In: Bulletin of the International Statistical Institute, Proceedings 50th Session, Beijing, Vol. 1. ISI: Voorburg

ON THE INTERPRETATION OF CHAIN GRAPHS

Nanny Wermuth, Psychological Institute, University of Mainz
D-55099 Mainz, Germany

1. Introduction

Chain graphs provide a flexible tool for representing complex relations among variables. These relations are marginal or conditional independencies and directed or symmetric associations. The graphs can be used to aid in the analysis of data from observational studies (Cox & Wermuth, 1993,1995), to update information in expert systems (Lauritzen & Spiegelhalter, 1988) and to represent alternatives in decision trees and decision networks (Smith, 1989).

Each variable in a chain graph is represented by a node and some pairs of nodes are connected by edges which indicate dependencies of subject-matter interest whenever it represents a substantive research hypothesis (Wermuth & Lauritzen (1990). Each edge missing in the graph means that the corresponding variables are conditionally independent, the precise conditioning set depending on the type of edges of the graph and on an order of the variables specified such that no variable is explanatory for itself. In these aspects chain graphs contrast with graphical representations of general simultaneous equation models (Bollen, 1989) and with so-called reciprocal graphs (Koster 1993; Spirtes, 1993).

We propose a two step modification of a special type of chain graph, namely a directed acyclic graph, to study its implications for a conditional distribution of a selected subset Y_S given another subset Y_C, that is, to construct the covariance and the concentration graph in any conditional distribution of interest. The concentration graph gives the information on which variable pairs are conditionally independent and which may be associated given all remaining variables of S and the covariance graph gives the information on which variable pairs are marginally independent and which may be marginally associated in the conditional distribution of Y_S given Y_C. We illustrate the procedure with variables from an observational study on university drop-outs. Thereby we utilize separation criteria of Pearl (1988), Lauritzen et al. (1990) and of Frydenberg (1990), as well as results on induced associations by Wermuth and

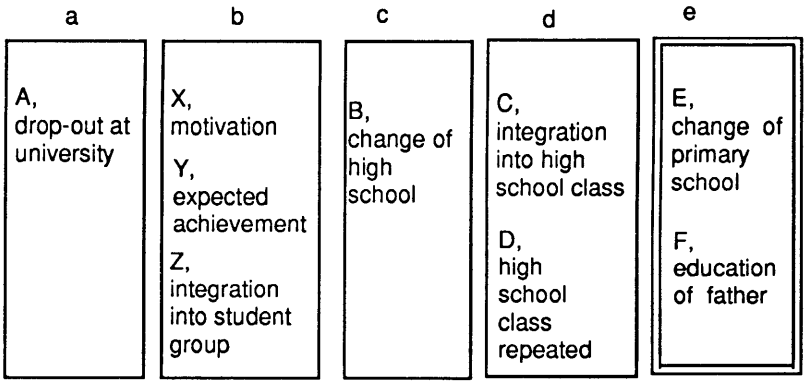


Figure 1: A first ordering of the variables with A, drop-out at university (box a), as the response variable of primary interest and with, for instance, B, change of high school (box c), as an intermediate variable being potentially explanatory to dropping out at university (box a) and to the student's attitudes towards his study situations (box b) and being a potential response to the other school career and demographic variables (boxes d, e). Several variables in a box indicate variables treated on equal footing, since we are at this stage not prepared to specify a single direction of dependence (boxes b, d) or since we consider them as purely explanatory variables (box e).

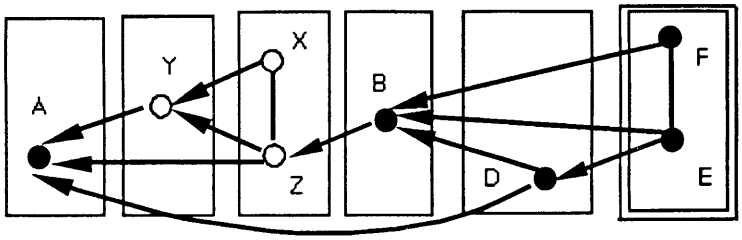
Cox (1995).

