Automatic Inference of the Contemporaneous Causal Order of a System of Equations

Kevin D. Hoover

Department of Economics
University of California
One Shields Avenue
Davis, California 95616-8578

Tel. (530) 752-2129
Fax (530) 752-9382
E-mail kdhoover@ucdavis.edu

First Draft, 12 February 2004

Automatic Inference of the Contemporaneous Causal Order of a System of Equations

When Stephen Perez and I first began our Monte Carlo studies of the efficacy of general-to-specific search methodologies in 1995, we were keenly aware of our limited ability to capture the tacit knowledge of the skilled time-series econometrician operating in the LSE tradition (Hoover and Perez 1999a, b). Econometrics, we believed, was an art, and our algorithm was not intended to replace the artist. David Hendry and Hans-Martin Krolzig's subsequent development of *PcGets* did not, in fact, eliminate the art of econometrics. Power tools did not eliminate the art of the cabinetmaker, but changed where his value-added lay and – importantly – made new things possible. *PcGets* is likewise a new, powerful tool, useful in the hands of skilled craftsman.

But no tool solves every problem. One open problem is briefly touched on in Hendry's answer to Question 16:

> When the reduced-form VAR has a diagonal covariance matrix, then all possible reductions of the system can be efficiently estimated by OLS, and model-selection procedures can operate equation-by-equation without any loss of efficiency. For a structural VAR (SVAR), with a recursive specification as in Wold (1949), a similar result holds for OLS being efficient.

The suggestion is that, if a recursive (or Wold causal order) is known for the contemporaneous variables in the SVAR, then *PcGets* can be applied equation by equation to find a parsimonious lag structure. But where is such knowledge to come from?

The SVAR can be written as:

$$\mathbf{A}_0\mathbf{Y}_t = \mathbf{A}(L)\mathbf{Y}_{t-1} + \mathbf{E}_t , \tag{1}$$

where $\mathbf{Y}_t$ is an $n \times 1$ vector of contemporaneous variables, $\mathbf{A}_0$ is an $n \times n$ matrix with ones on the main diagonal and possibly non-zero off-diagonal elements; $\mathbf{A}(L)$ is a polynomial

in the lag operator, L; and $\mathbf{E}_t$ is an $n \times 1$ vector of error terms with $\mathbf{E} = [\mathbf{E}_t]$, $t = 1, 2, \ldots,$ T and the covariance matrix $\mathbf{\Sigma} = \mathrm{E}(\mathbf{E}\mathbf{E}')$ diagonal. The individual error terms (shocks) can be assigned unequivocally to particular equations because $\mathbf{\Sigma}$ is diagonal. The matrix $\mathbf{A}_0$ defines the causal interrelationships among the contemporaneous variables. The system is identified provided that there are $n(n-1)/2$ zero restrictions on $\mathbf{A}_0$.[1] For any just-identified system, $\mathbf{A}_0$ can be rendered lower triangular by selecting the appropriate order of the variables $\mathbf{Y}$ along with the conformable order the rows of $\mathbf{A}_0$. This is the *recursive* (or *Wold causal*) *order*.

Starting with the SVAR as the data-generating process, premultiplying by $\mathbf{A}_0^{-1}$ yields the reduced-form or VAR:

$$\mathbf{Y}_t = \mathbf{A}_0^{-1}\mathbf{A}(L)\mathbf{Y}_{t-1} + \mathbf{A}_0^{-1}\mathbf{E}_t = \mathbf{B}(L)\mathbf{Y}_{t-1} + \mathbf{U}_t . \tag{2}$$

If we *know* $\mathbf{A}_0$, then recovery of the SVAR from the easily estimated VAR is straightforward. There are, however, a large number of $n \times n$ matrices, $\mathbf{P}_i$ that may be used to premultiply equation (2) such that the covariance matrix $\mathbf{\Omega} = E(\mathbf{P}_i^{-1}\mathbf{U}(\mathbf{P}_i^{-1}\mathbf{U})')$ is diagonal. Let $\mathbf{P} = \{\mathbf{P}_i\}$ be the set of all such orthogonalizing transformations.

For any causal ordering of $\mathbf{Y}$, there is a unique lower triangular $\mathbf{P}_i \in \mathbf{P}$ such that $\mathbf{P}_i\mathbf{P}_i' = \mathbf{\Omega}$. This is the Choleski decomposition of the covariance matrix and corresponds to a Wold causal ordering of the variables. Since the ordering of the variables in $\mathbf{Y}$ is arbitrary, there are as many such orderings as there are permutations of the elements of $\mathbf{Y}$. Each such ordering is just-identified and, therefore, observationally equivalent. There are also other overidentified causal orderings – that is $\mathbf{P}_i$ for which there are more than $n(n-1)/2$ zero restrictions.

