In: Highly Structured Stochastic Systems. P. Green, N. Hjort and S. Richardson (eds). Oxford University Press, 33-39.

Analysing social science data with graphical Markov models

Nanny Wermuth
University of Mainz

1 Introduction

The term graphical Markov models has been suggested by Michael Perlman, University of Washington, for multivariate statistical models in which a joint distribution satisfies independence statements that are captured by a graph. The study and the development of these models is such an active research area that some of their properties are not yet discussed in recent statistical books concentrating on them (Edwards, 2000; Lauritzen, 1996; Cox and Wermuth, 1996). We ask here
– how do they relate to models used more traditionally for data analysis?
– what do they offer in addition?
– are case studies available?

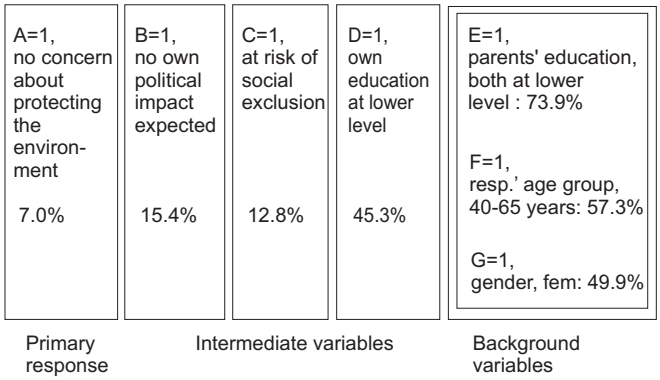
In independence graphs used to summarize aspects of detailed statistical analyses each vertex or node represents a variable feature of individuals under study. These features may be categorical. Then they are denoted by capital letters $A, B, C, \ldots$, they are modelled by discrete random variables and drawn as dots. Or, they may have numerical values of substantive meaning. Then they are denoted by capital letters $X, Z, U \ldots$, they are typically modelled by continuous variables and drawn as circles. If this distinction is not important, the individual components of a vector random variable Y may be denoted by $Y_1, Y_2, \ldots$, or, more compactly, just by integers $1, 2, \ldots$.

Often substantive knowledge is strong enough to specify a fully ordered sequence of the variables which starts with a background variable, ends with a response of primary interest, and has single intermediate variables, which are both, potential responses to variables of the past and potentially explanatory to variables of the future. Then, no variable is taken to be explanatory for itself and an independence graph fitted to the responses will be fully directed and acyclic. For these we give here examples of research questions and of theoretical results.

2 A motivating research example

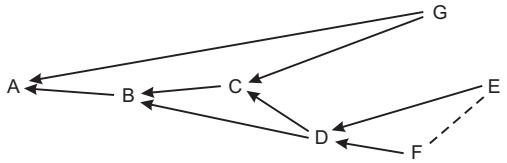
For the research questions: Who admits to be not concerned about protecting the environment? How does such an attitude develop? we use answers of 1228 respondents, aged between 18 and 65 years, from the General Social Survey in Germany in 1998.

The following Figure shows a first ordering of six variables which are categorized to be binary, together with the observed percentages for the stated category. Most variables are based on answers to a single question but, for instance, for risk of social exclusion is derived from several aspects such as no or incomplete vocational training and extended periods of unemployment.



After checking for interactive effects (Cox and Wermuth, 1994) and using a likelihood-ratio based model selection strategy we concluded that each of the univariate conditional distributions is here well described by logistic regressions having two main effects. For each response the important explanatory variables are shown in the graph by arrows pointing directly from the former to the latter. The factorization of the joint density f_V can be read off the graph to be

$$f_V = f_{A|B,G} f_{B|C,D} f_{C|D,G} f_{D|E,F} f_G f_{E,F}.$$



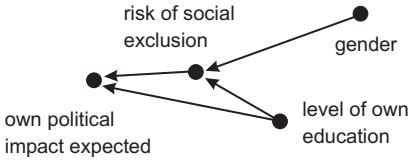
For each response the direction and strength of the dependencies can be read off the estimated conditional probabilities. The following table shows for which levels of the explanatory variables lowest and highest percentages are observed for level 1 of each response, together with these observed percentages.

