is some realization of $V \backslash \{X, Y\}$.

To make this statement fully meaningful requires applying Pearl's (2000) definitions 7.1.2 (Submodel) and 7.1.4 (Potential Response) to arrive at the *potential response,* $Y_{xr}(u)$. For brevity, we do not reproduce Pearl's definitions here. Instead, it suffices to map $Y_{xr}(u)$ and its elements to their settable system counterparts. Specifically, u corresponds to $(\mathbf{a}, z_0)$; x corresponds to z_i; r corresponds to the elements of $z_{(b)}$ other than z_0 and z_i, say $z_{(b)(i,0)}$; and, provided it exists, $Y_{xr}(u)$ corresponds to $r_j(z_{(b)}; \mathbf{a})$.

The caveat about the existence of $Y_{xr}(u)$ is significant, as $Y_{xr}(u)$ is not defined in the absence of a unique fixed point for the system. Further, even with a unique fixed point, the potential response $Y_{xr}(u)$ must also uniquely solve a set of equations denoted F_x (see Pearl, J., 2000, eq. (7.1)) for a submodel, and there is no general guarantee of such a solution. Fortunately, however, this caveat matters only for non-recursive PCMs. In the recursive case relevant for $G-$causality, the potential response is generally well defined.

Making a final identification between x' and z_i^*, and given the existence of potential responses $Y_{x'r}(u)$ and $Y_{xr}(u)$, we see that $Y_{x'r}(u) \neq Y_{xr}(u)$ corresponds to the settable systems requirement $r_j(z^*_{(b);i}; \mathbf{a}) - r_j(z_{(b)}; \mathbf{a}) \neq 0$.

Pearl, J. (2001, definition 1) gives a formal statement of the notion stated above, saying that if for given u and some r, x, and x' we have $Y_{xr}(u) \neq Y_{x'r}(u)$ then *X has a* ***controlled direct effect*** *on Y in model M and situation U = u*. In definition 2, Pearl, J. (2001) labels $Y_{x'r}(u) - Y_{xr}(u)$ the *controlled direct effect*, corresponding to the direct structural effect $r_j(z^*_{(b);i}; \mathbf{a}) - r_j(z_{(b)}; \mathbf{a})$ defined for settable systems.

Thus, although there are important differences, especially in non-recursive systems, the settable systems and PCM notions of direct causality and direct effects closely correspond in recursive systems. These differences are sufficiently modest that the results of WL linking direct structural causality to Granger causality, discussed next, also serve to closely link the PCM notion of direct cause to that of Granger causality.

5. $G-$Causality and Direct Structural Causality

In this section we examine the relation between direct structural causality and Granger causality, drawing on results of WL. See WL for additional discussion and proofs of all formal results given here and in Section 6.

5.1. Granger Causality

Granger, C.W.J. (1969) defined $G-$causality in terms of conditional expectations. Granger, C.W.J. and P. Newbold (1986) gave a definition using conditional distributions. We work with the latter, as this is what relates generally to structural causality. In what follows, we adapt Granger and Newbold's notation, but otherwise preserve the conceptual content.

For any sequence of random vectors $\{Y_t,\ t = 0,1,\ldots\}$, let $Y^t := (Y_0,\ldots,Y_t)$ denote its "$t-$history," and let $\sigma(Y^t)$ denote the sigma-field ("information set") generated by Y^t. Let $\{Q_t,S_t,Y_t\}$ be a sequence of random vectors. Granger, C.W.J. and P. Newbold (1986) say that Q_{t-1} does not G-cause Y_{t+k} with respect to $\sigma(Q^{t-1},S^{t-1},Y^{t-1})$ if for all $t = 0,1,\ldots,$

$$F_{t+k}(\,\cdot \mid Q^{t-1},S^{t-1},Y^{t-1}) = F_{t+k}(\,\cdot \mid S^{t-1},Y^{t-1}),\quad k = 0,1,\ldots, \tag{2}$$

where $F_{t+k}(\,\cdot \mid Q^{t-1},S^{t-1},Y^{t-1})$ denotes the conditional distribution function of Y_{t+k} given Q^{t-1},S^{t-1},Y^{t-1}, and $F_{t+k}(\,\cdot \mid S^{t-1},Y^{t-1})$ denotes that of Y_{t+k} given S^{t-1},Y^{t-1}. Here, we focus only on the $k = 0$ case, as this is what relates generally to structural causality.

As Florens, J.P. and M. Mouchart (1982) and Florens, J.P. and D. Fougère (1996) note, G non-causality is a form of conditional independence. Following Dawid (1979), we write $X \perp Y \mid Z$ when X and Y are independent given Z. Translating (2) gives the following version of the classical definition of Granger causality:

Definition 5.1 (Granger Causality) *Let $\{Q_t,S_t,Y_t\}$ be a sequence of random vectors. Suppose that*

$$Y_t \perp Q^{t-1} \mid Y^{t-1},S^{t-1} \quad t = 1,2,\ldots\,. \tag{3}$$

Then Q **does not** *$G-$***cause** *Y* **with respect to** *S. Otherwise, Q $G-$***causes** *Y* **with respect to** *S.*

As it stands, this definition has no necessary structural content, as Q_t, S_t, and Y_t can be any random variables whatsoever. This definition relates solely to the ability of Q^{t-1} to help in predicting Y_t given Y^{t-1} and S^{t-1}.

In practice, researchers do not test classical $G-$causality, as this involves data histories of arbitrary length. Instead, researchers test a version of $G-$causality involving only a finite number of lags of Y_t, Q_t, and S_t. This does not test classical $G-$causality, but rather a related property, *finite-order* $G-$causality, that is neither necessary nor sufficient for classical $G-$causality.

Because of its predominant practical relevance, we focus here on finite-order rather than classical $G-$causality. (See WL for discussion of classical $G-$causality.) To define the finite-order concept, we define the finite histories $\boldsymbol{Y}_{t-1} := (Y_{t-\ell},\ldots,Y_{t-1})$ and $\boldsymbol{Q}_t := (Q_{t-k},\ldots,Q_t)$.

Definition 5.2 (Finite-Order Granger Causality) *Let $\{Q_t,S_t,Y_t\}$ be a sequence of random variables, and let $k \geq 0$ and $\ell \geq 1$ be given finite integers. Suppose that*

$$Y_t \perp \boldsymbol{Q}_t \mid \boldsymbol{Y}_{t-1},S_t,\ \ t = 1,2,\ldots\,.$$

Then we say Q **does not finite-order** $G-$**cause** Y **with respect to** S. *Otherwise, we say Q* **finite-order** $G-$**causes** Y **with respect to** S.

We call $\max(k, \ell - 1)$ the "order" of the finite-order G non-causality.

Observe that $\boldsymbol{Q}_t$ replaces Q^{t-1} in the classical definition, that $\boldsymbol{Y}_{t-1}$ replaces Y^{t-1}, and that S_t replaces S^{t-1}. Thus, in addition to dropping all but a finite number of lags in Q^{t-1} and Y^{t-1}, this version includes Q_t. As WL discuss, however, the appearance of Q_t need not involve instantaneous causation. It suffices that realizations of Q_t precede those of Y_t, as in the case of contemporaneous causation discussed above. The replacement of S^{t-1} with S_t entails first viewing S_t as representing a finite history, and second the recognition that since S_t plays purely a conditioning role, there need be no restriction whatever on its timing. We thus call S_t "covariates." As WL discuss, the covariates can even include leads relative to time t. When covariate leads appear, we call this the "retrospective" case.

