Statistical Science
1993, Vol. 8, No. 3, 204–283

Linear Dependencies Represented by Chain Graphs

D. R. Cox and Nanny Wermuth

Abstract. Various special linear structures connected with covariance matrices are reviewed and graphical methods for their representation introduced, involving in particular two different kinds of edge between the nodes representing the component variables. The distinction between decomposable and nondecomposable structures is emphasized. Empirical examples are described for the main possibilities with four component variables.

Key words and phrases: Chain model, conditional independence, covariance selection, decomposable model, linear structural equation, multivariate analysis, path analysis.

1. INTRODUCTION

This paper has three broad objectives. The first is to illustrate the rich variety of special forms of association and dependence that can arise even with as few as three or four variables. The second is to show the value of graphical representation in clarifying these dependencies; for this we introduce graphs with two different kinds of edge and some further features which are also new. The third objective is to show the importance in interpretation of the distinction between decomposable and nondecomposable models.

A series of examples will be used in illustration, partly to show that many of the special structures do indeed arise in applications and partly to show in outline the implications for interpretation, although reference to the subject matter literature is necessary for a full account. Most of the examples arise from recent investigations at University of Mainz. For purposes of exposition we have chosen examples with at most four variables; that is, we have simplified by omitting mention of variables which analysis had shown to have no bearing on the points at issue.

We confine the discussion to those problems with essentially linear structure in which the interrelationships are adequately captured by the covariance matrix of the variables. Of course in applications, checks for nonlinearities and outliers are required, and these have been done for all examples whenever we had access to the raw data.

D. R. Cox is Warden, Nuffield College, Oxford OX1 1NF, United Kingdom. Nanny Wermuth is Professor, Psychologisches Institut, Johannes Gutenberg–Universität Mainz, Postfach 3980, D-55099 Mainz, Germany.

The need to discuss special structures arises partly because the relations of marginal independence and conditional independence expressed thereby are often of substantive interest and partly because in a *saturated model* with p component variables, that is, one in which the covariance matrix is unrestricted other than to being positive definite, there are $1/2p(p-1)$ correlations, and reduction of dimensionality may be desirable to avoid a superabundance of parameters.

There are strong connections with, in particular, the long history of work in path analysis in genetics, in simultaneous equations in econometrics and linear structural models in psychometrics and with the body of recent work applying graph-theoretic ideas to the study of systems of conditional independencies arising especially in the study of expert systems.

In Section 2 we review some general properties of linear regression systems as related to the covariance matrix of the variables and stress the distinction between multivariate regression and block regression and between decomposable and nondecomposable structures. In Section 3 we introduce the main conventions useful in a graph-theoretic representation of the independency relations that may hold; in Section 4 we discuss relations with previous work, and in Section 5 we give a series of empirical examples for four variables. The paper concludes in Section 6 with some general discussion. The emphasis throughout is on the structure and interpretation of the various models rather than on the procedures for fitting.

2. SOME PROPERTIES OF COVARIANCE MATRICES

It is convenient to set out some properties of systems of linear least squares regressions derivable from a covariance matrix. These are full regression equations

in a multivariate normal distribution. There is throughout the usual interplay between relatively weak second-order properties of least squares regression and the strong properties derivable from an assumption of multivariate normality, such as that zero correlation or zero partial correlation implies independence or conditional independence.

We consider first the $p \times 1$ vector $Y = (Y_1, \ldots, Y_p)^T$ with mean $E(Y) = \mu$. We denote the positive definite covariance matrix by $\text{cov}(Y) = \Sigma$, and its inverse, the concentration matrix, therefore by Σ^{-1}; the diagonal elements of Σ are the variances (σ_{ii}), those of Σ^{-1} are the precisions (σ^{ii}). The off-diagonal elements of Σ are the covariances (σ_{ij}), those of Σ^{-1} are the concentrations (σ^{ij}). A marginal correlation ρ_{ij} is expressible via elements of the covariance matrix, in a way similar to that in which a partial correlation, $\rho_{ij.k}$, given all of the remaining variables $k = \{1, \ldots, p\}\backslash\{i, j\}$, is expressible via elements of the concentration matrix:

$$\rho_{ij} = \sigma_{ij}(\sigma_{ii}\sigma_{jj})^{-1/2}, \rho_{ij.k} = -\sigma^{ij}(\sigma^{ii}\sigma^{jj})^{-1/2}.$$

This implies in particular that in the usual notation (Dawid, 1979a) for independence,

$$Y_i \perp\!\!\!\perp Y_j, \quad \text{if and only if } \sigma_{ij} = 0,$$
$$Y_i \perp\!\!\!\perp Y_j \mid Y_k, \quad \text{if and only if } \sigma^{ij} = 0,$$

where as above $k = \{1, \ldots, p\}\backslash\{i, j\}$.

To study regression models, we partition Y into Y_a and Y_b, $p_a \times 1$ and $p_b \times 1$, respectively, $p_a + p_b = p$, and call the two parts response and explanatory variables. Let the covariance matrix and the concentration matrix be conformally partitioned:

$$(1) \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \cdot & \Sigma_{bb} \end{pmatrix}, \Sigma^{-1} = \begin{pmatrix} \Sigma^{aa} & \Sigma^{ab} \\ \cdot & \Sigma^{bb} \end{pmatrix},$$

then the covariance matrix Σ_{bb} of the explanatory variables and correspondingly their concentration matrix $\Sigma^{-1}_{bb} = \Sigma^{bb.a} = \Sigma^{bb} - (\Sigma^{ab})^T(\Sigma^{aa})^{-1}\Sigma^{ab}$ do not contain parameters needed to specify a standard regression model of Y_a on Y_b. Instead, their observed counterparts are taken as fixed or indeed sometimes are fixed by sampling design.

We now distinguish between a multivariate regression and a block regression. To simplify the notation we shall without essential loss of generality take often $E(Y) = 0$. We describe the distinct parameters in the two types of regression models, that is, the two ways of parametrizing the conditional distribution of Y_a given Y_b. For a *multivariate regression* of Y_a on Y_b, that is, for $Y_a = \Pi_{a|b}Y_b + \varepsilon_a$ with $E(\varepsilon_a) = 0, E(\varepsilon_a Y_b^T) = 0$, the regression equation parameters $\Pi_{a|b}$ and the residual variance $\text{var}(\varepsilon_a)$ can be written in a matrix as $(\Sigma_{aa.b}, \Pi_{a|b})$, where

$$(2) \quad \Pi_{a|b} = \Sigma_{ab}\Sigma_{bb}^{-1},$$
$$\text{var}(\varepsilon_a) = \Sigma_{aa.b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ab}^T.$$

In a saturated multivariate regression (2) each component of Y_a is regressed separately on the full set of components Y_b.

On the other hand in a saturated *block regression* each component of Y_a is regressed not only on Y_b but also on all remaining components of Y_a. Then the regression equation parameters are instead proportional to the elements of the matrix $(\Sigma^{aa}, \Sigma^{ab})$ (Wermuth, 1992). The reason is that the expected value of a component Y_i of Y_a given all remaining variables of Y can be obtained by taking expectations in

$$(3) \quad \Sigma^{aa}Y_a + \Sigma^{ab}Y_b = \omega_a$$

where $E(\omega_a) = 0$, $\text{var}(\omega_a) = \Sigma^{aa}$ and dividing the ith equation by the concentration σ^{ii}. Equation (3) is derived from a block triangular decomposition of the concentration matrix, $\Sigma^{-1} = A^T T^{-1} A$, where

$$(4) \quad A = \begin{pmatrix} I_{aa} & (\Sigma^{aa})^{-1}\Sigma^{ab} \\ 0 & I_{bb} \end{pmatrix},$$
$$T^{-1} = \begin{pmatrix} \Sigma^{aa} & 0 \\ 0 & \Sigma^{bb.a} \end{pmatrix},$$

as the first p_a equations of $(T^{-1}A)(Y - E(Y)) = \omega$. The residuals ω have zero mean and covariance matrix T^{-1}.

For a block regression, the resulting coefficient of variable Y_j in the ith equation is minus a partial regression coefficient given all remaining variables of Y, that is, given all remaining response and explanatory variables. On the other hand, in a multivariate regression the coefficient of Y_j in the ith equation is a partial regression coefficient given all remaining variables of Y_b, that is, given all remaining explanatory variables. To express this distinction more formally, we write a partial regression coefficient $\beta_{ij.d}$ for $\{1, \ldots, p\} = a \cup b = \{\{i, j\}, d, g\}$ in terms of elements of the conditional covariance matrix of (Y_i, Y_j) given Y_d and of elements of the concentration matrix of (Y_i, Y_j), having marginalized over Y_g, as

$$\beta_{ij.d} = \frac{\sigma_{ij.d}}{\sigma_{jj.d}} = -\frac{\sigma^{ij.g}}{\sigma^{ii.g}}.$$

Note that in the case of a block regression g is empty and d is the set of all remaining variables of Y, that is, $d = (a \cup b)\backslash\{i, j\}$, while in the case of a multivariate regression $d = b\backslash\{j\}$, and $g = a\backslash\{i\}$. Note further that

$$(5) \quad Y_i \perp\!\!\!\perp Y_j \mid Y_d, \quad \text{if and only if } \beta_{ij.d} = 0.$$

To judge the relative strength of the dependence of a response on several explanatory variables, it is sometimes useful to compare the standardized regression coefficients, that is, $\beta^*_{ij.d} = \beta_{ij.d}\sigma_{jj}^{\frac{1}{2}}\sigma_{ii}^{-\frac{1}{2}}$.

One of the major distinctions between multivariate regression and block regression lies in the meaning of the relation between two components Y_i and Y_j, both within Y_a, and in the meaning of the relation of a

component Y_i from Y_a to a component Y_j from Y_b. To describe this in detail it is useful to recall how a partial regression coefficient relates to a partial correlation coefficient

$$\beta_{ij.d} = \rho_{ij.d}\sqrt{\frac{\sigma_{ii.d}}{\sigma_{jj.d}}} = \rho_{ij.d}\sqrt{\frac{\sigma^{jj.g}}{\sigma^{ii.g}}}.$$

Thus, in a block regression, that is, where $d = (a \cup b)\backslash\{i, j\}$, the relation between Y_i from Y_a and Y_j is measured essentially by the partial correlation given all remaining variables of Y, no matter whether Y_j is from Y_a or it is from Y_b. By contrast in a multivariate regression, that is, where $d = b\backslash\{j\}$, the measure of the relation of Y_i from Y_a to Y_j from Y_b is proportional to the partial correlation given the variables in Y_b other than Y_j; the correlation between Y_i and Y_j both within Y_a is given all variables in Y_b. Thus, a larger set of variables is considered simultaneously in block regression if compared with the corresponding multivariate regression. Written in matrix notation their parameters are related by

$$(6) \qquad \Pi_{a|b} = -(\Sigma^{aa})^{-1}\Sigma^{ab}, \Sigma_{aa.b} = (\Sigma^{aa})^{-1}$$

$$(7) \qquad \Sigma^{ab} = -(\Sigma_{aa.b})^{-1}\Pi_{a|b}, \Sigma^{aa} = (\Sigma_{aa.b})^{-1}.$$

Some of the special models we shall consider correspond to specifying some elements of regression equations to be zero, that is, to structures that appear simplified if compared with the saturated model. The choice between block regression and multivariate regression is then largely determined by the research questions and by a decision as to which of the two parametrizations permits a simpler description of the relations. For instance, in each of Examples 1, 2 and 7 of the empirical examples of Section 5 we can think of two variables as joint responses, $Y_a = (Y, X)^T$, and of two variables as explanatory, $Y_b = (V, W)^T$. A simplifying description is possible with block regression but not with multivariate regression in Example 1, while a simpler structure results with multivariate regression than with block regression in Examples 2 and 7.

If not only the conditional distribution of Y_a given Y_b is of interest, but the marginal relations among component variables within Y_b as well, we are led to a simple type of regression chain model: we specify the joint density via

$$f_{ab} = f_{a|b}f_b,$$

and make a choice for $f_{a|b}$ among a multivariate and a block regression.

A specification of the joint distribution of Y_a, Y_b by a saturated *multivariate regression chain model* has $(\Sigma_{aa.b}, \Pi_{a|b})$ as parameters for the conditional distribution of Y_a given Y_b and Σ_{bb} for the marginal distribution of Y_b. With a saturated *block regression chain model* the parameters are the regression coefficients obtained as described above from $(\Sigma^{aa}, \Sigma^{ab})$ and the concentration matrix $\Sigma^{bb.a} = \Sigma_{bb}^{-1}$.

Considering, for instance, a multivariate regression chain model instead of a multivariate regression model can lead to a simpler structure. This is the case in Example 7 but not in Example 2 of Section 5 since the explanatory variables can be taken to be marginally uncorrelated in the former but not in the latter.

In the next more complex regression chain model the joint density of three (vector) variables Y_a, Y_b and Y_c is specified via

$$f_{abc} = f_{a|bc}f_{b|c}f_c,$$

that is, via a regression of Y_a on Y_b and Y_c, a regression of Y_b on Y_c and the marginal distribution of Y_c. This would be an adequate approach if the components of Y_a are the response variables of primary interest having Y_b and Y_c as potential explanatory variables, if Y_b plays the role of an intermediate variable containing potentially explanatory components for Y_a and possible responses to Y_c and, finally, if Y_c consists of explanatory variables whose joint distribution is to be analyzed.

A particularly important family of regression chains are the *univariate recursive regressions* in which, for a given ordering of the components of $Y = (Y_1, \ldots, Y_q)^T$, we define the model via the regression of Y_r on $Y_{r+1}, \ldots, Y_p$ for $r = 1, \ldots, q$; $q \leq p-1$. An independence hypothesis is said to be *decomposable* if it specifies one or more of the regression coefficients in such a system to be zero. Early descriptions of univariate recursive regressions have been given by Wright (1921, 1923) with an emphasis on applications in genetics and by Tinbergen (1937) for the study of business cycles.

By contrast a *nondecomposable independence hypothesis* consists of a set of k independence relations for k distinct variable pairs that cannot, in its entirety, be reexpressed in terms of vanishing coefficients in the above form: that is, no ordering of the variables would produce a decomposable independence hypothesis with the same implications from the same distributional assumption. The following arguments apply provided that there are no so-called forbidden states, that is, states of zero probability (Dawid, 1979a).

For instance, for a trivariate normal distribution of Y, Z, X the hypothesis $Y \perp\!\!\!\perp X \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$ corresponds to zero concentrations for pairs (Y, X) and (X, Z) and it implies $X \perp\!\!\!\perp (Y, Z)$. This hypothesis can be reexpressed by $Y \perp\!\!\!\perp X \mid Z$ and $X \perp\!\!\!\perp Z$ corresponding to $\beta_{yx.z} = \beta_{xz} = 0$ in a univariate recursive system for $(Y, X, Z)^T$. Thus the hypothesis is decomposable even though initially not expressed in that form. On the other hand, no ordering of the variables would permit us to specify the hypothesis $Y \perp\!\!\!\perp X$ and $Z \perp\!\!\!\perp U$ as zero restrictions in a univariate recursive regression system. Thus the hypothesis is nondecomposable. Further examples for nondecomposable hypotheses are discussed in Section 5.

They arise in applications with four or more variables, as we shall see below, but suffer from a number of disadvantages both in terms of the difficulty of fitting, but more importantly, in terms of indirectness of interpretation. The need for such models was noted by Haavelmo (1943) who pointed out substantive research questions about relations which form a system of equations to be fulfilled simultaneously, but which are not a system of univariate recursive regressions. His subject matter example is as follows: consumption in an economy per year depends on total income, investment per year depends on consumption and total income is the sum of consumption and investment. A slightly simplified version of Haavelmo's argument for the simultaneous treatment of equations is given in Section 4. As a consequence of his results, the class of linear structural equations was developed to study simultaneous relations. It is mainly discussed in econometrics (Goldberger, 1964), in psychometrics (Jöreskog, 1973) and in sociology (Duncan, 1969); it includes univariate recursive regression systems and multivariate regressions as a subclass but, in general, a zero coefficient in a structural equation does not correspond to an independence relation. More generally the graphical representations to be introduced in Section 3 are equivalent to those used in path analysis and in discussions of structural equations only in rather special cases. We deal with this important point further in Section 4.

A representation in terms of univariate recursive regressions combines several advantages. First, and most importantly, it describes a stepwise process by which the observations could have been generated and in this sense may prove the basis for developing potential causal explanations. Second, each parameter in the system has a well-understood meaning since it is a regression coefficient: that is, it gives for unstandardized variables the amount by which the response is expected to change if the explanatory variable is increased by one unit and all other variables in the equation are kept constant. As a consequence, it is also known how to interpret each additional zero restriction: in the case of jointly normal variables, each added restriction introduces a further conditional independence, and it is known how parameters are modified if variables are left out of a system (Wermuth, 1989). Third, general results are available for interpreting structures, that is, for reading all implied independencies directly off a corresponding graph (Pearl, 1988; Lauritzen et al., 1990) and for deciding from the graphs of two distinct models whether they are equivalent (Frydenberg, 1990a). Fourth, an algorithm exists (Pearl and Verma, 1991; Verma and Pearl, 1992) which decides for arbitrary probability distributions and an almost arbitrary list of conditional independence statements whether the list defines a univariate recursive system; if it does, a corresponding directed acyclic graph is drawn. Fifth, the analysis of the whole association structure can be achieved with the help of a sequence of separate univariate linear regression analyses (Wold, 1954).

The word *causal* is used in a number of different senses in the literature; for a review see Cox (1992). Glymour et al. (1987) and Pearl (1988) have developed valuable procedures for finding relatively simple structures of conditional independencies which they define to be causal. We prefer to restrict the word to situations where there is some understanding of an underlying process. From this perspective it is unrealistic to think that causality could be established from a single empirical study or even from a number of studies of similar form. We aim, however, by introducing appropriate subject matter considerations into the empirical analysis, to produce descriptions and summaries of the data which point toward possible explanations and which in some cases of univariate recursive systems could be consistent with a causal explanation.

3. SOME GRAPHICAL REPRESENTATIONS

With only three component variables, the number of possible special independency models is fairly small but with four and more components there is a quite rich and potentially confusing variety of special cases to be considered. Graphical representation helps clarify the various possibilities, and it is convenient to introduce the key ideas and conventions in terms of three variables.

A systematic account of graphical methods by Whittaker (1990) emphasizes undirected graphs, that is, systems in which all variables are treated on an equal footing. Here we use largely directed graphs to emphasize relations of response and dependence; it is fruitful also to allow two different kinds of edge between the nodes of a graph and to introduce some additional special features.

First we introduce, where appropriate, a distinction between the response variables of primary interest, one or more levels of intermediate response variables, and explanatory variables, all in general with several component variables. The distinction between variable types is usually introduced on a priori subject matter considerations, for example via the temporal ordering of the variables. Sometimes, however, there are several such provisional interpretations and some may be suggested by the data under analysis. The distinction between variable types is expressed in the graphs via (c) below.

