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 $\underline{Y}_x$ 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.

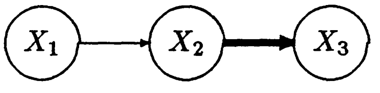
VALUE OF A GRAPHICAL REPRESENTATION

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.

exactly original they are not reinventing the wheel for the neural net exposition provides a deeper understanding contributing greatly to the upsurge in popularity of these methods.

There is therefore some pressure to embellish the conditional independence (CI) graph with additional information, on top of the essential iconographics for nodes, edges and directed edges; it is easy to understand the motivation of the authors in introducing further types of edges, such as the dashed edge. For instance, it is often suggested that the thickness of the edge should reflect the strength of the dependence and I agree that immediately conveys the information that the (2,3) dependence is stronger than the (1,2) dependence, thus helping the data analyst to make sense of possibly complex interactions.

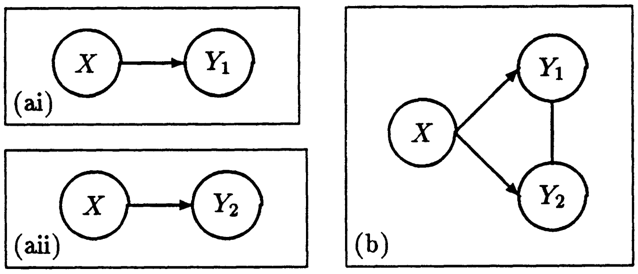


However, this is not a suggestion which I would support as it obscures the overriding defining feature of a conditional independence graph: the edge (1,3) is missing because $X_3 \perp\!\!\!\perp X_1|X_2$. It is the *absence* of an edge which generates the graph. Admittedly this is a subtle point and choosing to visually represent a defining feature by a blank space is perhaps unfortunate.

DISTINCTION BETWEEN MULTIVARIATE REGRESSION AND "BLOCK" REGRESSION

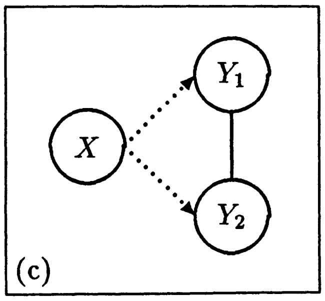
A particular contribution of the CW paper is to highlight the difference between multivariate regression and so-called "block" regression and to demonstrate that graphical modellers have some difficulty in portraying the former. The reason, of course, is that graphical modelling interests itself in the analysis of conditional relationships while multivariate regression focuses on marginal relationships.

For example, an idea of the distinction can be gained by asking what parameters have to be zero for an edge in a CI graph to vanish. In the multivariate regression of (Y_1, Y_2) on X, which essentially consists of computing separate univariate regressions of Y_1 on X and Y_2 on X, the regression coefficient $\beta_{Y_1X} = 0$ eliminates the edge connecting Y_1 with X in CI graph (ai). Similarly $\beta_{Y_2X} = 0$ eliminates the edge in (aii). Two separate CI graphs are required to represent these concepts.



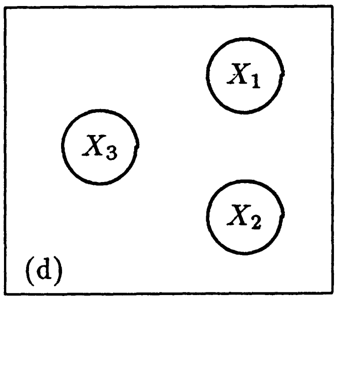
The "block" regression corresponds to CI graph (b). The edge connecting Y_1 with X in CI graph (b) vanishes if the partial regression coefficient $\beta_{Y_1X|Y_2} = 0$. The techniques may give the same numerical answers in certain special cases, for instance if $Y_1 \perp\!\!\!\perp Y_2|X$ or if $Y_2 \perp\!\!\!\perp X$, but in general they do not. The same issue of whether to parameterise in the conditional or in the marginal distribution arises in the analysis of discrete data, for example, see the papers of Liang, Zeger and Qaqish (1992), Laird and Ware (1982). There is no universal panacea.

