In: Computational Statistics. Y. Dodge & J. Whittaker (eds). Heidelberg: Physica.

Graphical Models for Dependencies and Associations

Nanny Wermuth and D.R. Cox

Psychological Institute, University of Mainz, Staudingerweg 9, 6500 Mainz, Germany

Nuffield College, Oxford, OX1 1NF, United Kingdom

Abstract
The role of graphical representations is described in distinguishing various special forms of independency structure that can arise with multivariate data, especially in observational studies in the social sciences. Conventions for constructing the graphs and strategies for analysing three sets of data are summarized. Finally some directions for desirable future work are outlined.

1. Introduction

In many fields of work, especially but not exclusively in observational studies in the social sciences, data are obtained on a considerable number of variables for each individual under investigation. The isolation of the dependencies and associations between these variables, which are typically of substantive interest, may sometimes be possible by relatively simple and direct methods of analysis but in many situations, especially where the relations are relatively subtle, there is a need for genuinely multivariate methods of statistical analysis into which subject-matter knowledge is to be integrated in an appropriate way.

The need to discuss special structures specified in terms of marginal independence and conditional independence arises partly because these independencies may be of substantive interest themselves but partly because they offer the possibility of simplifying potentially complex problems of analysing relations among component variables by splitting into a sequence of smaller analyses each involving possibly only few variables; thereby a superabundance of parameters is also avoided.

Already with four component variables there is a quite rich and potentially confusing variety of special structures to be considered and graphical representation can help to clarify the various possibilities. We use the graphical chain representations for dependencies and associations as they have been introduced by Lauritzen and Wermuth (1989), Wermuth and Lauritzen (1990) and Cox and Wermuth (1992a).

These graphs consist mainly of nodes for variables with circles for continuous variables and with dots for discrete variables, of at most one edge for the considered relation of each pair and of two kinds of edge to indicate different types of conditional analyses. As described below in detail, these graphs permit a precise identification of the meaning of each edge be it present or missing in the graph.

A systematic account of graphical methods by Whittaker (1990) emphasizes undirected graphs with all edges being undirected full lines, i.e. systems in which all variables are treated on an equal footing. Here we use largely directed graphs with edges being arrows in order to emphasize relations of response and dependence while undirected lines mainly indicate joint dependence of responses. We sometimes also use graphs with dashed instead of full edges: the dashed edges are to denote a smaller conditioning set for a relation than full edges.

There are strong connections with, in particular, the long history of work on path analysis in genetics (Wright, 1921, 1923), on simultaneous equations in econometrics (Goldberger, 1964) and on linear structural models in psychometrics (Jöreskog, 1973). But, simultaneous or structural equations do not, in general, permit a graphical chain representation as it is discussed here. The reason is that zero coefficients in a structural equation need not correspond to a relation of conditional independence. Exceptions are models which belong to the subclass of systems of univariate recursive regressions or which are multivariate regression equations.

Our use of graphical representations is intended to aid interpretation in situations with moderate numbers of variables, say four to ten, and it is to be contrasted with the use in connection with decision analysis via expert systems where very large numbers of variables may be involved (Lauritzen and Spiegelhalter, 1988; Pearl, 1988; Smith, 1988).

In this paper we shall give in Section 2 the conventions that are used to construct a chain graph which allow the precise interpretation of the independencies implied. In Section 3 we provide for three different sets of data of six to eight variables background information and outline general strategies that may be used to incorporate some of this substance matter knowledge into a model formulation. The paper concludes with a discussion of desirable future developments.

2. Conventions for constructing chain graphs

We distinguish 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 typically introduced by a priori subject matter considerations although we do not exclude it

being introduced in the light of initial data analysis.

The following conventions are used in constructing the chain graphs in this paper:

a) each continuous variable is denoted by a node which is a circle and each discrete variable by a node which is a dot;

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);

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 variables in the same box, i.e. marginally with respect to box variables of the remaining responses;

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

g) graphs drawn with boxes represent substantive research hypotheses in which the presence of an edge means that the corresponding partial association is large enough to be of substantive importance, corresponding to the notion that the model being represented is in some reasonable sense the simplest appropriate; graphs obtained by removing the boxes represent statistical models in which a connecting edge places no constraint on the association;

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, i.e. to indicate independence of several (joint) variables conditionally on all boxes to the right;

i) if a 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 conditional (regression) model instead of a regression chain model, the latter containing parameters also for those components which are exclusively explanatory.