In what follows, when we refer to $G-$causality, it will be understood that we are referring to finite-order $G-$causality, as just defined. We will always refer to the concept of Definition 5.1 as *classical* $G-$causality to avoid confusion.

5.2. A Dynamic Structural System

We now specify a canonical settable system that will enable us to examine the relation between $G-$causality and direct structural causality. As described above, in such systems "predecessors" structurally determine "successors," but not vice versa. In particular, future variables cannot precede present or past variables, enforcing the causal direction of time. We write $Y \Leftarrow X$ to denote that Y succeeds X (X precedes Y). When Y and X have identical time indexes, $Y \Leftarrow X$ rules out instantaneous causation but allows contemporaneous causation.

We now specify a version of the causal data generating structures analyzed by WL and White, H. and P. Kennedy (2009). We let $\mathbb{N}$ denote the integers $\{0, 1, \ldots\}$ and define $\bar{\mathbb{N}} := \mathbb{N} \cup \{\infty\}$. For given $\ell, m, \in \mathbb{N}$, $\ell \geq 1$, we let $\boldsymbol{Y}_{t-1} := (Y_{t-\ell}, \ldots, Y_{t-1})$ as above; we also define $\boldsymbol{Z}_t := (Z_{t-m}, \ldots, Z_t)$. For simplicity, we keep attributes implicit in what follows.

Assumption A.1 Let $\{U_t, W_t, Y_t, Z_t;\ t = 0, 1, \ldots\}$ be a stochastic process on $(\Omega, \mathcal{F}, P)$, a complete probability space, with U_t, W_t, Y_t, and Z_t taking values in $\mathbb{R}^{k_u}, \mathbb{R}^{k_w}, \mathbb{R}^{k_y}$, and $\mathbb{R}^{k_z}$ respectively, where $k_u \in \bar{\mathbb{N}}$ and $k_w, k_y, k_z \in \mathbb{N}$, with $k_y > 0$. Further, suppose that $Y_t \Leftarrow (Y^{t-1}, U^t, W^t, Z^t)$, where, for an unknown measurable $k_y \times 1$ function q_t, and for given $\ell, m, \in \mathbb{N}, \ell \geq 1$, $\{Y_t\}$ is structurally generated as

$$Y_t = q_t(\boldsymbol{Y}_{t-1}, \boldsymbol{Z}_t, U_t), \quad t = 1, 2, \ldots, \tag{4}$$

such that, with $Y_t := (Y'_{1,t}, Y'_{2,t})'$ and $U_t := (U'_{1,t}, U'_{2,t})'$,

$$Y_{1,t} = q_{1,t}(\boldsymbol{Y}_{t-1}, \boldsymbol{Z}_t, U_{1,t}) \qquad Y_{2,t} = q_{2,t}(\boldsymbol{Y}_{t-1}, \boldsymbol{Z}_t, U_{2,t}).$$

Such structures are well suited to representing the structural evolution of time-series data in economic, biological, or other systems. Because Y_t is a vector, this covers the case of panel data, where one has a cross-section of time-series observations, as in fMRI or EEG data sets. For practical relevance, we explicitly impose the Markov assumption that Y_t is determined by only a finite number of its own lags and those of Z_t and U_t. WL discuss the general case.

Throughout, we suppose that realizations of W_t, Y_t, and Z_t are observed, whereas realizations of U_t are not. Because U_t, W_t, or Z_t may have dimension zero, their presence is optional. Usually, however, some or all will be present. Since there may be a countable infinity of unobservables, there is no loss of generality in specifying that Y_t depends only on U_t rather than on a finite history of lags of U_t.

This structure is general: the structural relations may be nonlinear and non-monotonic in their arguments and non-separable between observables and unobservables. This system may generate stationary processes, non-stationary processes, or both. Assumption A.1 is therefore a general structural VAR; Example 3.3 is a special case.

The vector Y_t represents responses of interest. Consistent with a main application of $G-$causality, our interest here attaches to the effects on $Y_{1,t}$ of the lags of $Y_{2,t}$. We thus call $Y_{2,t-1}$ and its further lags "causes of interest." Note that A.1 specifies that $Y_{1,t}$ and $Y_{2,t}$ each have their own unobserved drivers, $U_{1,t}$ and $U_{2,t}$, as is standard.

The vectors U_t and Z_t contain causal drivers of Y_t whose effects are not of primary interest; we thus call U_t and Z_t "ancillary causes." The vector W_t may contain responses to U_t. Observe that W_t does not appear in the argument list for q_t, so it explicitly does not directly determine Y_t. Note also that $Y_t \Leftarrow (Y^{t-1}, U^t, W^t, Z^t)$ ensures that W_t is not determined by Y_t or its lags. A useful convention is that $W_t \Leftarrow (W^{t-1}, U^t, Z^t)$, so that W_t does not drive unobservables. If a structure does not have this property, then suitable substitutions can usually yield a derived structure satisfying this convention. Nevertheless, we do not require this, so W_t may also contain drivers of unobservable causes of Y_t.

For concreteness, we now specialize the settable systems definition of direct structural causality (Definition 4.1) to the specific system given in A.1. For this, let $\boldsymbol{y}_{s,t-1}$ be the sub-vector of $\boldsymbol{y}_{t-1}$ with elements indexed by the non-empty set $\boldsymbol{s} \subseteq \{1,\ldots,k_y\} \times \{t-\ell,\ldots,t-1\}$, and let $\boldsymbol{y}_{(s),t-1}$ be the sub-vector of $\boldsymbol{y}_{t-1}$ with elements of $\boldsymbol{s}$ excluded.

Definition 5.3 (Direct Structural Causality) *Given A.1, for given $t > 0$, $j \in \{1,\ldots,k_y\}$, and $\boldsymbol{s}$, suppose that for all admissible values of $\boldsymbol{y}_{(s),t-1}$, z_t, and u_t, the function $\boldsymbol{y}_{s,t-1} \to q_{j,t}(\boldsymbol{y}_{t-1}, z_t, u_t)$ is constant in $\boldsymbol{y}_{s,t-1}$. Then we say* $\boldsymbol{Y}_{s,t-1}$ **does not directly structurally cause** *$Y_{j,t}$ and write $\boldsymbol{Y}_{s,t-1} \overset{D}{\not\Rightarrow}_{\mathcal{S}} Y_{j,t}$. Otherwise, we say* $\boldsymbol{Y}_{s,t-1}$ **directly structurally causes** *$Y_{j,t}$ and write $\boldsymbol{Y}_{s,t-1} \overset{D}{\Rightarrow}_{\mathcal{S}} Y_{j,t}$.*

We can similarly define direct causality or non-causality of $Z_{s,t}$ or $U_{s,t}$ for $Y_{j,t}$, but we leave this implicit. We write, e.g., $\boldsymbol{Y}_{s,t-1} \overset{D}{\Rightarrow}_{\mathcal{S}} Y_t$ when $\boldsymbol{Y}_{s,t-1} \overset{D}{\Rightarrow}_{\mathcal{S}} Y_{j,t}$ for some $j \in \{1,\ldots,k_y\}$.

Building on work of White, H. (2006a) and White, H. and K. Chalak (2009), WL discuss how certain exogeneity restrictions permit identification of expected causal effects in dynamic structures. Our next result shows that a specific form of exogeneity enables us to link direct structural causality and finite order $G-$causality. To state this exogeneity condition, we write $\boldsymbol{Y}_{1,t-1} := (Y_{1,t-\ell},\ldots,Y_{1,t-1})$, $\boldsymbol{Y}_{2,t-1} := (Y_{2,t-\ell},\ldots,Y_{2,t-1})$, and, for given $\tau_1,\tau_2 \geq 0$, $\boldsymbol{X}_t := (X_{t-\tau_1},\ldots,X_{t+\tau_2})$, where $X_t := (W_t', Z_t')'$.

