In: Proceedings of the 55th Session of the International Statistical Institute, Sidney 2005.

On implications of recent results on graphical Markov models for the design and analysis of observational studies

Nanny Wermuth
Department of Mathematical Statistics, Chalmers/Gothenburg University

41296 Göteborg, Sweden
wermuth@math.chalmers.se

Graphical Markov models represent relations among random variables by combining simple yet powerful concepts: data generating processes, graphs and conditional independence. The origins can be traced back to independent work in genetics (S. Wright, 1921, [35]), in physics (W. Gibbs, 1902, [12]) and in probability theory (A. A. Markov, 1912, [22]). Wright used directed graphs to describe processes of how his genetic data could have been generated and to check consistency of such hypotheses with observed data. He called his method path analysis. Gibbs described total energy of systems of particles by the number of nearest neighbors for nodes in undirected graphs. Markov suggested how some seemingly complex structures can sometimes be explained in terms of a chain of simple dependencies using the notion of conditional independence.

A systematic development of graphical Markov models for representing multivariate statistical dependencies for both discrete and continuous variables started in the 1970's with work on fully undirected graph models for purely discrete and for Gaussian random variables and on linear models with graphs that are fully directed and have no cycles. This work was extended to models permitting sequences of response variables to be considered on equal footing, that is without specifications of a direction of dependence. Joint responses can be modeled in quite different ways, some define independence structures of distinct types of graphical chain model. Properties of corresponding types of graph have been studied intensively, so that, in particular, all independencies, implied by a given graph, can be derived by so-called separation criteria.

Several books give overviews of theory, analyses and interpretations of graphical Markov models in statistics, based on developments on this work during the first few decades, see [9], [17], [4], [34], and a wide range of different applications has been reported, see e.g. [13], [18]. For some compact descriptions and for references see [29], [30]. Applicability of directed acyclic graph (DAG) models to very large systems of units has been emphasized, see e.g. recently [8], and is simplified by free source computational tools within the framework of the R-project, see [21], [20], [3]. For deriving well-fitting Gaussian models a procedure distinguishing between significant, indeterminate and nonsignificant associations has been proposed [2] and for densities of arbitrary form compatible priors for DAG models have been derived [25].

Special extensions to time series have been developed ([7],[10],[11]) and relations to structural equation models (SEM) have been discussed with respect to independencies [15], and the interpretation of parameters [32]. Models which preserve exactly the independencies of the generating process after omitting some variables and conditioning on others form a slightly extended subclass of SEM models [24], [16]. Extensions to point processes and to multilevel models are work in progress. Graphical criteria for deciding on the identifiability of special linear models including hidden variables have been derived [26], [23], [28], [14], [27] and for proving identification of a large subclass of SEM models in only observed variables [1], [32].

One approach to studying properties and consequences of graphical Markov models is based on binary matrix forms of graphs [31]. This uses analogies between partial inversion of parameter matrices for linear systems and partial closing of directed and of undirected paths in graphs [33]. The starting point for this are stepwise generating processes either for systems of linear equations or for joint distributions.

In both cases the generating graph consists of a set of nodes, with node i representing random variable Y_i and a set of directed edges. Each edge is drawn as an arrow outgoing from what is called a parent node and pointing to an offspring node. The graph is acyclic if it is impossible to return to any starting node by following arrows pointing in the same direction. The set of parent nodes of node i is denoted by par_i and the graph is called a parent graph if there is a complete ordering of the variables as $(Y_1, \ldots, Y_d)$. Either a joint density is given by a recursive sequence of univariate conditional densities or a covariance matrix is generated by a system of recursive equations.

The joint density f_N, generated over a parent graph with nodes $N = (1, \ldots, d)$ and written in a compact notation for conditional densities in terms of nodes, is

$$(1) \qquad f_N = \prod_i f_{i|i+1,\ldots,d} = \prod_i f_{i|\mathrm{par}_i}.$$

The conditional independence statement $i \perp\!\!\!\perp j|\mathrm{par}_i$ is equivalent to the factorization $f_{i|\mathrm{par}_i,j} = f_{i|\mathrm{par}_i}$ and it is represented by a missing ij-arrow in the parent graph for $i < j$.