Level of response:	A=1		B=1		C=1		D=1	
Explanatory variables:	BG		CD		DG		EF	
levels with highest perc.:	1,2	26.6	1,1	40.7	1,1	28.2	1,1	65.2
levels with lowest perc.:	2,1	2.2	2,2	8.2	2,2	4.7	2,2	7.7

One main gain of the graph is the possibility of tracing developments. For instance the path from G to C to B to A describes that women are at higher risk for social exclusion than men, that those at higher risk for social exclusion are less likely to believe in having an own political impact and that those perceiving to have no own political impact are more likely to be unconcerned about the

environment.

Another important use of a directed acyclic graph is that its consequences can be derived if only subsets of the variables are considered and if subpopulations are selected. General answers have been given recently (Wermuth and Cox, 2001a). For instance, the independence graph implied if variables A, F, E are ignored, that is the graph for the joint distribution of all remaining variables, is - in this example - again a directed acyclic graph, the graph shown below.



This follows by summing over variables A, F, E in the joint density f_V, in which each response depends only on the directly important explanatory variables.

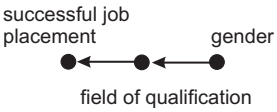
3 Some consequences of directed acyclic graphs

3.1 Consequences in simple cases

For relations among only three variables, we now show examples of consequences which have been described in the literature as spurious dependence, spurious association and selection bias. The graphs help here to visualize the concepts.

The first example for spurious dependence concerns the question of discrimination against women and data from the German labour market for academics, whose field of qualification was either mechanical enginering or home economics.

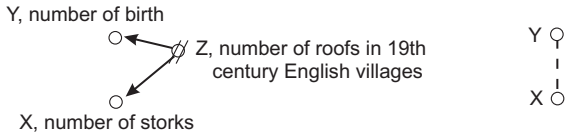
The well-fitting independence graph is



Ignoring the intermediate variable, i.e. marginalizing over B ($\not\emptyset$) leaves A dependent on C: the data appear to indicate discrimination, since men have a more than five times higher chance for successful job placement.

However, including the information of the field of qualification by fixing levels of the intermediate variable, i.e. conditioning on B ($\boxed{}$), shows A independent of C.

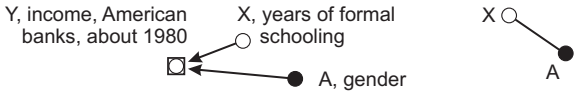
The second example for spurious asscociation was used by Y. Yule more than 100 years ago to argue that correlation is not causation. Ignoring the common explanatory variable, i.e. marginalizing over Z leaves Y and X associated.



Fixing levels of the common explanatory variable, i.e. conditioning on Z shows Y independent of X.

The third example for selection bias is due to H. Wainer. He showed how systematic differences in incomes of men and women having the same level of formal schooling get covered up when different scales are used for income in displays showing for both genders a systematic increase of income with higher levels of formal schooling.

Overall the level of formal schooling, X, is independent of gender, A, but after selecting levels of the common response variable, i.e. conditioning on Y, income, renders the explanatory variables to be associated: within given income groups women have a higher level of formal schooling.



Ignoring the common response variable, i.e. marginalizing over Y shows the two explanatory variables A and X to be independent.

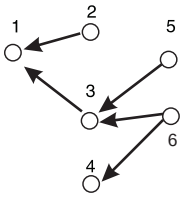
3.2 Some consequences of large graphs

In general, simple matrix calculations can be used to derive for all variable pairs whether a directed acyclic graph implies for instance marginal independence or not. An edge or incidence matrix is a way of storing the information in an independence graph with zeros for missing edges and ones for edges present. We let row i in an edge matrix correspond to node i in a graph and let the node set be ordered as $V = (1, \ldots, d_V)$ so that all ij-arrows for $j > i$ point from j to i. The edge matrix of a directed acyclic graph, $\mathcal{A}$, is then upper-triangular matrix with ones along the diagonal and a one in position (i, j) if and only if there is an ij-arrow present in the graph.

