GRAPHICAL MARKOV MODELS

Nanny Wermuth [1]

Professor of Statistics, Department of Mathematical Sciences
Chalmers Technical University/University of Gothenburg, Sweden

Graphical Markov models are multivariate statistical models which are currently under vigorous development and which combine two simple but most powerful notions, generating processes in single and joint response variables and conditional independences captured by graphs. The development of graphical Markov started with work by Wermuth (1976, 1980) and Darroch, Lauritzen and Speed (1980) which built on early results in 1920 to 1930 by geneticist Sewall Wright and probabilist Andrej Markov as well as on results for log-linear models by Birch (1963), Goodman (1970), Bishop, Fienberg and Holland (1973) and for covariance selection by Dempster (1972).

Wright used graphs, in which nodes represent variables and arrows indicate linear dependence, to describe hypotheses about stepwise processes in single responses that could have generated his data. He developed a method, called path analysis, to estimate linear dependences and to judge whether the hypotheses are well compatible with his data which he summarized in terms of simple and partial correlations. With this approach he was far ahead of his time, since corresponding formal statistical methods for estimation and tests of goodness of fit were developed much later and graphs that capture independences even much later than tests of goodness of fit.

It remains a primary objective of graphical Markov models to uncover graphical representations that lead to an understanding of data generating processes. Such processes are no longer restricted to linear relations but contain linear dependences as special cases. A probabilistic data generating process is a recursive sequence of conditional distributions in which response variables may be vector variables that contain discrete or continuous components. Thereby, each conditional distribution specifies both the dependence of response Y_a, say, on an explanatory variable vector Y_b and the undirected associations of the components of Y_a.

Graphical Markov models also generalize sequences in single responses and single explanatory variables that have been named Markov chains, after probabilist Markov. He recognized at the beginning of the 29th century that seemingly complex joint probability distributions may be radically simplified by using the notion of conditional independence and defined what are now called Markov chains.

In a Markov chain of random variables $Y_1, \ldots, Y_i, \ldots, Y_d$, the joint distribution is built up by starting with the density of f_d of $Y_d$and generating next $f_{d-1|d}$. Then, conditional independence of Y_{d-2} from Y_d given Y_{d-1} is taken into account with $f_{d-2|d-1,d} = f_{d-2|d-1}$. One continues such that, with $f_{i|i+1,\ldots d} = f_{i|i+1}$, response Y_i is conditionally independent of $Y_{i+2}, \ldots, Y_d$ given Y_{i+1}, written compactly

[1]For biography *see* the entry **Multivariate statistical analysis**.

in terms of nodes as $i \perp\!\!\!\perp \{i+2, \ldots, d\} | i+1$, and finally with $f_{1|2,\ldots,d} = f_{1|2}$ having just Y_2 as explanatory variable of response Y_1.

The directed graph that captures such a Markov chain is a single directed path of arrows. Thus, for $d = 5$ and node set $N = \{1, 2, 3, 4, 5\}$, the graph is

$$1 \prec\!\!\!- 2 \prec\!\!\!- 3 \prec\!\!\!- 4 \prec\!\!\!- 5.$$

The graph corresponds to the factorization of the joint density f_N given by

$$f_N = f_{1|2} f_{2|3} f_{3|4} f_{4|5} f_5.$$

The three defining local independence statements given directly by the above factorization or by the corresponding path of dependences are $1 \perp\!\!\!\perp \{2, 3, 4, 5\} | 2$, $2 \perp\!\!\!\perp \{4, 5\} | 3$ and $3 \perp\!\!\!\perp 5 | 4$. One also says that in the generating process, each response Y_i remembers of its past just the nearest past variable Y_{i+1}.

Directed acyclic graph are the most direct generalization of Markov chains. They have an ordered sequence of single nodes representing responses that may generate f_N, but each response may remember any subset or all of the variables in its past. Directed acyclic graphs are known as Bayesian networks when the node set does not only consist of random variables that correspond to varying features of observable units, but may include nodes for decisions or parameters.