The authors attempt to combine the graphs (ai, aii, b) and extract the best from both worlds by defining the dashed edges in the graph (c) by the interpretation that if such an edge is missing it should be concluded that $Y_1 \perp\!\!\!\perp X$ rather than $Y_1 \perp\!\!\!\perp X|Y_2$.



At this point I find I have to take up the cudgels and put the "purist" view that such an extension leads to difficulties and ambiguities and is even perhaps unnecessary. I make four points.

1. *Liability to misinterpretation:* Consider for example the graph

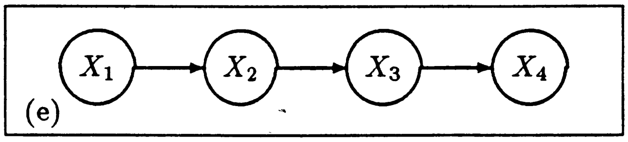


defined by missing *dashed* edges. To me, the only possible visual interpretation of graph (d) is that of complete (mutual) independence of X_1, X_2 and X_3. But of course, there are well-known counter examples to the assertion that $\{X_1 \perp\!\!\!\perp X_2, X_1 \perp\!\!\!\perp X_3, X_2 \perp\!\!\!\perp X_3\}$ implies the mutual independence of X_1, X_2, X_3. Only if (X_1, X_2, X_3) are jointly normal could such an assertion hold, which restriction would violate the attractive feature of graphical models that it unifies the theories of discrete and continuous variable dependence.

2. *Separation:* Key to the construction of CI graphs is the focus on the *joint* distribution and the mapping of the ternary conditional independence relation $X_a \perp\!\!\!\perp X_b | X_c$ to the, similarly ternary, separation property of subsets in a graph "a is separated by b from c." Technically this concept is defined by: all paths in the graph starting from a vertex in a and finishing at a vertex in b have a nonempty intersection with c.

Marginal independence is a *binary* relationship between random variables and so cannot easily map onto the separation property of nodes in a graph.

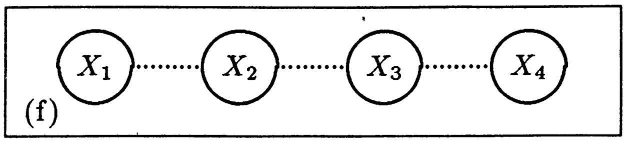
3. *Coherence:* To obtain a coherent picture CI graphs focus on a single joint distribution, $f_{123\ldots k}$ say, and analyse it in terms of conditional distributions of the form $f_{a|rest}$. Because $f_{12\ldots k} = f_{k|1\ldots k-1}\, f_{k-1|1\ldots k-2} \cdots f_{2|1} f_1$ this single joint distribution can be built up from a nested sequence of marginal distributions. For example, the missing (1,3) edge in the directed graph of a Markov chain



signifies the $X_3 \perp\!\!\!\perp X_1 | X_2$. However the graph (e) still refers to a single joint distribution of four random variables.

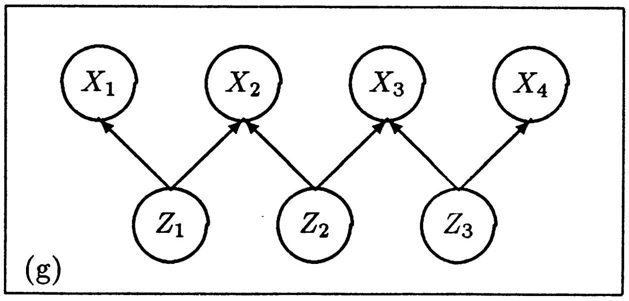
Unfortunately, a single joint distribution is not generally specified by all pairwise marginal distributions, and so a graph built from these may easily indicate ambiguities as in the mutual independence example above.