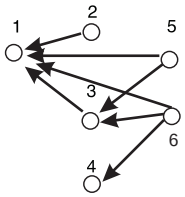
In the language for such directed graphs it has become a convention to call the node of a directly explanatory variable a parent and the node of a direct response variable a child. The node of an indirectly explanatory variable is named an ancestor, the node of an indirect response variable a descendant.

A directed acyclic graph is then often called the parent graph, G^V_{par}, with edge matrix $\mathcal{A}$. The graph obtained from it by adding a direction-preserving arrow for every ancestor-descendant relation is called the overall ancestor graph, G^V_{anc}, with edge matrix $\mathcal{B}$. An undirected graph of dashed or broken lines is induced by the parent graph G^V_{par} which has a missing ij-edge (and a missing ji-edge) if and only if Y_i is implied to be marginally independent of Y_j. It is called the induced overall covariance graph, G^V_{cov}. The name derives from joint Gaussian distributions for which marginal independencies are reflected as zeros

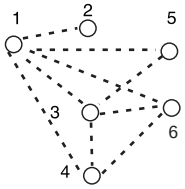
in the covariance matrix. We shall explain here why the edge matrix of the induced overall covariance graph is the indicator matrix of $\mathcal{B}\mathcal{B}^T$, where an indicator matrix $\mathrm{In}(M)$ of a matrix M has a one in position (i,j) if and only if the element of M in this position is nonzero. We give first an example with 6 nodes.



the parent graph G^V_{par} with edge matrix $\mathcal{A}$



the overall ancestor graph G^V_{anc} with edge matrix $\mathcal{B}$



the induced overall covariance graph G^V_{cov}, edge matrix $\mathrm{In}(\mathcal{B}\mathcal{B}^T)$

The corresponding edge matrices $\mathcal{A}$, $\mathcal{B}$ are

$$\mathcal{A} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 & 0 \\ & & 1 & 0 & 1 & 1 \\ & & & 1 & 0 & 1 \\ & \mathbf{0} & & & 1 & 0 \\ & & & & & 1 \end{pmatrix} \quad \mathcal{B} = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 1 \\ & 1 & 0 & 0 & 0 & 0 \\ & & 1 & 0 & 1 & 1 \\ & & & 1 & 0 & 1 \\ & \mathbf{0} & & & 1 & 0 \\ & & & & & 1 \end{pmatrix}.$$

In this small example it may be checked directly for which pairs the factorization of the joint density as given by the parent graph

$$f_{1,\ldots,6} = f_{1|2,3} f_2 f_{3|5,6} f_{4|6} f_5 f_6$$

implies that $f_{ij} = f_i f_j$ by intergating over all remaining variables.

The edge matrix of an overall ancestor graph, $\mathcal{B}$, is the indicator matrix of

$$I + \sum (\mathcal{A} - I)^r,$$

where I is the identity matrix and $(\mathcal{A} - I)^r$ counts for each $i < j$ the number of direction-preserving paths of length r present in the parent graph between them.

A matrix product BB^T has in position (i,j) the element $b_{ij} + \sum_{k>j} b_{ik} b_{jk}$. Therefore, if $\mathrm{In}(\mathcal{B}\mathcal{B}^T)$ is the induced the edge matrix, then there is an additional ij-one if and only if nonadjacent nodes i and j have a node k as a common parent in the ancestor graph. Thus, an additional ij-edge in G^V_{cov} compared to G^V_{par} arises if and only if either j is an ancestor of i or i and j have a common ancestor. This statement is equivalent to Pearl's (1988) separation criterion for directed acyclic graphs when the conditioning set is empty. The matrix result completes the search for the proper paths for all pairs at once.

By a similar simpler argument the induced overall concentration graph, G^V_{con}, can be shown to have edge matrix $\mathrm{In}(\mathcal{A}^T\mathcal{A})$. It is an undirected graph of full lines, where each edge concerns the conditional relation of two variables given all remaining ones, $i \perp\!\!\!\perp j \mid V \setminus \{i,j\}$. The name derives from joint Gaussian distribu-

tions for which these independencies are reflected as zeros in the concentration matrix which is the inverse of the covariance matrix.