Assumption A.2 For ℓ and m as in A.1 and for $\tau_1 \geq m, \tau_2 \geq 0$, suppose that $\boldsymbol{Y}_{2,t-1} \perp U_{1,t} \mid (\boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t)$, $t = 1,\ldots,T-\tau_2$.

The classical *strict exogeneity* condition specifies that $(\boldsymbol{Y}_{t-1}, \boldsymbol{Z}_t) \perp U_{1,t}$, which implies $\boldsymbol{Y}_{2,t-1} \perp U_{1,t} \mid (\boldsymbol{Y}_{1,t-1}, \boldsymbol{Z}_t)$. (Here, W_t can be omitted.) Assumption A.2 is a weaker requirement, as it may hold when strict exogeneity fails. Because of the conditioning involved, we call this *conditional exogeneity*. Chalak, K. and H. White (2010) discuss structural restrictions for canonical settable systems that deliver conditional exogeneity. Below, we also discuss practical tests for this assumption.

Because of the finite numbers of lags involved in A.2, this is a *finite-order* conditional exogeneity assumption. For convenience and because no confusion will arise here, we simply refer to this as "conditional exogeneity."

Assumption A.2 ensures that expected direct effects of $\boldsymbol{Y}_{2,t-1}$ on $Y_{1,t}$ are identified. As WL note, it suffices for A.2 that $U^{t-1} \perp U_{1,t} \mid (Y_0, Z^{t-1}, \boldsymbol{X}_t)$ and $\boldsymbol{Y}_{2,t-1} \perp (Y_0, Z^{t-\tau_1-1}) \mid (\boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t)$. Imposing $U^{t-1} \perp U_{1,t} \mid (Y_0, Z^{t-1}, \boldsymbol{X}_t)$ is the analog of requiring that serial correlation is absent when lagged dependent variables are present. Imposing $\boldsymbol{Y}_{2,t-1} \perp (Y_0, Z^{t-\tau_1-1}) \mid (\boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t)$ ensures that ignoring Y_0 and omitting distant lags of Z_t from $\boldsymbol{X}_t$ doesn't matter.

Our first result linking direct structural causality and $G-$causality shows that, given A.1 and A.2 and with proper choice of Q_t and S_t, $G-$causality implies direct structural causality.

Proposition 5.4 *Let A.1 and A.2 hold. If* $\boldsymbol{Y}_{2,t-1} \overset{D}{\not\Rightarrow}_S Y_{1,t}$, $t = 1,2,\ldots$, *then* $\boldsymbol{Y}_2$ *does not finite order* $G-$*cause* Y_1 *with respect to* $\boldsymbol{X}$, *i.e.,*

$$Y_{1,t} \perp \boldsymbol{Y}_{2,t-1} \mid \boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t, \quad t = 1,\ldots,T-\ \tau_2.$$

In stating G non-causality, we make the explicit identifications $\boldsymbol{Q}_t = \boldsymbol{Y}_{2,t-1}$ and $S_t = \boldsymbol{X}_t$.

This result leads one to ask whether the converse relation also holds: does direct structural causality imply $G-$causality? Strictly speaking, the answer is no. WL discuss several examples. The main issue is that with suitably chosen causal and probabilistic relationships, $\boldsymbol{Y}_{2,t-1}$ can cause $Y_{1,t}$, but $\boldsymbol{Y}_{2,t-1}$ and $Y_{1,t}$ can be independent, conditionally or unconditionally, i.e. Granger non-causal.

As WL further discuss, however, these examples are exceptional, in the sense that mild perturbations to their structure destroy the Granger non-causality. WL introduce a refinement of the notion of direct structural causality that accommodates these special cases and that does yield a converse result, permitting a characterization of structural and Granger causality. Let $supp(Y_{1,t})$ denote the support of $Y_{1,t}$, i.e., the smallest set containing $Y_{1,t}$ with probability 1, and let $F_{1,t}(\cdot \mid \boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t)$ denote the conditional distribution function of $U_{1,t}$ given $\boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t$. WL introduce the following definition:

Definition 5.5 *Suppose A.1 holds and that for given* $\tau_1 \geq m, \tau_2 \geq 0$ *and for each* $y \in$ *supp*$(Y_{1,t})$ *there exists a* $\sigma(\boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t)-$*measurable version of the random variable*

$$\int 1\{q_{1,t}(\boldsymbol{Y}_{t-1}, \boldsymbol{Z}_t, u_{1,t}) < y\}\; dF_{1,t}(u_{1,t} \mid \boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t).$$

Then $\boldsymbol{Y}_{2,t-1} \overset{D}{\not\Rightarrow}_{\mathcal{S}(\boldsymbol{Y}_{1,t-1},\boldsymbol{X}_t)} Y_{1,t}$ (**direct non-causality**$-\sigma(\boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t)$ *a.s.*). *If not,* $\boldsymbol{Y}_{2,t-1}$ $\overset{D}{\Rightarrow}_{\mathcal{S}(\boldsymbol{Y}_{1,t-1},\boldsymbol{X}_t)} Y_{1,t}$.

For simplicity, we refer to this as *direct non-causality a.s*. The requirement that the integral in this definition is $\sigma(\boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t)-$measurable means that the integral does not depend on $\boldsymbol{Y}_{2,t-1}$, despite its appearance inside the integral as an argument of $q_{1,t}$. For this, it suffices that $\boldsymbol{Y}_{2,t-1}$ does not directly cause $Y_{1,t}$; but it is also possible that $q_{1,t}$ and the conditional distribution of $U_{1,t}$ given $\boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t$ are in just the right relation to hide the structural causality. Without the ability to manipulate this distribution, the structural causality will not be detectable. One possible avenue to manipulating this distribution is to modify the choice of $\boldsymbol{X}_t$, as there are often multiple choices for $\boldsymbol{X}_t$ that can satisfy A.2 (see White, H. and X. Lu, 2010b). For brevity and because hidden structural causality is an exceptional circumstance, we leave aside further discussion of this possibility here. The key fact to bear in mind is that the causal concept of Definition 5.5 distinguishes between those direct causal relations that are empirically detectable and those that are not, for a given set of covariates $\boldsymbol{X}_t$.

We now give a structural characterization of $G-$causality for structural VARs:

Theorem 5.6 *Let A.1 and A.2 hold. Then* $\boldsymbol{Y}_{2,t-1} \overset{D}{\not\Rightarrow}_{\mathcal{S}(\boldsymbol{Y}_{1,t-1},\boldsymbol{X}_t)} Y_{1,t}$, $t = 1, \ldots, T - \tau_2$, *if and only if*

$$Y_{1,t} \perp \boldsymbol{Y}_{2,t-1} \mid \boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t, \quad t = 1, \ldots, T - \tau_2,$$

i.e., $\boldsymbol{Y}_2$ *does not finite-order* $G-$*cause* Y_1 *with respect to* $\boldsymbol{X}$.

Thus, given conditional exogeneity of $\boldsymbol{Y}_{2,t-1}$, *G non-causality implies direct non-causality a.s. and vice-versa*, justifying tests of direct non-causality *a.s.* in structural VARs using tests for $G-$causality.

This result completes the desired linkage between $G-$causality and direct causality in the PCM. Because direct causality in the recursive PCM corresponds essentially to direct structural causality in canonical settable systems, and because the latter is essentially equivalent to $G-$causality, as just shown, direct causality in the PCM is essentially equivalent to $G-$causality, provided A.1 and A.2 hold.