The following conventions have been used in constructing the graphs in this paper and are illustrated in their simplest form in Figures 1–3 for three variables:

(a) each continuous variable is denoted by a node, a circle;

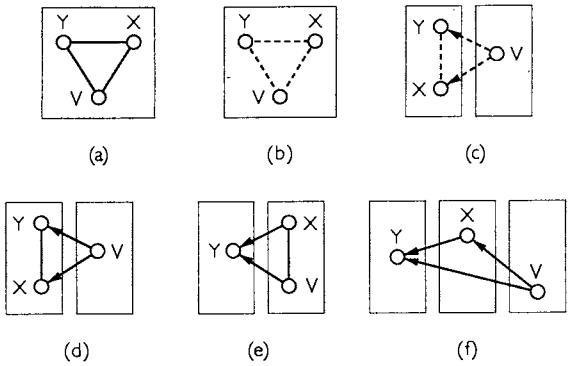


FIG 1. *Six distributionally equivalent ways of specifying a saturated model for three variables. (a) Joint distribution of Y, X, V with three substantial concentrations: (b) joint distribution of Y, X, V with three substantial covariances; (c) multivariate regression chain model with regressions of Y on V and of X on V and with correlated errors; (d) block regression chain model with regressions of Y on X, V and of X on Y, V; (e) univariate regression of Y on X, V and joint distribution of X, V; (f) univariate recursive regression system with Y as response to X, V; X as intermediate response to V. For instance, graph (e) with double lines round the right-hand box would represent the standard linear model for regression of Y on fixed explanatory variables X, V.*

(b) there is at most one connecting line between each pair of nodes, an edge;

(c) variables are graphed in boxes so that variables in one box are considered conditionally on all boxes to the right (in line with the notation $P(A \mid B)$ for the probability of A given B) so that the response variables of primary interest are in the left-hand box and its explanatory variables are in boxes to the right;

(d) if full lines are used as edges, each variable is considered conditionally on other variables in the same box (as well as those to the right), whereas if dashed lines are used variables are considered ignoring other response variables in the same box, that is, marginally with respect to response variables in the same box;

(e) the absence of an edge means that the corresponding variable pair is conditionally independent, the conditioning set being as specified in (d);

(f) variables in the same box are to be regarded

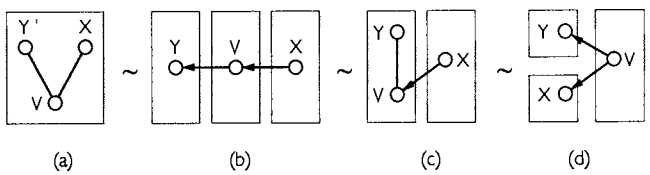


FIG. 2. *Four distributionally equivalent ways of specifying $Y \perp\!\!\!\perp X \mid V$; (a) covariance selection model for Y, X, V having parameters $\rho_{yv.x} \neq 0$, $\rho_{xv.y} \neq 0$, and $\rho_{yx.v} = 0$; (b) univariate recursive regression model with $\beta_{yv.x} \neq 0$, $\beta_{yx.v} = 0$, $\beta_{vx} \neq 0$; (c) block regression chain model with Y, V as joint responses to X and with independent parameters $\rho_{yv.x} \neq 0$, $\beta_{yx.v} = 0$, $\beta_{vx.y} \neq 0$; (d) two independent regressions of Y on V and of X on V with $\beta_{yv} \neq 0$, $\beta_{xv} \neq 0$, $\rho_{yx.v} = 0$.*

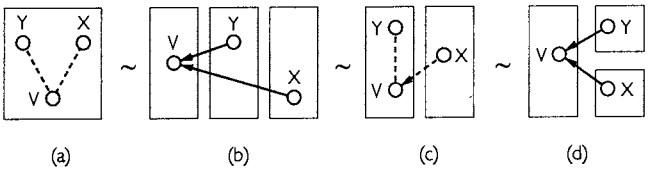


FIG. 3. *Four distributionally equivalent ways of specifying $Y \perp\!\!\!\perp X$; (a) linear structure in covariances with $\rho_{yv} \neq 0$, $\rho_{xv} \neq 0$, $\rho_{yx} = 0$; (b) univariate recursive regression model with $\beta_{vx.y} \neq 0$, $\beta_{vy.x} \neq 0$, $\beta_{yx} = 0$; (c) multivariate regression chain model with $\rho_{yv.x} \neq 0$, $\beta_{vx} \neq 0$, $\beta_{yx} = 0$; (d) multiple regression of V on two independent regressors Y, X, with $\beta_{vy.x} \neq 0$, $\beta_{vx.y} \neq 0$, $\rho_{yx} = 0$.*

in a symmetrical way, for instance as both response variables, and connected by undirected edges (lines without arrowheads, for correlations), whereas relations between variables in different boxes are shown by directed edges (arrows, for regression coefficients) such that an arrow points from the explanatory variable to the response;

(g) graphs drawn with boxes represent substantive research hypotheses (Wermuth and Lauritzen, 1990) in which the presence of an edge means that the corresponding partial correlation is large enough to be of substantive importance. This corresponds to the notion that the model being represented is the simplest appropriate one in the sense that relations considered to be unimportant are not part of the model; graphs obtained by removing the boxes represent statistical models in which a connecting edge places no such constraint on the correlation, that is, it could also be a zero correlation;

(h) a row of unstacked boxes implies an ordered sequence of (joint) responses and (joint) intermediate responses, each together with their explanatory variables. Boxes are stacked if no order is to be implied, in order to indicate independence of several (joint) variables conditionally on all boxes to the right;

(i) if the right-hand box has two lines around it, then the relations among variables in this box are regarded as fixed at their observed levels; this is to indicate a regression model instead of a regression chain model, the latter containing parameters also for those components which are exclusively explanatory.

In the present paper we use only graphs with edges of one type, that is, either all full lines or all dashed lines. It would be possible to have mixture of the two types of edge in the same graph, for example provided that all the edges within one block are of the same type and all the edges directed at a particular block are of the same type.

In a sense the distinction between full and dashed edges serves a double purpose. The distinction between full and dashed arrows from one box to another determines the different conditioning sets used in the various regression equations under consideration. The

distinction between full and dashed lines within a box specifies whether it is the concentration or the covariance matrix of the residuals that is the focus of interest. In this sense the nature of the edges corresponds to the parameters of interest.

The joint distribution of all variables is in the present context specified by the vector of means and the covariance or the concentration matrix. However any such given matrix may correspond to a number of models with quite different interpretations in the light of the distinction between types of variable as response, intermediate response or explanatory variable. A complete graph, that is, one in which all edges are present, represents a saturated model, that is, in the present context a model without any specified independence relations.

To stress the distinction between the multivariate regression and block regression contained in Figure 1, we write the corresponding equations explicitly. The multivariate regression equations implied by Figure 1c are

$$E(Y \mid V = v) - \mu_y = \beta_{yv}(v - \mu_v),$$
$$E(X \mid V = v) - \mu_x = \beta_{xv}(v - \mu_v),$$

with

$$\operatorname{cov}(\varepsilon_{y.v}, \varepsilon_{x.v}) = \rho_{yx.v}(\sigma_{yy.v}\sigma_{xx.v})^{1/2}.$$

By contrast the block regression equations implied by Figure 1d are

$$E(Y \mid X = x, V = v) - \mu_y = \beta_{yx.v}(x - \mu_x) + \beta_{yv.x}(v - \mu_v),$$
$$E(X \mid Y = y, V = v) - \mu_x = \beta_{xy.v}(y - \mu_y) + \beta_{xv.y}(v - \mu_v),$$

with

$$\beta_{yx.v} = \rho_{yx.v}(\sigma_{yy.v} / \sigma_{xx.v})^{1/2}, \beta_{xy.v} = \rho_{yx.v}(\sigma_{xx.v} / \sigma_{yy.v})^{1/2},$$
$$\operatorname{cov}(\varepsilon_{y.xv}, \varepsilon_{x.yv}) = -\rho_{yx.v}(\sigma_{yy.xv}\sigma_{xx.yv})^{1/2},$$

where the conditional variance of the variable given all remaining variables is the reciprocal value of a precision, for example, $\sigma_{yy.xv} = 1 / \sigma^{yy}$. Relations between the sets of parameters in the two types of regressions are given by Equations (6) and (7).

4. RELATIONS WITH PREVIOUS WORK

We illustrate the distinction between the graphical chain models of the present paper and structural equation models via two examples. Suppose first that X and Y are standardized to mean zero and variance one and denote their correlation coefficient by ρ. Then

$$Y = \rho X + \varepsilon_y, \; X = \rho Y + \varepsilon_x,$$

where $(\varepsilon_y, \varepsilon_x)$ are residuals from linear regression equations. That is, the coefficients ρ in these equations have an interpretation as regression coefficients. Direct calculation shows that

$$\operatorname{cov}(\varepsilon_y, \varepsilon_x) = \operatorname{cov}(Y - \rho X, X - \rho Y) = -\rho(1 - \rho^2),$$

which is nonzero unless $\rho = 0$. That is, the two regression equations imply correlated residuals except for degenerate cases.

On the other hand, if we were to adopt

$$Y - \rho X = \varepsilon_y, \; X - \rho Y = \varepsilon_x$$

as structural equations with uncorrelated residuals, then another direct calculation shows that the regression of Y on X is

$$E(Y \mid X = x) = \frac{E(YX)}{E(X^2)}x = \frac{\operatorname{var}(\varepsilon_y) + \operatorname{var}(\varepsilon_x)}{\operatorname{var}(\varepsilon_y) + \rho^2\operatorname{var}(\varepsilon_x)}\rho x$$

which is not ρx, again unless $\rho = 0$. That is, the coefficients in these structural equations do not have an interpretation as regression coefficients, as was noted by Haavelmo (1943).

To make the related point that missing edges in the graphical representation of linear structural equations (Van de Geer, 1971) do not in general have the independency interpretation of chain graphs, consider the following two structural equations

$$Y + \gamma_{yx}X + \gamma_{yv}V = \varepsilon_y,$$
$$\gamma_{xy}Y + X + \gamma_{xw}W = \varepsilon_x,$$

illustrated in Figure 4. For correlated errors $(\varepsilon_y, \varepsilon_x)$, a count of parameters shows that this represents a saturated model; that is, it allows an arbitrary covariance matrix for $(Y, X, V, W)^T$. That is, in particular, the missing edges between V and X, and between W and Y do not imply independencies, conditional or unconditional. For some further discussion of possibilities for interpreting the parameters in this model see Wermuth (1992) and Goldberger (1992). For linear structural equations in general, the interpretation of equation pa-

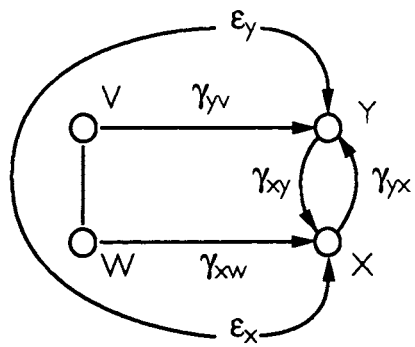


FIG. 4. *Graphical representation of two structural equations in which the missing edges for (V, X) and (W, Y) do not correspond to independencies and do not restrict the covariance matrix for* $(Y, X, W, V)^T$.

rameters, be they present or missing, has to be derived from scratch for each model considered.

However, an interpretation in terms of independencies is available also for structural equations, whenever such a model is *distributionally equivalent* to one of the chain graph models, that is, if the same joint distribution holds for the two types of models, possibly specified in two distinct ways, *and* the parameter vectors of the two models are in one-to-one correspondence.

Three classes or families of models can be identified to have this property. These are models that have a representation by a chain graph which is:

[1] a *covariance graph*, that is, a single box graph in which all present edges are undirected dashed lines, as in Figures 1b and 3a;

[2] a *multivariate regression graph*, that is, a two-box graph in which all present edges are dashed, being lines within and arrows between boxes, as in Figures 1c and 3c and in which the right-hand box has two lines around it, the distribution of its components being fixed.

[3] a *univariate recursive regression graph*, that is, a graph of $q + 1$ boxes, q of them with a single response variable and the right-hand box with $p - q$ additional explanatory variables, as in Figures 1f, 2b and 3b. In addition the right-hand box has two lines around it to indicate that only the conditional distribution of $Y_1, \ldots, Y_q$ given the remaining variables is the model of interest.

The conventions (a) to (i) for constructing chain graphs imply for univariate recursive regression graphs that arrows have the same interpretation no matter whether they are all dashed or whether they are all full arrows. That is whenever there are no proper joint responses in a model then dashed and full edge arrows are interpreted in the same way.

To distinguish better between dashed and full-edge graphs when their interpretation differs we suggest speaking further of:

[4] a *concentration graph*, that is, a single box graph in which all edges are undirected full lines, as in Figures 1a and 2a;

[5] a *block regression graph*, that is, a two-box graph in which all present edges are full, being lines within and arrows between boxes, as in Figures 1d and 2c and in which the right-hand box has two lines around it.

Then, a *multivariate regression chain graph* can be viewed as a combination of a (sequence of) graph(s) [2] with [1] and a *block regression chain graph* as a combination of a (sequence of) graph(s) [5] with [4]. More general chain graphs with both types of edges result as further combinations of these four building blocks.

Univariate recursive regression graphs are essentially identical to the *directed acyclic graphs* used in work on expert systems (Pearl, 1988). One of the latter results from one of the former by replacing the complete undirected graph of the explanatory variables by an acyclic orientation, that is, by a univariate recursive regression graph in arbitrary order of the nodes and by discarding all boxes.

To investigate distributional equivalence it is helpful to use the notion of a skeleton graph introduced by Verma and Pearl (1992). A *skeleton graph* is obtained from our Figures by removing boxes and arrows and ignoring the type of edge. For instance, the skeleton graphs in Figures 2a to 2d are all the same. If the skeletons differ then the corresponding models cannot be equivalent. But if the skeletons are the same, then the graphs may still imply different independencies, as in Figures 2 and 3.

Distributional equivalence to a model of univariate recursive regressions is closely tied to our notion of a nondecomposable independence hypothesis. We speak of a *decomposable model* if it is distributionally equivalent to a model of univariate recursive regressions and of a nondecomposable model otherwise. Thus, all saturated chain models for linear relations considered in this paper are decomposable, since they all specify the same joint distribution (Figure 1). A nonsaturated model is decomposable if and only if it contains *not* even one nondecomposable independence hypothesis. In complex cases, such a model may contain large sections that are decomposable and in analysis and interpretation account can be taken of that.

This notion of a decomposable model coincides with the notion of a decomposable graph when this graph has undirected full edges, that is, when it is a concentration graph. For variables with a joint normal distribution a concentration graph specifies a covariance selection model (Dempster, 1972). Such a model is decomposable if and only if the concentration graph is triangulated, that is, if it does not contain a chordless n-cycle for $n \geq 4$ (Wermuth, 1980; Speed and Kiiveri, 1986). A sequence of nodes $(a_1, \ldots, a_n)$ is said to form a *chordless n-cycle* in a chain graph if only consecutive nodes and the endpoints of the sequence are connected by edges and a chordless cycle in a sequence of four or more variables characterizes a nondecomposable independence hypothesis in concentrations. An example is Form (i) for (Y, X, V, W) discussed in Section 5. A special well-studied example of a decomposable covariance selection model is represented by a *chordless n-chain* in concentrations, that is, sequence of nodes $(a_1, \ldots, a_n)$ for which only consecutive nodes of the sequence are connected by edges. This is a Markov

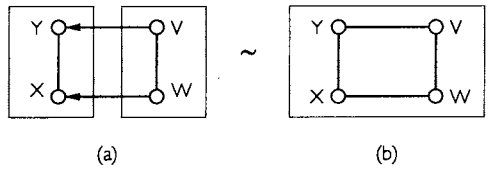


FIG. 5. *Block regression chain model (a) and covariance selection model (b) both specifying the nondecomposable hypothesis* (i): $Y \perp\!\!\!\perp W \mid (X, V)$ *and* $X \perp\!\!\!\perp V \mid (Y, W)$.

chain model. An example is Form (vi) for (Y, X, V, W) discussed in Section 5.

Figures 1–3 show that not only full-edge but also dashed-edge chain graph models can be decomposable, that is, distributionally equivalent to a model of univariate recursive regressions. We characterize situations in which this is not possible for four variables in the next section.

5. SOME EMPIRICAL EXAMPLES

We now introduce eight special kinds of independence hypothesis for four variables, together with their associated graphs, and illustrate most of them via empirical examples. All involve two or more independency conditions. The special structures we shall consider are as follows, the first three and the last two being nondecomposable:

(i) $$Y \perp\!\!\!\perp W \mid (X, V) \text{ and } X \perp\!\!\!\perp V \mid (Y, W),$$

(see Figures 5a and 5b) called the chordless four-cycle in concentrations and which correspond to the vanishing of two elements in the concentration matrix, and hence to a special case of the covariance selection models (Dempster, 1972). It can also be viewed as a chordless four-cycle in a block regression chain model with joint responses Y, X and joint explanatory variables V, W. Next we consider

(ii) $$Y \perp\!\!\!\perp W \mid V \text{ and } X \perp\!\!\!\perp V \mid W,$$

called a chordless four-cycle in a multivariate regression chain model (see Figure 6a) and which contains regressions of Y and X on V and W, being a special case of the seemingly unrelated regressions of Zellner (1962);

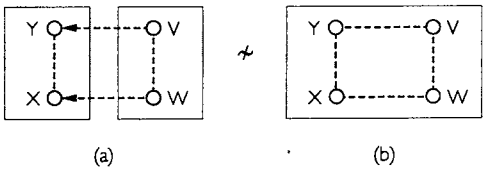


FIG. 6. *Multivariate regression chain model (a) specifying the nondecomposable hypothesis* (ii): $Y \perp\!\!\!\perp W \mid V$ *and* $X \perp\!\!\!\perp V \mid W$ *and a linear in covariances structure (b) specifying the nondecomposable hypothesis* (iii): $Y \perp\!\!\!\perp W$ *and* $X \perp\!\!\!\perp V$.

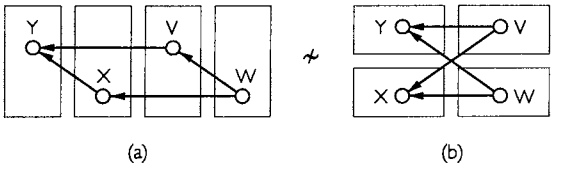


FIG. 7. *Univariate recursive regressions (a) specifying* (iv): $Y \perp\!\!\!\perp W \mid (X, V)$ *and* $X \perp\!\!\!\perp V \mid W$ *and independent multiple regressions with independent explanatory variables (b) specifying* (v): $Y \perp\!\!\!\perp X \mid (V, W)$ *and* $V \perp\!\!\!\perp W$.

(iii) $$Y \perp\!\!\!\perp W \text{ and } X \perp\!\!\!\perp V,$$

called the chordless four-cycle in correlations (see Figure 6b), a special case of covariance matrices with linear structure (Anderson, 1973).

These may be contrasted with a decomposable model based on a recursive sequence of univariate regressions with Y as response to X,V,W, with X as response to V,W and with V as response to W and having restrictions on the same two variable pairs (see Figure 7a)

(iv) $$Y \perp\!\!\!\perp W \mid (X, V) \text{ and } X \perp\!\!\!\perp V \mid W.$$

Four further cases, the first two decomposable, the last two not, are

(v) $$Y \perp\!\!\!\perp X \mid (V, W) \text{ and } V \perp\!\!\!\perp W,$$

two independent regressions of Y and X on two independent regressors V and W (see Figure 7b);

(vi) $$Y \perp\!\!\!\perp (V, W) \mid X \text{ and } X \perp\!\!\!\perp W \mid V,$$

called a chordless four-chain in concentrations or a Markov chain (see Figures 8a and 8b), that is, a chordless four-chain in a system of univariate recursive regressions again with Y as response to X,V,W, with X as response to V,W and with V as response to W and having response Y and explanatory variable W as chain endpoints;

(vii) $$Y \perp\!\!\!\perp W \text{ and } X \perp\!\!\!\perp V \text{ and } V \perp\!\!\!\perp W,$$

called a chordless four-chain in covariances (see Figures 9a and 9b) or a chordless four-chain in a multivariate regression chain model with Y, X as joint responses and having explanatory variables V,W as chain endpoints;

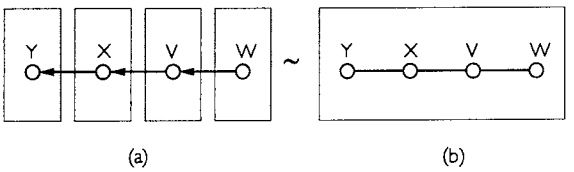


FIG. 8. *Univariate recursive regressions (a) and covariance selection model both specifying the decomposable hypothesis* (vi): $Y \perp\!\!\!\perp (V, W) \mid X$ *and* $X \perp\!\!\!\perp W \mid V$.

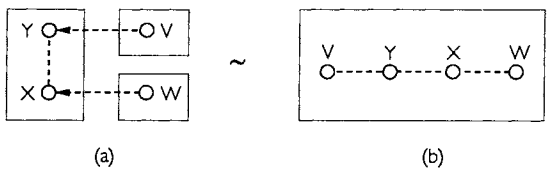


FIG. 9. *Multivariate regression chain model (a) and a linear in covariances structure (b) both specifying the nondecomposable hypothesis* (vii): $Y \perp\!\!\!\perp W$ *and* $X \perp\!\!\!\perp V$ *and* $V \perp\!\!\!\perp W$.

$$\text{(viii)} \qquad Y \perp\!\!\!\perp W \mid (X, V) \text{ and } X \perp\!\!\!\perp V \mid (Y, W) \text{ and } V \perp\!\!\!\perp W;$$

called a chordless four-chain in a block regression chain model with Y, X as joint responses and having explanatory variables V, W as chain endpoints. The corresponding chain graph has the same shape as the graph in Figure 9a, but dashed lines and arrows are replaced by full lines and arrows.

For our present purpose we give for each empirical example correlations and standardized concentrations showing these as the lower and upper triangle, respectively, such as in Table 1. This allows direct detection of linear marginal independencies between pairs of variables, as shown by very small marginal correlations, that is, standardized covariances, and linear conditional independencies between pairs of variables given all remaining variables, as shown by very small partial correlations, that is, standardized concentrations.

For a formal analysis, consistency of data with a particular structure would be examined via a likelihood ratio test or its equivalent, typically comparing a maximum likelihood fit of the constrained model with that of a saturated model. For the present purposes, however, it is enough to rely on informal comparisons of marginal correlations, partial correlations or standardized regression coefficients, although such dimensionless measures are not in general appropriate for comparing different studies.

Example 1 [*Table* 1, *Figure* 5, *Form* (i)]. Emotions as dispositions or traits of a person and emotions as states, that is, as evoked by particular situations, are notions central to research on stress and on strategies to cope with stressful events. Questionnaires with which the state-trait versions of the emotions anxiety and anger are measured have been developed by Spielberger et al. (1970, 1983). We obtained data for 684 female college students from C. Spielberger on the variables Y, state anxiety; X, state anger; V, trait anxiety and W, trait anger; summaries are displayed in Table 1.

The upper corner of Table 1 shows close agreement with the Form (i): $Y \perp\!\!\!\perp W \mid (X,V)$ and $X \perp\!\!\!\perp V \mid (Y,W)$, see also Figures 5a and 5b. This nondecomposable model has the simple interpretation that prediction of either state variable is not further improved by adding the other trait variable to the remaining two explanatory variables but it does not directly suggest a stepwise process by which the data might have been generated.