Some additional restrictions are needed to obtain well-defined statistical models corresponding to such graphs. These are at present (i) that each set of responses connected by full edges has only full arrows pointing to it, each set of responses connected by dashed edges has only dashed edges pointing to it and only arrows of the same kind point to each set of unconnected responses; (ii) that graphs with dashed

edges have no discrete response variables. We expect that some further mixtures of both types of edges are permissible, i.e. would for instance lead to Barndorff-Nielsen's (1978, p.122) mixed parametrisation of an exponential family; similarly, we expect that the models of Liang et al. (1991) have a graphical representation with dashed edges for discrete response variables, but more work is needed to obtain the precise details.

A large variety of traditional models is implied by special cases of the above graphs together with special distributional assumptions, for instance multiple linear regression, analysis of variance and covariance, multivariate regression, logistic regression, probit regression and log-linear models, but also more recent models such as seemingly unrelated regressions (Zellner, 1962), covariance selection (Dempster, 1972), independence hypotheses with linear structure in covariances (Anderson, 1973) or block regression (Wermuth, 1992).

The essential feature one needs to know for a chain graph representation is how conditional or marginal independences are reflected in the parametrisation. For instance, for the mixed case of both discrete and continuous variables this was derived for the Conditional Gaussian (CG) distribution and regressions by Lauritzen and Wermuth (1989) or if just linearity of regressions is assumed independence means just linear independence but not probabilistic independence.

Figure 1 shows the graphs of two mutually exclusive independence structures expressed with graphs for a special multivariate regression (dashed edges) and a special block regression (full edges). Independencies are expressed in the notation by Dawid (1979), for instance conditional independence of variables Y and V given Z is written as $Y \perp\!\!\!\perp V \mid Z$.

In general, sequences of boxes give a (partially) ordered partitioning of $\{1, \ldots, r\}$ and determines those sets of variables which are to be analysed simultaneously, called sometimes *concurrent variables*, i.e. all variables in the same box and in the boxes to the right and *concurrent explanatory variables*, i.e. all variables in boxes to the right of the considered set of responses. For a general chain $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_J)$ the concurrent variables are given by $\mathcal{C}^{(j)} = \mathcal{C}_j \cup \ldots \cup \mathcal{C}_J$ for $j = 1, \ldots, J$ while the concurrent explanatory variables for the responses in $\mathcal{C}^{(j)}$ are given by $\mathcal{C}^{(j+1)}$ except for $\mathcal{C}^{(J)}$, where it is the empty set. For instance, for the dependence chain $\mathcal{C} = (a, b, c)$ coresponding to the (vector) variables Y_a, Y_b, Y_c the chain elements a, b, and c define three sets of concurrent variables with: $a \cup b \cup c$, $b \cup c$, and c and three sets of concurrent explanatory variables with: $b \cup c$, c, and $\emptyset$, i.e. the empty set.

This provides the definitions needed to interpret edges in chain graphs be they present or missing:

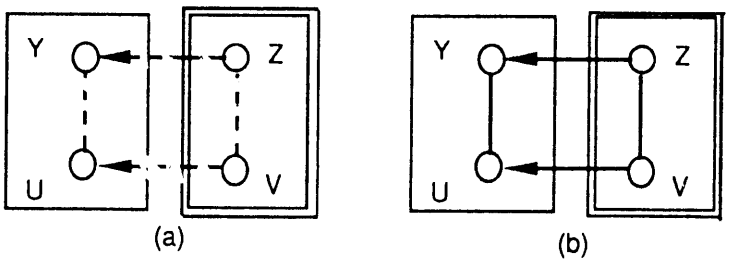

Figure 1: the left graph (a) specifies $Y \perp\!\!\!\perp V \mid Z$ and $U \perp\!\!\!\perp Z \mid V$ and and the right graph (b): $Y \perp\!\!\!\perp V \mid (Z, U)$ and $Z \perp\!\!\!\perp U \mid (Y, V)$. Under normal theory maximum likelihood estimates for hypothesis (a) on a multivariate regression of (Y, U) on (V, Z) can be directly obtained with an estimating algorithm for linear structural equations such as in LISREL (Jöreskog and Sörbom, 1984) but not with an algorithm for undirected graphical association models such as in MIM (Edwards, 1990, 1991) while those for hypothesis (b) on a block regression can be directly obtained from MIM but not from LISREL