It remains an important secondary objective of graphical Markov models to capture the independence structure of f_N by some type of graph. This is the set of all independence statements implied by the given graph and satisfied by f_V. In principle, all independence statements that arise from a given set of statements defining a graph, may be derived from basic laws of probability. Thus, the above Markov chain implies for instance

$$1 \perp\!\!\!\perp 4 | 3, \quad \{1, 2\} \perp\!\!\!\perp \{4, 5\} | 3, \quad \text{or} \quad 2 \perp\!\!\!\perp 4 | \{1, 3, 5\}.$$

But for many variables, methods defined for graphs simplify considerably the task of deciding whether an independence statement is implied by a given set of independence statements. These are called separation criteria; see Geiger, Verma and Pearl (1990), Lauritzen et al. (1990) and Marchetti and Wermuth (2009) for different but equivalent criteria on directed acyclic graphs.

For ordered sequences of vector variables, the graphs are directed acyclic in blocks which contain the joint responses. The undirected association of any two individual components of a response vector is represented by some type of line coupling two nodes. Thus, these types of graph contain undirected and directed edges but at most one edge for a node pair.

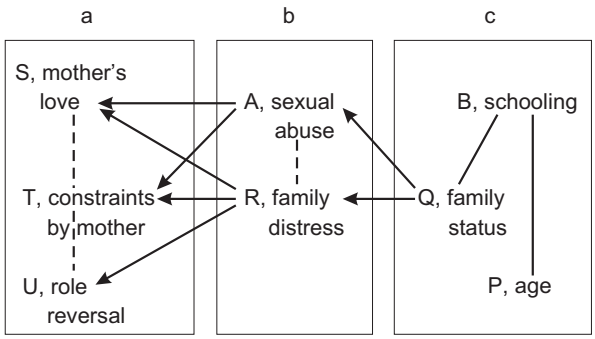
Four different types of such chain graphs for discrete variables have been classified and studied by Drton (2009), extending the three types that had been discussed before; see e.g. Wermuth and Cox (2004). Drton proves that two types have the desirable property that each model in the given class defines a curved exponential family; see e.g. Cox (2007) for the latter concept. The property holds for the blocked concentration graphs of Lauritzen and Wermuth (1989) and for the multivariate regression chain graphs of Cox and Wermuth (1993) provided the

latter joint distribution has some properties that it shares with a joint Gaussian distribution; see Wermuth (2010). In early books by statisticians on graphical Markov models, only blocked concentration graphs are discussed; see Edwards (2000), Lauritzen (1996), Whittaker (1990), an exception is Cox and Wermuth (1996).

The main difference among the four types of chain graph is the independence interpretation of missing edges. For undirected edges, the blocked concentration graphs and the chain graphs by Anderson, Madigan and Perlman (2001) use concentration graphs in which a missing ij-edge means $i \perp\!\!\!\perp j$ given all variables in the past and all remaining variables within the same block. Multivariate regression chain graphs and those named type III by Drton use covariance graphs in which a missing ij-edge means $i \perp\!\!\!\perp j$ given all variables in the past; see Marchetti and Lupparelli (2010) for parametrizations in terms of the multivariate logistic regressions which lead to the equivalence of the simple pairwise independences to the more complex defining independences given by Drton. The names remind one of the corresponding vanishing parameters joint Gaussian distributions, where the inverse of the covariance matrix is called the concentration matrix.

For the directed edges a missing ij-arrow, with i denoting the response at the arrowhead, means $i \perp\!\!\!\perp j$ given all remaining variables in the past of the block containing node i in both the multivariate regression chain graphs and those named type III by Drton, while the conditioning set includes in addition other components of the block containing node i. Thus, it is only in multivariate regression chains, that the conditional independence constraints defining the graph respect a given order of the vector variables. It can be shown that the separation criterion for multivariate regression chains do not change when a concentration graph is added as a last chain component.

The following small example of a well-fitting multivariate regression chain is for a set of data of Jochen Hardt, University of Mainz, on $n = 283$ adult, healthy females who agreed to be interviewed about different aspects of their childhood. Variables A, B are binary, the others are based on quantitative measurements. Each of Y_a and Y_c have three component variables and Y_b has two.