Example 2 [*Table* 2, *Figure* 6a, *Form* (ii)]. From a study of the status and reactions of patients awaiting a particular kind of operation (Slangen, Kleeman and Krohne, 1992) we obtained as basic information for 44 female patients: Y, the ratio of systolic to diastolic blood pressure; X, the diastolic blood pressure; both measured in logarithmic scale; V, body mass, that is, weight relative to height, and W, age. Table 2 shows substantial correlations except for a small marginal correlation of pair (Y,W) and a small partial correlation of pair (X,V). These are not to be directly interpreted if—as appears reasonable—each of the blood pressure variables is regarded as a potential response to body mass and age. Instead, the standardized regression coefficients in a saturated multivariate regression of Y, X on V,W display possible independencies of interest. They show close agreement with Form (ii): $Y \perp\!\!\!\perp W \mid V$ and $X \perp\!\!\!\perp V \mid W$, see also Figure 6a, with standardized regression coefficients

$$\begin{pmatrix} \hat{\beta}^*_{yv.w} & \hat{\beta}^*_{yw.v} \\ \hat{\beta}^*_{xv.w} & \hat{\beta}^*_{xw.v} \end{pmatrix} = \begin{pmatrix} 0.486 & 0.040 \\ 0.037 & -0.275 \end{pmatrix}$$

and from Table 2 correlated errors since $\hat{\rho}_{yx.vw} = -0.566$. This nondecomposable model gives as interpretation that diastolic blood pressure increases just with age

TABLE 1

Observed marginal correlations (lower half) and observed partial correlations given two remaining variables (upper half) means and standard deviations for $n = 684$ *students*

Variable	Y State anx	X State ang	V Trait anx	W Trait ang
Y: = State anxiety	1	0.45	0.47	−0.04
X: = State anger	0.61	1	0.03	0.32
V: = Trait anxiety	0.62	0.47	1	0.32
W: = Trait anger	0.39	0.50	0.49	1
Mean	18.87	15.23	21.20	23.42
Standard deviation	6.10	6.70	5.68	6.57

Data for Example 1 to Form (i): $Y \perp\!\!\!\perp W \mid (X,V)$ and $X \perp\!\!\!\perp V \mid (Y,W)$ and to Figures 5a and 5b.

TABLE 2
Observed marginal correlations (lower half), observed partial correlations given all remaining variables (upper half), means and standard deviations for n = 44 patients

Variable	Y Lratio bp	X Lsyst. bp	V Body mass	W Age
Y: = Log (syst/diast) bp	1	−0.566	−0.241	0.300
X: = Log diastolic bp	−0.544	1	−0.107	0.491
V: = Body mass	−0.253	0.336	1	0.572
W: = Age	−0.131	0.510	0.608	1
Mean	0.453	4.29	0.379	29.52
Standard deviation	0.091	0.13	0.060	10.59

Data for Example 2 to Form (ii): $Y \perp\!\!\!\perp W \mid V$ and $X \perp\!\!\!\perp V \mid W$ and to Figure 6a.

after controlling for an increase in body mass and that the ratio of systolic to diastolic blood pressure is higher the lower the body mass for persons of the same age. But again, the model does not directly suggest a stepwise process by which the data could have been generated.

Example 3 [*Table* 3, *Figure* 6b, *Form* (iii)]. In a study of strategies to cope with stressful events Kohlmann (1990) collected data for 72 students replying to a German and an American questionnaire. They are both intended to capture two similar strategies: Y, cognitive avoidance and V, blunting are thought of as strategies to reduce emotional arousal and X, vigilance and W, monitoring as strategies to reduce insecurity. The data in Table 3 agree well with Form (iii): $Y \perp\!\!\!\perp W$ and $X \perp\!\!\!\perp V$, see also Figure 6b, but not with (i) because in this case the marginal correlations but not the partial correlations are small.

It is plausible to see strong positive correlations between both pairs of similar strategies, a moderate negative correlation between each set of competing strategies measured one way and no correlation between a strategy measured with one questionnaire and the competing strategy measured with the other questionnaire. However, this structure again cannot be reexpressed with zero regression coefficients in any system of recursive univariate regressions; that is, it does not have a direct explanation as a process by which the data could have been generated.

Pairs of forms from the above special cases (i) to (iv) are mutually exclusive whenever the correlations of all variable pairs other than the two constrained pairs (Y, W) and (X, V) are substantial although with limited data it is of course possible that several different simplified structures are consistent with the data. An exception where two different sets of the above conditions may hold simultaneously is provided by (i) and (iii); that is, a chordless four-cycle in concentrations and in correlations can occur together if a very special structure is present, that is if the marginal correlations in the population satisfy orthogonalities such as

$$\begin{aligned} \rho_{yw} = 0, \rho_{xv} = 0, \\ \rho_{yv}\rho_{vw} + \rho_{yx}\rho_{xw} = 0, \\ \rho_{yv}\rho_{yx} + \rho_{vw}\rho_{xw} = 0. \end{aligned} \tag{8}$$

The next set of data is an example of this special case.

Example 4 [*Table* 4, *Figures* 5b *and* 6b, *Forms* (i) *and* (iii)]. In a study of effects of working conditions on the manifestation of hypertension, Weyer and Hodapp (1979) report the correlations among the four potential influencing variables displayed in Table 4 for 106 healthy employees. The variables, which are measured with questionnaires, are Y, nervousness; X, stress at work; V, satisfaction with work and W, hierarchical status at work. The observations agree well with both (i): $Y \perp\!\!\!\perp W \mid (X,V)$ and $X \perp\!\!\!\perp V \mid (Y,W)$ (see also Figure 5b) and with (iii): $Y \perp\!\!\!\perp W$ and $X \perp\!\!\!\perp V$ (see also Figure 6b). There is no immediate interpretation; however, one

TABLE 3
Observed marginal correlations (lower half) and observed partial correlations given two remaining variables (upper half), means and standard deviations for n = 72 students

Variable	Y Cogn. avoid.	X Vigilance	V Blunting	W Monitoring
Y: = Cognitive Avoidance	1	−0.30	0.49	0.21
X: = Vigilance	−0.20	1	0.21	0.51
V: = Blunting	0.46	**0.00**	1	−0.25
W: = Monitoring	**0.01**	0.47	−0.15	1
Mean	17.49	12.57	3.71	10.40
Standard deviation	6.77	6.39	2.12	3.07

Data for Example 3 to Form (iii): $Y \perp\!\!\!\perp W$ and $X \perp\!\!\!\perp V$ and to Figure 6b.

TABLE 4
Observed marginal correlations (lower half) and observed partial correlations given two remaining variables (upper half) for n = 106 healthy employees

Variable	Y Nervous	X Stress	V Satisf.	W Hier. Stat.
Y: = Nervousnous	1	0.33	0.26	**0.00**
X: = Stress at work	0.34	1	**0.06**	0.30
V: = Satisfaction with work	0.27	**0.04**	1	−0.35
W: = Hierarchical status	**0.01**	0.29	−0.34	1

Data for Example 4 to Forms (i) and (iii) and to Figures 5b and 6b simultaneously.

explanation for this special structure is that a different combination of the questionnaire items of X, V would lead to variables X^*, V^* such that the much simpler structure $(X^*, Y) \perp\!\!\!\perp (V^*, W)$ holds (Cox and Wermuth, 1992a). For the special structure (8) both the canonical correlations and the transformation matrix to obtain X^*, V^* can be expressed in closed form.

Example 5 [*Table* 5, *Figure* 7b, *Form* (v)]. For an analysis of aggregate economic data von der Lippe (1977) computed growth rates for 24 postwar years in Germany for Y, employment; X, capital gains; V, private consumption and W, exports. The correlation structure suggests that knowing the change in capital gain does not help in predicting the change in employment for given change levels of the demand side, that is, consumption and export (Wermuth, 1979); in addition, changes in consumption were not correlated with changes in exports. This implies two independent responses to two independent explanatory variables or close agreement to Form (v): $Y \perp\!\!\!\perp X \mid (V, W)$ and $V \perp\!\!\!\perp W$; see also Figure 7b.

Example 6 [*Table* 6, *Figure* 8, *Form* (vi)]. In a conditioning experiment with 48 subjects (Zeiner and Schell, 1971), one purpose was to examine discrimination between a noxious and an innocuous stimulus in two periods of a conditioning experiment with Y, a long-interval discriminatory response (6–10 seconds); X, a short-interval discriminatory response (1–5 seconds) in the light of earlier responses: V, the strongest response in the first interval and W, the response to an innocuous stimulus before the experiment itself; all responses are measured as skin resistance. The correlations displayed in Table 6 suggest (Hodapp and Wermuth, 1983, p. 384) a Markov structure (vi) in which $Y \perp\!\!\!\perp (V,W) \mid X$ and $X \perp\!\!\!\perp W \mid V$, see also Figures 8a and 8b, and thus in which the long-interval discriminatory response depends directly only on the short-interval discriminatory response; this short-interval response is directly dependent on the strongest response in the short interval and the latter is well predicted by just the response to an innocuous stimulus before the experiment.

Example 7 [*Table* 7, *Figure* 9, *Form* (vii)]. From an

TABLE 5
Observed marginal correlations (lower half) and observed partial correlations given two remaining variables (upper half) of growth rates for n = 24 postwar years in Germany

Variable	Y Employment	X Capital gain	V Consumption	W Export
Y: = Employment	1	**−0.11**	0.68	0.55
X: = Capital gain	0.47	1	0.50	0.43
V: = Consumption	0.67	0.55	1	−0.51
W: = Export	0.44	0.39	**0.04**	1

Data for Example 5 to Form (v): $Y \perp\!\!\!\perp X \mid (V, W)$ and $V \perp\!\!\!\perp W$ and to Figure 7b.

TABLE 6
Observed marginal correlations (lower half) and observed partial correlations given two remaining variables (upper half) for n = 48 subjects

Variable	Y Long	X Short	V Strong	W Innoc
Y: = Long int. discriminatory response	1	0.70	**−0.04**	**−0.12**
X: = Short int. discriminatory response	0.72	1	0.29	**0.14**
V: = Strongest short interval response	0.30	0.54	1	0.62
W: = Response to innocuous stimulus	0.19	0.43	0.71	1

Data for Example 6 to Form (vi): $Y \perp\!\!\!\perp (V,W) \mid X$ and $X \perp\!\!\!\perp V \mid W$ and to Figures 8a and 8b.

TABLE 7
Observed marginal correlations (lower half) and observed partial correlations given two remaining variables (upper half), means and standard deviations for n = 39 diabetic patients

Variable	Y GHb	X Knowledge	V Duration	W Fatalism
Y: = Glucose control, GHb	1	−0.431	−0.407	−0.262
X: = Knowledge, illness	−0.344	1	−0.111	−0.517
V: = Duration, illness	−0.404	**0.042**	1	−0.028
W: = Fatalism, illness	**−0.071**	−0.460	**0.060**	1
Mean	10.02	33.18	147.05	20.13
Standard deviation	2.07	7.86	92.00	5.75

Data for Example 7 to Form (vii): $Y \perp\!\!\!\perp V$, $Y \perp\!\!\!\perp W$, and $X \perp\!\!\!\perp V$ and to Figures 9a and 9b.

investigation of determinants of blood glucose control (Kohlmann et al., 1991), we have data for 39 diabetic patients, who had at most 10 years of formal schooling. The variables considered are Y, a particular metabolic parameter, the glycosylated hemoglobin GHb; X, a score for particular knowledge about the illness, V, the duration of illness in months, and W, a questionnaire score measuring the patients external attribution to "chance" of the occurence of events related to the illness; an attitude called external fatalism. The correlations in Table 7 suggest a structure of the Form (vii), that is, with $Y \perp\!\!\!\perp W, X \perp\!\!\!\perp V$, and $V \perp\!\!\!\perp W$, see also Figures 9a and 9b. One interpretation is that duration of illness and external fatalism are independent explanatory variables in two seemingly independent regressions, where metabolic adjustment is better (low values of GHb) the longer the duration of the illness, knowledge about the illness is lower the higher the external fatalism of a person, and after conditioning on duration and fatalism the metabolic adjustment is still better the higher the knowledge ($\beta_{yx.vw} = -0.431$).

6. DISCUSSION

There are a number of general issues arising from the special cases discussed in the previous section, especially the extension to more than four component variables and to models with other than only linear dependencies; for the latter see Cox and Wermuth (1993).

Graphs with, in our notation, full edges have an elegant connection with the theory of Markov random fields which allows general properties to be deduced. See Lauritzen (1989) for a survey of these topics and Isham (1981) for a review of Markov random fields in a broader context. Graphs with dashed edges, or possibly graphs with mixtures of dashed and full edges, do not have the same general features, and it is an open question as to what exactly can be said about them in generality.

There are four types of nondecomposable independence hypotheses illustrated in Section 4 for four variables, namely:

(a) *Nondecomposable hypotheses in block regression chain models* [*Form* (i), *Example* 1, *Table* 1, *Figure* 5a *and Form* (viii)]. In a block regression chain model the components, even in the simplest case, are divided into responses $Y_a = (Y, X)$ and explanatory variables $Y_b = (V, W)$ with a full directed arrow unless the corresponding regression coefficient in (3) is zero and a full undirected line for the explanatory variables unless they are marginally uncorrelated. For four variables a nondecomposable independence hypothesis in a block regression chain model is characterized by a chordless four-chain in the full edge chain graph, with the two ends of the sequence being explanatory variables, that is, for (V, Y, X, W) in our examples. Figure 5a with Form (i) gives an example of the four-cycle which contains the described four-chain, while Form (viii) leads to an example of the chordless four-chain;

(b) *Nondecomposable hypotheses in concentrations* [*Form* (i), *Example* 1, *Table* 1, *Figure* 5b]. Models of zero concentrations, that is, the covariance selection models of Dempster (1972), differ from block regression models – from (a) – in treating all variables on an equal footing, that is, having them in the same box where all edges are full undirected lines unless the corresponding variables are partially uncorrelated given the remaining component variables. For four variables a nondecomposable hypotheses in concentrations is characterized by a chordless four-cycle in the associated undirected graph of full edges, that is, in the concentration graph. Figure 5b with Form (i) gives an example of a chordless four-cycle in concentrations for (V, Y, X, W).

(c) *Nondecomposable hypotheses in multivariate regression chain models* [*Form* (ii), *Example* 2, *Table* 2, *Figure* 6a *and Form* (vii), *Example* 7, *Table* 7, *Figure* 9a]. In multivariate regression chain models the components are – as for (a) – even in the simplest case divided into responses $Y_a = (Y, X)$ and explanatory variables $Y_b = (V, W)$ with a dashed directed arrow unless the corresponding regression coefficient in (2) is zero, a dashed undirected line for the responses unless they are partially uncorrelated given the explanatory variables, and a dashed undirected line for the explanatory vari-

ables unless they are marginally uncorrelated. For four variables a nondecomposable independence hypothesis in a multivariate regression chain model is characterized by a chordless four-chain in the dashed edge chain graph with the two ends of the sequence being explanatory variables, that is, for (V, Y, X, W) in our examples. Figure 6a with Form (ii) gives an example of the four-cycle which contains the described four-chain, while Figure 9a with Form (vii) gives an example of the four-chain. Both are seemingly unrelated regressions (Zellner, 1962) together with a specification for the distribution of the explanatory variables.

(d) *Nondecomposable hypotheses in covariances [Form* (iii)*, Example* 3*, Table* 3*, Figure* 6b *and Form* (vii)*, Example* 7*, Table* 7*, Figure* 9b]. Models of zero covariances, that is, models for hypotheses linear in covariances (Anderson, 1973), have – as in (b) – a single block of variables. All edges are dashed undirected lines unless the corresponding variables are marginally uncorrelated. For four variables a nondecomposable independence hypothesis in covariances is characterized by a chordless four-chain in the associated undirected graph of dashed edges, that is, in the covariance graph. Figure 6b with Form (iii) gives an example of the four-cycle which contains a chordless four-chain, while Figure 9b with Form (vii) gives an example of the four-chain.

Models which contain even a single nondecomposable independence hypothesis cannot be distributionally equivalent to a model of univariate recursive regressions. Our examples illustrate that such nondecomposable structures arise in a number of different contexts. There is need to identify them and to find explanations of how they could have been generated. Criteria for establishing nondecomposability for more than four variables are not yet published for general dashed-edge chain graphs, while for full-edge chain graphs such criteria were given by Lauritzen and Wermuth (1989) and for undirected dashed line graphs by Pearl and Wermuth (1993).

We have in this paper concentrated on the kinds of special structure that can arise, especially on their specification and interpretation, rather than on the details of fitting and assessing model adequacy. Under normal-theory assumptions maximum-likelihood fitting and testing for nondecomposable models will call for iterative procedures. A rather general asymptotically efficient noniterative procedure based on embedding the model to be fitted in a saturated model is available (Cox and Wermuth, 1990) either for direct use or as a starting point for iteration (Jensen, Johansen and Lauritzen, 1991). Several issues are important for iterative algorithms. Is there a global maximum or are there several local maxima? Which conditions guarantee the existence of maximum-likelihood estimates? What are the convergence properties of an algorithm? Again, more is known for models represented by full-edged graphs (Speed and Kiiveri, 1986; Frydenberg and Edwards, 1989; Frydenberg and Lauritzen, 1989; Edwards, 1992) than for models with dashed edge graphs. Some of the latter may be fitted with algorithms suitable for linear structural equations; for a discussion of different alternatives see Lee, Poon and Bentler (1992).

For mixtures of discrete and continuous variables, models corresponding to chain graphs with full edges have been intensively studied (Lauritzen and Wermuth, 1989; Lauritzen, 1989; Frydenberg, 1990b; Wermuth and Lauritzen, 1990; Cox and Wermuth, 1992b; Wermuth, 1993), but for models corresponding to chain graphs with dashed edges or possibly mixtures of dashed and full edges the extensions to discrete and mixtures of discrete and continuous variables remain to be developed.

The issue of model choice in the analysis of data has too many ramifications to be discussed satisfactorily in the present paper; some different suitable strategies for analyses with a moderate number of variables are discussed in Wermuth and Cox (1992). In general, if there is sufficient substantive knowledge to give a firm indication both of the nature of the variables and of the independencies expected, then model choice consists largely of testing the adequacy of the proposed model, in particular in examining the supposedly zero correlations, concentrations and regression coefficients. The less the guidance from subject matter considerations, the more tentative will be the conclusions about model structure, but the broad principles of variable selection in empirical regression discussed, for example, by Cox (1968) and Cox and Snell (1974), will apply. In particular, where a number of different models of roughly equal complexity give satisfactory fits to the data, all should be incorporated in the conclusions, unless a choice can be made on subject matter grounds.

There are many aspects of the study of multiple dependencies and associations not addressed in the present paper. In particular the role of latent or hidden variables in clarifying the interpretation of relatively complex structures has not been dealt with, nor has the related matter of the effect of errors of observations in possibly distorting dependencies. Finally, we reemphasize the point made in Section 3 that a key argument for aiming for univariate recursive regressions consistent with subject matter knowledge is that they suggest a stepwise process by which the data might have been generated.

ACKNOWLEDGMENTS

We are grateful to Morten Frydenberg for his helpful comments, to the referees for very constructive sugges-

tions and to the British-German Academic Research Collaboration Programme for supporting our joint work.

REFERENCES

ANDERSON, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist.* **1** 135–141.

COX, D. R. (1968). Regression methods; notes on some aspects of regression analysis (with discussion). *J. Roy. Statist. Soc. Ser. A* **131** 265–279.

COX, D. R. (1992). Causality; some statistical aspects. *J. Roy. Statist. Soc. Ser. A* **155** 291–301.

COX, D. R. and SNELL, E. J. (1974). The choice of variables in observational studies. *J. Roy. Statist. Soc. Ser. C* **23** 51–59.

COX, D. R. and WERMUTH, N. (1990). An approximation to maximum-likelihood estimates in reduced models. *Biometrika* **77** 747–761.

COX, D. R. and WERMUTH, N. (1992a). On the calculation of derived variables in the analyses of multivariate responses. *J. Multivariate Anal.* **42** 167–172.

COX, D. R. and WERMUTH, N. (1992b). Response models for mixed binary and quantitative variables. *Biometrika* **79** 441–461.

COX, D. R. and WERMUTH, N. (1993). Some recent work on methods for the analysis of multivariate observational data in the social sciences. In *Conference Proceedings of the 7th International Conference on Multivariate Analysis, Pennsylvania State Univ., May 1992.* North-Holland, Amsterdam. To appear.

DAWID, A. P. (1979a). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 1–31.

DEMPSTER, A. P. (1972). Covariance selection. *Biometrics* **28** 157–175.

DUNCAN, O. D. (1969). Some linear models for two-wave, two-variable panel analysis. *Psychological Bulletin* **72** 177–182.

EDWARDS, D. (1992). *Graphical Modelling with MIM.* Manual, Univ. Copenhagen.

FRYDENBERG, M. (1990a). The chain graph Markov property. *Scand. J. Statist.* **17** 333–353.

FRYDENBERG, M. (1990b). Marginalization and collapsibility in graphical interaction models. *Ann. Statist.* **18** 790–805.

FRYDENBERG, M. and EDWARDS, D. (1989). A modified iterative proportional scaling algorithm for estimation in regular exponential families. *Comput. Statist. Data Anal.* **8** 143–153.

