erties that will terminate as a result of F_i from those that will persist despite of acting F_i. Such a model of persistence was invoked in (Pearl, 1993b); there, it was assumed that only those properties should persist that are not under any causal influence to terminate. This assumption yields formulas for the effect of *conditional interventions* (conditioned on the observation C) which, again, given Γ, can be estimated from nonexperimental data.

A more ambitious task has been explored by Spirtes, Glymour and Scheines, (1993) – estimation of the effect of intervention when the structure of Γ is not available and must also be inferred from the data. Recent developments in graphical models (Pearl and Verma, 1991; Spirtes, Glymour and Scheines, 1993) have produced methods that, under certain conditions, permit us to infer plausible causal structures from nonexperimental data, albeit with a weaker set of guarantees than those obtained through controlled randomized experiments. These guarantees fall into two categories: minimality and stability (Pearl and Verma, 1991). Minimality guarantees that any other structure compatible with the data is necessarily more redundant, and hence less trustworthy, than the one(s) inferred. Stability ensures that any alternative structure compatible with the data must be less stable than the one(s) inferred; namely, slight fluctuations in the distributions of the disturbances ε_i (2) will render that structure no longer compatible with the data.

When the structure of Γ is to be inferred under these guarantees, the formulas governing the effects of interventions and the conditions required for estimating these effects become rather complex (Spirtes, Glymour and Scheines, 1993). Alternatively, one can produce bounds on the effect of interventions by taking representative samples of inferred structures and estimating $P_{x_i'}(x_j)$ according to (10) for each such sample.

In summary, I hope my comments convince the reader that DAGs can be used not only for specifying assumptions of conditional independence but also as a formal language for organizing claims about external interventions and their interactions. I hope to have demonstrated as well that DAGs can serve as an analytical tool for predicting, from nonexperimental data, the effect of actions (given substantive causal knowledge), for specifying and testing conditions under which randomized experiments are not necessary and for aiding experimental design and model selection.

Comment

Michael E. Sobel

It is a pleasure to discuss these excellent papers. Spiegelhalter, Dawid, Lauritzen and Cowell nicely put together a number of themes, demonstrating, in a Bayesian context, the utility of graphical modelling in the construction of probabilistic expert systems. The authors show how graphs can be used heuristically to solicit expert opinion, and in Section 6, how the theory of conditional independence graphs can be used to make tractable (while maintaining reasonable substantive assumptions) the calculation of probabilistic features of the system (monitors). For example, the authors want to apply to the directed independence graph of their Figure 2 the decomposability theorem for undirected conditional independence graphs, which permits a full factorization of the probability distribution. To do so, they associate the graph of Figure 2 with its moral graph (an undirected conditional independence graph) and use the fact that the separation properties of the moral graph apply to the directed independence graph. They then embed the moral graph into a triangulated graph, enabling use of the desired theorem; further simplications come from organizing the cliques of the triangulated graph into junction trees.

My vantage point is that of a social statistician: as such, there is more for me to say about the paper by Cox and Wermuth. In particular, I want to expand on and further tie several themes in this paper to research in the social and behavioral sciences. Thus, discussion focuses primarily on this paper; I shall often freely borrow notation from there.

TYPES OF INDEPENDENCE GRAPHS

Cox and Wermuth nicely characterize various types of dependencies among random variables. Prior work has focused attention on two types of independence graphs. If no ordering is imposed on the variables, undirected graphs are used; here, the absence of an edge between two vertices denotes conditional independence of the variables associated with the vertices,

Michael E. Sobel is Professor of Sociology and of Applied Mathematics, Department of Sociology, University of Arizona, Tucson, Arizona 85721.

given all the remaining variables. In a normal theory context, this corresponds to a 0 in the concentration matrix of the variables; thus, Cox and Wermuth call this a concentration graph. If some variables are taken as ordered with respect to others, for example, a set of variables viewed as independent is temporally prior to a set viewed as dependent, a different type of theory is useful. For this case, vertices can be placed within blocks and blocks arrayed from left to right; there is no ordering within blocks, but given a vertex and its associated random variable, vertices in blocks to the right denote prior (temporally or otherwise) random variables. By virtue of this ordering, we are not typically interested in the distribution of a variable X in a block, conditioning on all other variables, but in the distribution of X, conditioning on variables in blocks to the right (prior variables), or conditioning on prior variables and other variables in the same block. The latter case has received a great deal of attention. Here, edges between variables within a block are undirected, and edges between variables in different blocks, denoted by arrows pointing to the left, are directed. The absence of an arrow (or undirected edge if $l = m$ below) from vertex i in block l to vertex j in block m denotes conditional independence of X_i and X_j, conditioning on all remaining variables in blocks 1, ... m; one might think of the conditioning set as containing prior and "present" variables. When X_j is viewed as dependent, it's relationship to X_i is measured by the partial regression coefficient $\beta_{x_i x_j \cdot \underline{x}_R}$, where R denotes all remaining vertices in blocks 1, ... , m; the regression is called a block regression.

