NeuroImage xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Comments and Controversies

Counterfactuals, graphical causal models and potential outcomes: Response to Lindquist and Sobel☆

Clark Glymour

Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213, Florida Institute for Human and Machine Cognition, Pensacola, Florida, 32507, USA

ARTICLE INFO

Article history:
Received 9 March 2011
Revised 28 June 2011
Accepted 22 July 2011
Available online xxxx

Keywords:
fMRI
Counterfactuals
Graphical causal models
Causal estimation
Potential outcomes

ABSTRACT

Lindquist and Sobel claim that the graphical causal models they call "agnostic" do not imply any counterfactual conditionals. They doubt that "causal effects" can be discovered using graphical causal models typical of SEMs, DCMs, Bayes nets, Granger causal models, etc. Each of these claims is false or exaggerated. They recommend instead that investigators adopt the "potential outcomes" framework. The potential outcomes framework is an obstacle rather than an aid to discovering causal relations in fMRI contexts.

In response to Ramsey et al. (2011a, 2011b), Lindquist and Sobel (in press) (LS) appear to make three claims and a recommendation. The claims are that (1) "causal effects" cannot be found by methods associated with a variety of directed graph representations of causal relations, including SEM, Granger causal models and Dynamic Causal Models (DCMs), all of which they doubt are "generally useful for 'finding causal effects' or estimating causal effects"; (2) the theory of graphical causal models developed by Spirtes et al. (1993) makes no counterfactual claims; and (3) that causal relations cannot be determined non-experimentally from samples that are a combination of systems with different propensities. Their recommendation is that fMRI researchers adopt the "potential outcomes" framework. Of these claims, (1) is mere assertion on unspecified grounds; (2) is false; (3) is false as a generalization, and distinguishing the cases in which it is true from those in which it is false is part of what is done in the paper to which LS respond. For empirical inquiry with large numbers of variables whose causal connections are unknown and with limited experimental control of the processes to be understood, the potential outcomes framework is an obstacle, not an aid, to discovery.

(1). The search for causal explanations of sample data is a form of statistical estimation. In statistical estimation one has a space of alternative hypotheses (in point-valued estimation of continuous parameters the set of alternatives is uncountably infinite). Each hypothesis determines a sampling distribution or set of sampling distributions. An estimator is a function from samples to subsets of the hypothesis space. Estimators are sought with various virtues, notably convergence (in one or another sense, under various assumptions) to the true hypothesis in the large sample limit. An estimator may have various finite sample error properties (e.g., unbiased). The asymptotic and finite sample properties of estimators under distribution assumptions are characterized theoretically, or nowadays estimated by simulation. One important facet of statistical research has been to develop new estimators for special problems with special assumptions about the distribution family.

Causal estimation has the very same structure. Causal relations are represented by a directed graph, or equivalently by a connection matrix, or set of such matrices (as for example, with time dependent structures). The space of alternative hypotheses is a set of such matrices together with various parameterizations of the connections among variables; the parameterizations transform each abstract graph into a statistical model. The statistical model determines a sampling distribution or set of sampling distributions. The goal of causal inference is usually to estimate features of the true connection matrix, and possibly parameter values for an associated statistical model. A causal estimator is just a function from sample data to a subset of connection matrices or graphs. Exactly as with conventional parameter estimation, the properties of estimators can be demonstrated mathematically or estimated by simulation. Exactly as with conventional statistics, the consistency and error properties of estimators will depend on the hypothesis space. And exactly as with conventional parameter estimation, research in causal inference consists in part in adapting estimators to special problems.

Causal estimation from fMRI data poses a very special, very interesting and difficult problem: the data are indirect, noisy, aggregated measurements of non-linear feedback processes, and the variables—the ROIs—are often built in whole or in part out of the sample data. Notwithstanding, the construction of estimators for causal structure from fMRI data is improving rapidly. For example, Smith et al. (2011)

☆ Research for this paper was supported by a grant from the James S. McDonnell Foundation.

E-mail address: cg09@andrew.cmu.edu.

1053-8119/$ – see front matter © 2011 Published by Elsevier Inc.
doi:10.1016/j.neuroimage.2011.07.071

have recently simulated fMRI data for a number of simple structures under a variety of realistic conditions on noise, measurement error and length of recording, and in a few unrealistic conditions (nearly deterministic systems; very small effects; canceling feedback). Work in press (Ramsey et al., 2011a) describes methods that recover the adjacencies in the graphs generating the Smith et al. data in simulated realistic conditions with nearly 100% precision and recall; the methods identify directions in most of the data-generating causal models with precision and recall ranging between 80 and 95%. And, contrary to some commentators, consistent causal estimators are available for classes of feedback systems or for cyclic graphs that represent them. Continuing research will undoubtedly improve on current causal estimators.

The potential outcomes framework, now standard in statistics, is essentially a special case of the graphical causal model framework but with twists that make causal estimation impossible except in very restricted contexts. That framework was developed and tailored for experimental trials with a small number of variables where the concern is to estimate the effect of a treatment, or treatment assignment, on an observed outcome variable in circumstances in which there is a great deal of prior information about which recorded variables are and are not causes of other variables. These are not the usual circumstances of fMRI research.