- a full edge denotes a partial association given all the remaining concurrent variables.
- a dashed edge denotes a partial association given all the remaining concurrent explanatory variables of the set of responses considered.
- a missing full edge means conditional independence given all the remaining concurrent variables.
- a missing dashed edge means conditional independence given all the remaining concurrent explanatory variables of the set of responses considered.

The graphs cannot reflect how a particular substantial conditional dependence looks like, i.e. for any edge present in the graph the type of the nonvanishing conditional relation needs further description.

Graphs with in our notation full edges have an elegant connection with the theory of Markov random fields which allow general properties to be deduced; for instance it is possible to read directly off the graphs all implied independencies and to decide from the graphs of two distinct models whether they are equivalent (Frydenberg, 1990). 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.

3. Outline of strategies for analysing three sets of data

We discuss in turn strategies that may be used in analysing the following three sets of data without giving details of analyses which have been or will be reported elsewhere. The first example has six binary variables with the probabilities of interest all not extreme so that linear in probability regressions can be used as an approximation to logistic regressions (Cox, 1966; Cox and Wermuth, 1992c, d). The second example is a mixture of six continuous and two discrete variables in which some of the dependencies are of a nonlinear kind. The third example with eight continuous variables contains six responses measuring different aspects of brain activity from which new responses can be derived as linear combinations which display special relations of independence to the explanatory variables (Cox and Wermuth, 1992b, Wermuth and Cox, 1992).

Example 1: From a cohort study of students who completed their first 13 years of formal schooling in the years 1973 to 1976 in Germany (Giesen et al., 1981) an analysis is reported by Weck (1991) for six binary variables observed for 2026 students. The variables are A, change of field of study (yes, 19%; no); B, change of high school (yes, 20%; no); C, integration into the high school class (poor, 10%; good); D, a high school class repeated (yes, 34%; no); E, change of primary school (yes, 20%; no); F, education of the father (at least 13 years of formal schooling, 43%; less than 13 years of formal schooling). Figure 2 shows the first classification into the response

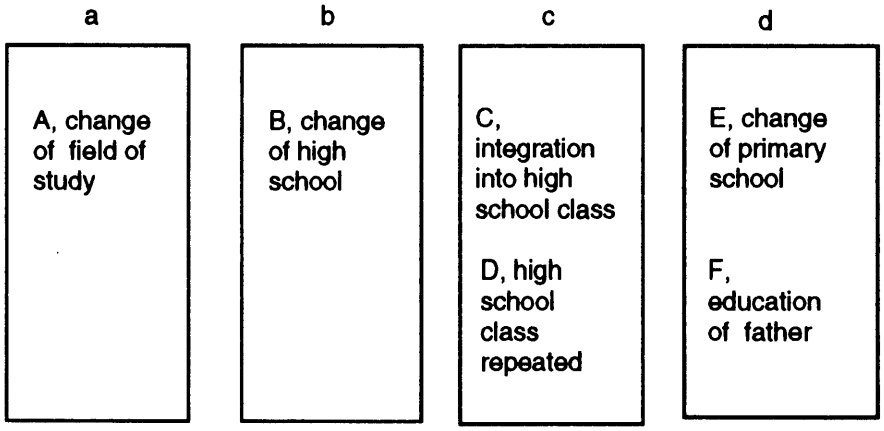

Figure 2: A first ordering of the variables in Example 1 with A as response variable of primary interest

variable of primary interest A, depending potentially on all other variables, into the intermediate response B to C, D, E, F, into the intermediate joint responses C, D