Cox and Wermuth also take up the case where the conditioning set consists of prior variables, using dashed edges in their graphs to distinguish this case from that above (where edges are full). With the same block structure and variables as above, the absence of a dashed arrow (dashed undirected edge if $l = m$) from i to j denotes independence of X_i and X_j, conditional on all remaining variables in blocks 1, ..., $m - 1$; if $l = m$, the $X_j - X_i$ relationship can be measured by the partial correlation, given the variables in blocks 1, ... , $m - 1$; otherwise, with X_j dependent, this relationship is measured by the partial regression coefficient $\beta_{x_i x_j \cdot \underline{x}_{R^*}}$, where R^* denotes all remaining vertices in blocks 1, ... , $m - 1$; the regression is called a multivariate regression.

The authors use the three types of independence graphs to illustrate the large number of ways in which the dependence structure of a set of random variables might be characterized. For example, their Figure 1 shows six different probabilistically equivalent ways of specifying a saturated model for just three variables. Subsequently, they exposit eight different types of dependence structures for four variables, using empirical examples to illustrate many of these. In each example, both substantive considerations and statistical evidence are used to select a model, but the data are not allowed to override substantive knowledge and/or interests. Example 2 features this nicely; the correlations and concentrations in Table 2 initially suggest a different model than that ultimately selected.

I look forward to seeing further developments in the theory of dashed independence graphs employed by Cox and Wermuth. This important case, apparently neglected in earlier work, is relevant to decision makers and planners, whose predictions depend on past information, not also on information contemporaneous with the time to which the prediction refers, and it is at least as relevant to social and behavioral scientists as the cases above. For example, multiple versions of the response are sometimes recorded in experiments (Winer, 1971). Here, a researcher typically wants to know the relationship between the response and the experimental variable, perhaps conditioning on a covariate vector, but certainly not also conditioning on the remaining versions of the response. Alternatively, in many studies, both experimental and nonexperimental, one measures a set of responses that are theoretically connected to a set of prior variables, but the responses are not so connected. For example, if interest centers on the educational attainments of siblings (or husbands and wives), one wants to know the partial regression coefficients relating the responses to family background variables. One might also want to know the relationship between the educational attainments, as measured by the partial correlation coefficient, conditioning on background variables. Again, the partial regression coefficients that also condition on the educational attainments of other siblings (or other spouse) are typically not of interest.

SIMULTANEOUS EQUATION MODELS

Cox and Wermuth have reservations about the use of simultaneous equation models featuring (see their Figure 4) coefficients γ_{xy} and γ_{yx} between "jointly determined" variables X and Y. For the model depicted in Figure 4, the authors point out that missing edges in the path diagram (graphical representation of the model) do not typically correspond to conditional independencies, and they argue that the interpretation of model parameters is problematic. (Note that their remarks would also hold if only one of the foregoing coefficients was nonzero and the errors were correlated.) They conclude that meaningful interpretations of the parameters of simultaneous equation models, when these exist, have to be developed on a case-by-case basis, a conclusion that challenges the conventional wisdom (in the social and behavioral sciences) on how such parameters are to be interpreted. Further examination of the conventional wisdom therefore seems worthwhile: the following look, while very brief, adds weight to Cox and Wermuth's conclusion.

In sociology and psychology (and also in some econometric work and papers on graphical models), it is not unusual to see the argument that γ_{xy} and γ_{yx} capture reciprocal causation. Becaue the concept of causation is asymmetrical, this does not make sense.

A more standard interpretation in economics is that structural parameters capture fundamental aspects of the behavior of economic agents. These parameters are preferred to the reduced form parameters; a single change in a structural parameter can change many reduced form parameters. Some economists, however, do not find this view compelling. For further criticism, as well as review of relevant literature, see Sobel (1994).