[1] I concentrate here on zero restrictions, although SVARs are sometimes identified in other ways.

The central identification problem for SVARs is to choose the one member of **P** that corresponds to the data-generating process: that is, to choose $\mathbf{P}_i = \mathbf{A}_0$ when $\mathbf{A}_0$ is unknown. The other elements can be thought of as defining pseudo-SVARs. But on what basis should we choose? There are at least two options. First, we can appeal to economic theory to tell us what the causal order should be. This is, in fact, what almost all practitioners of VAR methodologies do. Unfortunately, formal economic theory is rarely decisive about causal order. In reality, VAR practitioners follow one of two strategies: They choose the order arbitrarily, sometimes with an accompanying claim that their results are robust to alternative causal orderings – apparently unaware that such robustness really amounts to a claim that the contemporaneous terms do no real work at all, so that causal order is irrelevant. Sometimes they appeal, not so much to theory, as to "just so" stories. Intuition or commonsense tells them that, say, financial markets adjust more quickly than real markets, so that interest rates, for instance, ought to be causally ordered ahead of real GDP. It is usually easy, however, to tell a "just so" story to justify most any order – the time order of variables that are contemporaneously related at the given frequency of observation being especially unreliable. There is a special irony that this strategy should be so commonly accepted among VAR practitioners. After all, Sims's (1980) motivation in initiating the VAR program was to avoid the need to appeal to "incredible identifying restrictions."

A second method of choosing $\mathbf{P}_i$ is to try to extract more information out of the data. Graph-theoretic causal search is an approach (really a family of approaches) to this problem very much in the spirit of general-to-specific model selection. (Glymour, Spirtes, and Scheines 2000 and Pearl 2000 provide the most developed accounts of the

approach.) In a causal graph, arrows connecting causal variables to their effects represent causal relationships. The mathematics of graph theory can be used to analyze the causal structures. Importantly, it can be shown that there are isomorphisms between graphs and the probability distributions of variables. In particular, certain graphical patterns imply conditional independence and dependence relationships among the variables. The graph of the DGP can also be represented through the restrictions on $\mathbf{A}_0$. Working backwards from statistical measures of conditional independence and dependence, it is possible to infer the class of graphs compatible with the data. Sometimes that class has only a single member, and then $\mathbf{A}_0$ can be identified statistically.

The key idea of the graph-theoretic approach is simple. Suppose that $A \rightarrow B \rightarrow C$ (that is, A causes B causes C). A and C would be correlated, but conditional on B, they would be uncorrelated. Similarly for $A \leftarrow B \leftarrow C$. In each case, B is said to *screen* A from C. Suppose that $A \leftarrow B \rightarrow C$. Then, once again A and C would be correlated, but conditional on B, they would be uncorrelated. B is said to be the *common cause* of A and C. Now suppose that A and B are conditionally uncorrelated, $A \rightarrow C \leftarrow B$, and none of the variables that cause A or B directly cause C. Then, conditional on C, A and B are conditionally correlated. C is called an *unshielded collider* on the path ACB. (A *shielded* collider would have a direct link between A and B.)

Causal search algorithms use a statistical measure of independence, commonly a measure of conditional correlation to systematically check the patterns of conditional independence and dependence and to work backwards to the class of admissible causal

structures.[2] The PC algorithm is the most commonly used in the literature (Glymour *et al.* 2000, pp. 84-85, Pearl 2000, pp. 49-51, Cooper 1999, p. 45, figure 22).[3] It assumes that graphs are *acylical* – that is, there are no loops in causal chains such that an effect feeds back onto a direct or indirect cause. Acyclicality rules out simultaneous equations. There are six steps:

1. Start with a graph in which each variable is assumed to be connected by an undirected causal link.
2. Test for the unconditional correlation of each pair of variables, eliminating the link in the graph whenever the absence of correlation cannot be rejected.
3. Test for the correlation of each pair of variables conditional on a third variable, again eliminating the link if correlation is absent. Continue testing pairs conditional on pairs, triples, quadruples, and so on until the graph is pared down as far as the data permit.
4. For each conditionally uncorrelated pair of variables (i.e., ones without a direct link) that are connected through a third variable, test whether they become correlated conditional on that third variable. If so, the third variable is an unshielded collider. Orient the links as pointing into the unshielded collider.
5. If there are any pairs A and C that are not directly connected but are linked $A \rightarrow B$ — C, then orient the second link toward C, so that the triple is $A \rightarrow B \rightarrow C$.
6. If there is a pair of variables, A and B connected both by an undirected link and a directed path, starting at A, through one or more other variables to B (i.e., a path

[2] Absence of conditional correlation is a necessary, but not sufficient, condition for statistical independence.
[3] The name "PC algorithm" derives from the names of its authors, Peter and Clark (Pearl 2000, p. 50)

in which the arrows all orient in a chain), then orient the undirected link as $A \rightarrow B$.