We suggest that the possibility of deriving consequences of a given chain graph model and checking them against observations means that the general principle of making a hypothesis elaborate (attributed to Sir Ronald Fisher by William Cochran (1965)) can be applied to these multivariate structures.

2. A motivating example

For the variables shown in Figure 1 we shall illustrate which questions may arise regarding conditional distributions of subsets of variables. Figure 1 shows nine variables ordered in a chain of boxes which reflects our knowledge and judgement in this context about responses, intermediate variables and purely explanatory variables.

To study risk factors for dropping out from university, data for 3500 German high-school students were collected (Giesen et al., 1981; Gold 1988) one year before their graduation in years 1973 to 1975. Responses were recorded to a number of psychological questionnaires and tests and to questions regarding



A: Y+Z+D, Y: X+Z, X: Z, Z: X+B

B: D+E+F, C : D, D : C+E EF

Figure 2: A well fitting chain graph model for the variables of Figure 1; variable C has been deleted, since it is only related to another joint response, to variable D

school career and demographic background. About 73% (2544) of the students enrolled in university degree programs. They received questionnaires and tests as second and third year students and after having either successfully completed their studies or dropped out of university. The data collection ended in 1984 with 2375 students still in the study. For the analysis presented here we used data of 2162 students having complete records on nine variables to investigate developments which might increase the risk that a student stops studying without having received any university degree.

There are six binary variables variables: A, university drop-out (yes, 15.3%); B, change of high school (yes, 21.0%); C, integration into the high school class (poor, 9.9%); D, a high school class repeated (yes, 33.6%); E, change of primary school (yes, 19.8%); F, education of the father (at least 13 years of formal schooling, 42.8%) and three questionnaire scores: Y, achievement, the student's expectation of his achievement in the field of study ($y_{min} = 0$, $y_{max} = 8$, $\bar{y} = 6.16$, $s_y = 2.08$); X, motivation, the student's motivation towards high achievement in the field ($x_{min} = 10$, $x_{max} = 60$, $\bar{x} = 35.27$, $s_x = 8.67$); Z, integration, the student's perceived integration into his student group at university ($z_{min} = 0$, $z_{max} = 9$, $\bar{z} = 6.49$, $s_z = 2.42$).

After checking for outliers, nonlinear and interactive effects (Cox & Wermuth, 1994) and after combining the results of separate logistic and linear regression analyses the chain graph in Figure 2 was taken to be well compatible with the given observations.

The model notation for generalized linear models (McCullagh & Nelder, 1989) below Figure 2 shows that the regression components of this chain are main effect regressions; no higher than two factor interactions or nonlinear

relations are needed to describe the relations. The predicted drop-out rate is at 65% highest if the student's expectation of own achievement (Y) is low, his integration (Z) poor and he had already repeated a high school class (D). By contrast in the corresponding most favourable case the drop-out rate is as low as 6%. The most important of the three predictors is expected achievement, followed by integration and high school class repeated. This is reflected in the studentized regression coefficients, ordered correspondingly in absolute values as $| t |$: 7.59, 5.91, 3.39. For a description of further details of this and similar analyses see Streit (1995); Cox & Wermuth (1995).

In addition to the separate regression results the chain graph representation shows a number of indirect paths to the response of primary interest and permits additional interpretations, which are compatible with the observed structure. For instance, the path from X to A via Y is consistent with the following interpretation: motivation for high achievement (X) is likely to to increase the confidence in high achievement in the field (Y), which in turn reduces the risk of dropping out from university (A). The path from E to D to B to Z to A could be interpreted as follows: change of primary school (E) increases the risk that a high school class will have to be repeated (D), which in turn increases the risk that the student will change the high school at least once during his school career (B). Once a high school change has been experienced it becomes less likely that a student integrates well into his later student group (Z) and this in turn is a direct risk factor for dropping out from university (A). However the overall effect of a change in primary school on university drop-out will not be strong because the path from E to A is fairly long and because some of the dependencies along the path are rather weak.

Since none of the binary variables has levels which occur only with very low probabilities (below .05), the logistic regression may be well approximated with the linear in probability representations (Cox, 1966) and the whole system of relations can be regarded as having *quasi-linear dependencies* which means that any dependency present has a linear component. Then vanishing of least squares coefficients does imply an independence statement and any linear component of an overall effect is the product of the correlation coefficients along the path; see Wermuth & Cox (1995) for some further discussion of this notion.

