Alessio Moneta
(Scuola Superiore Sant'Anna)

Causal inference in empirical economics:
How to cope with complexity?

This paper is concerned with the issue of whether it is possible to search for valid and reliable causal models on the basis of observational data when the phenomena under investigation are characterized by complexity features such as the potential influence of many interacting variables, nonlinearity, feedbacks or the capacity of adaptation. These features are often exhibited by economic phenomena addressed, for example, in the study of the business cycles, industrial dynamics or the stock market. Certainly, causal inference starting from observational data is a difficult and debated task even in a very simple setting, as the one in which there are few variables interacting in a linear way and the underlying probability distribution is normal. But a strand of literature developed at the intersection between philosophy of science, statistics and computer science, pioneered by the publications of P. Spirtes, C. Glymour and R. Scheines (1993, 2000) and J. Pearl (2000), has created a rigorous mathematical framework for causal inference. "Graphical Causal Models" have been developed in this framework which permit the researcher to analyze the problem of causal inference as the problem of recovering a good approximation of the structure of the data generating process. Important features of the generative model are recovered under precise assumptions about (i) the probability distribution of the data (e.g. normality), (ii) the restrictions that the causal structure imposes on the probability distribution (e.g. Causal Markov condition) and (iii) the structure of the causal relationships (e.g. causal sufficiency or acyclicity). Not all of these conditions can be tested vis-à-vis the data, so that considerations about the compatibility of the assumptions with substantive knowledge (e.g. about the possible economic mechanism) play an important role here. The machine learning literature has developed asymptotically reliable algorithms that search for observational equivalent features of the generative model under a variety of conditions about (i), (ii) and (iii). The first step of such algorithms typically consists of tests of conditional independence or analysis of independent components. In this paper I analyze the possibilities that causal search methods developed in this framework offer for analyzing economic phenomena characterized by complex features such as those mentioned above. I focus my analysis on three features. The first is the case in which the underlying probability distribution among the variables of interest is non-Gaussian. The second is the case in which the causal interactions are nonlinear. The third is the case in which there are many interacting variables at the micro level but the system can be described as the co-movements of few macro variables that are aggregates (i.e. sums or averages) of sets of micro-variables. I will argue that if only one of the three features is present, the available causal search algorithms are able to cope with that. If the setting under study exhibits two out of the three features (e.g. non-Gaussianity and aggregation or, alternatively, nonlinearity and aggregation) it is still possible to envisage asymptotically reliable causal search algorithms. These algorithms, however, are going to be practically difficult to be implemented because of the enormous quantity of data needed to get consistent and efficient tests about conditional independence. If all the three features are present, it is not known at the present state of the art whether reliable causal search algorithms can be designed.

I will argue that from a certain level of complexity onwards (for example in the setting which simultaneously includes the three aforementioned cases) an alternative route can be taken to analyze the generative causal model. Agent-Based Models are simulation models in which the interactions of autonomous agents are investigated with the aim of analyzing the effects of the interactions on the system as a whole. This modeling framework has gained increased popularity in the recent years but the issue about the empirical reliability of such Agent-Based Models remains an open problem, because there are many ways of calibrating a simulation models, and alternative models can equally well replicate the data, so that there is no simple way of adjudicating among competitive models using observational data. A way to address the problem of empirical validation, I argue, is to find approximation (e.g. linearization) of both the simulated and real data generating process, so that causal search algorithms can be applied to both sets of data and their outputs can be confronted. I conclude proposing a taxonomy of the different settings encountered in economic analysis from the most simple to the most complex. While it is hard to find a solution for causal inference in each setting, for many of them I propose possible routes which involve methods taken from the causal model search literature.