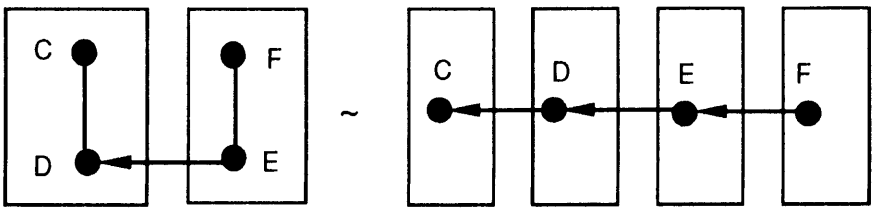

Figure 3: Two distributionally equivalent models implying $C \perp\!\!\!\perp (E,F) \mid D$ and $D \perp\!\!\!\perp F \mid E$

to E, F, and into the joint explanatory variables E, F. The strategy we use for the analysis consists of several steps:

(i) for each given response we check whether one of the potential explanatory variables brings no additional predictive effect in addition to all of the remaining concurrent variables. This means that we test e.g. $A \perp\!\!\!\perp D \mid (B,C,E,F)$ or $D \perp\!\!\!\perp E \mid (C,F)$;

(ii) whenever several variables appear to have no such additional predictive effect, we check whether they may be jointly removed. Thus, we test for instance $A \perp\!\!\!\perp B,D \mid (C,E,F)$. If there appears to be a good fit (in such a test with typically many degrees of freedom) we take the corresponding model as a basis and check in turn whether the main effect of each variable removed does still not lead to an improvement in prediction; this is to reassure us that we have not overlooked a substantial relation as hidden in an overall test;

(iii) we check whether the test results for independence of any set of joint responses suggests any simplifications. Figure 3 shows such a result for variables C, D: the structure with C, D as joint responses to the joint explanatory variables E, F is distributionally equivalent to the displayed special univariate recursive system.

(iv) the results of these checks are summarized in a chain graph, as in Figure 4.

(v) we check whether any logistic regression can be well approximated in terms of a linear in probability regression because the predicted probabilities for a response are not too extreme, i.e. they are all between .10 and .90; the object being to present the results in a way which is most directly interpretable. For the dependencies in Figure 4 this is possible for all responses, and, in addition, there are exclusively main effects. For instance we can write (Cox and Wermuth, 1992c)

$$\hat{\pi}^{A|B,C,D,E,F}_{1|j,k,l,m,n} = \hat{\pi}^{A|C,E,F}_{1|k,m,n} = .25 + .05k^* + .02m^* + .02n^*,$$

where e.g. $k^* = +1$ if $k = 1$ (poor integration into high school class) and $k^* = -1$ if $k = 0$. This representation permits to read off directly that the probability to change the field of study is estimated as highest (34%) for a student, who integrated

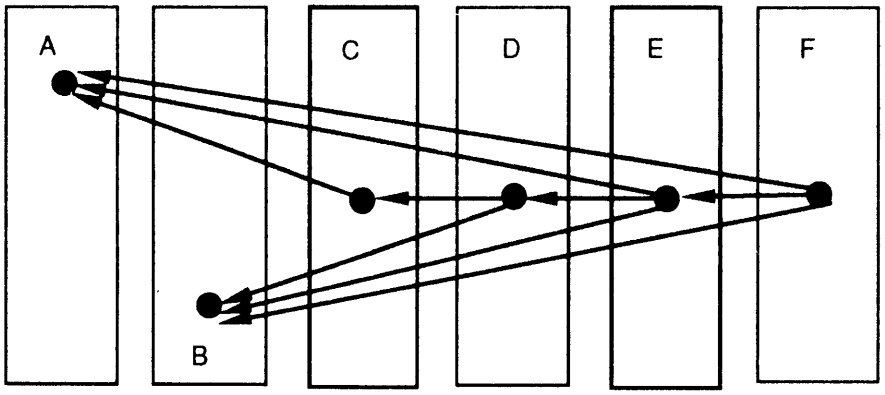

Figure 4: Chain graph of the simplest model well compatible with the data of Example 1: a special system of univariate recursive logistic regressions

poorly into his high school class, who had changed primary school, and whose father had reached a higher educational level; it is estimated as lowest (16 %) for a student who integrated well into his high school class, who did not change primary school, and whose father had less than 13 years of formal schooling.