Applications of the PO framework seem to assume that (1) most of the causal relations are known; (2) that the causal ("treatment") variables are categorical; (3) that the number of actual variables is quite small; and (4) that the variables whose causal relations are of interest are directly measured. None of this is true of many fMRI studies. In order to apply the PO framework in a concrete case, one must know which variables are potentially direct causes of which others, and which variables cannot be direct causes of others. Applying the PO framework thus presupposes exactly what is not known in fMRI contexts (and many other scientific contexts). That may help to explain why provably consistent search methods relevant to fMRI are not available for PO models. It may also explain why Lindquist and Sobel disparage the very possibility of systematic scientific search despite long-standing proofs of the existence of consistent estimators of causal relations in well-defined and testable circumstances, and despite any number of simulation and empirical examples of successful applications of such estimators (Spirtes, et al., 2010).

Point (2). The implications for experimental manipulations that may or may not have ever been done are what make a causal model *causal*. Claims about what the outcome would be of a hypothetical experiment that has not been done are one form of *counterfactual* claims. They say that if such and such were to happen then the result would be thus and so—where the such and such has not happened or has not yet happened. (Of course, if the experiment is later done, then the proposition becomes factually true or factually false.) Thus it is a very serious charge to say, as LS do, that the graphical model framework does not represent or entail any counterfactual claims. The charge is quite false. The systematization of the connection between graphical representations of causal relations and predictions of outcomes of experimental interventions has a long history, but its non-parametric form was inaugurated in Spirtes et al., 1993. Extending those results, Pearl (2000) developed a complete algorithm for computing when an acyclic graphical causal model implies a testable prediction and for estimating the predicted effect when that is possible. LS say that Spirtes et al. (1993) do not consider counterfactuals, and indeed the word "counterfactual" is not used in that book. Which only illustrates that before drawing conclusions about the content of a work one should read more than the index.

The potential outcomes framework posits a set of "counterfactual variables"—each variable, X, that is a relatively direct effect of a variable, Y, has a shadow counterfactual variable X(y) for each possible assignment of a value y to Y. A joint probability distribution is introduced over values of the actual and the counterfactual variables. That joint distribution permits the formal expression of a variety of counterfactual relations that are not defined in what LS call the "agnostic" framework of graphical causal models that LS attribute to Spirtes et al. (1993). For example, PO assumes the following is well defined: "the joint probability of Y were X is forced to have value 1 and of Y were X forced have the value 0, given that actually $X = 1$." This is the sort of quantity denoted in potential outcomes notation as $f(Y(1),Y(0)|X = 1)$. In contrast, in the same circumstance the "agnostic" graphical causal model framework only defines "the probability distribution of Y were X forced to equal 1, given that X actually equals 1," or in PO notation $f(Y(1)|X = 1)$, and "the probability distribution of Y were X forced to equal 0, given that X actually equals 1," or in PO notation, $f(Y(0)|X = 1)$. But I emphasize that no counterfactual *variables* are used or needed in the graphical causal model framework. In the potential outcomes framework, if nothing is known about which of many variables are causes of the others, then for each variable, and for each value of the other variables, a new counterfactual variable is required. In practice that would require an astronomical number of counterfactual variables for even a few actual variables.

Point (3). In what appears to be intended as a criticism of the possibility of recovering causal structure from observational data, LS sketch an example in which the sample is a mixture of units with different propensities for an effect, i.e., different probability distributions. When the sample is a mixture of units with differing causal structures and/or probability distributions, predictions about the effects of an experimental distribution may still hold, not for every individual in the population but for the distribution that would result if the intervention were to be applied to the entire population. In the paper to which LS respond, Ramsey et al. (2011a, 2011b) provide a general theory of when that is possible. LS ignore the general theory in favor of sketching an example that fails to distinguish prediction of individual or sub-group effects from prediction of population effects.

References

Lindquist, M. and M. Sobel (2010). Graphical models, potential outcomes and causal inference: Comment on Ramsey, Spirtes and Glymour. NeuroImage in press.

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

Ramsey, J., Spirtes, P., Glymour, C., 2011a. On meta-analyses of imaging data and the mixture of records. NeuroImage. doi:10.1016/j.neuroimage.2010.07.065.

Ramsey, J., Hanson, S., Glymour, C., 2011b. Multi-subject search correctly identifies causal connections and most causal directions in the DCM models of the Smith et al. Simulation Study. NeuroImage. doi:10.1016/j.neuroimage.2011.06.068.

Spirtes, P., Glymour, C., Scheines, R., 1993. Causation, Prediction and Search, Springer Lecture Notes in Statistics, New York, 2nd edition. MIT Press, Cambridge, MA.

Spirtes, P., Glymour, C., Scheines, R., Tillman, R., 2010. Automated search for causal relations: theory and practice. In: Dechter, R., Geffner, H., Halpern, J. (Eds.), Heuristics, Probability and Causality: Honoring Judea Pearl. College Publications (No place of publication given), Chapter 27.

Please cite this article as: Glymour, C., , NeuroImage (2011), doi:10.1016/j.neuroimage.2011.07.071