As we shall show in this paper, it is possible to work out the implications, that is the consequences of a chain graph model, for conditional distributions of selected subsets of variables. Such consequences are independence statements implied by the graph and, provided the graph describes a generating process of

of a quasi-linear system (see Wermuth and Cox; 1995), such consequences can be statements about conditional associations, in addition.

Typical questions that can arise for a model as the one of Figure 2 are as follows:
– if this structure holds, which relations should we expect among the remaining variables in a different study of only academics, that is after conditioning on variable A, when there is no information on the self-judgements of the students, that is after marginalizing over variables Y, X, Z? Or,
– if this structure holds, but information on the background of the students prior to taking up their studies (on variable B to F) is not available, what changes in the relations among the remaining observed variables should result?

Answers to such questions can be helpful in gaining confidence in a hypothesized model. A model is more likely to be a good description of the investigated relations if predictions derived from it for results of different studies or for results of different analyses of the same data are consistent with what is observed.

3. Components of chain graphs

Nodes $V = \{1, \ldots, p\}$ in an independence graph represent variables $Y_1, \ldots, Y_p$; each pair is connected by at most one edge. We assume in this paper that each edge that is present in a given independence graph represents a particular nonvanishing (conditional) association. This means that for the purpose of this paper the distinction between an independence graph drawn without boxes and with boxes is dropped. Usually, in a statistical model represented by the former, an edge that is present corresponds to a free paramter, like for example to a regression coefficient, which may take any value including zero, the value for a particular independence.

Quantitative variables are modelled by continuous random variables and are represented by nodes which are circles; qualitative variables are modelled by discrete random variables and drawn in the graphs as dots. If one variable is considered to be a response to another, then the edge is an arrow pointing to the response from the explanatory variable; for symmetric associations edges are lines. Symmetric asociations are of interest among variables considered to be on equal footing, like for instance among symptoms of a disease, among personality characteristics or among different strategies employed by a person to cope with stressful events.

A *path* between nodes i and j is a succession of edges connecting the two

nodes, irrespective of the orientation of the edges. A graph constructed from a given independence graph by keeping the edges within a selected subset of nodes and by deleting edges to nodes outside the set is called an *induced subgraph.* Induced subgraphs are *complete* if all nodes within it are joined by an edge.

Three types of nodes t along a directed path can be distinguished among consecutive nodes i, t, j, that is for nodes i, j having node t as a *common neighbour.* From a *source node* t two arrows point to i and j; a *transition node* t has an incoming arrow from j, say, and an outgoing arrow to i; and a *sink node* t has two arrows pointing at it from each of i and j. If the paths shown in Figure 3 are subgraphs of an independence graph induced by the three nodes, then they are called a *source node oriented,* a *transition node oriented* and a *sink node oriented* v-*configuration*, respectively.

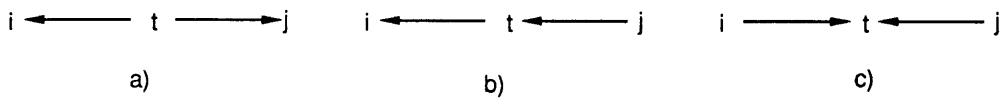


Figure 3: Types of common neighbour nodes t in a directed graph: a) a source node; b) a transition node; c) a sink or collision node

Illustration to a): An independence graph as in Figure 3a) is obtained for variables Y_t, age; Y_i, diastolic blood pressure; Y_j, body mass, that is weight relative to height, observed for healthy female adults. The reason is that the risks for higher blood pressure and for larger body mass both increase with age, but for groups of healthy persons of the same age knowledge of body mass does not improve prediction of the diastolic blood pressure. If age is not recorded there is an association between body mass and diastolic blood pressure. ◇

Illustration to b): An independence graph as in Figure 3b) is obtained for variables Y_i, job offer for an academic in Germany; Y_t, field of study (engineering or home economics); Y_j, gender. The arrow to t from j means that almost only males tend to choose engineering as their field of study while almost only women tend to choose home economics as their field of study. The arrow from t to i means there are many more job offers in engineering than in home economics. The unjoined nodes i, j mean that given the field of study males and females have equal chances for getting a job offer. If the field of study is not recorded it appears as if there were discrimination against women on this job market, that is after marginalizing over Y_t. ◇