Another interpretation, due essentially to Strotz and Wold (1960), used in econometrics (e.g., Fisher, 1970) and psychometrics (Sobel, 1990), is that the underlying model is recursive:

$$\begin{aligned} Y_t &= \gamma_{yv} V + \gamma_{yx} X_{t-1} + \varepsilon_y , \\ X_t &= \gamma_{xw} W + \gamma_{xy} Y_{t-1} + \varepsilon_x . \end{aligned} \tag{1}$$

This is a linear dynamical system in discrete time with fixed coefficients; under suitable conditions Y_{t+r} and X_{t+r} converge, as r gets large, to values Y and X respectively. Under this interpretation, both γ_{yx} and γ_{xy} are regression coefficients in (1). However, note the errors are constant over time, which seems substantively unreasonable.

The foregoing supports Cox and Wermuth's view that despite frequent use, parameters of simultaneous equation models tend to elude meaningful interpretation. To balance the discussion a bit, without denying the general point, I can think of occasional examples where one would clearly want to use such a model to get the right interpretation. Let

$$\begin{aligned} Y &= \gamma_{yv} V + \gamma_{yx} X^* + \tilde{\varepsilon}_y , \\ X &= \gamma_{xw} W + \gamma_{xy} Y^* + \tilde{\varepsilon}_x , \end{aligned} \tag{2}$$

with $(V, W, X^*, Y^*) \perp\!\!\!\perp (\tilde{\varepsilon}_y, \tilde{\varepsilon}_x)$, and $\perp\!\!\!\perp$ denotes independence. To fix ideas, suppose that (V, W, X^*, Y^*) are temporally prior to (X, Y), and X^* and Y^* are anticipated (and unfortunately unobserved) values of X and Y, respectively. Thus, the researcher considers:

$$\begin{aligned} Y &= \gamma_{yv} V + \gamma_{yx} X + \varepsilon_y , \\ \mathrm{X} &= \gamma_{xw} W + \gamma_{xy} Y + \varepsilon_x , \end{aligned} \tag{3}$$

where $\varepsilon_y = \tilde{\varepsilon}_y - \gamma_{yx}\delta_x$, $\delta_x = X - X^*$, $\varepsilon_x = \tilde{\varepsilon}_x - \gamma_{xy}\delta_y$, $\delta_y = Y - Y^*$. Suppose that $(V, W, X^*, Y^*) \perp\!\!\!\perp (\delta_x, \delta_y)$. Under the setup above, X is correlated with ε_y, Y is correlated with ε_x, and block regression gives inconsistent estimates for the parameters of (2); an exception is the case where anticipations are perfect, that is, $X = X^*$, $Y = Y^*$. Consistent estimates of the regression coefficients can be obtained by using W and V as instruments in the first and second equations of (3), respectively. In this example, note that simultaneity arises from measurement error and simultaneous equation methods are needed to estimate the parameters of the relevant conditional expectation.

Given the problems above, it is useful to recall that a simultaneous equation model specifies a conditional distribution $f(\underline{x_2}|\underline{x_1})$; from this it is evident that the dependencies can be characterized either by a multivariate regression (called the reduced form in econometrics) or, if an ordering is imposed on the dependent variables, by means of a sequence of univariate recursive regressions (called the recursive form in econometrics). Following Wold, Cox and Wermuth emphasize the value of this recursive form.

GRAPHICAL MODELS AND SOCIAL SCIENCE RESEARCH

Graphical models could be useful in the social sciences, but I am not sure social scientists will pay them much attention; certainly the review article by Kiiveri and Speed (1982) in *Sociological Methodology* went unnoticed. There are probably several reasons for this. First, social scientists do not typically think in terms of probabilistic dependence and independence, conditional or otherwise. In statistical modeling, the social scientist's goal is to test hypotheses and arrive at quantitative estimates of relationships; if a model in use permits an interpretation in terms of the foregoing probabilistic concepts, for example, the univariate recursive regressions, that is well and nice, but secondary. In many cases, comparisons across groups are sought; here one typically wants to compare estimates of various quantities, and knowledge that within group conditional independence structures are identical (or not) across groups does not fully answer the primary questions. Second, following the lead of econometricians, quantitative social scientists argue that they are modeling processes and testing theories, as opposed to exploring data structures, and that tools appropriate for the latter are inappropriate for the former. In that vein, while Cox and Wermuth demonstrate, via their examples, the value of using graphical models especially in exploratory work, quantitative social scientists, who actually do a fair amount of exploratory work before hitting upon the desired confirmatory model, often do not acknowledge this exploratory process.