FRYDENBERG, M. and LAURITZEN, S. L. (1989). Decomposition of maximum-likelihood in mixed interaction models. *Biometrika* **76** 539–555.

GLYMOUR, C., SCHEINES, R., SPIRTES, P. and KELLY, K. (1987). *Discovering Causal Structure.* Academic, New York.

GOLDBERGER, A. S. (1964). *Econometric Theory.* Wiley, New York.

GOLDBERGER, A. S. (1992). Models of substance; comment on "On block recursive linear regression equations," by N. Wermuth. *Revista Brasileira de Probabilidade e Estatística* **6** 46–48.

HAAVELMO, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11** 1–12.

HODAPP, V. and WERMUTH, N. (1983). Decomposable models: A new look at interdependence and dependence structures in psychological research. *Multivariate Behavioral Research* **18** 361–390.

ISHAM, V. (1981). An introduction to spatial point processes and Markov random fields. *Internat. Statist. Rev.* **49** 21–43.

JENSEN, S. T., JOHANSEN, S. and LAURITZEN, S. L. (1991). Globally convergent algorithms for maximizing a likelihood function. *Biometrika* **78** 867–878.

JÖRESKOG, K. G. (1973). A general method for estimating a linear structural equation system. In *Structural Equation Models in the Social Sciences* (A. S. Goldberger and O. D. Duncan, eds.) 85–112. Seminar Press, New York.

KOHLMANN, C.-W. (1990). *Streßbewältigung und Persönlichkeit.* Huber, Bern.

KOHLMANN, C. W., KROHNE, H. W., KÜSTNER E., SCHREZENMEIR, J., WALTHER, U. and BEYER, J. (1991). Der IPC-Diabetes-Fragebogen: ein Instrument zur Erfassung krankheitsspezifischer Kontrollüberzeugungen bei Typ-I-Diabetikern. *Diagnostica* **37** 252–270.

LAURITZEN, S. L. (1989). Mixed graphical association models (with discussion). *Scand. J. Statist.* **16** 273–306.

LAURITZEN, S. L., DAWID, A.P., LARSEN, B. and LEIMER, H. G. (1990). Independence properties of directed Markov fields. *Networks* **20** 491–505.

LAURITZEN, S. L. and WERMUTH, N. (1989). Graphical models for association between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17** 31–57.

LEE, S.-Y., POON, W.-Y. and BENTLER, P. M. (1992). Structural equation models with continuous and polytomous variables. *Psychometrika* **57** 89–105.

PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufman, San Mateo, CA.

PEARL, J. and VERMA, T. S. (1991). A theory of inferred causation. In *Principles of Knowledge Representation and Reasoning* (J. A. Allen, R. Fikes and E. Sandewall, eds.). Morgan Kaufman, San Mateo, CA.

PEARL, J. and WERMUTH, N. (1993). When can an association graph admit a causal interpretation? In *Conference Proceedings of the 4th International Workshop on Artificial Intelligence and Statistics, Fort Lauderdale, Florida.* To appear.

SLANGEN, K., KLEEMANN, P. P. and KROHNE, H. W. (1992). Coping with surgical stress. In *Attention and Avoidance; Strategies in Coping with Aversiveness* (H. W. Krohne, ed.) 321–348. Springer, New York.

SPEED, T. P. and KIIVERI, H. T. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* **14** 138–150.

SPIELBERGER, C. D., GORSUCH, R. L. and LUSCHENE, R. E. (1970). *Manual for the State-Trait Anxiety Inventory.* Consulting Psychologists Press, Palo Alto, CA.

SPIELBERGER, C. D., RUSSELL, S. and CRANE, R. (1983). Assessment of anger. In *Advances in Personality Assessment* (J. N. Butcher and C. D. Spielberger, eds.) **2** 159–187. Erlbaum, Hillsdale, NJ.

TINBERGEN, J. (1937). *An Econometric Approach to Business Cycle Problems.* Hermann, Paris.

VAN DE GEER, J. P. (1971). *Introduction to Multivariate Analysis for the Social Sciences.* Freeman, San Francisco.

VERMA, T. S. and PEARL, J. (1992). An algorithm for deciding if a set of observed independencies has a causal explanation. In *Uncertainty in Artificial Intelligence* (D. Dubois, M. P. Wellman, B. D'Ambrosio and P. Smets, eds.) 8 323–330. Morgan Kaufmann, San Mateo, CA.

VON DER LIPPE, P. (1977). Beschäftigungswirkung durch Umverteilung? *WSI-Mitteilungen* **8** 505–512.

WERMUTH, N. (1979). Datenanalyse und multiplikative Modelle. *Allgemeines Statistisches Archiv* **63** 323–339.

WERMUTH, N. (1980). Linear recursive equations, covariance selection, and path analysis. *J. Amer. Statist. Assoc.* **75** 963–972.

WERMUTH, N. (1989). Moderating effects in multivariate normal distributions. *Methodika* **3** 74–93.

WERMUTH, N. (1992). On block-recursive regression equations

(with discussion). *Revista Brasileira de Probabilidade e Estatística* **6** 1–56.

WERMUTH, N. (1993). Association structures with few variables: characteristics and examples. In *Theory and Methods for Population Health Research* (K. Dean, ed.) 181–202. Sage, London.

WERMUTH, N. and COX, D. R. (1992). Graphical models for dependencies and associations. In *Computational Statistics, Proceedings of the 10th Symposium on Computational Statistics, Neuchâtel.* (Y. Dodge and J. Whittaker, eds.) **1** 235–249. Physica, Heidelberg.

WERMUTH, N. and LAURITZEN, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *J. Roy. Statist. Soc. Ser. B* **52** 21–72.

WEYER, G. and HODAPP, V. (1979). Job-stress and essential hypertension. In *Stress and Anxiety* (I. G. Sarason and C. D. Spielberger, eds.) **6** 337–349. Hemisphere, Washington, D.C.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Wiley, Chichester.

WOLD, H. O. (1954). Causality and econometrics. *Econometrica* **22** 162–177.

WRIGHT, S. (1921). Correlation and causation. *Journal of Agricultural Research* **20** 557–585.

WRIGHT, S. (1923). The theory of path coefficients: A reply to Niles' criticism. *Genetics* **8** 239–255.

ZEINER, A. R. and SCHELL, A. M. (1971). Individual differences in orienting, conditionality, and skin resistance responsitivity. *Psychophysiology* **8** 612–622.

ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* **57** 348–368.

in Artificial Intelligence (L. N. Kanal, J. Lemmer and T. S. Levitt, eds.) **3** 199–208. North-Holland, Amsterdam.

SPIEGELHALTER, D. J. and COWELL, R. G. (1992). Learning in probabilistic expert systems. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 447–466. Clarendon Press, Oxford.

SPIEGELHALTER, D. J., DAWID, A. P., HUTCHINSON, T. A. and COWELL, R. G. (1991a). Probabilistic causality assessment after a suspected adverse drug reaction: a case study in Bayesian network modelling. *Philos. Trans. Roy. Soc. London Ser. A* **337** 387–405.

SPIEGELHALTER, D. J., HARRIS, N. L., BULL, K. and FRANKLIN, R. C. G. (1991b). Empirical evaluation of prior beliefs about frequencies: methodology and a case study in congenital heart disease. Technical Report 91-4. MRC Biostatistics Unit, Cambridge.

SPIEGELHALTER, D. J. and LAURITZEN, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks* **20** 579–605.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (1993). *Causation, Prediction, and Search.* Springer, New York.

SRINIVAS, S. and BREESE, J. (1990). IDEAL: a software package for the analysis of influence diagrams. In *Uncertainty in Artificial Intelligence* (L. N. Kanal, J. Lemmer and T. S. Levitt, eds.) **6** 212–219. North-Holland, Amsterdam.

TARJAN, R. E. and YANNAKAKIS, M. (1984). Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM J. Comput.* **13** 566–79.

THIESSON, B. (1991). (G)EM algorithms for maximum likelihood in recursive graphical association models. Master's thesis, Dept. Mathematics and Computer Science, Aalborg Univ.

THOMAS, A., SPIEGELHALTER, D. J. and GILKS, W. R. (1992). BUGS: A program to perform Bayesian inference using Gibbs sampling. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 837–842. Clarendon Press, Oxford.

THOMPSON, E. A. (1986). Genetic epidemiology: a review of the statistical basis. *Statistics in Medicine* **5** 291–302.

TITTERINGTON, D. M., MURRAY, G. D., MURRAY, L. S., SPIEGELHALTER, D. J., SKENE, A. M., HABBEMA, J. D. F. and GELPKE, G. J. (1981). Comparison of discrimination techniques applied to a complex data-set of head-injured patients (with discussion). *J. Roy. Statist. Soc. Ser. A* **144** 145–175.

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Wiley, Chichester.

VAN DER GAAG, L. (1991). Computing probability intervals under independency constraints. In *Uncertainty in Artificial Intelligence* (P. P. Bonissone, M. Henrion, L. N. Kanal and J. F. Lemmer, eds.) **6** 457–466. North-Holland, Amsterdam.

WALLEY, P. (1990). *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London.

WARNER, H. R., TORONTO, A. F., VEASEY, L. G. and STEPHENSON, R. (1961). A mathematical approach to medical diagnosis—application to congenital heart disease. *Journal of the American Medical Association* **177** 177–184.

WERMUTH, N. (1976). Model search among multiplicative models. *Biometrics* **32** 253–263.

WERMUTH, N. and LAURITZEN, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika* **70** 537–552.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Analysis.* Wiley, Chichester.

WRIGHT, S. (1934). The method of path coefficients. *Ann. Math. Statist.* **5** 161–215.

ZADEH, L. A. (1983). The role of fuzzy logic in the management of uncertainty in expert systems. *Fuzzy Sets and Systems* **11** 199–228.

Comment: Assessing the Science Behind Graphical Modelling Techniques

A. P. Dempster

These papers, labelled here CW (Cox and Wermuth) and SDLC (Spiegelhalter, Dawid, Lauritzen and Cowell), are welcome reviews of extensive collaborations. CW are the more limited of the pair in their aims, making a few points convincingly, most notably (1) that covariance-based regression models are conceptually distinct from the simultaneous causal models of econometrics, even when both varieties are expressed through identical linear equations, and (2) that models with covariance matrices corresponding to restricted graphical structures often give good fits to empirical matrices. The SDLC paper by contrast is a tour de force that aims to leave no relevant topic unmentioned.

Both sets of authors intend their formal models and computations to speak to issues of scientific knowledge and science-based decision making, and in particular both are concerned about the informal scientific understanding that motivates their formal models. CW are reluctant to use the term "causal," viewing it as too ambiguous, but the authors substitute nonspecific language such as "appropriate subject matter considerations." SDLC, in contrast, discuss "influence" and "relevance" that take "account of one's understanding of causal structure." The difference appears to be that CW wish to hold to the idea that informal prior knowl-

A. P. Dempster is Professor of Statistics, Harvard University, Statistics Department, Science Center, 1 Oxford Street, Cambridge, Massachusetts 02138.

edge of the phenomena is limited to descriptive aspects, while SDLC accept that understanding of causal processes is part of normal scientific thinking. My own view is closer to the latter: I do not see that informal causal understanding can or should be suppressed (Dempster, 1990). On the other hand, the formal statistical models of these papers need not to be interpreted causally, because the essential role of probability models is to produce inferences and predictions derived from uncertainty relations. Probabilistic causation strikes me as an oxymoron, since probability quantifies progressions of internal uncertain knowledge while causation identifies external mechanics of change. The graphical model restrictions proposed in both papers depend for support and credibility on the quality of the underlying science, including causal interpretations, and on how aptly the formal models capture that science, but the uses of the models are mainly inferential and decision oriented.

CW choose to make their arrows point away from explanatory variables and toward response variables, whereas SDLC make their arrows point away from the disease node about which predictions are desired and toward the predictive variables. It is interesting that the medically driven "algorithmic approach" to the problem of telephone diagnosis of blue babies reported in Franklin et al. (1991) also reverses the statistically driven SDLC choice. My sense is that the CW and Franklin et al. strategies represent mainline scientific thinking, and that SDLC may weaken their claim to be able to represent genuine medical expertise by bucking the tide. It is normal for a clinician to absorb seriatim relatively simple pieces of information about a patient and to attempt to reason from these data to disease states. Disease states are relatively complex and sometimes amorphous constructs, but when they are well defined and separated, as in the blue baby example, it is relatively easy to mentally cycle through looking for a match of each disease to a list of symptoms. The cleverness of the Franklin et al. (1991) algorithm appears to derive from the skill of the senior coauthor in putting together a logical sequence of tests for matches involving subsets of the predictors, in such a way as to tease out, with a low error rate, which of 26 disease categories is the true one. Why do SDLC not attempt to directly probabilize the clinical expertise displayed in Franklin et al. (1991)?

The tree structure displayed in Franklin et al. (1991) is an event tree including decision nodes, of a kind commonly found in elementary decision analysis texts. By contrast, the expertise required in the modelling phase of the direct acyclic graph (DAG) approach of SDLC, with its conditioning on disease states, and its request for discernments of Markov structures, seems far from the practical expertise of clinicians. The technology seems rigid as implemented because the DAG structure is required to be the same whatever the disease. In principle, SDLC and Franklin et al. (1991) are attempting the same task of constructing a set of logical constraints on multivariate outcomes, and SDLC should have the advantage because they have the more powerful tools of probabilistic logic, whereas Franklin et al. (1991) use simpler and unrealistic deterministic logic with failures of diagnoses transparently labelled on the graph. Nevertheless the brief numerical comparisons in subsection 5.3 of SDLC indicate that in the current state of the art the deterministic approach does better. It is also troubling that in Table 6 the CHILD model with assessed conditional independencies is minimally better, if at all, than the naive model that is sometimes called idiot's Bayes. Is their technology the right medicine?

Both papers surprised me by a near absence of discussion of what is known about sample selection, in contrast with their softer heuristic assessments of relations or nonrelations among variables, whether derived from subject matter considerations or causal insights. In their examples, both papers ultimately assume that samples exist from which multivariate models can legitimately be estimated. Obviously statistical relations among variables are strongly influenced by processes that select the units making up a sample. These processes operate in a social realm operationally separated from the biological processes of disease, whence the selection mechanisms can be considered causally independent of the biological mechanisms. SDLC do mention at one point the effect of an original medical judgment on the flow of referrals in examples given, but such considerations appear to play no role in discussions of what conditional independence assumptions to adopt, at least in a first pass. Can this be justified? I think not. Technical papers in applied statistics by tradition and habit move quickly to formal assumptions and in doing so hide large areas of subjective choices of (unit, variable) pairs for consideration and analysis. By implication these choices are traditionally viewed as made for good reasons, usually by the statistician's client or substantive collaborator and hence are removed from the statistical analysis. Consequences are that statisticians tend to express distorted views of the mix of subjective and objective elements in the mosaics to which their analyses contribute and pay little heed to the creative experiences of carrying analyses back to their elemental sources in accrued informal scientific knowledge. My critical attitude about the scientific limitations of many statistical models and analyses rests on a perception that their ties to overall contexts are too often too loosely tied down.

The pessimistic cast of the foregoing remarks applies to the present state of applied statistics, not the future. The understanding of complex multivariate relations and of associated computational strategies, as detailed

especially by SDLC, bodes well for ongoing development. Although critics will point to the dangers, there is also great promise in the strategy of constructing stochastic models that adequately represent uncertainties about the states of hidden aspects of complex phenomena, and hence they are bases for credible probabilistic inferences. Since SDLC are more pointed to complexity than are CW, most of my subsequent remarks are directed to their enterprise. These experienced colleagues scarcely need my advice to push on to more ambitious and intensive modelling efforts, constructing formal models of population incidence, of disease signs and their rates of progression, and of prospects for intervention and cure, before proceeding to combine such submodels into larger systems capable of supporting informed and sound decisions. As part of this process, many specific concerns may arise and lead to profitable debates, some of which I will attempt to stimulate.

A major topic of concern is "eliciting subjective judgments." Expert judgment enters model construction at the successive stages of choosing variables, choosing graphical structures and choosing numerical probabilities. The soundness of each stage merits questioning, not least as they relate to social responsibilities and public purposes, but most critical attention is focused on the last stage of probability assessment that often draws on teams of experts selected and coached for the purpose, generally by those responsible for the preceding stages of choice. A fascinating case study of elicitation is to be found in the major "NUREG-1150" study of five large nuclear power plants (U.S. Nuclear Regulatory Commission, 1990). Extensive external reviews in this case forced a delay of more than a year while elicitations were redone to improve the quality of the panels and the credibility of their judgments. The point to stress is that subjectivity does not imply freedom to choose the first numbers that come to mind, nor to choose handily available local staff as experts. Analyses depend on subjective evidence to bridge gaps and supplement inadequate empirical data bases, but such evidence is not of a wholly different character from empirical evidence, as many discussions of systems and technologies for elicitation might suggest. Numbers provided by an expert are acceptable only if there is credible evidence that they are distillations of the expert's accumulated knowledge and experience. When final inferences have sensitive dependence on expertise, the information and analyses on which expert judgments depend need to be clearly set forth so they can be challenged, debated and revised, much as statistical data analysis and models are subject to criticism and model revisions. My sense is that much remains to be done by way of developing and testing quality control standards for evidential inputs from experts. There is no subjective nirvana, just as there is no objective nirvana, but real expertise exists and substantial payoffs can be expected from using it well.

An obvious way to decrease the influence of elicited expertise on diagnosis, or on risk assessment in general, is to increase the weight of documented empirical studies, including quality-checked data bases. Again it is instructive to compare and contrast the NUREG-1150 engineering example, where the sample size is 5 from a world-wide population of order 1,000, and where the physical description of each plant, though dauntingly complex, is more easily accessible and decomposable into independent subsystems than is the human body. My sense is that in the engineering analysis vastly more data sets on components, such as data on reliabilities of various types of critical elements, were assembled than is typical of medical expert systems. In place of data on physical components, medical statistics tends to rely on collections of studies and associated meta-analyses. These give clues to variation among different patient populations, and so may help inform expert prior assessments that depend on patient flows through treatment systems. A remaining difficulty with population thinking is that any specific patient under diagnosis automatically belongs to many cross classified subpopulations, depending on age, sex, ethnic identity, social class and so forth. Available data comes from margins and complex mixtures of these populations. There is large scope to jointly model networks and mixtures to facilitate combination of data from multiply interrelated varieties of populations.

SDLC correctly adopt a general approach to treating the complexity of an individual patient as demanding stochastic modelling of interacting biological systems such as heart and lungs. The currently fashionable alternative approach to complexity through deterministic chaos theory may be promising when one subsystem operates under controlled circumstances, but is scarcely capable of faithfully representing the complex social and biological processes routinely encountered by risk assessors (cf. Casdagli, 1992). I am uncomfortable, however, with the basic principle of simplicity used by both CW and SDLC that performs radical surgery on the proliferating parameter sets of highly multivariate statistical models. DAG models are mathematically transparent, have relatively few parameters and suggest elegant and fast computational strategies. Less felicitously, however, the models are expressed through variables that are rarely better than crude proxies for hidden variables that actually express underlying causal mechanisms, whence the substantive understanding that could justify DAGs if the hidden variables were observable is less than compelling and may be misleading when used to justify DAG assumptions for simple (e.g., dichotomous) variables. I believe that wholesale parameter reduction as widely practiced

in contemporary statistics is not the only way to achieve simplicity. The main lesson that I took away from Wermuth's doctoral research (cf. Dempster, Schatzoff and Wermuth, 1977) is that smooth systems of declining parameter values are usually a more efficient way to simplify statistical complexity than sharp cutoffs that set most parameter values to zero. Computational strategies of choice then become radically different. Classical estimation techniques that are adequate with relatively few parameters must be replaced with Bayesian or similar methods that reflect prior assessments of patterns of smooth decline. Donoho et al. (1992) illustrate a notable non-Bayesian approach. My own preference is for Bayesian models with many more hidden variables and many more dependence parameters than SDLC allow, to have a reasonable possibility of capturing actual mechanisms. I believe that rapidly developing computing power and algorithms that sample posteriors should be used to implement and test more complex Bayesian models.

Beyond the elicitation of priors and beyond the problem of simplifying the complex structures of highly multivariate and selectively filtered populations encountered in real practice, there remains a gray area that SDLC address briefly in two sentences as situations where "the number of assessments made is insufficient to specify a joint distribution uniquely." The use of maximum entropy or other arbitrary prior generation principles typically leads to exactly the unrealistic procedures that the smoothing of large parameter sets is designed to avoid. SDLC fail to mention the belief function approach (Shafer, 1976) that Dempster and Kong (1988) show fits naturally into network modelling built on decompositions of evidence into independent sources similar in spirit to the "graphical modelling" approach of SDLC. It is my view as a coinventor of the BEL theory that it is a near cousin of the Bayesian strategy that descends directly from classical subjective probability and is not a foreign interloper from distant tribes of semicoherent formal systems. Unlike the naive upper and lower probability models that have been studied by Good, Walley and others, the BEL system constructs models from judgmentally independent assessments on knowledge spaces and combines the components by a simple precise rule that reduces to the Bayesian rule for combining likelihood and prior in the special Bayesian case. The chief hindrance to developing and testing BEL models for probabilistic expert systems has been computational difficulties. Shafer, Kong and others showed in the mid-1980s how to decompose BEL computations coincidentally with the parallel demonstrations of Lauritzen and Spiegelhalter (1988) that SDLC feature. But these clever algorithms only stave off computational complexity temporarily. The future of both Bayesian and BEL approaches depends on the revolution that has been gathering speed for the past five years on Monte Carlo posterior sampling.