Illustration to c): An independence graph as in Figure 3c) is obtained for 24 postwar years in Germany and variables Y_t, growth rate in capital gain; Y_i, growth rate in consumption; Y_j, growth rate in exports. The arrows to t from i and j mean that the increase in capital gain growth rates are larger the larger the growth rates in consumption and in exports are, respectively. The unjoined nodes i, j mean that changes in consumption within Germany could at that time not be predicted from changes in export growth rates. There is a substantial negative partial correlation between growth rates in consumption and export given the growth rates in capital gain, that is Y_i and Y_j are associated after conditioning on Y_t. $\diamond$

Because two arrows meet head-on at a sink node it is also called a *collision node.* A path containing a collision node is a *collision path* and a path is said to be *collisionless*, otherwise. A path with only transition nodes and arrows leading to i from j is called a *direction preserving* path, where node i is called a *descendant* of node j and node j an *ancestor* of i.

In a *directed acyclic graph*, G^V_{da}, all edges are directed, i.e. they are arrows, and there is no direction preserving path from a node back to itself. The nodes can be numbered $1, \ldots, p$, possibly in more than one way, without changing the independencies implied by this type of graph, so that the variables form a system of univariate recursive regressions. Typically the order of the variables is specified from subject-matter knowledge. Given such an order each edge present in the graph corresponds to a particular conditional association and each edge missing to a conditional independence statement in one of Y_1 regressed on $Y_2, \ldots, Y_p$; Y_2 regressed on $Y_3, \ldots Y_p; \ldots$; Y_{p-1} regressed on Y_p; thus to an independence of the form $Y_i \perp\!\!\!\perp Y_j \,|\, Y_{\{i+1,\ldots p\}\setminus\{j\}})$, for $i < j$. If the system of univariate recursive regressions is regarded as describing a process by which the data are generated then the corresponding directed acyclic graph is called a *generating graph.*

The concentration and covariance graphs of a subset of variables Y_S of Y_V satisfying the independencies of a given directed acyclic graph G^V_{da}, may be considered conditionally given another subset Y_C. Such a *covariance graph* will have an (undirected, dashed line) edge i, j if and only if $Y_i \perp\!\!\!\perp Y_j \,|\, Y_C$ is not implied by the generating graph and such a *concentration graph* will have an (undirected, full line) edge i, j if and only if $Y_i \perp\!\!\!\perp Y_j \,|\, Y_{C\cup S\setminus\{i,j\}}$ is not implied by the generating graph. These graphs may be derived directly from a separation criterion for directed acyclic graphs or with the help of the modified directed graphs described in the next section.

A *general chain graph* has both directed and undirected edges, like the graph

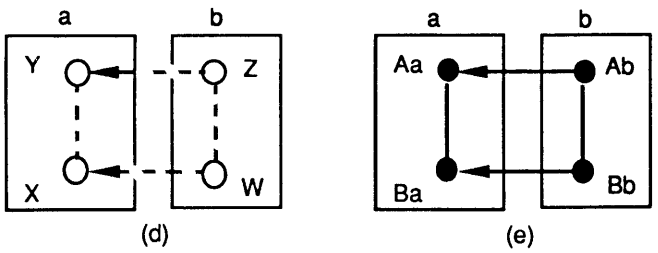


Figure 4: Two distinct small chain graphs, the left graph (d) specifies $Y \perp\!\!\!\perp W \mid Z$ and $X \perp\!\!\!\perp Z \mid W$ and the right graph (e): $A_a \perp\!\!\!\perp B_b \mid (B_a, A_b)$ and $B_a \perp\!\!\!\perp A_b \mid (A_a, B_b)$

of Figure 2, and it may have dashed arrows pointing to nodes connected either by undirected full lines or by undirected dashed lines and it may have full arrows pointing to nodes connected by undirected full lines. It does not contain any directed cycles in the following sense: if we start at a node i and move along a path in the graph respecting the directions of the arrows we cannot come back to node i after having passed an arrow. This implies that the set of nodes V can be partioned into subsets $a, b, c, d, \ldots$ such that the graph can be arranged in a sequence, i.e. a chain of boxes such that there are only undirected edges within each box and only arrows betwen boxes. As has been described for Figure 1 the chain order has to be specified from subject-matter considerations.