The graph is constructed after checking for nonlinear and interactive effects by using the results of a sequence of linear and logistic regressions. These show that

the estimated dependencies, not displayed here, are in the direction hypothesized by the researchers and that the background variable Y_c does not improve prediction of Y_a given the more specific information about childhood of Y_b

The resulting factorization is $f_N = f_{a|b}f_{b|c}f_{c\cdot}$. The independences defining the multivariate regression chain graph are $S \perp\!\!\!\perp U|\{a,b\}$, $a \perp\!\!\!\perp c|b$ and $Q \perp\!\!\!\perp P|B$, where relations within a are modeled using a covariance graph, those within b using a concentration graph.

An important feature of multivariate regression chains is that they can be used to formulate hypotheses on development in joint responses but that the goodness-of-fit of a model to data can be well judged in terms of univariate regressions, provided that either the last block with a concentration graph is missing or that this graph is triangulated that is without any chordless cycles of size four or larger as in the example above.

The outstanding feature of multivariate regression chains is that consequences of a given family of densities f_N can be derived when marginalizing over some variables, in set M and conditioning on others, in set C. In particular, graphs can be obtained for node set $N' = N \setminus \{C, M\}$ which capture precisely the independence structure implied by a generating graph in node set N for $f_{N'|C}$ the family of densities of Y'_N given Y_C. Such graphs are named independence-preserving, when they can be used to derive the independence structure that would have resulted from the generating graph by conditioning on a larger node set $\{C, c\}$ or by marginalising over a larger node set $\{M, m\}$.

From a given generating graph and by using the same sets C, M three corresponding graphs result. These are in a subclass of the much larger class of MC-graphs of Koster (2002), a maximal ancestral graph (MAGs) of Richardson and Spirtes (2002) and a summary graph of Wermuth (2010); see Sadeghi (2009) for a proof of Markov equivalence that is for showing that the three corresponding, but different types of graph capture the same independence structure.

To derive consequences of multivariate regression chains in f_N not only for independences but also for conditional dependences in f'_N, the generating family of joint densities f_N has to share some properties with the family of joint Gaussian distributions. These result with specific, but not very restrictive requirements for the generating process; see Wermuth (2010). The results builds on previous discussions of such special properties; see Dawid (1979), Lauritzen (1996), Studený (2005), Kang and Tian (2009), San Martin, Mochart and Rolin (2005), Wermuth and Cox (2004) and proofs use properties of two corresponding matrix operators, one for transforming Gaussian parameter matrices and one for transforming matrix representations of graphs; see Wermuth, Wiedenbeck and Cox (2006).

In that case, the summary graph shows when a generating conditional dependence of Y_i on Y_k, say, in f_N remains undistorted in $f_{N'|C}$, parametrized in terms of conditional dependences and when it be may become severely distorted; see also Wermuth and Cox (2008). Some of the distortions may also occur in randomized intervention studies but can be avoided by changing C or M. Thus, these results are relevant for controlled clinical trials, for comparing or combining results from

different studies on a core set of variables and, more generally, for the planning stage of follow-up studies designed to replicate some results of a given larger study by using a subset of the variables and studying a subpopulation.

In the near future, more results on estimation and goodness of fit tests are expected; see also Drton, Eichler and Richardson (2009), Cox (2007) and more discussions of causal interpretations; see Cox and Wermuth (2004), Pearl (2009). Comparative evaluations will be needed of alternative computational methods that are in use now for very large sets of data; see e.g. Edwards, deAbreu and Labouriau (2010), Dobra (2009), Wang and Leng (2007).

REFERENCES

[1] Andersson, S.A., Madigan, D., Perlman, M.D. (2001). Alternative Markov properties for chain graphs. *Scand. J. Statist.* **28**, 33–85.

[2] Birch, M.W. (1963) Maximum likelihood in three-way contingency tables. *J. Roy. Statist. Soc. B* **25**, 220–233.

[3] Bishop, Y.M.M., Fienberg, S.F. and Holland, P.W. (1975). *Discrete multivariate analysis*. MIT Press. Cambridge.

[4] Cox, D.R. (2007). *Principles of statistical inference*. Cambridge University Press. Cambridge.

[5] Cox, D.R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statist. Science*, **8**, 204–218; 247–277.

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

[7] Cox, D.R. and Wermuth, N. (2004). Causality: a statistical view. *Internat. Statist. Review* **72**, 285–305.