ACKNOWLEDGMENTS

I thank Emery Brown for helpful discussions on both medical and statistical issues. My work is partially supported by ARO Grant DAAL03-91-0089 and NSF Grant DMS-90-03216.

Comment: Conditional Independence and Causal Inference

Clark Glymour and Peter Spirtes

Fourteen years ago, in an essay on conditional independence as a unifying theme in statistics, Philip Dawid wrote that "Causal inference is one of the most important, most subtle, and most neglected of all the problems of Statistics" (Dawid, 1979a). Only shortly later, several statisticians (Wermuth and Lauritzen, 1983; Kiiveri and Speed, 1982) introduced frameworks that connect conditional independence, directed acyclic graphs (hereafter DAGs) and causal hypotheses. In these models the vertices of a DAG G represent variables, and a directed edge $X \rightarrow Y$ expresses the proposition that some change in variable X will produce a change in Y even if all other variables represented in G are prevented from changing. The power and generality of DAG models derive from their dual role in representing both causal or structural claims and

Clark Glymour is Alumni Professor of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, and Adjunct Professor of History and Philosophy of Science, University of Pittsburgh. Peter Spirtes is Associate Professor of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

also patterns of independence and conditional independence constraints on distributions. The paper by Spiegelhalter, Dawid, Lauritzen and Cowell (SDLC) provides a valuable review of the current state of the art in using and constructing statistical and causal hypotheses represented by DAGs. The paper by Cox and Wermuth (CW) lays out interesting problems concerning how to generalize DAG models. Our remarks concern four issues raised – explicitly or implicitly – by these papers:

1. Do other graphical objects with a plausible causal or structural interpretation represent sets of conditional independence relations that cannot be represented by DAGs? We will briefly describe generalizations of DAG models representing marginals of distributions with latent variables and generalizations representing feedback; graphical chain models do not represent such processes.
2. Are classifications or diagnoses using Bayesian networks or DAG models more reliable than those made by other existing classification techniques? We believe the question is unsettled.
3. Besides classification, what other uses do DAG models have? We think the essential use of such models is in predicting the effects of interventions – experiments, policies, etc. – that change the joint distribution of variables in a population, and this use connects these models with analyses by Rubin (1974, 1977) and others of the invariance of conditional probabilities under interventions and with a wealth of issues in experimental design.
4. What is the state of the art of automatic techniques for constructing DAG models? We will briefly note properties of several procedures that appear to be more generally applicable than the automated search illustrated in the SDLC review.

1. DIRECTED ACYCLIC GRAPHS AND GENERALIZATIONS

After introducing a variety of graphical structures to represent patterns of conditional independence relations not represented by any DAG – "nondecomposable" sets of conditional independence relations – and illustrating these patterns in empirical examples, CW say: "Our examples illustrate that such nondecomposable structures arise in different contexts. There is need to identify them and to find explanations of how they could have been generated." This question, as we understand it, asks what sorts of causal processes might lead to nondecomposable patterns of conditional independence relations; that the issue is posed near the end of their paper suggests that when nondecomposable patterns are found, the various graphical representations CW consider have no clear interpretation as causal hypotheses. To address their question, we first briefly consider the connection between causal structure and conditional independence in DAG models, then in graphical chain models and finally in alternative generalizations of DAG models.

In various frameworks, each DAG can be paired with any member of families of probability distributions over variables represented by vertices in the graph. The frameworks differ in their selection of restrictions on graph/distribution pairs, $\langle G, P \rangle$. Common restrictions include: (1) the Markov condition (Kiiveri and Speed, 1982): for admissible $\langle G, P \rangle$ X is independent of its nondescendants in G given its parents in G; (2) the "recursive diagram" or "directed independence graph" condition (Wermuth and Lauritzen, 1983): for admissible $\langle G, P \rangle$ and a given ordering of variables, $X \rightarrow Y$ is in G if and only if Y is after X in the ordering, and Y is dependent on X conditional on the set U of all vertices (excluding X) that precede Y in the ordering; (3) the Minimality condition (Pearl, 1988): for $\langle G, P \rangle$ satisfying the Markov condition, if H is a proper subgraph of G then $\langle H, P \rangle$ does not satisfy the Markov condition; (4) positivity of distributions; (5) the DAG isomorph or Faithfulness condition (Pearl, 1988): for admissable $\langle G, P \rangle$, vertices X, Y are independent conditional on set U of vertices only if the Markov condition applied to G entails that conditional independence. The restrictions on graph/distribution pairs are related. Directed independence graphs + positivity is equivalent to Markov + Minimality + positivity. Markov + Faithfulness + positivity entails the other conditions but is strictly stronger than Markov + the other conditions.

The Markov condition, the directed independence graph condition and the Minimality condition are directly motivated by intuitions about causality reflected in statistical practice throughout the century and in philosophy of science for almost half a century. [A few examples: a special case of the Markov condition is essential to Fisher's (1951) arguments in *The Design of Experiments* and throughout subsequent work on experimental design; the Markov condition is the guiding idea of latent variable models, as Bartholomew's (1987) recent review notes (without mentioning directed graphs explicitly); the Markov and Faithfulness conditions are tacitly assumed in the arguments about model selection developed by Simon (1954) and by Blalock (1961) early in the 1960s. In philosophy of science, aspects of the directed independence graph condition, for example, were given in a condition for "probabilistic causality" proposed by Suppes (1970) and aspects of the Markov condition were given by Reichenbach (1956).]

The Faithfulness condition can be viewed as requiring stability of conditional independence over small variations in parameters in models; in other terms,

conditional independence facts are to be explained by structure alone. Under a natural parameterization of linear normal models satisfying the Markov condition for a DAG, G, the set of unfaithful distributions has zero Lebesgue measure (Spirtes, Glymour and Scheines, 1993). The Markov and Faithfulness conditions are realized – sometimes without explicit graphical representations – in a wide array of models with causal interpretations in the social sciences, epidemiology and elsewhere and in the design of experiments and derivation of null hypotheses.

For every distribution P over a set of variables V and every ordering of the variables there exists a DAG compatible with the ordering such that P satisfies the Markov and Minimality conditions and a DAG compatible with the ordering such that P satisfies the directed independence graph condition. In contrast, there are many distributions that satisfy both the Markov and Faithfulness conditions for no DAG whatsoever; even if for some orderings of the variables there is a DAG for which P satisfies the Markov and Faithfulness conditions, there may not be such a DAG for every ordering of the variables. Unlike the other combinations of assumptions, the Markov and Faithfulness conditions jointly enable independencies to give some information about the directions of edges. The distributions that CW call "nondecomposable" do not satisfy the Markov and Faithfulness conditions for any DAG. Their introduction of other graphical representations for nondecomposable distributions therefore suggests that CW are implicitly imposing the Faithfulness condition on the set of distributions represented by a DAG.

1.1 Graphical Chain Models

CW describe a number of different kinds of "block" graphs, some of which represent sets of conditional independence relations that cannot be represented by DAGs unless the Faithfulness condition is violated. Their structures include graphical chain models in the sense of Lauritzen and Wermuth (1989) (hereafter, LW), structures also discussed by SDLC. These objects contain directed edges, undirected edges and variables grouped into blocks. The blocks of variables are linearly ordered; a directed edge $X \rightarrow Y$ occurs only if X is in a block previous to Y; undirected edges can only join variables in the same block. An edge $A \rightarrow B$ or $A - B$ occurs if and only if A and B are dependent conditional on the set of all variables occurring in the same block as B or in previous blocks.

The terminology of "explanatory" and "response" variables, and other remarks in the review papers, strongly suggest that directed edges in graphical chain models are given a causal interpretation, but the causal or structural significance of blocks and undirected edges is problematic. Wermuth and Lauritzen (1990) say little more than that variables joined by undirected edges in the same block are "on an equal footing." SDLC suggest undirected edges $X - Y$ represent reciprocal causation; in some units of the population X influences Y and in other units Y influences X. Under this interpretation, the chain graph represents a mixture of two subpopulations, each represented by a different DAG. We doubt that such mixtures generally exhibit the conditional independencies represented by a graphical chain model with undirected edges, but in any case the SDLC suggestion remains to be demonstrated. If feedback processes are represented by directed cyclic graphs, then it follows from LW that graphical chain models cannot represent them. Neither do graphical chain models represent the marginal conditional independence relations among observed variables that follow by the Markov condition from DAG models with latent variables (although other sorts of graphs that CW describe, but whose causal interpretation is not clear, can represent some marginal distributions of this kind). Graphical chain models could be used to represent a collection of alternative DAG models when one is unsure as to which structure is correct and the structures share certain conditional independence properties, but SDLC and CW and the papers they review do not unequivocally offer this interpretation.

The question of how the various "nondecomposable" forms of conditional independence relations described in CW could have been generated receives a straightforward answer using different generalizations of DAG models. Rather than starting with sets of conditional independence relations, finding a graphical formalism to represent them and then asking what causal process could have generated the constraints, we start with various sorts of causal processes represented by directed graphs and ask what sort of sets of conditional independence relations or marginal conditional independence relations they generate. It is important to be willing to abandon the idea, characteristic of graphical chain models, that the absence of an edge between two variables X and Y (which has a clear causal interpretation, namely that X does not directly cause Y) must always represent some conditional independence between X and Y; otherwise one excludes the natural representation of feedback processes. Two relevant generalizations of DAG models have been investigated.

1.2 Feedback and Reciprocal Causation

For many pairs of variables, A influences B and B influences A, whether directly or through some other set of variables considered in the system. Feedback processes can be represented by time series, but for linear systems they are often represented as well by finite directed cyclic graphs (DCGs). Methods for calculating correlations for cyclic systems flow from the

work of Haavelmo (1943) and Mason (1956). Despite this pedigree, even in the linear case very little is known about the connections between DCGs and conditional independence properties. The various conditions we have mentioned extend naturally to cyclic graphs, but the relationships among the conditions are different in cyclic and acyclic graphs. In the acyclic case, it is possible to define a graphical condition, *d*-separation, (Pearl, 1988) between three disjoint sets of variables **X**, **Y** and **Z** in a DAG G, such that **X** is *d*-separated from **Y** given **Z** if and only if the Markov condition applied to G entails that **X** is independent of **Y** given **Z**. In cyclic graphs the natural extension of the Markov condition does not capture all of the atomic independencies entailed by the natural extension of *d*-separation, and some formulations of the Markov condition are uninformative when extended to cyclic graphs (at least in the linear case).

In the case of linear normal models with unspecified values of some linear coefficients, there is a clear association of families of probability distributions with cyclic graphs, but we do not know in general how to characterize the conditional independence relations a linear normal cyclic system entails for all values of its free parameters. There is a purely graphical necessary and sufficient condition for a cyclic graph to require (for all linear models associated with it) that $\rho_{XY.U} = 0$, where U is a single variable (Glymour et al., 1987). The condition is in fact equivalent to a special case of *d*-separation for cyclic graphs. We have examined several four-variable cyclic graphs, and we find that the vanishing partial correlations of second order they require (again assuming linearity) also agree with the generalization of *d*-separation to cyclic graphs. There is no established convention for association of probability distributions with DCGs in the nonlinear case, but the linear case suggests that given the "right" association *d*-separation may correctly characterize the set of conditional independence relations common to all of the distributions associated with the graph. [Added in proof: The Markov condition in fact *fails* for some linear models (with correlated errors) for DCGs. For example, $x_3 = a\,x_1 + b\,x_4 + \varepsilon_3$ and $x_4 = c\,x_2 + d\,x_3 + \varepsilon_4$ does not entail that $P_{23.14} = 0$, as required by the Markov condition for the graph $x_1 \rightarrow x_3 \rightleftarrows x_4 \leftarrow x_2$. Spirtes has proven that *d*-separation does characterize the vanishing partial correlations implied by all linear models (with corrected errors) associated with any DCGs. See Directed Cyclic Graphs, Conditional Independence, Non-Recursive Linear Structural Equation Models, Carnegie Mellon Univ. Technical Report Phil-35, Dept. of Philosophy, 1993.]

1.3 Latent Variables

Consider a DAG G representing a causal process and any associated probability distribution P, where $\langle G,P\rangle$ satisfy Markov condition. Suppose that only a proper subset **O** of variables in the graph are measured or recorded. What conditional independence relation among variables in **O** is required by the Markov condition applied to G? What graphical object represents those marginal conditional independence relations and also represents information about G? A nice answer to both questions is given in Verma and Pearl (1990). They introduce the notion of the *inducing path graph* for G which contains only measured variables in G, encodes all of the marginal conditional independence relations G entails (by the Markov condition) and includes some of the causal information represented in G.

An undirected path U between X and Y is an *inducing path* over **O** in G if and only if (i) every member of **O** on U except for the endpoints occurs at the collision of two arrowheads on the path, and (ii) for every vertex V on U where two arrowheads collide, there is a directed path from V to X or from V to Y. There is an inducing path between X and Y in G over **O** if and only if X and Y are not independent conditional on any subset of $\mathbf{O}\backslash\{X,Y\}$. For variables X, Y in **O**, in the inducing path graph H for G over **O**, $X \leftrightarrow Y$ in H if and only if there is an inducing path between X and Y over **O** in G that is directed into X and also directed into Y; there is an edge $X \rightarrow Y$ in G if and only if there is no edge $X \leftrightarrow Y$ in H, and there is an inducing path between X and Y over **O** in G that is out of X and into Y. (It is easy to show that there are no inducing paths connecting X, Y in G over **O** that are not directed into X or into Y.) The two kinds of edges in an inducing path graph H have a straightforward causal interpretation: A directed edge $X \rightarrow Y$ occurs in H only if there is a directed path from X to Y in G, that is, X is a cause of Y; a double-headed edge $X \leftrightarrow Y$ occurs in H only if there is an unmeasured T and a directed path from T to X and a directed path from T to Y, the two paths intersecting only at T, that is, only if X and Y have an unmeasured common cause.

Unfortunately, observed conditional independence relations do not generally determine a unique inducing path graph, and so both for the purpose of studying causal inference and for characterizing indistinguishability of latent variable DAG models, another structure is required. A *partially oriented inducing path graph* (or POIPG for brevity) over a subset of variables **O**, represents a class of inducing path graphs over **O** that share the same adjacencies. A POIPG looks like an inducing path graph, but with the presence or absence of some arrowheads left unspecified. A directed edge in a POIPG indicates that all inducing path graphs in the class have that edge; a bidirected edge indicates that all inducing path graphs in the class have that bidirected edge. POIPGs can have edges ending in a mark, an "o," as in X o$\rightarrow$ Y, allowing some of the inducing path graphs represented to have $X \leftrightarrow$

Y and some to have $X \rightarrow Y$. Similarly, a POIPG may contain an edge X o-o Y. Two edges sharing a vertex, each with a mark at that vertex, can be underlined, as in -o $\underline{X}$ o-, indicating that the two "o" marks cannot simultaneously be arrowheads in any inducing path graph it represents. For some latent variable causal structures and sets of measured variables, the hypothesis that one measured variable does (or does not) cause another measured variable, or that two measured variables are affected by a latent common cause, can be read from the POIPG constructed from the conditional independence relations among the measured variables.

Spirtes (1992) describes a procedure for constructing a POIPG from conditional independence relations among observed variables and optional background knowledge, and Spirtes and Verma (1992) adapt this result to provide a polynomial time procedure to decide indistinguishability (by conditional independence) of any two DAGs with latent variables, assuming the Markov condition. Three examples of POIPGs are given in Figure 1 (ii), (iii) and (vii).

The DCG models and the POIPGs provide representations of most of the nondecomposable sets of independence hypotheses discussed by CW and explain how such independence properties could be generated. Of the five nondecomposable sets of independence hypotheses CW describe, four can be generated by a feedback process or a process with unmeasured common causes and represented by a DCG or POIPG. The fifth set of nondecomposable independencies can be generated by a cyclic graph but only with special parameter values (i.e., unfaithfully). Referring to CW's eight cases:

(i) $Y \perp\!\!\!\perp W$ (X,V) and $X \perp\!\!\!\perp V \mid (Y,W)$: DCG with $Y \rightarrow X \rightarrow W \rightarrow V \rightarrow Y$ or all arrows reversed [represented by the cyclic graph in Figure 1 (i)].

(ii) $Y \perp\!\!\!\perp W \mid V$ and $X \perp\!\!\!\perp V \mid W$ [represented by the POIPG in Figure 1 (ii)].

(iii) $Y \perp\!\!\!\perp W$ and $X \perp\!\!\!\perp V$: [represented by the POIPG in Figure 1 (iii)].

(iv) (v) and (vi) are represented by DAGs.

(vii) $Y \perp\!\!\!\perp W$ and $X \perp\!\!\!\perp V$ and $V \perp\!\!\!\perp W$ [represented by the POIPG in Figure 1 (vii)]. The POIPG in Figure 1 (vii) actually represents these independence relations only under the assumption of composition; that is, that for any four disjoint sets of random variables, **X**, **Y**, **Z**, **W**, the relations $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ and $\mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z}$ entail $\mathbf{X} \perp\!\!\!\perp (\mathbf{Y},\mathbf{W}) \mid \mathbf{Z}$. Composition holds for normal distributions.

(viii) $Y \perp\!\!\!\perp W \mid (X,V)$, $X \perp\!\!\!\perp V \mid (Y,W)$ and $V \perp\!\!\!\perp W$. This set of conditional and unconditional independence relations is not represented exactly by any DCG or POIPG unless the Faithfulness condition is violated.

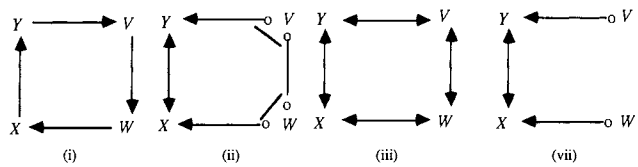


FIG. 1.

Most of the empirical examples CW give have small sample sizes, and the independence decisions are informal. In assessing the value of DCG and POIPG representations, it does not therefore seem important to consider whether feedback or latent variables are in these particular cases likely to be the correct substantive interpretations of the conditional independence relations.

2. CLASSIFICATION AND BAYESIAN NETWORKS

The construction of a Bayesian network expert system can be expensive and time consuming. Why bother? One use we can imagine is as a kind of personal calculator, a device an expert—or anyone who wishes to defer to and emulate that expert—can use to find out what her degrees of belief ought to be given various pieces of evidence. The expert, or expert emulator, can then use that information however she chooses in making decisions. In some contexts this seems to us a perfectly sensible purpose. Another conceivable purpose is to provide a system that combines prediction with explanations of how and why a prediction was obtained. Updating a Bayesian network resembles a course of reasoning, and perhaps some people may want such accounts of how predictions are obtained. But these are mostly advantages of computer-side manner. What advantages do Bayesian networks have as tools for furthering our knowledge and control of empirical domains?

Consider predictions (which we will refer to as classifictions) of a variable or variables **Y** using a set of variables **X** as predictors, for new individuals or samples drawn from a fixed distribution. There are a variety of automatic classification methods now available: neural networks, automatically constructed Bayesian networks, various forms of regression, automatically constructed decision trees and combinations of these (Shaffer, 1993). There are also a number of methods that rely on expert knowledge, such as hand-crafted decision trees and hand-crafted expert Bayesian networks. In such problems, there is a good deal of psychological evidence that computerized models of experts make better predictions in many domains than do the experts themselves, but so also do simple algorithmic prediction methods—for example, linear or logistic regression—when there is a relevant database. Do expert system Bayesian networks (or automatically constructed

Bayesian networks) have any advantages in reliability or computational ease over these other methods of classification, and if so, under what conditions?

Research has just begun on these questions, and the jury is still out on whether a Bayesian network constructed by consulting an expert makes superior classifications. SDLC note that all versions of the CHILD network with graphical structure extracted from an expert do less well at diagnostic prediction than does a "simple algorithmic" method (a hand-crafted decision tree). Moreover, SDLC compare the predictive accuracy of the network with fixed parameters estimated by the expert and with parameters changed by conditioning on data from new cases—unsurprisingly, the latter is superior—but they give no comparison with the predictive accuracy of the network when the parameters are estimated as much as possible entirely from the data. We wonder whether the model using parameter estimates based as much as possible on frequencies would (at least for some sample sizes) in this case do better than either of the methods of estimating parameters which they compare. The application of a Bayesian network constructed by consultation with an expert appears even more dubious in domains, such as psychology and sociology, in which rather less is known about causal mechanisms.

The graphical structure of Bayesian networks typically entails constraints on the joint distribution of measured variables. We expect a predictor that entails conditional independence constraints satisfied by the population distribution to have a smaller expected squared error than a predictor that does not, but the value of this advantage depends on our capacity to identify those constraints correctly: a predictor entailing a constraint false in the population will be biased. It seems to us a dicey question whether reductions in the variance of estimates are worth the risks of bias occasioned by assuming special conditional independence constraints on a distribution. Whatever the final result, it appears to us that while the method of constructing Bayesian networks with the aid of experts shows promise and is certainly worthy of further research, no decisive case has yet been made for the value of building Bayesian networks or causal models for the purpose of predicting within samples from a fixed distribution.