Steps 1-4 are based in statistical inference. Step 5 follows logically, because orienting the undirected link in the other direction would turn the pattern into an unshielded collider, which would have already been identified in Step 4. Step 6 follows because orienting the undirected link in the other direction would, contrary to assumption, render the graph cyclical.

The PC algorithm can be illustrated with the example in Figure 1. Panel A shows the graph of the data-generating process (DGP). It determines just what the tests should find, small-sample problems to one side. The graph corresponds to a particular matrix $\mathbf{A}_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ a_{31} & a_{32} & 1 & 0 \\ a_{41} & 0 & a_{43} & 1 \end{bmatrix}$, where the variables are ordered *WXYZ*, the rows correspond to effects and the columns to causes, and the a_{ij} to the non-zero elements.

Step 1 starts with panel B in which all the variables are connected. It is analogous to the general unrestricted model (GUM) of *PcGets*. Step 2 eliminates link 1, because *W* and *X* are unconditionally uncorrelated in the DGP. Step 3 eliminates link 5 (*X* and *Z* are uncorrelated conditional on *Y*). Step 4 orients links 3 and 4 toward *C* (*W* and *X* are correlated conditional on *Y* – i.e., *Y* is an unshielded collider on *WYX*). Step 5 orients link 6 towards *Z*. Step 6 orients link 2 toward *Z*. The algorithm is able to recover the DGP.

Not every DGP can be recovered uniquely. A graph and a probability distribution are *faithful* when the independence relationships in the graph stand in one-to-one correspondence with those implied by the probability distribution. The *skeleton* of a

graph is the pattern of its causal linkages ignoring their direction. The *observational equivalence theorem* (Pearl 2000, p. 19, Theorem 1.2.8) states that any probability distribution that can be faithfully represented by an acyclical graph, can equally well be represented by another acyclical graph with the same skeleton and the same unshielded colliders. A graph identical to panel A of Figure 1 except that link 6 was reversed would not be observationally equivalent to panel A because it would add an unshielded collider (*Y* on *XYZ*). A graph that reversed link 2 would be observationally equivalent to the graph in panel A because it would have the same skeleton and neither add nor subtract unshielded colliders. While the graph with link 2 reversed illustrates the observational equivalence theorem, Step 6 of the algorithm rules it out, since it possess a cycle ($W \rightarrow Y \rightarrow Z \rightarrow W$), which violates the antecedent of the observational equivalence theorem.

Swanson and Granger (1997) were the first to introduce graph-theoretic search into the analysis of the contemporaneous causal order of VARs. Swanson and Granger restrict the class of orderings to causal chains – that is, to orders in which the matrix $\mathbf{A}_0$ is diagonal. Graph-theoretic methods were generally not conceived with time-series data in mind. Granger and Swanson realized that the relevant information for the contemporaneous causal ordering of the SVAR is actually contained in the covariance matrix of the VAR error terms in equation (2). They estimate to (2) and calculate $\hat{\Omega}$, from which all the conditional correlations needed by the search algorithm can be calculated. Demiralp, Selva. (2000), Bessler and Lee (2002), Moneta (2003) and Demiralp and Hoover (2004) have extended their strategy to the less restricted class of structures compatible with the PC algorithm. Demiralp and Hoover (2004) provide

Monte Carlo evidence that shows that the PC algorithm is highly effective at recovering the skeleton of the DGP graph, moderately effective at recovering the directions of individual links provided that signal-to-noise ratios are high enough.

The observational equivalence theorem implies that some structures cannot in principle be recovered. For example, if the DGP really displays a Wold causal order ($\mathbf{A}_0$ is lower triangular), then there are no unshielded colliders, so no links can be directed, and all possible Wold causal orders are observationally equivalent. Even when the DGP cannot be recovered, the class of data-admissible models will generally be narrowed. Theory may in some instances permit some links to be oriented, which may, according to Steps 5 and 6, imply other orderings. Undirected links might also be ordered by exploiting information about regime changes.[4]

Acyclical graphs are not fully adequate to economics, as so much of economics is represented in the form of simultaneous systems. Some economists, including Wold (1949) and Granger (1969) argue that there is no true simultaneity. Hoover (2001, ch. 6) argues that an adequate account of causality most permit simultaneity. Developing the causal analysis of *cyclical* graphs stands at the forefront of this research (see Pearl 2000, pp. 95-96, 142-143 and Richardson 1996).