5.3. The Central Role of Conditional Exogeneity

To relate direct structural causality to $G-$causality, we maintain A.2, a specific conditional exogeneity assumption. Can this assumption be eliminated or weakened? We show that the answer is no: A.2 is in a precise sense a necessary condition. We also give a result supporting tests for conditional exogeneity.

First, we specify the sense in which conditional exogeneity is necessary for the equivalence of $G-$causality and direct structural causality.

Proposition 5.7 *Given A.1, suppose that* $\boldsymbol{Y}_{2,t-1} \overset{D}{\not\Rightarrow}_{\mathcal{S}} Y_{1,t}$, $t = 1,2,\ldots$ *. If A.2 does not hold, then for each t there exists* $q_{1,t}$ *such that* $Y_{1,t} \perp \boldsymbol{Y}_{2,t-1} \mid \boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t$ *does not hold.*

That is, if conditional exogeneity does not hold, then there are always structures that generate data exhibiting $G-$causality, despite the absence of direct structural causality. Because $q_{1,t}$ is unknown, this worst case scenario can never be discounted. Further, as WL show, the class of worst case structures includes precisely those usually assumed in applications, namely separable structures (e.g., $Y_{1,t} = q_{1,t}(\boldsymbol{Y}_{1,t-1}, \boldsymbol{Z}_t) + U_{1,t}$), as well as the more general class of invertible structures. Thus, in the cases typically assumed in the literature, the failure of conditional exogeneity *guarantees* $G-$causality in the absence of structural causality. We state this formally as a corollary.

Corollary 5.8 *Given A.1 with* $\boldsymbol{Y}_{2,t-1} \overset{D}{\not\Rightarrow}_{\mathcal{S}} Y_{1,t}$, $t = 1,2,\ldots$, *suppose that* $q_{1,t}$ *is invertible in the sense that* $Y_{1,t} = q_{1,t}(\boldsymbol{Y}_{1,t-1}, \boldsymbol{Z}_t, U_{1,t})$ *implies the existence of* $\xi_{1,t}$ *such that* $U_{1,t} = \xi_{1,t}(\boldsymbol{Y}_{1,t-1}, \boldsymbol{Z}_t, Y_{1,t})$, $t = 1,2,\ldots$ *. If A.2 fails, then* $Y_{1,t} \perp \boldsymbol{Y}_{2,t-1} \mid \boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t$ *fails,* $t = 1,2,\ldots$.

Together with Theorem 5.6, this establishes that in the absence of direct causality and for the class of invertible structures predominant in applications, *conditional exogeneity is necessary and sufficient for G non-causality.*

Tests of conditional exogeneity for the general separable case follow from:

Proposition 5.9 *Given A.1, suppose that* $E(Y_{1,t}) < \infty$ *and that*

$$q_{1,t}(\boldsymbol{Y}_{t-1}, \boldsymbol{Z}_t, U_{1,t}) = \zeta_t(\boldsymbol{Y}_{t-1}, \boldsymbol{Z}_t) + \upsilon_t(\boldsymbol{Y}_{1,t-1}, \boldsymbol{Z}_t, U_{1,t}),$$

where ζ_t *and* υ_t *are unknown measurable functions. Let* $\varepsilon_t := Y_{1,t} - E(Y_{1,t}|\boldsymbol{Y}_{t-1}, \boldsymbol{X}_t)$*. If A.2 holds, then*

$$\begin{gathered}\varepsilon_t = \upsilon_t(\boldsymbol{Y}_{1,t-1}, \boldsymbol{Z}_t, U_{1,t}) - E(\upsilon_t(\boldsymbol{Y}_{1,t-1}, \boldsymbol{Z}_t, U_{1,t}) \mid \boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t) \\ E(\varepsilon_t|\boldsymbol{Y}_{t-1}, \boldsymbol{X}_t) = E(\varepsilon_t|\boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t) = 0 \qquad \textit{and} \\ \boldsymbol{Y}_{2,t-1} \perp \varepsilon_t \mid \boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t. \end{gathered} \tag{5}$$

Tests based on this result detect the failure of A.2, given separability. Such tests are feasible because even though the regression error ε_t is unobserved, it can be consistently estimated, say as $\hat{\varepsilon}_t := Y_{1,t} - \hat{E}(Y_{1,t}|\boldsymbol{Y}_{t-1}, \boldsymbol{X}_t)$, where $\hat{E}(Y_{1,t}|\boldsymbol{Y}_{t-1}, \boldsymbol{X}_t)$ is a parametric or nonparametric estimator of $E(Y_{1,t}|\boldsymbol{Y}_{t-1}, \boldsymbol{X}_t)$. These estimated errors can then be used to test (5). If we reject (5), then we must reject A.2. We discuss a practical procedure in the next section. WL provide additional discussion.

WL also discuss dropping the separability assumption. For brevity, we maintain separability here. Observe that under the null of direct non-causality, $q_{1,t}$ is necessarily separable, as then ζ_t is the zero function.

6. Testing Direct Structural Causality

Here, we discuss methods for testing direct structural causality. First, we discuss a general approach that combines tests of *G* non-causality (GN) and conditional exogeneity (CE). Then we describe straightforward practical methods for implementing the general approach.

6.1. Combining Tests for GN and CE

Theorem 5.6 implies that if we test and reject GN, then we must reject either direct structural non-causality (SN) or CE, or both. If CE is maintained, then we can directly test SN by testing GN; otherwise, a direct test is not available.

Similarly, under the traditional separability assumption, Corollary 5.8 implies that if we test and reject CE, then we must reject either SN or GN (or both). If GN is maintained, then we can directly test SN by testing CE; otherwise, a direct test is not available.

When neither CE nor GN is maintained, no direct test of SN is possible. Nevertheless, we can test structural causality indirectly by combining the results of the GN and CE tests to isolate the source of any rejections. WL propose the following indirect test:

(1) *Reject* SN if either:

(*i*) the CE test *fails to reject* and the GN test *rejects*; or

(*ii*) the CE test *rejects* and the GN test *fails to reject*.

If these rejection conditions do not hold, however, we cannot just decide to “accept" (i.e., fail to reject) SN. As WL explain in detail, difficulties arise when CE and GN both fail, as failing to reject SN here runs the risk of Type II error, whereas rejecting SN runs the risk of Type I error. We resolve this dilemma by specifying the further rules:

(2) *Fail to reject* SN if the CE and GN tests both *fail to reject;*

(3) *Make no decision* as to SN if the CE and GN tests both *reject.*

In the latter case, we conclude only that CE and GN both fail, thereby obstructing structural inference. This sends a clear signal that the researcher needs to revisit the model specification, with particular attention to specifying covariates sufficient to ensure conditional exogeneity.

Because of the structure of this indirect test, it is not enough simply to consider its level and power. We must also account for the possibility of making no decision. For this, define

$$\begin{aligned}
p \ &: = P[\text{ wrongly make a decision }] \\
&= \ P[\text{ fail to reject CE or GN} \mid \text{CE is false and GN is false }] \\
q \ &: = P[\text{ wrongly make no decision }] \\
&= \ P[\text{ reject CE and GN} \mid \text{CE is true or GN is true }].
\end{aligned}$$

These are the analogs of the probabilities of Type I and Type II errors for the "no decision" action. We would like these probabilities to be small. Next, we consider

$$\begin{aligned}\alpha^* &:= P[\text{ reject SN or make no decision} \mid \text{CE is true and GN is true }] \\ \pi^* &:= P[\text{ reject SN} \mid \text{exactly one of CE and GN is true }].\end{aligned}$$

These quantities correspond to notions of level and power, but with the sample space restricted to the subset on which CE is true or GN is true, that is, the space where a decision can be made. Thus, α^* differs from the standard notion of level, but it does capture the probability of taking an incorrect action when SN (the null) holds in the restricted sample space, i.e., when CE and GN are both true. Similarly, π^* captures the probability of taking the correct action when SN does not hold in the restricted sample space. We would like the "restricted level" α^* to be small and the "restricted power" π^* to be close to one.