3. OTHER USES OF BAYESIAN NETWORKS AND CAUSAL MODELS

In the preceding section we used the qualifier "within a fixed distribution" because we believe the special value of DAG causal models is in predicting the results of interventions that change the distribution of variable values in a population. Predictions of this sort are not considered in the SDLC paper, but they are often the very point of causal models in studies that aim to influence policy. Such predictions can be made if one knows the causal structure of the systems in the population and understands the direct effects of the intervention. Unlike prediction within a fixed distribution, predictions of the outcomes of interventions absolutely require the use of the causal relations represented in the directed graph. Regression or other methods which take no account of causal structure will not suffice.

In a Bayesian network, given values for X on a new unit, we estimate the value of Y by computing the conditional probability of Y given X and doing whatever with the result. For a trivial example, suppose the network is Figure 2 (i) with binary variables, value 1 indicating the condition and 0 indicating its absence. The parameters of the network are $P(Smoking)$, $P(Yellow\ fingers \mid Smoking)$ and $P(Cancer \mid Smoking)$. If someone presents without yellow fingers we can compute $P(Cancer \mid Yellow\ fingers = 0)$; much of the SDLC review is devoted to how to perform such calculations in more complex cases. But what if, after constructing the network, we were to adopt a policy that prevents yellow fingers? Suppose we make everyone wash their hands twice a day and wear gloves in between, convenient gloves that do not make smoking more difficult and that are not carcinogenic. Assume our Bayesian network correctly describes the distribution of yellow fingers, smoking and cancer in the population before the new policy. Can the network be used to predict the probability of cancer in someone without yellow fingers after the policy is effected? Not by computing $P(Cancer \mid Yellow\ fingers = 0)$ as we did before. Instead we compute $P_{new}(Cancer \mid Yellow\ fingers = 0) = P_{new}(Cancer) = P(Cancer \mid Smoking)P(Smoking)$ (assuming after the policy is adopted no one has yellow fingers.) This is exactly the computation appropriate for the different network shown in Figure 2 (ii) with parameters $P_{new}(Yellow\ fingers)$, $P(Smoking)$, $P(Cancer \mid Smoking)$. The new network is obtained from the old by removing the directed edge from *Smoking* into *Yellow fingers*, giving *Yellow fingers* a new exogenous distribution and leaving the other parameters unchanged. The relation between the new network describing the distribution after the intervention and the original network describing the distribution before the intervention perfectly reflects the hypothetical facts: with the policy in place,

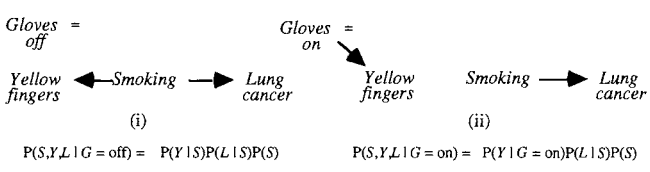


FIG. 2.

smoking no longer causes yellow fingers; the policy changes the probability of yellow fingers (to 0, or pretty close), but because yellow fingers do not cause either smoking or cancer, the new policy does not alter the joint distribution of these two variables.

Interventions to exogenously determine the distribution of values of a variable X, and that affect other variables only through X, break whatever edges into X originally obtained in the graph or graphs describing the causal structure(s) in the population and reparameterize the joint distribution accordingly. (Other kinds of interventions, which we do not consider here, may introduce as well as break edges.) We say X is "directly manipulated" by the intervention. The analysis is not ad hoc. When an intervention variable is introduced (*Gloves*, in the example) and the original distribution is understood to be conditional on a particular value of the intervention variables (e.g., *Gloves* = 0), the rule just illustrated follows from the Markov condition. A general proof is given in Spirtes, Glymour and Scheines, (1993).

This simple principle is at the center of experimental design. In graphical terms, Fisher wanted to randomize because he believed determining treatment by randomization guarantees that the structures describing the experiment will then contain no edges from causes of the outcome variable into the treatment variable. Rubin's proposals for causal inference in experimental designs (Rubin, 1974, 1977) and their extension by Pratt and Schlaifer (1988), are all consequences of the Markov condition for the special cases in which the intervention entirely determines the distribution of the variable or variables directly manipulated. The principle also explains features of distributions assumed in Bayesian discussions of experimental design (Kadane and Seidenfeld, 1992).

So it is easy in principle to determine the effects of a policy intervention provided one has a correct description of causal structure and a parameterization of the population distribution, and one knows the distribution of the directly manipulated variables that will result from the policy. Prediction of the outcomes of interventions is not so obvious if only a POIPG is available – and a POIPG is the best way we know of to characterize causal structure (without feedback) from observed conditional independence relations. There is, however, an algorithm that, given a POIPG and a set of measured variables to be directly manipulated, gives sufficient conditions and necessary conditions under which other variables can be predicted, and computes the new distribution (of a predictable variable) given the original joint distribution and the postpolicy distribution of the directly manipulated variables. (Spirtes and Glymour, 1993; Spirtes, Glymour and Scheines, 1993).

4. MODEL DISCOVERY

Extracting causal and probability information from experts can be time consuming and difficult even when the experts have real knowledge. Worse, in many problems the real knowledge of experts is quite limited, and according to a considerable psychological literature experts in many subjects know substantially less than they think they do. So we should be interested in fast, reliable procedures that can combine fragmentary prior knowledge with data to specify or partially specify causal or structural models. Few topics are more controversial in statistics, or, in our experience, more apt to draw scorn rather than research, although explicit arguments against the very idea (as opposed to arguments against particular procedures that have been proposed) tend to be feeble. For example, that "any data can be fit by several alternative models" (Rodgers and Maranto, 1989), or that there is no mechanical way to tell whether statistical dependencies are generated by an unknown causal process or by chance. Were the first objection sound a parallel would apply to all of statistical estimation. The second objection overlooks that humans can have some conviction that statistical dependencies are due to some causal process without knowing what that process is, and that even absent experimental manipulations, the very existence of a sensible model that explains puzzling features of a sample, may reasonably increase our conviction that the data are not a chance artifact.

Especially when it can be assumed that there are no latent factors at work, in our view directed graphical model specification is essentially a form of set valued estimation involving unfamiliar parameters, but subject to the same concerns for asymptotic reliability, error probabilities, variation of estimates and so on, as is ordinary parameter estimation. In the absence of strong prior information, model estimates should be set valued exactly because of indistinguishability classifications noted by SDLC. A classical version of the estimation theory should provide computable, consistent estimators; a Bayesian version should show how to compute at least the posterior mode and show that in the large sample limit the procedure yields the correct model – or class of models – almost surely.

A rudimentary theory of this kind already exists. SDLC note the results of Cooper and Herskovits (1992), which, given a linear ordering of discrete valued variables, for Dirichlet priors find the DAG compatible with the ordering and distribution that is the posterior mode on the sample evidence. Substituting a heuristic greedy search algorithm for the correct procedure, which is computationally intractable, their K2 algorithm is fast even for quite large numbers of variables and performs extremely well on simulated large sam-

ples. A non-Bayesian procedure, the PC algorithm, provably generates the set of all DAGs that represent (according to the Markov and Faithfulness conditions) a set of conditional independence facts in a distribution (assuming such a DAG exists). Prior ordering or partial ordering is optional, and the output may direct some or even all edges, depending on the structure of the correct DAG, even if no ordering information is input. The procedure minimizes the number of conditional independence tests required and the size of the set of variables conditioned on in each required test. PC has been implemented for multinormal and for multinomial variates in the TETRAD II program (Spirtes et al., forthcoming). The computational demands of the procedure depend on the sparseness of G. For fixed maximal degree, computation increases in the worst case as a polynomial function of the number of vertices. The procedure can be readily integrated with prior knowledge restricting G, and its error probabilities, as functions of sample size and average degree, have been investigated in extensive simulation studies with random graphs and randomly generated multinormal distributions. Wedelin (1993) has recently reported a procedure, so far implemented only for binary variables, that uses a parametrization related to the Fourier transform and an iterative algorithm for approximate maximum likelihood estimation of DAG models. The estimation is interleaved with an algorithm using Minimum Description Length criteria to construct a DAG, or an indistinguishability class of DAGs, from the data. The procedure is asymptotically correct for DAGs paired with faithful multinomial distributions. It does not require prior information about the ordering of the variables and has produced excellent results on simulated data with large numbers of variables.

SDLC briefly discuss the BIFROST program which generates chain graphs, described in more detail by Lauritzen, Thiesson and Spiegelhalter, (1992) (LTS) and illustrated again with data for the CHILD network. The program requires as input a partial ordering of the variables by blocks. It is not clear from this description whether the algorithm is practical for large numbers of variables, whether it is asymptotically correct, and to what extent the correct output depends on correctly specifying the block structure. We would like to know how the procedure performs on larger problems such as the ALARM network (Beinlich et al., 1989) for emergency medicine, which contains 37 variables, and has been used in tests of the reliability of the three procedures previously mentioned.

All of the algorithms so far described assume there are no latent common causes of measured variables. In real problems we often do not know at the outset whether statistical dependencies may be due to unmeasured factors affecting two or more measured variables. Absent some bound on the number of variables, there is an infinity of alternative DAGs that may accord with a set of observed conditional independence facts assuming the Markov and Faithfulness conditions, and there is no possibility of estimating a finite indistinguishability class of DAGs. What might be wanted instead are inference procedures that will describe features common to all DAGs admitting distributions yielding features of the observed marginal distribution, that is, POIPGs. It is often suggested that absent experimental interventions these kinds of inference cannot be correctly made even in principle, but with reasonable background assumptions that is not true. A correct algorithm for inferring POIPGs from conditional independence relations among observed variables is the FCI procedure given in Spirtes (1992), whose output is a POIPG. The procedure has been implemented for multinomial and multinormal distributions. The Spirtes and Verma algorithm, noted earlier, for deciding indistinguishability(by conditional independence) of DAGs with unobserved variables depends on the fact that POIPGs obtained by the FCI algorithm completely characterize the observed marginal conditional independence constraints entailed for the subset of observed variables by a DAG with latent variables. The procedure recovers each of the POIPGS (ii), (iii) and (vii) in Figure 1 from the corresponding conditional independence relations CW provide and also the undirected version of the cyclic graph in (i) (although we have no general proof that the algorithm correctly recovers cyclic graphs), as well as much more complicated structures in other cases.

Comment

Joe R. Hill

The authors of these two papers are among the most active nodes in an ever growing hypergraph of interesting papers on statistical applications of graph theory. It is an honor to discuss these two new hyperedges.

My discussion is divided into four parts. Section 1 discusses statistical applications of graph theory. Section 2 briefly describes ways of leveraging parallels between probability and database theory. Section 3 highlights two important points made in each of the papers. Finally, Section 4 asks some specific questions.

1. STATISTICAL APPLICATIONS OF GRAPH THEORY

Graph theory has a lot to offer statisticians. Consequently, graph theory is quickly becoming an integral part of modern statistics. Graphs, both directed and undirected, and hypergraphs can be used to (a) represent qualitative multivariate relationships, (b) specify and visualize multivariate statistical models, (c) determine statistical properties of multivariate models and (d) develop computationally efficient algorithms for dealing with large multivariate models. The first two of these contribute to effective communication between applications experts and statisticians. The third helps statisticians develop appropriate statistical theory. The fourth makes computing feasible for more complicated problems.

Graphical models provide a flexible paradigm for describing multivariate statistical models. They can have *discrete* variables (as in Bayesian networks, graphical and recursive loglinear models for contingency tables, and influence diagrams for applied decision analysis), or *continuous* variables (as in covariance selection and structural equation models). Conditional Gaussian models (Lauritzen and Wermuth, 1989; Wermuth and Lauritzen, 1990) provide a framework for having both kinds of variables in a single graphical model. Graphical models can have *directed* edges (as in Bayesian networks, influence diagrams and regression models) or *undirected edges* (as in graphical and decomposable loglinear models, covariance selection models and Markov random field models for image restoration). Chain graphs provide a framework for having both kinds of edges in a single graphical model.

In their paper, Cox and Wermuth (CW) introduce, for multivariate normal models, the concept of *dashed* edges as a way to represent constraints on covariance matrices (i.e., to represent marginal independencies), complementing the use of *full* edges to represent constraints on concentration matrices (i.e., to represent conditional independencies). They illustrate the use of the new enriched class of models with a number of empirical examples.

Spiegelhalter, Dawid, Lauritzen and Cowell (SDLC) give a status report on their ongoing development of Bayesian networks for expert systems. They have carefully combined a number of methods. They elicit Bayesian graphical models from medical experts. They use graphical ideas to convert the model into a computationally efficient form. They apply Bayesian estimation techniques to "learn" probability parameters as additional data are observed, and they use significance testing methods to monitor and critique the model.

SDLC provide an effective method for eliciting the qualitative, the probabilistic and the initial quantitative aspects of an expert-defined model. The key to their method is to use a directed acyclic graph to represent the qualitative relationships between variables. Nearly everything else follows from this graph.

This graph determines a recursive factorization of the joint distribution with, for each variable, a factor that is the conditional distribution of that variable, given its parents. This representation of the joint distribution has two advantages. First, the number of probabilities that the expert has to specify is considerably less than for a general joint distribution that does not encode the implied conditional independencies as efficiently. Second, these probabilities are "easy" for an expert to specify for three reasons: (a) the expert has to think about the distribution of only one variable at a time, (b) each distribution is conditioned on the parents of the variable, which are the variables that directly influence it and (c) the conditioning events can be thought of as fixed scenarios. In short, it is easy for an expert to think about the probability distribution of a single "effect" given its immediate "causes." This second advantage contrasts sharply with the problems associated with directly specifying an overall joint distribution. In that case, the expert would not be able to think conditionally but would have to think in multiple dimensions simultaneously and would typically have to specify many very small probabilities.

Once the model has been specified, it is converted to a junction tree representation for efficient computation. This conversion is carried out in a series of steps

Joe R. Hill is R & D Manager, EDS Research, 5951 Jefferson St. NE, Albuquerque, New Mexico 87109.

guided and justified by three important ideas: (a) graph separation in the moral graph of an ancestral set determines conditional independence, (b) the cliques of a chordal graph form an acyclic hypergraph and only acyclic hypergraphs have junction trees [later in the paper, their method for specifying hyper-Markov prior distributions depends on the fact, proved by Vorob'ev (1962), that a consistent set of marginal distributions has an extension iff the margins they are defined on form an acyclic hypergraph] and (c) large problems can be made computationally more tractible by decomposing them into smaller, component problems that require communication between neighboring components only.

The major point of this section has been to emphasize the important role that graph theory is playing in both of these papers. It has helped in communicating with substantive experts. It has helped in specifying and understanding multivariate statistical models. And it has helped with the computational aspects of those models. It is time that we started teaching graph theory in statistics courses of all levels.

2. LEVERAGING PARALLELS TO DATABASE THEORY

2.1 A Problem

Not everything is bliss in the world of graphical models. They have some rather subtle properties. They also lack some properties that seem at first to be trivially true. Some of the more important of these problems arise when probabilities can be zero. Although this situation does not arise in either of the papers, newcomers to graphical models might be misled into thinking that some statements made in SDLC and CW are valid in more general settings.

For example, CW state that "for a trivariate normal distribution of Y, Z, X the hypothesis $Y \perp\!\!\!\perp X \mid Z$ *and* $X \perp\!\!\!\perp Z \mid Y$ corresponds to zero concentrations for pairs (Y, X) and (X, Z) and it implies $X \perp\!\!\!\perp (Y, Z)$." Nothing could be simpler. The conditional independency $Y \perp\!\!\!\perp X \mid Z$ splits X and Y and the conditional independency $X \perp\!\!\!\perp Z \mid Y$ splits X and Z, so the two of them together split X and (Y, Z), hence they imply the conditional independency $X \perp\!\!\!\perp (Y, Z)$. For multivariate normal models, which CW are dealing with, this reasoning is fine; in fact, it is valid for any family of strictly positive probability distributions. However, if probabilities can be zero, then the result is not true! For example, the distribution $p(0, 0, 0) = p(1, 1, 1) = 1/2$, $p(x, y, z) = 0$ otherwise, satisfies the first two of these conditional independencies, but does not satisfy the third. See Moussouris (1974) and Dawid (1979b) for other examples.

The problem is that the Gibbs–Markov theorem requires strictly positive probability distributions. This positivity condition limits the possible applications of the equivalence of graph-generated conditional independence models and factorizations of joint distributions. In particular, the theorem cannot be applied to Bayesian networks with functional constraints (Lauritzen and Spiegelhalter, 1988) or to contingency tables with structural zeros or to statistical mechanics systems with forbidden states (Moussouris, 1974).

In his discussion of Besag's paper on Markov random fields in spatial statistics, Hammersley (1974) explained why he and Clifford did not publish the result when they first discovered it in 1971. He wrote (pp. 230–231),

> In proving this result, we assumed a *positivity condition*, namely that no probability should be zero. . . . In many of the most important practical applications to statistical mechanics, the physical system is subject to constraints which prevent the system from assuming certain *forbidden states*. . . . So it seemed to us not only aesthetically desirable but also practically important to amend our proof in order to make the theorem independent of the positivity condition . . . The very good reason for our failure [to do so] was the unexpected discovery by a graduate student, Mr John Moussouris, of a counter-example!

In short, Hammersley and Clifford did not publish the result because they thought the positivity condition limited the theorem too much for it to be useful in practice. Now no one doubts the importance of the theorem even with the positivity condition. But it is still quite inconvenient that no result exists for distributions with zero probabilities.

2.2 A Solution

Here is a solution that was suggested by parallels to relational database theory. Table 1 summarizes basic database/probability parallels; see Hill (1991) for more details. To state the results, we need some terminology from graph theory. A *hypergraph* is a set of nodes together with a set of hyperedges; each *hyperedge* is a subset of the nodes of the hypergraph. The *2-section* of a hypergraph is an undirected graph with the same set of nodes as the hypergraph and an edge between each pair of nodes that belong to a common hyperedge. A hypergraph is *conformal* if its set of hyperedges equals the set of cliques of the edge set of its 2-section. A hypergraph is *acyclic* if it is conformal and its 2-section is chordal. It can be shown that a hypergraph is acyclic iff it has the running intersection property iff it has a junction tree.

We also need some terminology adapted from database theory. Graph separation in an undirected graph determines a set of conditional independencies. A set

TABLE 1
Basic database and probability parallels

Probability concepts	Database concepts
Set of random variables V	Set of attributes (column names) R
Distribution for V, $p[V]$, a probability function	Relation (table) over R, $r[R]$, an indicator function for a set of tuples (rows)
Marginal distribution of $X \subseteq V$, $p[X]$	Projection of r onto $X \subseteq R$, $r[X]$
Conditional distribution $p[V \mid X = x]$	Selection $r[R \mid X = x]$
Factorization constraint $\otimes\{V_1, \ldots, V_k\}$, $V_j \subseteq V$	Join dependency $\bowtie\{R_1, \ldots, R_k\}$, $R_j \subseteq R$
Conditional independency $X \perp\!\!\!\perp Y \mid Z$, binary factorization constraint $\otimes\{X \cup Z, \ Y \cup Z\}$	Multivalued dependency $Z \twoheadrightarrow X \mid Y$, binary join dependency $\bowtie\{X \cup Z, Y \cup Z\}$

of conditional independencies is said to be *graph-generated* if there exists a graph that generates it. A conditional independency $X \perp\!\!\!\perp Y \mid Z$ *splits* variables in X from variables in Y; the variable set Z is called the *kernel* of this conditional independency. The *split graph* generated by a set of conditional independencies has an edge between every pair of variables that is not split by any of the conditional independencies in the set. The *closure* of a set of conditional independencies is the set of conditional independencies implied by the original set. Two sets of conditional independencies are said to *cover* each other if their closures are equal. A set of conditional independencies is said to be *conflict-free* if it is graph-generated and it does not split any of its kernels. Two sets of constraints are said to be *equivalent* if the sets of probability distributions that satisfy them are equal. Similar definitions have been given for databases.

The Gibbs-Markov theorem can be stated in the following three ways, each providing insight into the relationships between graphs, sets of conditional independencies and factorization constraints.

THEOREM 1+. *Let $\mathcal{G}$ be an undirected graph over V. The set of conditional independencies generated by $\mathcal{G}$ is equivalent, for strictly positive distributions, to the factorization constraint generated by the cliques of $\mathcal{G}$.*

THEOREM 2+. *Let $\mathcal{V}$ be a hypergraph over V. The set of conditional independencies implied by the factorization constraint generated by $\mathcal{V}$ is equivalent, for strictly positive distributions, to the factorization constraint generated by $\mathcal{V}$ if and only if $\mathcal{V}$ is conformal.*

THEOREM 3+. *Let $\mathbf{C}$ be a set of conditional independencies defined on V. $\mathbf{C}$ is equivalent, for strictly positive distributions, to the factorization constraint generated by the cliques of the split graph of $\mathbf{C}$.*

Fagin, Mendelzon and Ullman (1982) and Berri et al. (1983) proved the following database theorems, which, after accounting for the different terminology, look a lot like the three theorems stated above. In fact, however, because relations are indicator functions (therefore allowing zero values), these theorems, which have stronger requirements on the underlying graphical structure, suggest a way to relax the positivity condition.