[8] Darroch, J.N., Lauritzen, S.L. and Speed, T.P. (1980) Markov fields and log-linear models for contingency tables. *Ann. Statist.* **8**, 522–539.

[9] Dawid (1979). Some misleading arguments involving conditional independence *J. Roy. Statist. Soc. B*, **41**, 249–252.

[10] Dempster, A.P. (1972) Covariance selection *Biometrics* **28**, 157–175.

[11] Dobra, A. Variable selection and dependency networks for genomewide data. *Biostatistics* **10**, 621–639.

[12] Drton, M. (2009). Discrete chain graph models. *Bernoulli* **15**, 736–753.

[13] Drton, M., Eichler, M. and Richardson, T.S. (2009). Computing maximum likelihood estimates in recursive linear models. *J. Mach. Learn. Res.* **10**, 2329–2348.

[14] Edwards, D. (2000). *Introduction to graphical modelling.* 2nd ed. Springer, New York.

[15] Edwards, D., de Abreu, G.C.G, and Labouriau, R. (2010). Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. *BMC Bioinformatics 2010*, 11–18.

[16] Geiger, D., Verma, T.S. and Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks* **20**, 507–534.

[17] Goodman, L.A.(1970) The multivariate analysis of qualitative data: interaction among multiple classifications. *J. Amer. Statist. Assoc.* **65**, 226–256.

[18] Kang C. and Tian, J. (2009) Markov properties for linear causal models with correlated errors. *J. Mach. Learn. Res.* **10**, 41–70.

[19] Koster (2002). Marginalising and conditioning in graphical models. *Bernoulli* **8**, 817–840.

[20] Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press, Oxford.

[21] Lauritzen, S.L., Dawid, A.P., Larsen, B. and Leimer, H.G. (1990). Independence properties of directed Markov fields. *Networks* **20**, 491–505.

[22] 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.

[23] Marchetti, G.M. and Lupparelli, M. (2010). Chain graph models of multivariate regression type for categorical data. Submitted and available on ArXiv, http://arxiv.org/abs/0906.2098v2.

[24] Marchetti, G.M. and Wermuth, N. (2009). Matrix representations and independencies in directed acyclic graphs. *Ann. Statist.* **47**, 961–978.

[25] Pearl, J. (2009) Causal inference in statistics: An overview *Stat. Surveys* **3**, 96–146.

[26] Richardson, T.S. and Spirtes, P. (2002). Ancestral Markov graphical models. *Ann. Statist.* **30**, 962–1030.

[27] Sadeghi, K. (2009). Representing modified independence structures. *Transfer thesis*, Oxford University.

[28] San Martin E., Mochart M. and Rolin, J.M. (2005). Ignorable common information, null sets and Basu's first theorem. *Sankya* **67**, 674–698.

[29] Studený, M. (2005). *Probabilistic conditional independence structures*. Springer, London.

[30] Wang, H. and Leng, C. (2007). Unified LASSO estimation via least squares approximation. *J. Amer. Statist. Assoc.* **102**, 1039–1048.

[31] Wermuth, N. (1976). Analogies between multiplicative models for contingency tables and covariance selection. *Biometrics* **32**, 95–108.

[32] Wermuth, N. (1980). Linear recursive equations, covariance selection, and path analysis. *J. Amer. Statist. Assoc.* **75**, 963–97.

[33] Wermuth, N. (2010). Probability distributions with summary graph structure. Submitted and available in ArXiv, http://arxiv.org/abs/1003.3259.

[34] Wermuth, N. and Cox, D.R. (2004). Joint response graphs and separation induced by triangular systems. *J.R. Stat. Soc. Ser. B Stat. Methodol.* **66**, 687-717.

[35] Wermuth, N. and Cox, D.R. (2008). Distortions of effects caused by indirect confounding. *Biometrika* **95**, 17–33.

[36] Wermuth, N., Wiedenbeck, M. and Cox, D.R. (2006). Partial inversion for linear systems and partial closure of independence graphs. BIT, Numerical Mathematics, **46**, 883–901.

[37] Whittaker, J. (1990). *Graphical models in applied multivariate statistics.* Chichester: Wiley.