4. *Latent variable embedding:* It may be unnecessary to invent new types of graphs. For example, consider an analysis of the undirected dashed edge chain graph



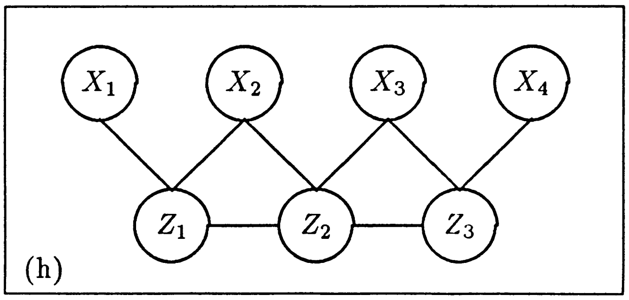
defined by $\{X_1 \perp\!\!\!\perp X_3, X_1 \perp\!\!\!\perp X_4, X_2 \perp\!\!\!\perp X_4\}$ and ask if information on X_1 is needed to predict X_3 when X_2 is known.

Now the graph (f) is a consequence of the directed CI graph (g)

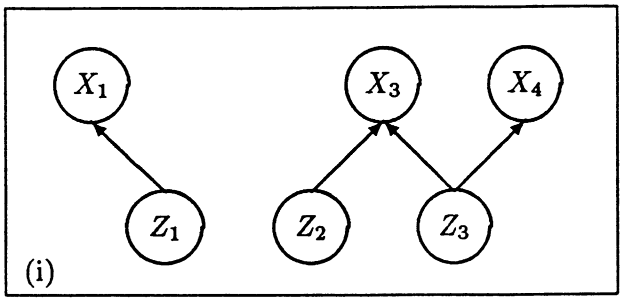


in which Z_1, Z_2 and Z_3 are mutually independent and the X's are conditionally independent given the Z's ($X_1 \perp\!\!\!\perp X_3$ as they have no Z's in common). The CI graph (f) is a "consequence" in the sense that the marginal distribution of (X_1, X_2, X_3, X_4) is obtained from that of $(X_1, X_2, X_3, X_4, Z_1, Z_2, Z_3)$ by integrating out (Z_1, Z_2, Z_3) and has the requisite properties of marginal independences indicated by missing dashed lines.

The moralisation procedure of Lauritzen and Speigelhalter (1988) indicates that (g) is embedded in the undirected CI graph (h) for the joint distribution of $(X_1, X_2, X_3, X_4, Z_1, Z_2, Z_3)$.



Since X_2 does not separate X_1 from X_3 in (h), the answer is that X_1 cannot be discounted if X_2 is observed. However, ironically if X_2 is not observed, the graph reflecting the distribution $(X_1, X_3, X_4, Z_1, Z_2, Z_3)$ is (i)



and clearly, X_1 is uninformative about X_3.

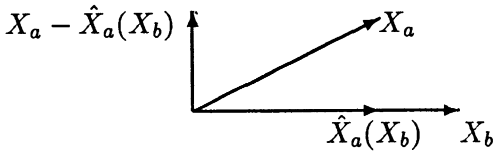
This device of embedding the dashed graph into a CI graph with "latent variables" certainly solves some problems. It also indicates why latent variables in highly structured graphs allow marginal empirical dependences to determine the statistical analysis. A prime example of this is the graphical analysis of the state space model underlying the Kalman filter.