Two small chain graphs are shown in Figure 4. For these the independence structures cannot be represented in terms of a directed acyclic graph.

Illustration to d): An independence graph as in Figure 4d) is obtained in a particular study of healthy female adults and variables Y, log (systolic/diastolic) blood pressure; X, log diastolic blood pressure; Z, body mass (weight in kg relative to height in cm); W, age in years. The body mass is likely to be higher the older a person is and the two blood pressure measures remain correlated after regression on body mass and age. High body mass but not high age is a direct risk factor for a low ratio of systolic to diastolic blood pressure and, similarly, high age but not high body mass is a direct risk factor for a high level of diastolic blood pressure. ◇

Illustration to e): An independence graph as in Figure 4e) is obtained in a study of 350 children selected so that there were roughly equal numbers of children with organic and psychosocial risk factors at the time of birth. Two variables are recorded at age four years (a) and at age two years (b).

These are A, psychic disorder (yes, no) and B, motoric handicap (yes, no). At age two years the two developmental aspects A_b, B_b are associated marginally. If psychic disorder at age four years is to be predicted from the other three variables, then the information about motoric handicap at age two years does not improve prediction given the information on the remaining two variables. Similarly if motoric handicap at age four years is to be predicted from the other three variables, then the information about psychic disorder at age two years does not improve prediction given the information on the remaining two variables. ◇

In a general chain graph the interpretation of any pairwise relation of nodes within a box is conditionally on variables in all boxes to the right for dashed (undirected) lines and, in addition, given the remaining variables within the box for full (undirected) lines. A pairwise relation of nodes between boxes means a regression of the response variable (to which the arrow points) on variables in all boxes to the right for dashed arrows and, in addition, on the remaining variables within the box of the response variable for full arrows. In case each box contains only a single variable a chain graph is a directed acyclic graph and in case there is only a single box the chain graph is degenerated to a concentration or a covariance graph for all variables.

4. Constructing conditional covariance and concentration graphs

We assume first that the joint distribution of variables Y_V, satisfies the independencies of a given directed acyclic graph G^V_{da}, where $V = \{1, \ldots, p\}$ are the nodes of the graph and that each edge present corresponds to a substantial regression coefficients. From such a graph we construct the covariance and the concentration graph in the conditional distribution of Y_S given Y_C. The selected set S, the conditioning set C and the set M of variables, over which we implicitly marginalize, partition V, i.e $V = S \cup C \cup M$. If Y_V has a joint normal distribution these two graphs reflect the implied pattern of zeros in the conditional covariance matrix of Y_S given Y_C, in $\Sigma_{SS.C}$, and in the conditional concentration matrix of Y_S given Y_C, in $(\Sigma_{SS.C})^{-1}$, where in the usual notation for matrices Σ and Σ^{-1}, partitioned with respect to S, C, M we can write

$$\Sigma_{SS.C} = \Sigma_{SS} - \Sigma_{SC}\Sigma_{CC}^{-1}\Sigma_{CS},$$

$$(\Sigma_{SS.C})^{-1} = \Sigma^{SS.M} = \Sigma^{SS} - \Sigma^{SM}(\Sigma^{MM})^{-1}\Sigma^{MS},$$

and $\Sigma^{SS.M}$ denotes the concentration matrix of Y_S obtained after marginalizing over variables Y_M.

In order to study such derived structures more generally we begin by modifying a directed acyclic graph. A directed graph modified at node t, $G^{V.t}$ is obtained from G^V_{da} in the following way. Within the subgraph of all ancestors A_t of t we join any sink-oriented v-configuration, that is any pair (r,s) of A_t which is unjoined but has a collision node as common neighbour in A_t. By repeating this for each node of a given set N the directed graph modified on a set of nodes N, $G^{V.N}$, is obtained.