WL provide useful bounds on the asymptotic properties ($T \to \infty$) of the sample-size T values of the probabilities defined above, p_T, q_T, α^*_T, and π^*_T:

Proposition 6.1 *Suppose that for $T = 1,2,\ldots$ the significance levels of the CE and GN tests are α_{1T} and α_{2T}, respectively, and that $\alpha_{1T} \to \alpha_1 < .5$ and $\alpha_{2T} \to \alpha_2 < .5$. Suppose the powers of the CE and GN tests are π_{1T} and π_{2T}, respectively, and that $\pi_{1T} \to 1$ and $\pi_{2T} \to 1$. Then*

$$\begin{aligned} p_T &\to 0, \qquad \limsup q_T \leq \max\{\alpha_1, \alpha_2\}, \\ |\alpha_1 - \alpha_2| &\leq \liminf \alpha^*_T \leq \limsup \alpha^*_T \leq \alpha_1 + \alpha_2 + \min\{\alpha_1, \alpha_2\}, \quad \textit{and} \\ \min\{1-\alpha_1, 1-\alpha_2\} &\leq \liminf \pi^*_T \leq \limsup \pi^*_T \leq \max\{1-\alpha_1, 1-\alpha_2\}. \end{aligned}$$

When $\pi_{1T} \to 1$ and $\pi_{2T} \to 1$, one can also typically ensure $\alpha_1 = 0$ and $\alpha_2 = 0$ by suitable choice of an increasing sequence of critical values. In this case, $q_T \to 0$, $\alpha^*_T \to 0$, and $\pi^*_T \to 1$. Because GN and CE tests will not be consistent against every possible alternative, weaker asymptotic bounds on the level and power of the indirect test hold for these cases by Proposition 8.1 of WL. Thus, whenever possible, one should carefully design GN and CE tests to have power against particularly important or plausible alternatives. See WL for further discussion.

6.2. Practical Tests for GN and CE

To test GN and CE, we require tests for conditional independence. Nonparametric tests for conditional independence consistent against arbitrary alternatives are readily available (e.g. Linton, O. and P. Gozalo, 1997; Fernandes, M. and R. G. Flores, 2001; Delgado, M. A. and W. Gonzalez-Manteiga, 2001; Su, L. and H. White, 2007a,b, 2008; Song, K., 2009; Huang, M. and H. White, 2009). In principle, one can apply any of these to consistently test GN and CE.

But nonparametric tests are often not practical, due to the typically modest number of time-series observations available relative to the number of relevant observable variables. In practice, researchers typically use parametric methods. These are convenient, but they may lack power against important alternatives. To provide convenient procedures for testing GN and CE with power against a wider range of alternatives, WL propose augmenting standard tests with neural network terms, motivated by the “QuickNet" procedures introduced by White, H (2006b) or the extreme learning machine (ELM) methods of Huang, G.B., Q.Y. Zhu, and C.K. Siew (2006). We now provide explicit practical methods for testing GN and CE for a leading class of structures obeying A.1.

6.2.1. Testing Granger Non-Causality

Standard tests for finite-order $G-$causality (e.g., Stock, J. and M. Watson, 2007, p. 547) typically assume a linear regression, such as[3]

$$E(Y_{1,t}|\boldsymbol{Y}_{t-1},\boldsymbol{X}_t) = \alpha_0 + \boldsymbol{Y}'_{1,t-1}\rho_0 + \boldsymbol{Y}'_{2,t-1}\beta_0 + \boldsymbol{X}'_t\beta_1.$$

For simplicity, we let $Y_{1,t}$ be a scalar here. The extension to the case of vector $Y_{1,t}$ is completely straightforward. Under the null of GN, i.e., $Y_{1,t} \perp \boldsymbol{Y}_{2,t-1} \mid \boldsymbol{Y}_{1,t-1},\boldsymbol{X}_t$, we have $\beta_0 = 0$. The standard procedure therefore tests $\beta_0 = 0$ in the regression equation

$$Y_{1,t} = \alpha_0 + \boldsymbol{Y}'_{1,t-1}\rho_0 + \boldsymbol{Y}'_{2,t-1}\beta_0 + \boldsymbol{X}'_t\beta_1 + \varepsilon_t. \qquad \text{(GN Test Regression 1)}$$

If we reject $\beta_0 = 0$, then we also reject GN. But if we don’t reject $\beta_0 = 0$, care is needed, as not all failures of GN will be indicated by $\beta_0 \neq 0$.

Observe that when CE holds and if GN Test Regression 1 is correctly specified, i.e., the conditional expectation $E(Y_{1,t}|\boldsymbol{Y}_{t-1},\boldsymbol{X}_t)$ is indeed linear in the conditioning variables, then β_0 represents precisely the direct structural effect of $\boldsymbol{Y}_{2,t-1}$ on $Y_{1,t}$. Thus, GN Test Regression 1 may not only permit a test of GN, but it may also provide a consistent estimate of the direct structural effect of interest.

To mitigate specification error and gain power against a wider range of alternatives, WL propose augmenting GN Test Regression 1 with neural network terms, as in White’s (2006b, p. 476) QuickNet procedure. This involves testing $\beta_0 = 0$ in

$$Y_{1,t} \;=\; \alpha_0 + \boldsymbol{Y}'_{1,t-1}\rho_0 + \boldsymbol{Y}'_{2,t-1}\beta_0 + \boldsymbol{X}'_t\beta_1 + \sum_{j=1}^{r}\psi(\boldsymbol{Y}'_{1,t-1}\gamma_{1,j} + \boldsymbol{X}'_t\gamma_j)\beta_{j+1} + \varepsilon_t. \qquad \text{(GN Test Regression 2)}$$

Here, the activation function ψ is a generically comprehensively revealing (GCR) function (see Stinchcombe, M. and H. White, 1998). For example, ψ can be the logistic cdf $\psi(z) = 1/(1+\exp(-z))$ or a ridgelet function, e.g., $\psi(z) = (-z^5 + 10z^3 - 15z)\exp(-.5z^2)$ (see, for example, Candès, E. (1999)). The integer r lies between 1 and $\bar{r}$, the maximum number of hidden units. We randomly choose (γ_{0j},γ_j) as in White, H (2006b, p. 477).

[3] For notational convenience, we understand that all regressors have been recast as vectors containing the referenced elements.

