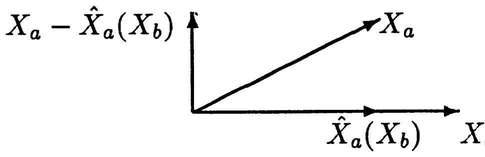
This device of embedding the dashed graph into a CI graph with "latent variables" certainly solves some problems. It also indicates why latent variables in highly structured graphs allow marginal empirical dependences to determine the statistical analysis. A prime example of this is the graphical analysis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

$$\sum^{aa} = \text{var}(X_a|\text{X}_b)^{-1}. \qquad (1)$$

Here $\text{var}(X_a|X_b)$ is the partial or residual variance of X_a having regressed out X_b, and defined by $\text{var}(X_a - \hat{X}_a(X_b))$ 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

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

where $R(i)$ is the multiple correlation coefficient of X_i with the remaining variables. In consequence, the larger Σ^{ii} in relation to $\text{var}(X_i)$ 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 $\Sigma^{ij} / \sqrt{\Sigma^{ii}\Sigma^{jj}} = -\text{corr}\,(X_i, X_j|X_{rest})$.

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

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

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 $\text{var}(AX_a|X_b) = A\text{var}(X_a|X_b)A'$ where A is a fixed linear transform, and if B is invertible, $\text{var}(X_a|BX_b) = \text{var}(X_a|X_b)$ 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 not only when these are observational. We agree that the graphs suggested by Glymour and Spirtes could possibly be chosen as another description of our nondecomposable models but we do not regard them as indicating useful potential processes to generate the data, the point of our distinction.

In a recent paper, Stone (1993) elucidates requirements for particular causal interpretations. He also examines critically strongly ignorable treatment allocation. Pearl in his contribution gives an important graphical interpretation exactly of this assumption, this facilitating the judgement of the effects of interventions in a hypothesized causal process.

Several contributors mention the role of latent variables, including as a special case the occurrence of measuring errors. We agree that their use, preferably sparingly, especially in elucidating nondecomposable models, needs further study. For instance, the tetrad conditions studied by Spirtes, Glymour and Scheines (1993) for linear relations become relevant as well for binary variables having a quadratic exponential distribution. This distribution has some of the properties of the multivariate normal distribution and provides exact or approximate answers to Hill's question about graphical theory for binary distributions and to Whittaker's comments on complete independence.

Dempster favours shrinking estimates toward zero as opposed to setting parameters exactly to zero. We agree when empirical prediction is the objective, but not where essentially qualitative understanding via simple representations is involved, and the latter is our main concern.

The issue, raised by Whittaker, of labelling the edges of a graph can be solved in various ways if a single degree of freedom is attached to each edge (by partial correlation coefficients or by standardized regression coefficients, for instance). The introduction of graphs with dashed edges has, however, a different objective, because it leads to structures of independence different from those discussed by Whittaker, thus enriching the class of graphical chain models, as pointed out by Hill. Whittaker's graphs (ai) and (aii) do not represent the multivariate regression of our Figure 1c because the essential association between the two responses is omitted.

Whittaker points out the relation of the Schur complement to partial correlations and inverse covariance matrices. An early treatment of this in the statistical literature is by Cramér (1946, subsections 22.7, 23.4 and 23.5). The connection between partial correlation and canonical parameters in the exponential family has opened the road to defining analogous independence structures for discrete variables and for mixed discrete and continuous variables, known now as block regression (full edge) chain models.

In general distributional assumptions are necessary, in addition to the independence graph, for a full specification of a statistical model. Indeed some research hypotheses may not be possible for a particular joint distribution of specified form. For example, $X \perp\!\!\!\perp Y|A$ cannot hold without additional independences if the joint distribution is given by the linear logistic regression of the binary variable A on the bivariate normal variable (X, Y). Similarly if (X, Y) are conditionally bivariate normal given the discrete variable A, then marginal independence of X and Y is possible only with additional independences. See Cox and Wermuth (1992b) for further details.

We were glad to see that Sobel regards our introduction of multivariate regression (dashed edge) chain graphs as a step toward more traditional analyses in the social sciences. In fact, it was one of our purposes to provide simple examples which help one to recognize similarities and distinctions between different approaches, the latter being explicitly appreciated by both Sobel and Dempster.

Because of the particular focus of our paper, we have put little emphasis on such issues as description of sample selection, checking data quality, testing model adequacy, examining the need of data transformation and comparison of the fits of different kinds of models. All of these are a normal if often difficult part of applied statistical work. From our present perspective, whether the formal aspects to the analysis are in frequentist or Bayesian terms is a secondary issue.

A special topic for further work concerns the role of graphs with both kinds of edge, for example, in representing the regression for multivariate binary data studied by Zhao and Prentice (1990) and by Fitzmaurice and Laird (1993).

Rejoinder

David J. Spiegelhalter, A. Philip Dawid, Steffen L. Lauritzen and Robert G. Cowell

We are grateful to the discussants for their thoughtful comments: since our paper is already quite long enough we shall try to restrict our responses. We shall first deal with representations of causality, followed by some technical points on zero probabilities. Automatic model construction will then be considered, and whether a