Having given a few reasons for doubting that social scientists will pay much attention to graphical modeling, I nevertheless give several examples of how such models can be useful. First, in many areas of social science, not that much is known, and it is often useful to start with an exploratory analysis. Researchers who take advantage of graphical models could be led to systematically explore dependence structures that they would not otherwise have considered. This may lead to a model which attempts to pin down the relationships of interest more precisely. Consideration of these models could also be useful in so-called confirmatory work; the following examination of a typical modeling exercise in covariance structure analysis should illustrate the point. A researcher begins with a model of interest. (The case where a nested sequence of models of interest is entertained at the outset is similar and thus will not be exposited separately.) One aim of the analysis is to select a preferred model. Perhaps the initial model fits the data adequately, using conventional statistical criteria (e.g., the likelihood ratio). In this case, the analysis is terminated. But now suppose, as often happens, that this model does not fit the data. In that event, a researcher who nevertheless prefers this initial model may shop around for a goodness-of-fit index (there are many) that suggests the fit is really good enough after all. If such an index cannot be found or if the researcher did not look for one, the initial model is rejected, and typically a search for a better fitting model begins. There are many ways to conduct such a search, but typically modification indices, which tell the user the constrained parameters in the analysis to free up, are used. After a sequence of such modifications, an unsaturated model that fits the data by conventional criteria is found, or one of the many possible versions of the saturated model is obtained. Now of course this search procedure is nothing but exploratory analysis, and when used poorly, it leads to a model that is at best not to be taken seriously. Instead of looking around for goodness-of-fit indices and modification indices (or at least in addition to), a natural alternative at this stage is to ask whether it is reasonable to widen the class of searches, and if so, whether it is reasonable to use graphical models to see if alternative types of structures, perhaps not initially contemplated, may account for the data. If the answer is yes, with intelligent use, we might find out something new; of course, if used like some of the indices above, this will not be the case.

CAUSATION AND CAUSAL INFERENCE

I use the facts that Cox and Wermuth disassociate their work from causal concepts and Spiegelhalter et al. use the term "direct influence" to refer to intuitive judgements of relevance as a license to close with some remarks on causation and irrelevance; these remarks are more general in nature, not particularly addressed to either paper.

There is a large philosophical literature on causation, and numerous views have been espoused (including the view that probabilistic relations have nothing to do with causation). Thus, the merits of an inference about causation (hereafter causal inferences) cannot be evaluated unless the concept of causation under consideration has been made clear. Undaunted by this problem, many researchers in artificial intelligence, decision science, philosophy and statistics who write on graphical models often simply equate the absence (presence) of a directed edge or a path in an independence graph with the absence (presence) of causation; in many instances they neither formally define causation by conditional independence nor attempt to say what it is. Their counterparts in the social and behavioral sciences utilize path diagrams in a similar way, equating the presence or absence of parameters or functions of these with the presence or absence of causation.

Although social and behavioral scientists do not typically say what causation is, at least among users of structural equation models there appears to be an implicit commitment to a manipulative account of the causal relation, evidenced in the interpretation of model parameters as unit (or average) effects. For example, in the context of a univariate regression, $\beta_{yx\cdot \underline{x}_{R^*}}$ is interpreted as the amount Y would increase for any unit (or on average) if the value of X, say x, were increased to $x + 1$, and all remaining variables (in the conditioning set) were "held" constant. Of course, these variables are not actually held constant, but merely conditioned upon, a point I shall ignore here [but see Sobel (1990)]. If this value is 0, one might say that X does not cause Y. In the normal theory context, this is equivalent to conditional independence; this ties the discussion to treatments in the literature on graphical models which use conditional independence and dependence relations to make causal inferences, arguing that the inferences so obtained will sustain a manipulative account.

The foregoing types of interpretations are very strong, and one wonders when these are warranted. To that end, such interpretations hinge on comparing, for any unit, its values on the dependent variable(s) as the unit takes on all values of the independent variable(s). The averages when all units in the population take on the same value can then be compared with one another, by looking, for example, at average differences. Readers familiar with Rubin's (1974, 1977, 1978, 1980) work on causal inference or the review by Holland (1986) will realize that I have just defined an average effect. Of course, in practice a unit can be administered only one value of the causal variable. Nevertheless, when treatment assignment is random, or random conditional on a vector of covariates, valid causal inferences can be obtained by calculating the usual sample quantities (valid in the sense that the estimator is unbiased and/or consistent for the desired population quantities).