Parallel to our comment above about estimating direct structural effects of interest, we note that given A.1, A.2, and some further mild regularity conditions, such effects can be identified and estimated from a neural network regression of the form

$$\begin{aligned} Y_{1,t} &= \alpha_0 + \boldsymbol{Y}'_{1,t-1}\rho_0 + \boldsymbol{Y}'_{2,t-1}\beta_0 + \boldsymbol{X}'_t\beta_1 \\ &\quad + \sum_{j=1}^{r} \psi(\boldsymbol{Y}'_{1,t-1}\gamma_{1,j} + \boldsymbol{Y}'_{2,t-1}\gamma_{2,j} + \boldsymbol{X}'_t\gamma_{3,j})\beta_{j+1} + \varepsilon_t. \end{aligned}$$

Observe that this regression includes $\boldsymbol{Y}_{2,t-1}$ inside the hidden units. With r chosen sufficiently large, this permits the regression to achieve a sufficiently close approximation to $E(Y_{1,t}|\boldsymbol{Y}_{t-1}, \boldsymbol{X}_t)$ and its derivatives (see Hornik, K. M. Stinchcombe, and H. White, 1990; Gallant, A.R. and H. White, 1992) that regression misspecification is not such an issue. In this case, the derivative of the estimated regression with respect to $\boldsymbol{Y}_{2,t-1}$ well approximates

$$\begin{aligned} &(\partial/\partial \boldsymbol{y}_2)E(Y_{1,t} \mid \boldsymbol{Y}_{1,t-1}, \boldsymbol{Y}_{2,t-1} = \boldsymbol{y}_2, \boldsymbol{X}_t) \\ &\qquad = E[\ (\partial/\partial \boldsymbol{y}_2)q_{1,t}(\boldsymbol{Y}_{1,t-1}, \boldsymbol{y}_2, \boldsymbol{Z}_t, U_{1,t}) \mid \boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t]. \end{aligned}$$

This quantity is the *covariate conditioned expected marginal direct effect of* $\boldsymbol{Y}_{2,t-1}$ *on* $Y_{1,t}$.

Although it is possible to base a test for GN on these estimated effects, we do not propose this here, as the required analysis is much more involved than that associated with GN Test Regression 2.

Finally, to gain additional power WL propose tests using transformations of $Y_{1,t}$, $\boldsymbol{Y}_{1,t-1}$, and $\boldsymbol{Y}_{2,t-1}$, as $Y_{1,t} \perp \boldsymbol{Y}_{2,t-1} \mid \boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t$ implies $f(Y_{1,t}) \perp g(\boldsymbol{Y}_{2,t-1}) \mid \boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t$ for all measurable f and g. One then tests $\beta_{1,0} = 0$ in

$$\begin{aligned} \psi_1(Y_{1,t}) &= \alpha_{1,0} + \psi_2(\boldsymbol{Y}_{1,t-1})'\rho_{1,0} + \psi_3(\boldsymbol{Y}_{2,t-1})'\beta_{1,0} + \boldsymbol{X}'_t\beta_{1,1} \\ &\quad + \sum_{j=1}^{r} \psi(\boldsymbol{Y}'_{1,t-1}\gamma_{1,1,j} + \boldsymbol{X}'_t\gamma_{1,j})\beta_{1,j+1} + \eta_t. \qquad \text{(GN Test Regression 3)} \end{aligned}$$

We take ψ_1 and the elements of the vector ψ_3 to be GCR, e.g., ridgelets or the logistic cdf. The choices of $\gamma, r,$ and ψ are as described above. Here, ψ_2 can be the identity $(\psi_2(\boldsymbol{Y}_{1,t-1}) = \boldsymbol{Y}_{1,t-1})$, its elements can coincide with ψ_1, or it can be a different GCR function.

6.2.2. Testing Conditional Exogeneity

Testing conditional exogeneity requires testing A.2, i.e., $\boldsymbol{Y}_{2,t-1} \perp U_{1,t} \mid \boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t$. Since $U_{1,t}$ is unobservable, we cannot test this directly. But with separability (which holds under the null of direct structural non-causality), Proposition 5.9 shows that $\boldsymbol{Y}_{2,t-1} \perp U_{1,t} \mid \boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t$ implies $\boldsymbol{Y}_{2,t-1} \perp \varepsilon_t \mid \boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t$, where $\varepsilon_t := Y_{1,t} - E(Y_{1,t}|\boldsymbol{Y}_{t-1}, \boldsymbol{X}_t)$. With correct specification of $E(Y_{1,t}|\boldsymbol{Y}_{t-1}, \boldsymbol{X}_t)$ in either GN Test Regression 1 or 2 (or some other appropriate regression), we can estimate ε_t and use these estimates to test $\boldsymbol{Y}_{2,t-1} \perp$

$\varepsilon_t \mid \boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t$. If we reject this, then we also must reject CE. We describe the procedure in detail below.

As WL discuss, such a procedure is not "watertight," as this method may miss certain alternatives to CE. But, as it turns out, there is no completely infallible method. By offering the opportunity of falsification, this method provides crucial insurance against being naively misled into inappropriate causal inferences. See WL for further discussion.

The first step in constructing a practical test for CE is to compute estimates of ε_t, say $\hat{\varepsilon}_t$. This can be done in the obvious way by taking $\hat{\varepsilon}_t$ to be the estimated residuals from a suitable regression. For concreteness, suppose this is either GN Test Regression 1 or 2.

The next step is to use $\hat{\varepsilon}_t$ to test $\boldsymbol{Y}_{2,t-1} \perp \varepsilon_t \mid \boldsymbol{Y}_{1,t-1}, \boldsymbol{X}_t$. WL recommend doing this by estimating the following analog of GN Test Regression 3:

$$\begin{aligned}\psi_1(\hat{\varepsilon}_t) &= \alpha_{2,0} + \psi_2(\boldsymbol{Y}_{1,t-1})'\rho_{2,0} + \psi_3(\boldsymbol{Y}_{2,t-1})'\beta_{2,0} + \boldsymbol{X}_t'\beta_{2,1} \\ &\quad + \sum_{j=1}^{r} \psi(\boldsymbol{Y}_{1,t-1}'\gamma_{2,1,j} + \boldsymbol{X}_t'\gamma_{2,j})\beta_{2,j+1} + \eta_t. \qquad \text{(CE Test Regression)}\end{aligned}$$

Note that the right-hand-side regressors are identical to those of GN Test Regression 3; we just replace the dependent variable $\psi_1(Y_{1,t})$ for GN with $\psi_1(\hat{\varepsilon}_t)$ for CE. Nevertheless, the transformations ψ_1, ψ_2, and ψ_3 here may differ from those of GN Test Regression 3. To keep the notation simple, we leave these possible differences implicit. To test CE using this regression, we test the null hypothesis $\beta_{2,0} = 0$: if we reject $\beta_{2,0} = 0$, then we reject CE.

As WL explain, the fact that $\hat{\varepsilon}_t$ is obtained from a "first-stage" estimation (GN) involving potentially the same regressors as those appearing in the CE regression means that choosing $\psi_1(\hat{\varepsilon}_t) = \hat{\varepsilon}_t$ can easily lead to a test with no power. For CE, WL thus recommend choosing ψ_1 to be GCR. Alternatively, non-GCR choices may be informative, such as

$$\psi_1(\hat{\varepsilon}_t) = |\hat{\varepsilon}_t|, \quad \psi_1(\hat{\varepsilon}_t) = \hat{\varepsilon}_t(\lambda - 1\{\hat{\varepsilon}_t < 0\}), \ \lambda \in (0,1), \quad \text{or} \quad \psi_1(\hat{\varepsilon}_t) = \hat{\varepsilon}_t^2.$$

Significantly, the asymptotic sampling distributions needed to test $\beta_{2,0} = 0$ will generally be impacted by the first-stage estimation. Handling this properly is straightforward, but somewhat involved. To describe a practical method, we denote the first-stage (GN) estimator as $\hat{\theta}_{1,T} := (\hat{\alpha}_{1,T}, \hat{\rho}_{1,T}, \hat{\beta}_{1,0,T}', \hat{\beta}_{1,1,T}', \ldots, \hat{\beta}_{1,r+1,T})'$, computed from GN Test Regression 1 ($r = 0$) or 2 ($r > 0$). Let the second stage (CE) regression estimator be $\hat{\theta}_{2,T}$; this contains the estimated coefficients for $\boldsymbol{Y}_{2,t-1}$, say $\hat{\beta}_{2,0,T}$, which carry the information about CE. Under mild conditions, a central limit theorem ensures that

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, C_0),$$

where $\hat{\theta}_T := (\hat{\theta}_{1,T}', \hat{\theta}_{2,T}')'$, $\theta_0 := plim(\hat{\theta}_T)$, convergence in distribution as $T \to \infty$ is denoted $\xrightarrow{d}$, and $N(0, C_0)$ denotes the multivariate normal distribution with mean zero and covariance matrix $C_0 := A_0^{-1} B_0 A_0^{-1\prime}$, where

$$A_0 := \begin{bmatrix} A_{011} & \mathbf{0} \\ A_{021} & A_{022} \end{bmatrix}$$

is a two-stage analog of the log-likelihood Hessian and B_0 is an analog of the information matrix. See White, H. (1994, pp. 103–108) for specifics.[4] This fact can then be use to construct a well behaved test for $\beta_{2,0} = 0$.