In the meantime, a natural extension of general-to-specific single-equation modeling would be to use graph-theoretic algorithms to select the contemporaneous causal ordering of the SVAR and then to apply algorithms like *PcGets* to the individual equations.

[4] See Hoover (1990, 1991, 2001, chs. 8-10), Hoover and Sheffrin (1992), and Hoover and Siegler (2000).

References

Bessler, David A. and Seongpyo Lee. (2002). 'Money and prices: U.S. data 1869-1914 (a study with directed graphs)', *Empirical Economics*, Vol. 27, pp. 427-46.

Cooper, Gregory F. (1999). 'An overview of the representation and discovery of causal relationships using Bayesian networks', in Clark Glymour and Gregory F. Cooper (eds) *Computation, Causation, and Discovery*, American Association for Artificial Intelligence, Menlo Park, CA and MIT Press, Cambridge, MA, pp. 3-64.

Demiralp, Selva. (2000). 'The structure of monetary policy and transmission mechanism', unpublished Ph.D dissertation, Department of Economics, University of California, Davis.

Granger, C.W.J. (1969) "Investigating causal relations by econometric models and cross-spectral Methods," *Econometrica* 37(3), 424-38.

Hoover, Kevin D. (1990) "The logic of causal inference: Econometrics and the conditional analysis of causality," *Economics and Philosophy* 6(2), 207-234.

Hoover, Kevin D. (1991) "The causal direction between money and prices: An alternative approach," *Journal of Monetary Economics* 27(3), 381-423.

Hoover, Kevin D. (2001). *Causality in Macroeconomics*, Cambridge University Press, Cambridge.

Hoover, Kevin D. and Stephen J. Perez. (1999a). 'Data mining reconsidered: encompassing and the general-to-specific approach to specification search', *Econometrics Journal*, Vol. 2, pp. 167-91.

Hoover, Kevin D. and Stephen J. Perez. (1999b). "Reply to our discussants," *Econometrics Journal*, Vol. 2, pp. 244-247.

Hoover, Kevin D. and Selva Demiralp. (2004) "Searching for the causal structure of a vector autoregression," *Oxford Bulletin of Economics and Statistics* 65 (Supplement), 745-767.

Hoover, Kevin D. and Steven M. Sheffrin. (1992) "Causation, spending and taxes: Sand in the sandbox or tax collector for the welfare state?" *American Economic Review* 82(1), 225-248.

Hoover, Kevin D. and Mark Siegler "Taxing and spending in the long view: The causal structure of U.S. fiscal policy after 1791," *Oxford Economic Papers*, vol. 52, no. 4, December 2000, pp. 745-773.

Moneta, Alessio. (2003) "Graphical models for structural vector autoregressions," unpublished typescript, S. Anna School of Advanced Studies, Pisa, Italy.

Pearl, Judea. (2000). *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge.

Richardson, Thomas. (1996) "A discovery algorithm for directed cyclical graphs," in F. Jensen and E. Horwitz (eds.) *Uncertainty in Artificial Intelligence: Proceedings of the Twelfth Congress*. San Francisco: Mogan Kaufman, pp. 462-469.

Sims, Christopher A. (1980) 'Macroeconomics and reality', *Econometrica*, Vol. 48, pp. 1-48.

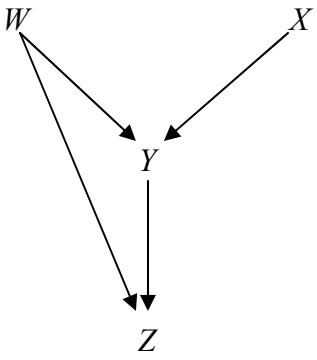
Spirtes, Peter, Clark Glymour, and Richard Scheines. (2000) *Causation, Prediction, and* Search, 2nd edition. Cambridge, MA: MIT Press.

Swanson, Norman R. and Clive W.J. Granger. (1997). 'Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions', *Journal of the American Statistical Association*, Vol. 92, pp. 357-67.

Wold, Herman O. A. (1949). "Statistical estimation of economic relationships," *Econometrica*, Vol. 17 (Supplement), pp. 1-21..

Figure 1

Panel A



Panel B