THEOREM DB1. *Let $\mathcal{G}$ be an undirected graph over R. The set of multivalued dependencies generated by $\mathcal{G}$ is equivalent to the join dependency generated by the cliques of $\mathcal{G}$ if and only if $\mathcal{G}$ is chordal.*

THEOREM DB2. *Let $\mathcal{R}$ be a hypergraph over R. The set of multivalued dependencies implied by the join dependency generated by $\mathcal{R}$ is equivalent to the join dependency generated by $\mathcal{R}$ if and only if $\mathcal{R}$ is acyclic.*

THEOREM DB3. *Let $\mathbf{M}$ be a set of multivalued dependencies defined on R. $\mathbf{M}$ is equivalent to the join dependency generated by the cliques of the split graph of $\mathbf{M}$ if and only if $\mathbf{M}$ has a conflict-free cover.*

By translating database terms into probability terms (Table 1) in these three database theorems, we get the following three probability theorems, the proofs of which will be given elsewhere.

THEOREM 1*. *Let $\mathcal{G}$ be an undirected graph over V. The set of conditional independencies generated by $\mathcal{G}$ is equivalent to the factorization constraint generated by the cliques of $\mathcal{G}$ if and only if $\mathcal{G}$ is chordal.*

THEOREM 2*. *Let $\mathcal{V}$ be a hypergraph over V. The set of conditional independencies implied by the factorization constraint generated by $\mathcal{V}$ is equivalent to the factorization constraint generated by $\mathcal{V}$ if and only if $\mathcal{V}$ is acyclic.*

THEOREM 3*. *Let $\mathbf{C}$ be a set of conditional independencies defined on V. $\mathbf{C}$ is equivalent to the factorization constraint generated by the cliques of the split graph of $\mathbf{C}$ if and only if $\mathbf{C}$ has a conflict-free cover.*

Although Theorems 1*, 2* and 3* do not require strictly positive distributions, they do impose stricter constraints on the underlying graphical structures than do Theorems 1+, 2+ and 3+. Theorem 1* re-

quires the graph to be chordal for there to be equivalence, whereas Theorem 1+ puts no requirements on it. Theorem 2* requires the hypergraph to be acyclic for there to be equivalence, whereas Theorem 2+ requires only that it be conformal. Theorem 3* requires the set of conditional independencies to have a conflict-free cover for there to be equivalence, whereas Theorem 3+ puts no requirements on it (actually, the closure with respect to strictly positive distributions of a set of conditional independencies is always graph-generated).

As far as I know, Theorems 1*, 2* and 3* are new, although, by now, they are probably not unexpected.

Parallel developments in the two fields have occurred in the past, with neither aware of the other, apparently. For example, Vorob'ev's (1962) results on extending consistent marginal distributions parallel similar results for the extension of consistent databases (Beeri et al., 1983). And Beeri and Kifer's (1986a, 1986b, 1987) work on fixing sets of multivalued dependencies that have intersection anomalies parallels Dawid's (1979b) method for fixing up sets of conditional independencies.

3. MODELS AND DATA

Two simple but important points, each mentioned in both papers and neither having to do directly with graph theory, deserve to be emphasized. First, both papers take the position that a model represents the substantive knowledge that an expert brings to the problem prior to seeing specifically relevant data. One practical consequence of such a position is that statisticians cannot work in a vacuum; rather, they must interact and communicate effectively with domain specialists. And, on a more philosophical note, this position highlights the fact that a scientifically meaningful model for the data is as much a subjective prior assessment of the relative likelihood of possible values as is a scientifically meaningful model for the parameters of such a model. Second, SDLC stress and CW mention that observed data allow us not only to estimate parameters in the model but also to monitor and, if need be, to critique the model. It is refreshing to see frequentists concerned about representing expert knowledge and Bayesians worried about model criticism.

4. SOME QUESTIONS FOR THE AUTHORS

Can you have discrete variables in chain graphs with dashed edges? Can you explain why the diagnostic ability of the Bayesian network was not as good as that of the CART-like algorithm? From Table 6, it appears that for 110 cases (of 168) the Bayesian network assigned the correct diagnosis the highest probability; what were the ranks of the correct diagnoses for the other 58 cases? Has anyone created Bayesian networks with both discrete and continuous variables? Of course, with mixed models the number of parameters in each distribution will not stay fixed after updating. Has anyone considered creating a "Bayesian chip" that could be used to create truly parallel "Bayesian machines"?

Reading and thinking about these papers has been a real pleasure.

Comment: What's Next?

David Madigan

These papers represent two of the many different graphical modeling camps that have emerged from a flurry of activity in the past decade. The paper by Cox and Wermuth falls within the statistical graphical modeling camp and provides a useful generalization of that body of work. There is, of course, a price to be paid for this generality, namely that the interpretation of the graphs is more complex. I cannot resist complementing the authors on the remarkable feat of finding an example for each of the different graphical models they propose.

The paper by Spiegelhalter, Dawid, Lauritzen and Cowell falls within the probabilistic expert system camp. This is a tour de force by researchers responsible for much of the astonishing progress in this area. Ten years ago, probabilistic models were shunned by the artificial intelligence community. That they are now widely accepted and used is due in large measure to the insights and efforts of the authors, along with other pioneers such as Judea Pearl and Peter Cheeseman.

I will confine my remaining comments to the Spiegelhalter et al. paper and explore some open questions that I believe will rapidly become important, now that

David Madigan is Assistant Professor, Department of Statistics, GN-22, University of Washington, Seattle, Washington 98195.

many basic technical issues are being successfully solved.

WHAT CAN YOU DO WITH A GRAPHICAL MODEL?

My primary concern is with the apparent mismatch between the informal, qualitative character of human reasoning and the rigorous, formal, quantitative approach of graphical models (Henrion, Breese and Horvitz, 1991). Knowledge-based system builders now have access to knowledge representation tools of considerable expressive power and flexibility (e.g., Skuce, 1991) while the poor graphical modeler has to make do with nodes, links and probability distributions. These concerns are practically motivated. At the University of Washington we are constructing an intelligent tutoring system (ITS) for basic statistics. At the heart of any ITS is an explicit model of the student's knowledge. Acknowledging the inherent uncertainty, we use a Bayesian graphical model for this purpose. However, a second ITS component concerns instructional strategy—the procedural knowledge of experienced teachers. Graphical models fail dismally to represent this knowledge, yet a simple rule-based system does a reasonable job. In a project at the Fred Hutchinson Cancer Research Center in Seattle, we are constructing a knowledge-based system to assist nurses who handle telephone calls from bone marrow transplant patients and their physicians (Bradshaw et al., 1993). Graphical models can calculate the probabilities of various complications, but cannot represent the heuristic knowledge of experienced nurses as they manage the call. In general, the range of potential applications for graphical models is considerably smaller than for knowledge-based systems.

There may be a way out of this dilemma: a number of authors have suggested combining conventional knowledge-based systems with probabilistic models. The key to the success of such hybrid systems is that each component contributes to the portion of the process that it does best: the knowledge-based components guide the interaction by using rough rules-of-thumb that can help to quickly scope, categorize, gather information about, structure and interpret important aspects of the problem; the probabilistic components rely on carefully crafted assessments of uncertainty to provide specific answers about particular situations in a rigorous manner (Bradshaw et al., 1993; Szolovits and Pauker, 1978). Control rests with the knowledge-based component, which calls the probabilistic component as required.

Closely related to this is the emerging area of "knowledge-based model construction" (KBMC). The effective application of belief network tools requires a relatively high level of modeling sophistication, and model construction has proven to be a serious bottleneck. These tools contain some of the algorithms of probabilistic modeling, but cannot embody the experience and intuition of the skilled modeler. KBMC seeks to combine probabilistic modeling tools (including belief networks and influence diagrams) with a knowledge-based system that helps domain experts without extensive training in probabilistic modeling to build, evaluate and refine probabilistic models (Breese, 1989; Goldman and Breese, 1992; Holtzman, 1989). For complex problem domains, sharing and re-use of model components is vital: the knowledge base could dynamically assemble a probabilistic model, tailored to the problem at hand, from model fragments (Almond, Bradshaw and Madigan, 1993). Notable applications of KBMC technology include the Boeing Company's DDUCKS tool, a knowledge-based influence diagram workbench (Bradshaw et al., 1991) and the text understanding application of Goldman and Charniak (1992).

In short, it seems likely that in the future, graphical models will not exist as stand-alone applications, but rather will be embedded in larger systems, encompassing a variety of knowledge bases, databases and models.

MODEL UNCERTAINTY

An alternative to KBMC is to automatically induce models from existing databases. This is discussed by the authors in subsection 5.4. They begin by stating that "An approach that takes model comparison to its full consequence is to induce the network directly from data ... ignoring the prior structural and quantitative information available." Why does the "full consequence" involve the absence of prior information? One of the great advantages of the Bayesian graphical model approach is that prior knowledge, both structural and quantitative, can *realistically* be elicited and incorporated into both model selection and subsequent inference (Madigan and York, 1993). Indeed, with even a modest number of nodes, the graphical model space is vast, and there is a concern that in the absence of *some* prior knowledge, model selection procedures may fail (Draper, 1993).

Historically, model selection procedures have focused on finding the single "best" model. However, this ignores model uncertainty, leading to poorly calibrated predictions: it will often be seen in retrospect that one's uncertainty bands were not wide enough (Draper, 1993). A Bayesian solution to this problem involves averaging over all plausible models when making inferences about quantities of interest (see, for example, Raftery, 1988, and Kass and Raftery, 1993). Indeed Hodges (1987) comments that "what is clear is that when the time comes for betting on what the future holds, one's uncertainty about that future should be fully represented, and model [averaging] is the only

tool around." In many applications, however, because of the size of the model space and awkward integrals, this averaging will not be a practical proposition, and approximations are required. Draper (1993) describes "model expansion": averaging over all plausible models in the neighborhood of a "good" model. Madigan and Raftery (1991) describe an approach for Bayesian graphical models that involves seeking out the most plausible models and averaging over them. Raftery (1993) applies this to structural equation models. Madigan and York (1993) suggest a Markov Chain Monte Carlo approach that provides a workable approximation to the complete solution. These methods can also be applied to incomplete data (Madigan and Kong, in preparation). The point is that with Bayesian graphical models, correctly accounting for model uncertainty is entirely possible.

Model averaging in the context of expert systems raises special problems: displaying multiple models requires careful software design; enhanced explanation facilities are required; software for model prior elicitation is needed. The issue of compatible priors in alternative models, addressed by the authors in Section 8, is of considerable importance. While the procedure suggested seems reasonable, a more general framework is required. Certainly, when precisely specified probabilities are involved, the procedure should be used with extreme caution.

INTERCAMP COMMUNICATION

Other (independence) graphical modeling camps are to be found within decision analysis, philosophy of science and statistics. Several different camps are located in computer science. To date, these camps have communicated remarkably effectively with each other, fostering rapid progress. The challenge we face is to maintain the communication. The gulf between the two papers here demonstrates both the diversity of the progress and the extent of the challenge.

ACKNOWLEDGMENTS

I am indebted to Russell Almond, Adrian Raftery, Jeremy York and especially Jeff Bradshaw and David Draper for helpful discussions. This work was supported in part by a grant from the NSF.

Comment

Sharon-Lise Normand

1. INTRODUCTION

The authors of these two highly complementary articles are to be congratulated on their timely contributions to the readership of *Statistical Science* and to statisticians in general. The article by Spiegelhalter and colleagues provides a comprehensive review of the most recent *statistical* developments in expert systems, guiding us through a complete analysis in the expert system domain. Cox and Wermuth present a pointed discussion on the interpretation and graphical representation of linear dependencies for continuous valued random variables. In this discussion I will expand upon the range of applications of graphical models and emphasize some specific areas discussed by the authors. Specifically, my comments will address (1) the role of graphical models in statistical inference, (2) data propagation in graphs and (3) limitations of graphical models.

2. THE ROLE OF GRAPHICAL MODELS

Graphical models can play an important role in structuring statistical analyses, in performing complicated computations and in communicating results. Thus the motivation for creating a graphical representation of a statistical model is threefold: (1) the graph provides an effective vehicle for communication among researchers, (2) the graph displays a knowledge map of the dependency structure posited in the model and finally (3) the graph can be transformed into a static secondary structure that can be used for efficient probability calculations. Professor Spiegelhalter and his colleagues touch on all three reasons with emphasis placed on calculating probabilities while Professors Cox and Wermuth stress the value of the graph as a knowledge map. It is particularly important to note that one may choose to exploit any or all three reasons for using a graphical model.

The term *graphical model* has a very precise definition in the contingency table literature (Darroch, Laurit-

Sharon-Lise Normand is Assistant Professor of Biostatistics, Department of Health Care Policy, Harvard Medical School, 25 Shattuck Street, Parcel B, 1st Floor, Boston, Massachusetts 02115.

zen and Speed, 1980; Edwards and Kreiner, 1983; Wermuth and Lauritzen, 1983). In this discussion I will, however, use the term more generally to refer to statistical models that host some conditional independence properties. Hierarchical models (Lindley and Smith, 1972; Morris, 1987) are a class of statistical models that immediately come to my mind when discussing graphical models. Inherent in hierarchical models is the notion of conditional independence across observations at one stage and across parameters at another stage. Consider for example a two-stage normal hierarchical model used to combine information across experiments. The observed data will consist of a summary measure from each experiment, y_i, and an associated measure of precision, V_i. In a random effects model, it is assumed that for $i = 1, 2, \ldots, k$ studies

$$y_i|\theta_i \overset{\text{indep.}}{\sim} N(\theta_i, V_i), \tag{1}$$

$$\theta_i|\mu,\tau^2 \overset{\text{indep.}}{\sim} N(\mu,\tau^2), \tag{2}$$

where θ_i represents the underlying study effect for the ith experiment and μ and τ^2 are the hyperparameters of the mixing distribution governing the generation of each underlying study effect. The directed graph corresponding to this model will have k separate nodes for each summary measure, k separate nodes for each underlying study effect and a node for each of the hyperparameters. Unlike the CHILD network discussed by Spiegelhalter and colleagues (Figure 2 in their article) and the examples considered in Cox and Wermuth's paper, only a subset of the nodes in the graph representing this hierarchical model will ever be observed. Substantially more complicated hierarchical models, those with more stages and more dependency structure such as the multiprocess models of Harrison and Stevens (1976), can be represented graphically.

The value of displaying the qualitative structure of statistical models has been vastly underutilized by statisticians but appreciated in other branches of science. In the medical arena, we frequently encounter graphical representations of decision models, namely *decision trees*. In its simplest form, the decision tree is a singly connected graph in which some nodes represent risk factors such as age and gender, some nodes represent complications and symptoms and some nodes represent decisions. For example, researchers may be interested in investigating whether older patients who are suspected of having an acute myocardial infarction will benefit from thrombolytic therapy. A decision-analytic model is then built using information from the experts (cardiologists) and from the results of clinical trials (e.g., the rate of incapacitating complications from thrombolytic therapy for older patients). Some statisticians are investigating methods of quantifying uncertainty in medical decision analysis (Katz and Hui, 1989) because, typically, statistical error is not incorporated in most decision analyses. Clearly, the expert system methodology could play a substantial role in this effort—propagation of the uncertainty attached to the decision tree inputs is naturally accommodated within the graphical framework.

More recently we have witnessed in the statistical literature the use of graphical representations to understand the dependency structure in order to perform the "correct" computations. For example, Bernardinelli and Montomoli (1992) use a graphical representation of a hierarchical model of relative risk mortality to display the qualitative structure of the data but also to indicate which conditional distributions must be specified to calculate the joint distribution. Gilks et al. (1993) construct a graphical model for modeling precursors of cervical cancer in an application of Gibbs sampling in medicine for a similar specification purpose.

Finally, as Professor Spiegelhalter and his colleagues have indicated, the graphical model can be used to perform efficient probability calculations in high dimensional problems. The main goal is to have queries regarding certain sets of variables answered quickly. This is achieved through local computations performed through an algorithm designed to capitalize on the dependency structure embedded in the statistical model. In the expert system setting, the computational efficiency of the propagation algorithm is obvious. However, it has been shown that even in standard models, computation within a graphical framework can be beneficial. Normand and Tritchler (1992) discuss the use of a graphical model as the computational device for updating parameter estimates in a hierarchical model and show that the graphical model characterizes the hierarchical model and its computations in a unified way.

3. DATA PROPAGATION IN GRAPHS

Because one of the central roles for the expert system is that of updating the system once evidence has been realized, I will recast for the reader the essence of how this is achieved. The task at hand is the following: information is observed and consequently, the joint distribution needs to be updated in light of this new information. Essentially the problem becomes one of conditioning and a brute force approach is clearly undesirable in high dimensional problems. It is worth recalling that there were over a billion possible configurations in the CHILD network. Propagation refers to the transmission of hereditary features to or through offspring and this is the "divide and conquer" strategy employed in graphical models: the joint set of random variables is divided into subgroups, a source subgroup is identified, a marginal is taken in the source subgroup

and then that marginal multiplies a function on a destination subgroup. A marginal is then taken in the destination subgroup and that marginal multiplies a function on its destination subgroup and so on and so on. Professor Spiegelhalter and his colleagues refer to the subgroups as belief universes and equate these universes to the cliques of the relevant undirected graph. A clique is a set of random variables such that no further factorization of the probability function characterizing the distribution is possible; that is, there are no further independence constraints on the elements in the clique. Ideally the state space of the cliques should be small otherwise the efficiency gains through the use of the propagation algorithm will be lost. The propagation algorithm described in the article by Professor Spiegelhalter and colleagues is based on the junction tree. The junction tree may be thought of as a singly connected graph in which each node consists of sets of random variables (the *cliques*). In the case of multinomial random variables, any node in the junction tree may be used as a root for propagation. The steps necessary to transform the original directed graph into the junction tree (referred to as compilation by Professor Spiegelhalter and his colleagues) are many and sometimes nontrivial.

4. LIMITATIONS OF GRAPHICAL MODELS

The (potential) limitations of graphical models that I envision are related only to one of my three motivating reasons for using graphical models, and these have to do with efficient computation. First, Professor Spiegelhalter and his colleagues have indicated the importance of the size of the state spaces of the cliques obtained after triangulation in measuring the computational benefits of a graphical model approach. In preserving all the induced dependence relationships in a model through the "moralizing" procedure, the dimensionalities of the cliques are increased. These dimensionalities are further increased after triangulation. It is not immediately clear in which statistical models the computational advantages of a graphical model approach will be realized. Further research into identifying classes of statistical models that could benefit from the computational efficiency of graphical models needs to be undertaken.

Second, the junction tree algorithm for propagation in graphical models works well in models in which the random variables arise from a multinomial distribution. There does exist an algorithm that mimics the junction tree algorithm in models for which some of the variables are multinomial and some are Gaussian. However in these latter graphs (mixed graphs), only means and variances are propagated. Moreover, there is an additional requirement in the compilation process for marked graphs, that of strong decomposability, that further increases the dimensionalities of the cliques. In addition, for graphs that host other distributions, Monte Carlo methods have to replace exact methods.

Third, I am not satisfied with how well one can assess model fit in graphical models. I have a difficult time assessing model adequacy in a logistic regression model with more than five covariates! I will not equate my model-checking capabilities to those of Professor Spiegelhalter and his colleagues but surely model assessment involving the number of variables typified in the expert system domain requires a tremendous amount of skill. In the CHILD network, how can one assess whether age at presentation is related quadratically to disease or whether age at presentation is related quadratically to lung disease but only linearly to the remaining five diseases? How important is correct specification of the functional form of the model variables and how important are "missing links" in predicting the state of a particular configuration? The node monitors proposed by Professor Spiegelhalter and colleagues are admittedly using a prequential approach but I hope research will extend to model diagnostic methods using other endpoints.

Graphical models will play an increasingly important role in the structuring of statistical analyses for complex problems. These models enhance communication among researchers, thereby facilitating scientific modeling, and provide a unifying approach to computation. Research into automating algorithms for distributions other than the multinomial and the Gaussian distributions should be explored. More examples of graphical models need to identified and analyzed, and the effects of model misspecification on prediction quantified. In closing, I thank the authors for presenting their valuable ideas.

ACKNOWLEDGMENT

This research was partially supported by AHCPR Grant RO1 HS07118-01.

Comment: Graphical Models, Causality and Intervention

Judea Pearl

I am grateful for the opportunity to respond to these two excellent papers. Although graphical models are intuitively compelling for conceptualizing statistical associations, the scientific community generally views such models with hesitancy and suspicion. The two papers before us demonstrate the use of graphs – specifically, directed acyclic graphs (DAGs) – as a mathematical tool of great versatility and thus promise to make graphical languages more common in statistical analysis. In fact, I find my own views in such close agreement with those of the authors that any attempt on my part to comment directly on their work would amount to sheer repetition. Instead, as the editor suggested, I would like to provide a personal perspective on current and future developments in the areas of graphical and causal modeling. A complementary account of the evolution of belief networks is given in Pearl (1993a).

I will focus on the connection between graphical models and the notion of causality in statistical analysis. This connection has been treated very cautiously in the papers before us. In Lauritzen and Spiegelhalter (1988), the graphs were called "causal networks," for which the authors were criticized; they have agreed to refrain from using the word "causal." In the current paper, Spiegelhalter et al. deemphasize the causal interpretation of the arcs in favor of the "irrelevance" interpretation. I think this retreat is regrettable for two reasons: first, causal associations are the primary source of judgments about irrelevance, and, second, rejecting the causal interpretation of arcs prevents us from using graphical models for making legitimate predictions about the effect of actions. Such predictions are indispensable in applications such as treatment management and policy analysis. I would like to supplement the discussion with an account of how causal models and graphical models are related.