Constructing this test is especially straightforward when the regression errors of the GN and CE regressions, ε_t and η_t, are suitable martingale differences. Then B_0 has the form

$$B_0 := \begin{bmatrix} E[\,\mathcal{Z}_t \varepsilon_t \varepsilon_t' \mathcal{Z}_t'\,] & E[\,\mathcal{Z}_t \varepsilon_t \eta_t' \mathcal{Z}_t'\,] \\ E[\,\mathcal{Z}_t \eta_t \varepsilon_t' \mathcal{Z}_t'\,] & E[\,\mathcal{Z}_t \eta_t \eta_t' \mathcal{Z}_t'\,] \end{bmatrix},$$

where the CE regressors $\mathcal{Z}_t$ are measurable-$\sigma(\mathcal{X}_t)$, $\mathcal{X}_t := (vec[\boldsymbol{Y}_{t-1}]', vec[\boldsymbol{X}_t]')'$, $\varepsilon_t := Y_{1,t} - E(Y_{1,t} \mid \mathcal{X}_t)$, and $\eta_t := \psi_1(\varepsilon_t) - E[\psi_1(\varepsilon_t) \mid \mathcal{X}_t]$. For this, it suffices that $U_{1,t} \perp (Y^{t-\ell-1}, X^{t-\tau_1-1}) \mid \mathcal{X}_t$, as WL show. This memory condition is often plausible, as it says that the more distant history $(Y^{t-\ell-1}, X^{t-\tau_1-1})$ is not predictive for $U_{1,t}$, given the more recent history $\mathcal{X}_t$ of $(Y^{t-1}, X^{t+\tau_2})$. Note that separability is not needed here.

The details of C_0 can be involved, especially with choices like $\psi_1(\hat{\varepsilon}_t) = |\hat{\varepsilon}_t|$. But this is a standard $m-$estimation setting, so we can avoid explicit estimation of C_0: suitable bootstrap methods deliver valid critical values, even without the martingale difference property (see, e.g., Gonçalves, S. and H. White, 2004; Kiefer, N. and T. Vogelsang, 2002, 2005; Politis, D. N., 2009).

An especially appealing method is the *weighted bootstrap* (Ma, S. and M. Kosorok, 2005), which works under general conditions, given the martingale difference property. To implement this, for $i = 1, \ldots, n$ generate sequences $\{\mathcal{W}_{t,i}, t = 1, \ldots, T\}$ of IID positive scalar weights with $E(\mathcal{W}_{t,i}) = 1$ and $\sigma^2_{\mathcal{W}} := var(\mathcal{W}_{t,i}) = 1$. For example, take $\mathcal{W}_{t,i} \sim \chi^2_1/\sqrt{2} + (1 - 1/\sqrt{2})$, where χ^2_1 is chi-squared with one degree of freedom. The weights should be independent of the sample data and of each other. Then compute estimators $\hat{\theta}_{T,i}$ by weighted least squares applied to the GN and CE regressions using (the same) weights $\{\mathcal{W}_{t,i}, t = 1, \ldots, T\}$. By Ma, S. and M. Kosorok (2005, theorem 2), the random variables

$$\sqrt{T}(\hat{\theta}_{T,i} - \hat{\theta}_T), \quad i = 1, \ldots, n$$

can then be used to form valid asymptotic critical values for testing hypotheses about θ_0.

To test CE, we test $\beta_{2,0} = 0$. This is a restriction of the form $\mathbb{S}_2\, \theta_0 = 0$, where $\mathbb{S}_2$ is the selection matrix that selects the elements $\beta_{2,0}$ from θ_0. Thus, to conduct an asymptotic level α test, we can first compute the test statistic, say

$$\mathcal{T}_T := T\, \hat{\theta}_T' \, \mathbb{S}_2' \, \mathbb{S}_2 \, \hat{\theta}_T = T \hat{\beta}_{2,0,T}' \hat{\beta}_{2,0,T}.$$

[4] The regularity conditions include plausible memory and moment requirements, together with certain smoothness and other technical conditions.

We then reject CE if $\mathcal{T}_T > \hat{c}_{T,n,1-\alpha}$, where, with n chosen sufficiently large, $\hat{c}_{T,n,1-\alpha}$ is the $1-\alpha$ percentile of the weighted bootstrap statistics

$$\mathcal{T}_{T,i} := T\,(\hat{\theta}_{T,i} - \hat{\theta}_T)'\,\mathbb{S}_2'\,\mathbb{S}_2\,(\hat{\theta}_{T,i} - \hat{\theta}_T) = T\,(\hat{\beta}_{2,0,T,i} - \hat{\beta}_{2,0,T})'\,(\hat{\beta}_{2,0,T,i} - \hat{\beta}_{2,0,T}), \quad i = 1,\dots,n.$$

This procedure is asymptotically valid, even though $\mathcal{T}_T$ is based on the "unstudentized" statistic $\mathbb{S}_2\,\hat{\theta}_T = \hat{\beta}_{2,0,T}$. Alternatively, one can construct a studentized statistic

$$\mathcal{T}_T^* := T\,\hat{\theta}_T'\,\mathbb{S}_2'\,[\mathbb{S}_2\,\hat{C}_{T,n}\,\mathbb{S}_2']^{-1}\,\mathbb{S}_2\,\hat{\theta}_T,$$

where $\hat{C}_{T,n}$ is an asymptotic covariance estimator constructed from $\sqrt{T}(\hat{\theta}_{T,i} - \hat{\theta}_T)$, $i = 1,\dots,n$. The test rejects CE if $\mathcal{T}_T^* > c_{1-\alpha}$, where $c_{1-\alpha}$ is the $1-\alpha$ percentile of the chi-squared distribution with $\dim(\beta_{0,2})$ degrees of freedom. This method is more involved but may have better control over the level of the test. WL provide further discussion and methods.

Because the given asymptotic distribution is joint for $\hat{\theta}_{1,T}$ and $\hat{\theta}_{2,T}$, the same methods conveniently apply to testing GN, i.e., $\beta_{1,0} = \mathbb{S}_1\,\theta_0 = 0$, where $\mathbb{S}_1$ selects $\beta_{1,0}$ from θ_0. In this way, GN and CE test statistics can be constructed at the same time.

WL discuss three examples, illustrating tests for direct structural non-causality based on tests of Granger non-causality and conditional exogeneity. A matlab module, *testsn*, implementing the methods described here is available at `http://ihome.ust.hk/~xunlu/code.htm`.