Example 2: For 68 diabetic patients we have data from an investigation of determinants of blood glucose control (Kohlmann et al., 1991). The variables considered are Y, a particular metabolic parameter, the glycosylated haemoglobin GHb; X, a score for particular knowledge about the illness; three different attitudes of the patient measured with sum scores of questionnaires which are to capture how the patient attributes what is happening in relation to his illness: Z, social externality (powerful others are responsible); U fatalistic externality (mere chance determines what occurs); V, internality (the patient sees himself as mainly responsible); W, duration of illness in months; and two intrinsic characteristics of the patient: A, level of education (less than 13 years, 56%; at least 13 years) and B, gender (males, 51%; females).

For an analysis we proceed similarly as in Example 1, but take into account some of the special features of the present data. Figure 5 shows a first ordering of the variables into the response of primary interest, Y, into two sets of intermediate responses, X and Z, U, V, and into a set of purely explanatory variables W, A, B.

With a total of only 68 observations we cannot expect to get useful estimates of the relations in all four cells defined by the two dichotomous variables A, B, thus we collapse over any discrete explanatory variable whenever it has little predictive effect on the variable of primary interest. This is possible for gender, but not for the level of education. We further delete any continuous variable with no direct relation

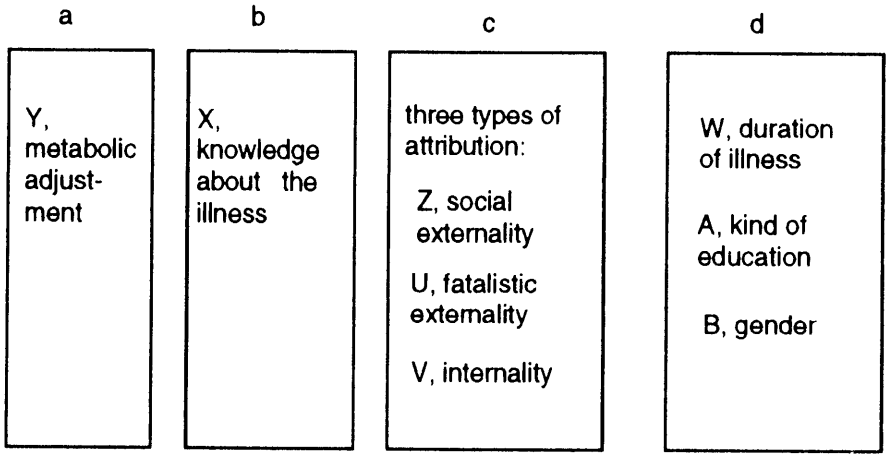

Figure 5: A first classification of the variables in Example 2 with Y as response variable of primary interest

to the responses of main interest, i.e. to either Y or X, so that the independence graph shown in Figure 6 just shows explanatory variables important for Y or X. To

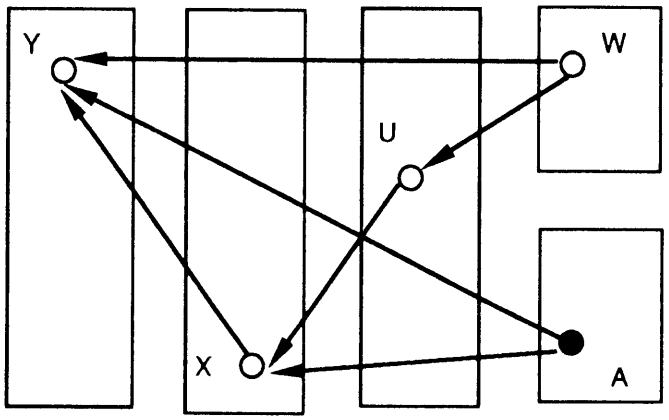

Figure 6: Chain graph for a a subset of variables of Example 2 of the simplest model well compatible with the data

describe the effects of a discrete explanatory variable seperate independence graphs for the different levels of the discrete variable may sometimes be shown (Højsgaard and Skjøth, 1992) or further detailed descriptions may be given.