In Sobel (1992), I introduce the concept of causation in distribution and use the ideas in Rubin's model to examine the issue of spurious causation. Since spurious causation is typically defined as a case in which certain marginal dependencies vanish upon conditioning, the results are relevant to literature in graphical modeling that equates the absence of causation with conditional independence. The idea behind causation in distribution is to examine the distribution of the response $\underline{Y}_{\underline{x}}$ when every element of the population has the same value $\underline{x}$ on the causal vector ($\underline{X}$) and to compare the distributions as $\underline{x}$ varies. If the distributions do not change as $\underline{x}$ varies, one says $\underline{X}$ does not cause $\underline{Y}$ in distribution and otherwise one says $\underline{X}$ causes $\underline{Y}$ in distribution. For a conditioning set $\underline{X}_{R^*}$, I show (1) $\underline{X} \perp\!\!\!\perp \underline{Y} \mid \underline{X}_{R^*}$ does not imply $\underline{X}$ does not cause $\underline{Y}$ in distribution, and (2) $\underline{X}$ does not cause $\underline{Y}$ in distribution, does not imply $\underline{X} \perp\!\!\!\perp \underline{Y} \mid \underline{X}_{R^*}$. For example, if $\underline{X}_{R^*}$ is prior to variable X, and X prior to variable Y, with no variables intervening between X and Y, the results state that X may (or may not) "directly influence" Y (using the sense of directly influence in the graphical modelling literature), but X may not (may) cause Y in distribution. Note also there is no path connecting X to Y in this example. This should suggest that causal inferences based on the usual conditional independence relations do not generally sustain a manipulative account of the causal relation. Sobel (1992) also gives necessary and sufficient conditions for equivalence of conditional independence and causation in distribution.

The foregoing suggests more cautious use of the term "causation" in future work. Not surprisingly, I do not like the terms "causal network" and "influence diagrams"; is not influence just another synonym for causation? The terms employed by Spiegelhalter et al. (directed graphical model, belief networks) seem preferable. Finally, I want to briefly take up the term "irrelevance," sometimes defined via structures that satisfy the axioms of generalized conditional independence (Smith, 1988). (Smith uses the term "uninformative" and is always careful to mention the conditioning set.) From my view, scientists often allow the connotative aspects of words to creep into their use of technical terms, and this can be detrimental. Thus, one might want to choose terms whose connotative aspects are in accord, as much as possible, with the technical definition. In that vein, relevance seems to encompass many things, including causation; for example, the phrase "causally irrelevant" describes one form of irrelevance. Even leaving aside causation, adding information to the conditioning set of marginalizing over this set can make "irrelevant" variables become "relevant"; should these variables have been called irrelevant to begin with?

Comment

Joe Whittaker

It gave me great pleasure to read these articles. Here we have two papers on the application of conditional independence: one to the specification of a graphical model for assessing association in multivariate responses and the other to message passing on a directed graph, in a paper which expertly summarises the probabilistic view of dealing with uncertainty in expert systems. Right at the outset, let me state my own belief that it is not so much the graphic display but the notion of conditional dependence and independence and the idea of a ternary relationship that X_1 affects (or is irrelevant to) X_2 in the presence of X_3, which constitutes the fundamental contribution of graphical models to statistical analysis.

I particularly want to focus on the Cox and Wermuth (CW) paper, which I believe raises some unresolved issues, and discuss three topics in more detail: the value of a graphical representation, the distinction between multivariate and "block" regression and the role of the Schur complement as a partial variance.

VALUE OF A GRAPHICAL REPRESENTATION

Few practising statisticians can be unaware of the immediate and powerful impact of visual display in conveying the results of a statistical analysis to a consulting client. A tremendous selling point of graphical models is the graph: a fact which is well known to statistical researchers in related areas such as path analysis, causal modelling, factor analysis and structural equation modelling. The same lesson can be learnt from the recently expanding field of neural networks, where statisticians [for instance, Ripley (1993) and Cheng and Titterington (1993)] are discovering that neuroscientists and computer scientists have been busy proposing neural network formulations of nonlinear statistical classification methods. While perhaps not

Joe Whittaker is Senior Lecturer, Mathematics Department, Lancaster University, LA1 4YF, United Kingdom.