It is generally accepted that, because they provide information about the dynamics of the system under study, causal models, regardless of how they are discovered or tested, are more useful than associational models. In other words, whereas the joint distribution tells us how probable events are and how probabilities would change with subsequent observations, the causal model also tells us how these probabilities would change as a result of external interventions in the system. For this reason, causal models (or "structural models" as they are often called) have been the target of relentless scientific pursuit and, at the same time, the center of much controversy and speculation. What I would like to discuss in this commentary is how complex information about external interventions can be organized and represented graphically and, conversely, how the graphical representation can be used to facilitate quantitative predictions of the effects of interventions.

The basic idea goes back to Simon (1977) and is stated succinctly in his foreword to Glymour et al. (1987): "The advantage of representing the system by structural equations that describe the direct causal mechanisms is that if we obtain some knowledge that one or more of these mechanisms has been altered, we can use the remaining equations to predict the consequences – the new equilibrium." Here, by "mechanism" Simon means any stable relationship between two or more variables that remains invariant to external influences until it falls directly under such influences.

This mechanism-based model was adapted in Pearl and Verma (1991) for defining probabilistic causal theories; each child-parent family in a DAG Γ represents a deterministic function $X_i = f_i(\mathbf{pa}_i, \varepsilon_i)$, where $\mathbf{pa}_i$ are the parents of variable X_i in Γ, and ε_i, $0 < i < n$, are mutually independent, arbitrarily distributed random disturbances. Characterizing each child-parent relationship as a deterministic function, instead of the usual conditional probability $P(x_i \mid \mathbf{pa}_i)$, imposes equivalent independence constraints on the resulting distributions and leads to the same recursive decomposition

$$P(x_1, \ldots, x_n) = \prod_i P(x_i \mid \mathbf{pa}_i) \tag{1}$$

that appears in Eq. (1) of Spiegelhalter et al.'s article. However, the functional characterization also specifies how the resulting distribution would change in response to external interventions, since, by convention, each function is presumed to remain constant unless specifically altered. This formulation is merely a nonlinear generalization of the usual structural equation models, where function constancy (or stability) is implicitly

Judea Pearl is Professor of Computer Science and Director of the Cognitive Systems Laboratory, University of California Los Angeles, 405 Hilgard Avenue, Los Angeles, California 90024.

assumed. Moreover, the nonlinear character of f_i permits us to treat changes in the function f_i itself as a variable, F_i, by writing

$$X_i = f_i'(\mathbf{pa}_i, F_i, \varepsilon_i) \tag{2}$$

where

$$f_i'(a, b, c) = f_i(a, c) \text{ whenever } b = f_i.$$

Thus, any external intervention F_i that alters f_i can be represented graphically as an added parent node of X_i, and the effect of such an intervention can be analyzed by Bayesian conditionalization, that is, by simply setting this added parent variable to the appropriate value f_i.

The simplest type of external intervention is one in which a single variable, say X_i, is forced to take on some fixed value x_i'. Such intervention, which we call *atomic*, amounts to replacing the old functional mechanism $X_i = f_i(\mathbf{pa}_i, \varepsilon_i)$ with a new mechanism $X_i = x_i'$ governed by some external force F_i that sets the value x_i'. If we imagine that each variable X_i potentially could be subject to the influence of such an external force F_i, then we can view the causal network Γ as an efficient code for predicting the effects of atomic interventions and of various combinations of such interventions.

The effect of an atomic intervention $set(X_i = x_i')$ is encoded by adding to Γ a link $F_i \rightarrow X_i$ (Figure 1), where F_i is a new variable taking values in $\{set(x_i'), idle\}$, x_i' ranges over the domain of X_i, and *idle* represents no intervention. Thus, the new parent set of X_i in the augmented network is $\mathbf{pa}_i' = \mathbf{pa}_i \cup \{F_i\}$, and it is related to X_i by the conditional probability

$$P(x_i \mid \mathbf{pa}_i') = \begin{cases} P(x_i \mid \mathbf{pa}_i), & \text{if } F_i = idle, \\ 0, & \text{if } F_i = set(x_i') \text{ and } x_i \neq x_i', \\ 1, & \text{if } F_i = set(x_i') \text{ and } x_i = x_i'. \end{cases} \tag{3}$$

The effect of the intervention $set(x_i')$ is to transform the original probability function $P(x_1, \ldots, x_n)$ into a new function $P_{x_i'}(x_1, \ldots, x_n)$, given by

$$P_{x_i'}(x_1, \ldots, x_n) = P'(x_1, \ldots, x_n \mid F_i = set(x_i')), \tag{4}$$

where P' is the directed Markov field dictated by the augmented network $\Gamma' = \Gamma \cup \{F_i \rightarrow X_i\}$ and (3), with an arbitrary prior distribution on F_i. In general, by adding a hypothetical intervention link $F_i \rightarrow X_i$ to each node in Γ, we can construct an augmented probability function $P'(x_1, \ldots, x_n; F_1, \ldots, F_n)$ that contains information about richer types of interventions. Multiple interventions would be represented by conditioning P' on a subset of the F_i's (taking values in their respective $set(x_i')$), while the preintervention probability function P would be viewed as the posterior distribution induced by conditioning each F_i in P' on the value *idle*.

This representation yields a simple and direct transformation between the preintervention and the postintervention distributions:

$$P_{x_i'}(x_1, \ldots, x_n) = \begin{cases} \dfrac{P(x_1, \ldots, x_n)}{P(x_i \mid \mathbf{pa}_i)}, & \text{if } x_i = x_i', \\ 0, & \text{if } x_i \neq x_i'. \end{cases} \tag{5}$$

This transformation reflects the removal of the term $P(x_i|\mathbf{pa}_i)$ from the product decomposition of (1), since $\mathbf{pa}_i$ no longer influence X_i. Transformations involving conjunctive and disjunctive actions can be obtained by straightforward applications of (4) (Goldszmidt and Pearl, 1992; Pearl, 1993b; Spirtes, Glymour and Scheines 1993). The transformation exhibits the following properties:

1. An intervention $set(x_i')$ can affect only the descendants of X_i in Γ.
2. For any set $\mathbf{S}$ of variables, we have

$$P_{x_i'}(\mathbf{S} \mid \mathbf{pa}_i) = P(\mathbf{S} \mid x_i', \mathbf{pa}_i). \tag{6}$$

In other words, given $X_i = x_i'$ and $\mathbf{pa}_i$, it is superfluous to find out whether $X_i = x_i'$ was established by external intervention or not. This can be seen directly from the augmented network Γ' (Figure 1), since $\{X_i\} \cup \mathbf{pa}_i$ d-separates F_i from the rest of the network, thus legitimizing the conditional independence $\mathbf{S} \perp\!\!\!\perp F_i \mid (X_i, \mathbf{pa}_i)$.

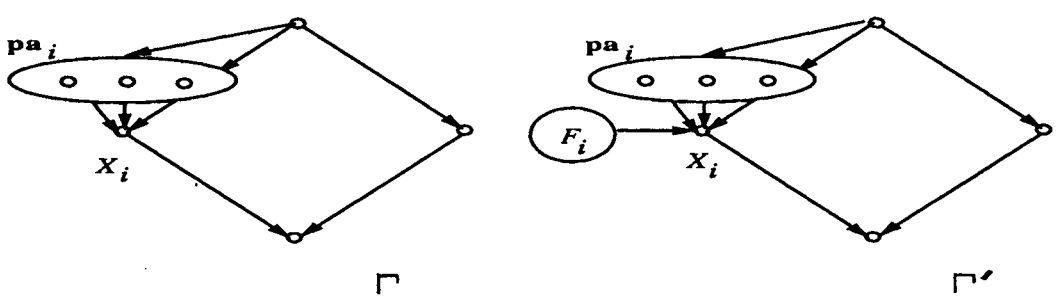


FIG. 1. *Representing external intervention, F_i, by an augmented network $\Gamma' = \Gamma \cup \{F_i \rightarrow X_i\}$.*

3. A necessary and sufficient condition for an external intervention $set(X_i = x_i')$ to have the same effect on X_j as the passive observation $X_i = x_i'$ is that X_i d-separates $\mathbf{pa}_i$ from X_j, that is,

$$P_{x_i'}(x_j) = P(x_j \mid x_i') \text{ iff } X_j \perp\!\!\!\perp \mathbf{pa}_i \mid X_i. \tag{7}$$

Equation (4) explains why randomized experiments are sufficient for estimating the effect of interventions even when the causal network is not given: because the intervening variable F_i enters the networks as a root node (i.e., independent of all other ancestors of X_i) it is equivalent to a treatment-selection policy governed by a random device.

The immediate implication of (5) is that, given the structure of the causal network Γ, one can infer postintervention distributions from preintervention distributions; hence, we can reliably estimate the effects of interventions from passive (i.e., nonexperimental) observations. Of course, (5) does not imply that we can always substitute observational studies for experimental studies, as this would require an estimation of $P(x_i \mid \mathbf{pa}_i)$. The mere identification of $\mathbf{pa}_i$ (i.e., the direct causal factors of X_i) requires substantive causal knowledge of the domain which is often unavailable. Moreover, even when we have sufficient substantive knowledge to structure Γ, some members of $\mathbf{pa}_i$ may be unobservable, or *latent*. Fortunately, there are conditions for which an unbiased estimate of $P_{x_i}(x_j)$ can be obtained even when the $\mathbf{pa}_i$ variables are latent and, moreover, a simple graphical criterion can tell us when these conditions are satisfied.

Assume we are given a causal network Γ together with nonexperimental data on a subset $\mathbf{X}_o$ of observed variables in Γ and we wish to estimate what effect the intervention $set(X_i = x_i')$ would have on some response variable X_j. In other words, we seek to estimate $P_{x_i'}(x_j)$ from a sample estimate of $P(\mathbf{X}_o)$. Applying (4), we can write

$$\begin{aligned} P_{x_i'}(x_j) &= P'\left(x_j \mid F_i = set(x_i')\right) \\ &= \sum_{\mathbf{S}} P'(x_j \mid \mathbf{S}, X_i = x_i', F_i = set(x_i'))\, \\ &\quad \times P'\left(\mathbf{S} \mid F_i = set(x_i')\right), \end{aligned} \tag{8}$$

where $\mathbf{S}$ is any set of variables. Clearly, if $\mathbf{S}$ satisfies

$$\mathbf{S} \perp\!\!\!\perp F_i \text{ and } X_j \perp\!\!\!\perp F_i \mid (X_i, \mathbf{S}), \tag{9}$$

then (8) can be reduced to

$$\begin{aligned} P_{x_i'}(x_j) &= \sum_{\mathbf{S}} P(x_j \mid \mathbf{S}, x_i') P(\mathbf{S}) \\ &= E_{\mathbf{S}}[P(x_j \mid \mathbf{S}, x_i')]. \end{aligned} \tag{10}$$

Thus, if we find a set $\mathbf{S} \subseteq \mathbf{X}_o$ of observables satisfying (9), we can estimate $P_{x_i'}(x_j)$ by taking the expectation (over $\mathbf{S}$) of $P(x_j \mid \mathbf{S}, x_i')$, and the latter can easily be estimated from nonexperimental data. It is also easy to verify that (9) is satisfied by any set $\mathbf{S}$ that meets the following *back-door criterion*:

1. No node in $\mathbf{S}$ is a descendant of X_i, and
2. $\mathbf{S}$ d-separates X_i from X_j along every path containing an arrow into X_i.

The name "back-door" echos condition 2, which requires that only indirect paths from X_i to X_j be d-separated; these paths can be viewed as entering X_i through the back door.

In Figure 2, for example, the sets $\mathbf{S}_1 = \{X_3, X_4\}$ and $\mathbf{S}_2 = \{X_4, X_5\}$ would qualify under the back-door criterion, but $\mathbf{S}_3 = \{X_4\}$ would not because X_4 does not d-separate X_i from X_j along the path $(X_i, X_3, X_1, X_4, X_2, X_5, X_j)$. Thus, we have obtained a simple graphical criterion for finding a set of observables for estimating (by conditioning) the effect of interventions from purely nonexperimental data.

It is interesting that the conditions formulated in (9) are equivalent to those known as *strongly ignorable treatment assignment* (SITA) conditions in Rubin's model for causal effect (Pearl, 1993c; Rosenbaum and Rubin, 1983). [The graphical translation of Rubin's model invokes the mechanism $X_i \rightarrow X_j \leftarrow \mathbf{r}$, where X_i represents the treatment-assignment, X_j the observed response, and $\mathbf{r}$ represents the causal-effect variable. Indeed, following the counterfactual interpretation of $\mathbf{r}$, X_j is a deterministic function of X_i and $\mathbf{r}$, and $\mathbf{r}$ plays the role of f_i in (2) (Pearl, 1993c)]. Reducing the SITA conditions to the graphical back-door criterion facilitates the search for an optimal conditioning set $\mathbf{S}$ and significantly simplifies the judgments required for ratifying the validity of such conditions in practical situations.

Equation (4) was derived under the assumption that the preintervention probability P is given by the product of (1), which represents a joint distribution prior to making any observations. To predict the effect of action F_i *after* observing C, we must also invoke assumptions about persistence, so as to distinguish prop-

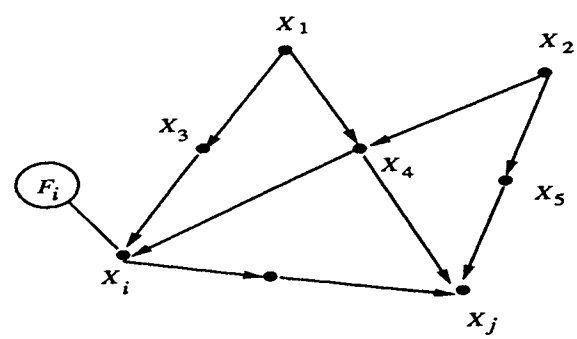


FIG. 2. *A DAG representing the back-door criterion, adjusting for variables* $\{X_3, X_4\}$ *(or* $\{X_4, X_5\}$*) yields an unbiased estimate of* $P(x_j \mid set(x_i'))$.

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 $1, \ldots m$; 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 $1, \ldots, m$; 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 $1, \ldots, m - 1$; if $l = m$, the $X_j - X_i$ relationship can be measured by the partial correlation, given the variables in blocks $1, \ldots, m - 1$; 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 $1, \ldots, m - 1$; 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:

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

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.

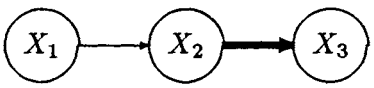
Joe Whittaker is Senior Lecturer, Mathematics Department, Lancaster University, LA1 4YF, United Kingdom.

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

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



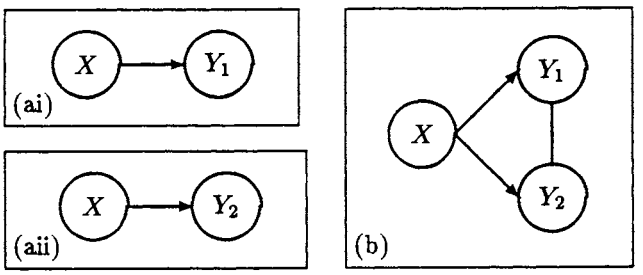
immediately conveys the information that the (2,3) dependence is stronger than the (1,2) dependence, thus helping the data analyst to make sense of possibly complex interactions.

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

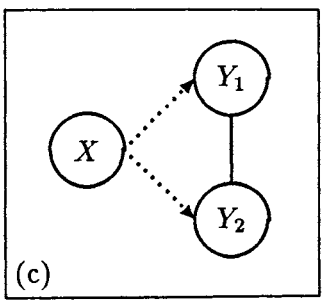
A particular contribution of the CW paper is to highlight the difference between multivariate regression and so-called "block" regression and to demonstrate that graphical modellers have some difficulty in portraying the former. The reason, of course, is that graphical modelling interests itself in the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

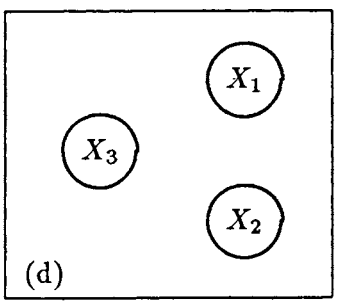
The authors attempt to combine the graphs (ai, aii, b) and extract the best from both worlds by defining the dashed edges in the graph (c)



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

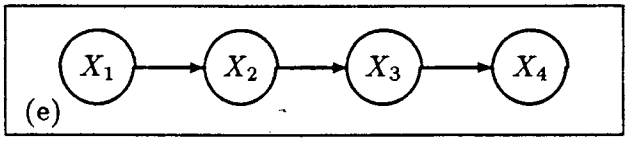


defined by missing *dashed* edges. To me, the only possible visual interpretation of graph (d) is that of complete (mutual) independence of X_1, X_2 and X_3. But of course, there are well-known counter examples to the assertion that $\{X_1 \perp\!\!\!\perp X_2, X_1 \perp\!\!\!\perp X_3, X_2 \perp\!\!\!\perp X_3\}$ 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.

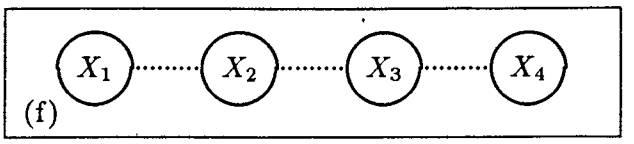
3. *Coherence:* To obtain a coherent picture CI graphs focus on a single joint distribution, $f_{123\ldots k}$ say, and analyse it in terms of conditional distributions of the form $f_{a|rest}$. Because $f_{12\ldots k} = f_{k|1\ldots k-1} f_{k-1|1\ldots k-2} \cdots f_{2|1} f_1$ this single joint distribution 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.

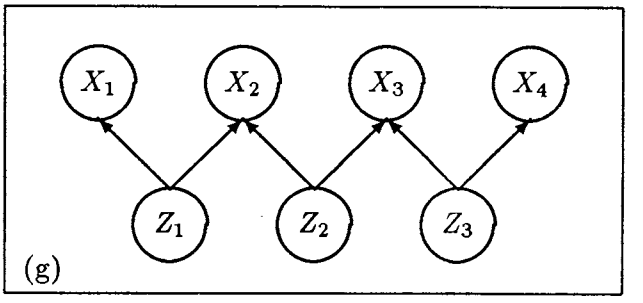
Unfortunately, a single joint distribution is not generally specified by all pairwise marginal distributions, and so a graph built from these may easily indicate ambiguities as in the mutual independence example above.

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



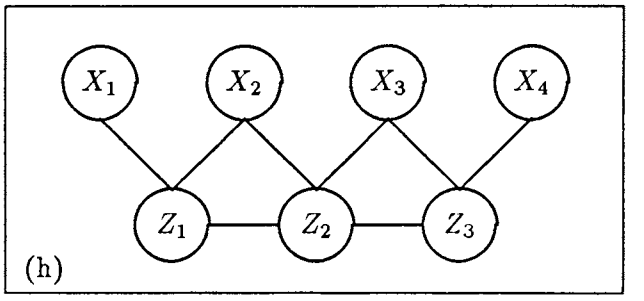
defined by $\{X_1 \perp\!\!\!\perp X_3, X_1 \perp\!\!\!\perp X_4, X_2 \perp\!\!\!\perp X_4\}$ and ask if information on X_1 is needed to predict X_3 when X_2 is known.

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

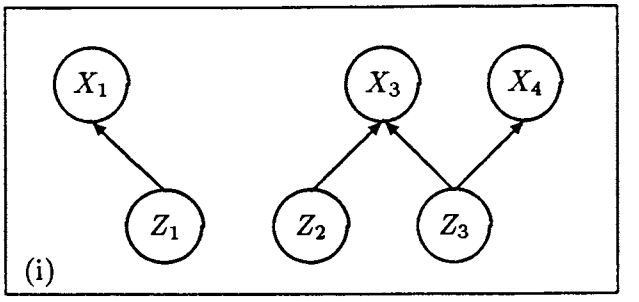


in which Z_1, Z_2 and Z_3 are mutually independent and the X's are conditionally independent given the Z's ($X_1 \perp\!\!\!\perp X_3$ as they have no Z's in common). The CI graph (f) is a "consequence" in the sense that the marginal distribution of (X_1, X_2, X_3, X_4) is obtained from that of $(X_1,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

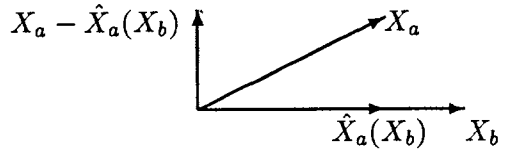
This device of embedding the dashed graph into a CI graph with "latent variables" certainly solves some problems. It also indicates why latent variables in highly structured graphs allow marginal empirical dependences to determine the statistical analysis. A prime example of this is the graphical analysis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

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

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

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

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

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

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

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 $\mathrm{var}(AX_a|X_b) = A\mathrm{var}(X_a|X_b)A'$ where A is a fixed linear transform, and if B is invertible, $\mathrm{var}(X_a|BX_b) = \mathrm{var}(X_a|X_b)$ 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