ROLE OF THE PARTIAL VARIANCE (SCHUR COMPLEMENT)

The technical conditions for conditional independence in multivariate normal distributions, for instance, that $X_1 \perp\!\!\!\perp X_2|X_3$ is characterised by a zero in the inverse variance matrix of (X_1, X_2, X_3), appear somewhat bizarre at a first acquaintance. A good understanding requires an interpretation of the elements of this inverse variance matrix, and I found it useful in writing Chapter 5 of my book (Whittaker, 1990) to use the concept of the partial variance as the vehicle for this explanation. For instance, slightly extending the notation of the CW paper, when a vector X with variance Σ is partitioned into (X_a, X_b) the block in the inverse variance Σ^{-1} corresponding to X_a is $\Sigma^{aa} = (\Sigma^{-1})_{aa}$ (and *not* $(\Sigma_{aa})^{-1}$), the essential content of the inverse variance lemma is that

$$\Sigma^{aa} = \operatorname{var}(X_a|X_b)^{-1}. \tag{1}$$

Here $\operatorname{var}(X_a|X_b)$ is the partial or residual variance of X_a having regressed out X_b, and defined by $\operatorname{var}(X_a - \hat{X}_a(X_b))$ where $\hat{X}_a(X_b)$ is the fitted (multivariate) regression of X_a on X_b. These entities can be represented in the Pythagorean vector diagram



The notion of a partial variance permits the diagonal elements of the inverse variance matrix to be interpreted as functions of the multiple correlation coefficient: if $a = \{i\}$ is 1-dimensional, so that b denotes the $p - 1$ remaining variables, then (1) becomes

$$\Sigma^{ii} = \operatorname{var}(X_i|X_{rest})^{-1} = \operatorname{var}(X_i)^{-1} / (1-R^2(i))$$

where $R(i)$ is the multiple correlation coefficient of X_i with the remaining variables. In consequence, the larger Σ^{ii} in relation to $\operatorname{var}(X_i)$ the more predictable is X_i from the other variables. By choosing $a = \{i, j\}$ to be 2-dimensional, formula (1) enables an explicit expression for the off-diagonal elements of the inverse variance in terms of the partial correlation of X_i and X_j given the remaining variables. In point of fact $\Sigma^{ij} / \sqrt{\Sigma^{ii}\Sigma^{jj}} = -\operatorname{corr}(X_i, X_j|X_{rest})$.

The inverse variance lemma, which is by no means new, is really just statistical interpretation of inverting a partitioned matrix. In fact $\operatorname{var}(X_a|X_b)$ can be computed from $\operatorname{var}(X_a) - \operatorname{cov}(X_a, X_b)\operatorname{var}(X_b)^{-1}\operatorname{cov}(X_b, X_a)$ which in the mathematical literature is well known as the Schur complement of the matrix

$$\begin{bmatrix} \operatorname{var}(X_a) & \operatorname{cov}(X_a, X_b) \\ \operatorname{cov}(X_b, X_a) & \operatorname{var}(X_b) \end{bmatrix}.$$

The determinant represents the squared length (volume) of the residual vector in the Pythagorean vector diagram above. This quantity is denoted by $\Sigma_{aa|b}$ in CW as in many books on the multivariate normal distribution, but such a notation obscures various elementary properties such as $\operatorname{var}(AX_a|X_b) = A\operatorname{var}(X_a|X_b)A'$ where A is a fixed linear transform, and if B is invertible, $\operatorname{var}(X_a|BX_b) = \operatorname{var}(X_a|X_b)$ expressing the invariance of the partial variance to a change of units in the regressor variables.

Various forms of the lemma exist and a frequent application is to Bayesian analysis for instance, in the analysis of linear models by Lindley and Smith (1972), in standard treatments of factor analysis, and in Kalman filtering.

Rejoinder

D. R. Cox and Nanny Wermuth

We are grateful to all the contributors for their thoughtful and constructive contributions. There is rather little with which we disagree so that our reply is brief.

While to some extent the use of the word *causal* is a matter of convention, we much prefer to restrict the word to situations in which we have knowledge of some underlying process. We reassure Dempster that we are deeply concerned with the elucidation of processes that might have generated the data, but are cautious about what conclusions can be drawn from single investigations or even repeated investigations, especially but