4 Relations to traditional methods and case studies

In the social sciences structural equation models (SEM) and, more generally, linear structural relation models (Bollen, 1989) have been used extensively for analysing multivariate data. They have been developed as extensions of path analysis models (Wright, 1934), which in the econometric literature are better known as linear recursive equations with uncorrelated residuals (Goldberger, 1964). Graphical Markov models provide a different extension in which both categorical and numerical features can be modelled. In the subclass of chain graph models joint distributions are decomposed recursively into conditional joint distributions and simplified by conditional independencies. There are no theoretical restrictions on the form of the conditional distributions, however algorithms for computing estimates under each specified model are not yet available generally.

It has recently been shown (Koster, 1999) how an independence graph can be associated with each Gaussian structural equation model to read off the graph all independence statements implied by the model. But, while in chain graphs every missing edge corresponds to an independence statement and every edge present can be associated with a specific conditional or marginal association of the variable pair, this does not hold in general for structural equation models. An edge present in the graph relates directly to a parameter in an equation but may be connected in a complicated way to any statement about the conditional or marginal association of the variable pair. Similarly, a variable pair with an edge missing may be associated no matter which conditioning set is chosen.

Results about fitting chain graphs approximately with the help of univariate conditional regressions and results for deriving chain graphs induced by directed acyclic graphs (Wermuth and Cox, 2001a) can be viewed as supplementing existing, useful data analysis tools. Local fitting of univariate conditional distributions permits to break up seemingly complex structures into tractable subcomponents. These components may then be directly related to substantive knowledge available about subsets of the variables under study.

Some case studies using chain graph models are by Klein, Keiding, and Kreiner (1995), Hardt (1995), Cox and Wermuth (1993; 1996, Chapter 6; 2001), Pigeot, Caputo, and Heinicke (1999), Stanghellini, McConway and Hand (1999), Pigeot, Heinicke, Caputo and Brüderl (2000), Wermuth and Cox (2002), and Cheung and Andersen (2002).

Acknowledgement

Support by the HSSS-program of the European Science Foundation and the Radcliffe Institute for Advanced Study at Harvard University is gratefully acknowledged.

References

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

Cheung, S.Y. & Andersen, R. (2002). Time to Read: Family Resources and Educational Inequalities, *Journal of Comparative Family Studies.* To appear.

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

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

Cox, D.R. & Wermuth, N. (1996). *Multivariate Dependencies – Models, Analysis and Interpretation.* London: Chapman & Hall.

Cox, D.R. & Wermuth, N. (2001) Some statistical aspects of causality. *European Sociological Review*, **17**, 65-74.

Edwards, D. (2000) Introduction to graphical modelling. New York: Springer.

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

Hardt, J. (1995) *Chronifizierung und Bewältigung bei Schmerzen.* Lengerich: Pabst.

Klein, J.P., Keiding, N. & Kreiner, S. (1995). Graphical models for panel studies, illustrated on data from the Framingham heart study. Statistics in Medicine 14, 1265-1290.

Koster, J. (1999). On the validity of the Markov interpretation of path diagrams of Gaussian structural equation systems of simultaneous equations. *Scand. J. Statist.*, **26**, 413-431.

Lauritzen, S.L. (1996). *Graphical models.* Oxford: Oxford University Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* San Mateo, CA: Morgan Kaufmann.

Pigeot, I., Caputo, A. & Heinicke, A. (1999) A graphical chain model derived from a model selection strategy for the sociologists graduates study. *Biometrical Journal*, **41**, 217-234.

Pigeot, I., Heinicke, A., Caputo, A. & J. Brüderl, J. (2000) The professional career of sociologists: a graphical chain model reflecting early influences and associations. Allgemeines Statistisches Archiv, **84**, 3-21.

Stanghellini, E., McConway, K.J. & Hand, D.J. (1999), A discrete variable chain graph for applicants for credit, Journal of the Royal Statistical Society , Series C, Applied Statistics, 48, 239-251.

Wermuth, N. & Cox, D.R. (2001a). Joint response graphs and separation induced by triangular systems. Research Report, Australian National University. http://www.maths.anu.edu.au/research.reports/01srr.html

Wermuth, N. & Cox, D.R. (2002) Graphical models: an overview. *Encyclopedia of Behavioral Sciences.* Elesevier. To appear.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, **5**, 161-215.