(1) The covariance graph G^S_{cov} given C implied by G^V_{da} is a graph of dashed lines in nodes of S. It has an edge i,j if and only if in $G^{V.C}$ there is a collisionless path from i to j outside C.

(2) The concentration graph G^S_{con} given C implied by G^V_{da} is a graph of full lines in nodes of S. It has an edge i,j if and only if in $G^{V.C\cup S}$ there is an edge i,j or a collisionless path from i to j in M, that is outside $C \cup S$.

Example 1: If the directed acyclic graph is as given in Figure 5, the selected set is $S = \{3,4,5,6\}$ and the conditioning set is $C = \{1\}$, then $M = \{2,7,8,9,10,11\}$, the generating graph is modified in two steps to give $G^{V.C}$ and $G^{V.C\cup S}$ of Figure 6

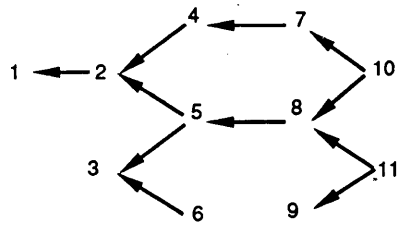


Figure 5: A given directed acyclic graph in 11 nodes

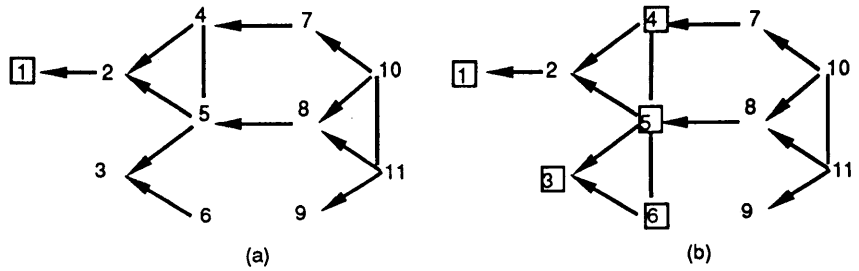


Figure 6: The modified directed graphs (a) $G^{V.C}$ and (b) $G^{V.C\cup S}$ obtained from the graph of Figure 5, where $C = \{1,8\}$ and $S = \{3,4,5,6\}$

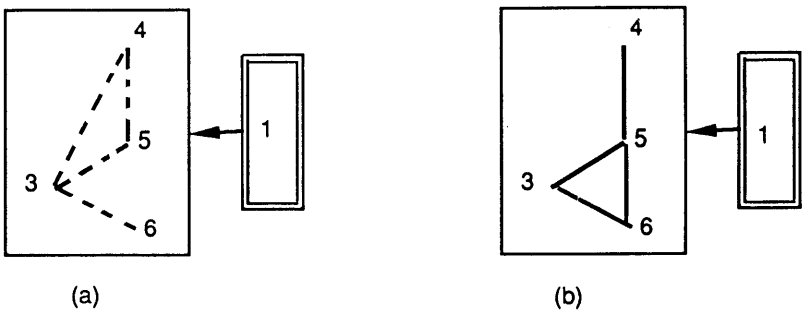


Figure 7: The conditional covariance graph (a) and concentration graph (b) of Y_S given Y_C, where $S = \{3,4,5,6\}$ and $C = \{1\}$, as implied by the generating graph of Figure 5

and the desired conditional covariance and concentration graphs are as given in Figure 7. Each has in this case more edges than the subgraph induced by nodes S in the generating graph.

Example 2: If the directed acyclic graph is again as given in Figure 5, the selected set is $S = \{2,3,4,5,6\}$ and the conditioning set is $C = \{10,11\}$, then $M = \{1,7,8,9\}$, the modified directed graphs $G^{V.C}$ and $G^{V.C \cup S}$ are in Figure 8 and the desired conditional covariance and concentration graph are as given in Figure 9.