7. Summary and Concluding Remarks

In this paper, we explore the relations between direct structural causality in the settable systems framework and direct causality in the PCM for both recursive and non-recursive systems. The close correspondence between these concepts in recursive systems and the equivalence between direct structural causality and $G-$causality established by WL enable us to show the close linkage between $G-$causality and PCM notions of direct causality. We apply WL's results to provide straightforward practical methods for testing direct causality using tests for Granger causality and conditional exogeneity.

The methods and results described here draw largely from work of WC and WL. These papers contain much additional relevant discussion and detail. WC provide further examples contrasting settable systems and the PCM. Chalak, K. and H. White (2010) build on WC, examining not only direct causality in settable systems, but also notions of indirect causality, which in turn yield implications for conditional independence relations, such as those embodied in conditional exogeneity, which plays a key role here. WL treat not only the structural VAR case analyzed here, but also the "time-series natural experiment" case, where causal effects of variables D_t, absorbed here into Z_t, are explicitly analyzed. The sequence $\{D_t\}$ represents external stimuli, not driven by $\{Y_t\}$, whose effects on $\{Y_t\}$ are of interest. For example, $\{D_t\}$ could represent passively observed visual or auditory stimuli, and $\{Y_t\}$ could represent measured neural activity. Interest may attach to which stimuli directly or indirectly affect which neurons or groups of neurons. WL also examine the structural content of classical Granger causality and a variety of related alternative versions that emerge naturally from different versions of Assumption A.1.

Acknowledgments

We express our deep appreciation to Sir Clive W.J. Granger for his encouragement of the research underlying the work presented here.

References

E. Candès. Ridgelets: Estimating with Ridge Functions. *Annals of Statistics*, 31:1561–1599, 1999.

K. Chalak and H. White. Causality, Conditional Independence, and Graphical Separation in Settable Systems. Technical report, Department of Economics, Boston College, 2010.

A. P. Dawid. Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society, Series B*, 41:1–31, 1979.

A. P. Dawid. Beware of the DAG! *Proceedings of the NIPS 2008 Workshop on Causality, Journal of Machine Learning Research Workshop and Conference Proceedings*, 6:59–86, 2010.

M. A. Delgado and W. Gonzalez-Manteiga. Significance Testing in Nonparametric Regression Based on the Bootstrap. *Annals of Statistics*, 29:1469–1507, 2001.

M. Eichler. Granger Causality and Path Diagrams for Multivariate Time Series. *Journal of Econometrics*, 137:334-353, 2007.

M. Eichler and V. Didelez. Granger-causality and the Effect of Interventions in Time Series. *Lifetime Data Analysis*, forthcoming.

R. Engle, D. Hendry, and J.F. Richard. Exogeneity. *Econometrica* 51:277–304, 1983.

M. Fernandes and R. G. Flores. Tests for Conditional Independence, Markovian Dynamics, and Noncausality. Technical report, European University Institute, 2001.

J.P. Florens and D. Fougère. Non-causality in Continuous Time. *Econometrica*, 64:1195–1212, 1996.

J.P. Florens and M. Mouchart. A Note on Non-causality. *Econometrica*, 50:583–591, 1982.

A.R. Gallant and H. White. On Learning the Derivatives of an Unknown Mapping with Multilayer Feedforward Networks. *Neural Networks*, 5:129–138, 1992.

D. Galles and J. Pearl. Axioms of Causal Relevance. *Artificial Intelligence*, 97:9–43, 1997.

R. Gibbons. Game Theory for Applied Economists. Princeton University Press, Princeton, 1992.

S. Gonçalves and H. White. Maximum Likelihood and the Bootstrap for Nonlinear Dynamic Models. *Journal of Econometrics*, 119:199–219, 2004.

C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37:424–438, 1969.

C. W. J. Granger and P. Newbold. Forecasting Economic Time Series (2nd edition). Academic Press, New York, 1986.

J. Halpern. Axiomatizing Causal Reasoning. *Journal of Artificial Intelligence Research*, 12:317-337, 2000.

K. Hornik, M. Stinchcombe, and H. White. Universal Approximation of an Unknown Mapping and its Derivatives Using Multilayer Feedforward Networks. *Neural Networks*, 3:551–560, 1990.

G. B. Huang, Q.Y. Zhu, and C.K. Siew. Extreme Learning Machines: Theory and Applications. *Neurocomputing*, 70:489–501, 2006.

M. Huang and H. White. A Flexible Test for Conditional Independence. Technical report, Department of Economics, University of California, San Diego.

N. Kiefer and T. Vogelsang. Heteroskedasticity-autocorrelation Robust Testing Using Bandwidth Equal to Sample Size. *Econometric Theory*, 18:1350–1366, 2002.

N. Kiefer and T. Vogelsang. A New Asymptotic Theory for Heteroskedasticity-autocorrelation Robust Tests. *Econometric Theory*, 21:1130–1164, 2005.

O. Linton and P. Gozalo. Conditional Independence Restrictions: Testing and Estimation. Technical report, Cowles Foundation for Research, Yale University, 1997.

S. Ma and M. Kosorok. Robust Semiparametric M-estimation and the Weighted Bootstrap. *Journal of Multivariate Analysis,* 96:190-217, 2005.

J. Pearl. Causality. Cambridge University Press, New York, 2000.

J. Pearl. Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411-420, 2001.

D.N. Politis. Higher-Order Accurate, Positive Semi-definite Estimation of Large-Sample Covariance and Spectral Density Matrices. Technical report, Department of Economics, University of California, San Diego, 2009.

A. Roebroeck, A.K. Seth, and P. Valdes-Sosa. Causality Analysis of Functional Magnetic Resonance Imaging Data. *Journal of Machine Learning Research*, (this issue), 2011.

K. Song. Testing Conditional Independence via Rosenblatt Transforms. *Annals of Statistics*, 37:4011-4015, 2009.

M. Stinchcombe and H. White. Consistent Specification Testing with Nuisance Parameters Present Only Under the Alternative. *Econometric Theory*, 14:295-324, 1998.

J. Stock and M. Watson. Introduction to Econometrics. Addison-Wesley, Boston, 2007.

L. Su and H. White. A Consistent Characteristic Function-Based Test for Conditional Independence. *Journal of Econometrics*, 141:807-834, 2007a.

L. Su and H. White. Testing Conditional Independence via Empirical Likelihood. Technical report, Department of Economics, University of California, San Diego, 2007b.

L. Su and H. White. A Nonparametric Hellinger Metric Test for Conditional Independence. *Econometric Theory*, 24:829–864, 2008.

H. Varian. Intermediate Microeconomics (8th edition). Norton, New York, 2009.

H. White. Estimation, Inference, and Specification Analysis. Cambridge University Press, New York, 1994.

H. White. Time-series Estimation of the Effects of Natural Experiments. *Journal of Econometrics*, 135:527-566, 2006a.

H. White. Approximate Nonlinear Forecasting Methods. In G. Elliott, C.W.J. Granger, and A. Timmermann, editors, *Handbook of Economic Forecasting*, pages 460–512, Elsevier, New York, 2006b.

H. White and K. Chalak. Settable Systems: An Extension of Pearl's Causal Model with Optimization, Equilibrium, and Learning. *Journal of Machine Learning Research*, 10:1759-1799, 2009.

H. White and P. Kennedy. Retrospective Estimation of Causal Effects Through Time. In J. Castle and N. Shephard editors, *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, pages 59–87, Oxford University Press, Oxford, 2009.

H. White and X. Lu. Granger Causality and Dynamic Structural Systems. *Journal of Financial Econometrics*, 8:193-243, 2010a.

H. White and X. Lu. Causal Diagrams for Treatment Effect Estimation with Application to Selection of Efficient Covariates. Technical report, Department of Economics, University of California, San Diego, 2010b.