From the data of Example 2 we can conclude that for all patients the metabolic adjustment is better (low levels of GHb) the better the knowledge (X) about the illness. The variables level of education (A) and fatalistic externality (U) have an

additive effect on knowledge about the illness (X): knowledge is higher the lower the fatalistic attitude and it is higher the higher the eduational level. But, there is a significant interactive effect of level of education (A) and duration of illness (W) on metabolic adjustment (Y), i.e. there are different determinants for patients with lower ($A = 1$) and with a higher educational level ($A = 2$). Note that it is impossible to judge from the graph alone whether the effect of two explanatory on a response is additive or interactive. The interactive effect of A and W on Y is revealed by analyzing the data for levels $A = 1$ and $A = 2$ separately: for the patients with a lower educational level the metabolic adjustment is rather poor for short durations of illness but quite good for long durations of illness. One tentative interpretation is that poor adjustment in the early years of the illness becomes continuously better the longer the patient has had experience with the illness. By contrast, for patients with a higher educational level the metabolic adjustment is rather good in the early years of the illness, but poorer for long durations of illness. Possibly these patients use their experience with the illness to take more calculated risks the longer the illness lasts.

There is of course a danger of overinterpreting the associations of just a single sample: the observed relations could instead be the consequence of some unknown selection effects. A decision on this will be possible once changes in metabolic adjustment of individual patients have been documented over time.

Example 3: For 72 students, who participated in an experiment designed to observe the effects of stress inducing conditions (Hänsel, 1992) we obtained six measurements of brain activity (CNV: Contingent negative variation), i.e. CNV measured at three locations (frontal, central and parietal) and under two experimental conditions, one in which the participant has to prepare for a motor activity at a stimulus four seconds after a signal has forwarned him (g: the 'go'-situation) and another in which he is not to react (n: the 'nogo'-situation). This defines six responses $Y(f,n), Y(c,n), Y(p,n), Y(f,g), Y(c,g), Y(p,g)$ all measured during a fixed first part of an early interval after the stimulus (Glanzmann and Fröhlich, 1986). In addition, we have observations for two potential explanatory variables: X_1, the personality characteristic anxiety measured as sum score with Spielberger's Trait-State inventory and X_2, attention or arousal measured as the difference in eye movements under the two experimental conditions (EOG, g-n: Electrooculogram, difference between 'go'- and 'nogo'-situation). Extreme values in EOG are taken as an indication that no unconfounded measurement of brain activities is possible, hence persons with such values are excluded. No simplyfied structure could be deduced from the observed correlations between the eight variables.

In this example we proceed quite differently than before since we have a set of

six joint responses (see also Figure 7) which are similar, in the sense that they are

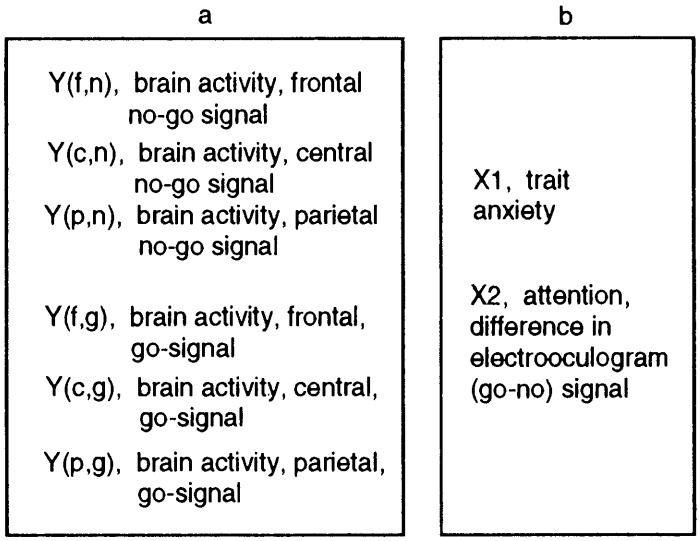

Figure 7: Responses Y and explanatory variables X in Example 3

measured in the same way and they are thought of as capturing different aspects of some underlying phenomenon. This means that we have a situation in which we can expect that new responses might lead to simplified interpretations if they are derived as linear combinations of the original response variables in such a way that each new response Y_i^* has linear conditional independence of all explanatory variables except one (Cox and Wermuth, 1992b).