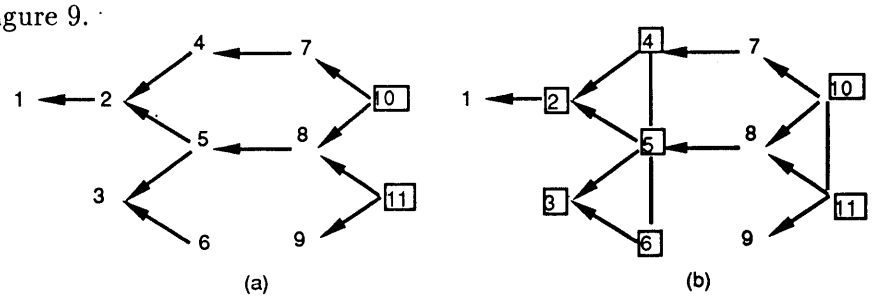


Figure 8: The modified directed graphs (a) $G^{V.C}$ and (b) $G^{V.C \cup S}$ obtained from the graph of Figure 5, where $S = \{2,3,4,5,6\}$ and $C = \{10,11\}$

Example 3: The directed acyclic graph of Figure 5 implies as covariance graph for $S = \{7,8,9\}$ marginally, that is with an empty conditioning set, $C = \emptyset$, a path from unjoined nodes 7 and 9 via node 8 and a complete concentration graph for this trivariate distribution.

Example 4: The directed acyclic graph of Figure 5 implies as concentration

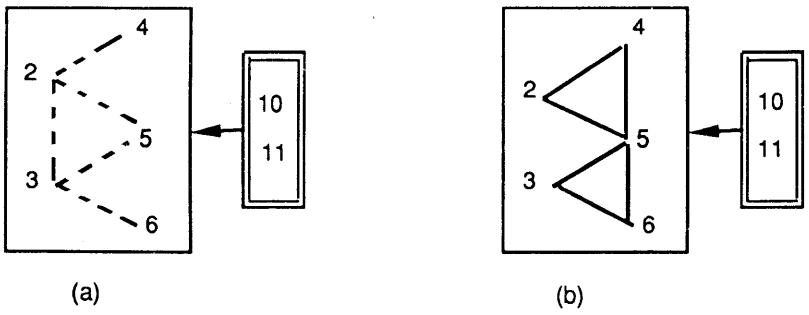


Figure 9: The conditional covariance graph (a) and concentration graph (b) of Y_S given Y_C, where $S = \{2,3,4,5,6\}$ and $C = \{10,11\}$, as implied by the generating graph of Figure 5

graph for $S = \{4,5,6\}$ conditionally given all remaining variables a path from unjoined nodes 4 and 6 via node 5 and a complete covariance graph for this trivariate distribution.

These results are direct consequences of the separation criterion for directed acyclic graphs (Pearl, 1988). There are two quite different routes of generalizing them. Either, the more general separation criterion for a full-edge chain graph (Frydenberg, 1990) could be adapted to obtain modified graphs with the necessary information on implied covariance and concentration graphs or, one may attempt to reinterpret a chain graph with the help of a directed acyclic graph.

For the former the notions of an ancestor, of a collision node and of an edge-inducing path would have to be generalized appropriately. In general however, stronger assumptions on a corresponding joint distribution are necessary to derive statements about implied associations from a graph containing undirected edges than from a fully directed graph; see Wermuth & Cox (1995) for a discussion.

For the latter, it follows for instance from results on the Markov equivalence of independence graphs (Frydenberg, 1990) that the set of independencies implied by a full-edge chain graph remains unchanged if each (partially) undirected v-configuration in it can be oriented to have a source or a transition node configuration and no directed cycle results. More general results are available but will not be given here.

For instance, if in Figure 2 the undirected edge XY is replaced by an arrow

pointing to X from Z and the undirected edge FE is replaced by an arrow pointing to E from F and the boxes are removed, then the resulting graph is fully directed and acyclic; it implies the same independence structure for the eight variables as the chain graph of Figure 2.

After modifying this directed acyclic graph in the way described above it follows that in the conditional distribution of B,D,F,E given A the covariance and the concentration graph are complete, i.e. all edges are present, and in the marginal distribution of A,Y,X,Z the covariance graph is complete but in the concentration graph edge AX is missing, that is, the edges present in it coincide with the edges present in the subgraph induced by A,Y,X,Z in Figure 2. Thus, for instance, an observed strong association between A, drop-out rate, and X, student's motivation, given Y, expected achievement, and Z, integration into student group, would be evidence against the structure of Figure 2 and, similarly, were the associations for any edge present in the graph or for an edge present in one of the implied covariance or concentration graphs weak this would need an explanation.

Bibliography

Bollen, K.A. (1989). *Structural equations with latent variables.* New York: Wiley.

Cochran, W.G. (1965). The planning of observational studies of human populations. *J. Roy. Statist. Soc. A*, **128**, 234-265.

Cox, D. R. (1966). Some procedures connected with the logistic qualitative response curve. In: *Research Papers in Statistics: Essays in Honour of J. Neyman's 70th Birthday* (ed. F. N. David), pp. 55-71. Chichester: Wiley.

Cox, D.R. & Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statist. Science*, **8**, 204-218; 247-277.

Cox, D.R. & Wermuth, N. (1994). Tests of linearity, multivariate normality and adequacy of linear scores. *Applied Statistics, J. Roy. Statist. Soc. C*, **43**, 347-355.

Cox, D.R. & Wermuth, N. (1995). *Multivariate dependencies, models, analysis and interpretation.* In preparation.

Frydenberg, M. (1990). The chain graph Markov property. *Scand. J. Statist*, **17**, 333-353.

Giesen, H., Böhmeke, W., Effler, M., Hummer, A., Jansen, R., Kötter, B. Krämer, H.-J. Rabenstein, E. & Werner, R.R. (1981) *Vom Schüler zum Studenten. Bildungslebensläufe im Längsschnitt.* Reihe: Monografien zur Pädagogischen Psychologie, 7. München: Reinhardt.

Gold, A. (1988). *Studienabbruch, Abbruchneigung und Studienerfolg: Vergleichende Bedingungsanalysen des Studienverlaufs.* Frankfurt, Main: Lang.

Koster, J.T.A. (1993). Markov properties of nonrecursive causal models. Department of Sociology. Erasmus University, Rotterdam. Technical Report.

Lauritzen, S.L., Dawid, A.P., Larsen, B. & Leimer, H.-G. (1990). Independence properties of directed Markov fields. *Networks*, **20**, 491-505.

Lauritzen, S.L. & Spiegelhalter, D.J. (1988). Local computation with probabilities on graphical structures and their application to expert systems (with discussion). *J. Roy. Statist. Soc. B*, **50**, 157-224.

McCullagh, P.& Nelder, J. A. (1989). *Generalized Linear Models, 2nd ed.*. London: Chapman and Hall.

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

Smith, J.Q. (1989). Influence diagrams for statistical modelling. *Ann. Statist.*,**17**, 654-672.

Spirtes, P. (1993). Directed cyclic graphs, conditional independence and nonrecursive linear nstructural equation models. Department of Philosophy, Carnegie-Mellon University, Pittsburg. Technical Report.

Streit, R. (1995). *Graphische Kettenmodelle mit binären Zielgrößen: Modellierung und Datenbeispiele in psychologischer Forschung.* Lengerich: Pabst.

Wermuth, N. & Lauritzen, S.L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *J. Roy. Statist. Soc. B*, **52**, 21-72.

Wermuth, N. & Cox, D.R. (1995). A note on association models defined over independence graphs. Berichte zur Stochastik und verwandten Gebieten. Universität Mainz, 95-2.

Acknowledgement

The paper is based on joint work with Sir David Cox, University of Oxford; support of this joint work by the Humboldt- and the Max-Planck-Society is gratefully acknowledged.

Summary

In observational studies a general principle of empirical research is 'to make a hypothesis elaborate', that is, to study implications of a hypothesis under systematically varied conditions. We show that this principle can be applied to multivariate structures represented by chain graphs since some of the consequences of such a hypothesized structure can be derived for any joint conditional distribution of a subset of variables.

Résumé

Dans des études d'observations un principe general consiste à "l'élaboration d'une hypothèse", c'est-à-dire d'étudier l'implications d'une hypothèse dans des conditions variées systematiquement. Nous démontrons que ce principe peut ètre appliqué à des structures multivariable, lesquelles on peut représenter avec des images des chaines; c'est parce que on peut d'eriver les conséquences d'une telle structure pour chaque distribution conditionelle d'un subgroupe des variables.