Without going into much of the detail reported in Wermuth and Cox (1992) we note that the derived responses were calculated for three corresponding sets of four variables $[Y(f,n), Y(f,g), X_1, X_2]$, $[Y(c,n), Y(c,g), X_1, X_2]$, and $[Y(p,n), Y(p,g), X_1, X_2]$, in which we have the same explanatory variables and the type of responses under the two experimental conditions; only the location of the measurement is different for the three variable sets. The squared canonical correlations are (.02, .24), (.10, .23), and (.07, .22), respectively, and the calculated transformation matrices ((1, .54) & (1, -.65); (1, .78) & (1 -.91); (1, .21) & (1 -.71)) suggest that for all three locations of measurement we can take as derived responses the differences and the sums of the measurements under the two experimental conditions. This leads to the association structure displayed in Figure 8 as well compatible with the observations. The standardized regression coefficients in the two correponding multivariate regression analyses are the following sets of simple correlations (.15, .33, .34) for the dependence of the first set of derived responses on trait anxiety and (-.48, -.48, -.40)

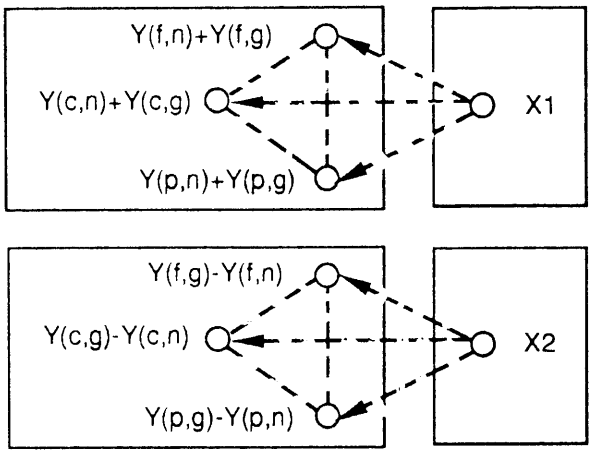

Figure 8: Chain graph showing separate dependence of derived responses on just one of the explanotory variables of Example 3

for the dependence of the second set on attention. They indicate a certain replication of the results at the different sites since the correlations differing between sites show the same direction of dependence just different strengths.

4. Discussion

There are a number of open problems for future research connected with the issues discussed in this paper. For instance, techniques for relatively routine fitting of data to complex models including the calculation of standard errors are needed, possibly utilizing previous work by Frydenberg and Edwards (1989) Cox and Wermuth, (1990, 1991) and Jensen et al. (1991). The role of latent variables needs more study, both in connection with errors in measurement and with the incorporation of hidden variables. Properties of graphs with what we have termed dashed edges need further study. Models with discrete variables and with mixtures of discrete and continuous variables corresponding to graphs with dashed edges will have to be developed.

While our account here emphasized the interpretation of models and strategies of analysis, formal statistical problems of estimation and testing of model adequacy have been addressed elsewhere.

Acknowledement

We are grateful to the British German Academic Research Collaboration Programme for supporting our work.

References

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

Barndorff-Nielsen, O. (1978). *Information and exponential families in statistical theory.* Chichester: Wiley.

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. and Wermuth, N. (1990). An approximation to maximum-likelihood estimates in reduced models. *Biometrika*, **77**, 747-761.

Cox, D.R. and Wermuth, N. (1991). A simple approximation for bivariate and trivariate normal integrals. *International Statistical Review*, **59** 263-269.

Cox, D.R. and Wermuth, N. (1992a). Chain graph representations of linear dependencies. Berichte zur Stochastik und verw. Gebiete,\92-3. Universität Mainz.

Cox, D.R. and Wermuth, N. (1992b). On the calculation of derived variables in the analyses of multivariate responses. To appear in: *J. Multivariate Analysis.*

Cox, D.R. and Wermuth, N. (1992c). Response models for mixed binary and quantitative variables. To appear in: *Biometrika.*

Cox, D.R. and Wermuth, N. (1992d). A comment on the coefficient of determination for binary responses. *The American Statistician*, **46**, 1-4.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. B* **41**, 1-31.

Dempster, A.P. (1972). Covariance selection. *Biometrics*, **28**, 157-175.

Edwards, D. (1990). Hierarchical mixed interaction models (with discusssion). *J. Roy. Statist. Soc. Ser. B* **52**, 3-20, 51-72.

Edwards, D. (1991). A guide to MIM. Manual.

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

Frydenberg, M. and Edwards, D. (1989). A modified iterative proportional scaling algorithm for estimation in regular exponential families. *Comp. Stat. Data Anal.*, **8**, 143-153.

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