The joint covariance matrix Σ of mean-centered and continuous variables Y_i, generated over a parent graph with nodes $N = (1, \ldots, d)$, is given by a system of linear recursive equations with uncorrelated residuals, written as

$$(2) \qquad AY = \varepsilon,$$

where A is an upper-triangular matrix with unit diagonal elements and ε is a residual vector of zero-mean uncorrelated random variables ε. A diagonal form of the residual covariance matrix $\mathrm{cov}(\varepsilon) = \Delta$ is equivalent to specifying that each row of A in (2) defines a linear least squares regression equation ([6], p. 302) for response Y_i regressed on $Y_{i+1}, \ldots, Y_d$. The vanishing contribution of Y_j to the linear regression of Y_i on $Y_{i+1}, \ldots, Y_d$ is represented by zero value in position (i, j) in the upper triangular part of A with corresponding direct consequences for $\Sigma^{-1} = A^{\mathrm{T}}\Delta^{-1}A$.

Sequences of joint responses occur in different types of chain graphs. All these chain graphs have in common that the nodes are arranged in a sequence of say d_{CC} chain components g, each containing one or more nodes. For partially ordered nodes $N = (1, \ldots, g, \ldots, d_{CC})$ a joint density is considered in the form

$$(3) \qquad f_N = \prod_g f_{g|g+1,\ldots,d_{CC}}.$$

The types of question that can be answered now for these types of joint response models induced by a stepwise generating process are: Which independencies (either linear or probabilistic) are preserved if the ordering the variables is modified or if some of the variables are considered as joint instead of univariate responses or if some of variables are explicitly omitted or if a subpopulation is selected? Which of the associations contained in a new parametrization are merely induced? [31]. Which types of confounding may occur and, can one correct for them at least in linear systems? [32].

Stepwise generating processes in univariate responses arise both in observational and in intervention studies. With an intervention the probability distribution is changed so that the intervening variable is decoupled from all variables in the past that relate directly to it in an observational setting, see [19]. Two main assumptions distinguish "causal models with potential outcomes" (or counterfactual models) from general generating processes in univariate responses. These are (1) unit-treatment additivity and (2) a notional intervention. These two assumptions taken together assure that there are no unobserved confounders and that there is no interactive effect on the response by an unobserved variable and the intervening variable. One consequence of these assumptions for linear models is that the effect of the intervening variable on the response averaged over past variables coincides with its conditional effects given past unobserved variables. Different definitions of causality have recently been compared from a statistical viewpoint [5].

As more results become available on independence equivalence and on parameter equivalence of different models, on identification of latent variable models, on different types of confounding and different types of selection bias, the more likely it becomes that we can better design and draw conclusions from observational studies with many variables.

REFERENCES

[1] Brito, C. & Pearl, J. (2002). A new identification condition for recursive models with correlated errors *Structural equation modeling*, **9 (4)**, 459-474.

[2] Drton, M. & Perlman, M. (2005). A SINful approach to Gaussian graphical model selection. *Statistical Science*. To appear.

[3] Badsberg, J.H. (2004). DynamicGraph: interactive graphical tool for manipulationg graphs. URL: http://cran.r-project.org.

[4] Cox, D. R. & Wermuth, N. (1996). *Multivariate dependencies: models, analysis, and interpretation*. Chapman and Hall, London.

[5] Cox, D.R. & Wermuth, N. (2004). Causality a statistical view. *Int. Statist. Rev.*, **72**, 285–305.

[6] Cramér, H. (1946). *Mathematical methods of statistics*. Princeton, N.J.: Princeton University Press.

[7] Dahlhaus, R. (2000). Graphical interaction models for multivariate time series. *Metrika*, 157–172.

[8] Dobra, A. (2003). Markov bases for decomposable graphical models. *Bernoulli*, **9**, 1093–1108.

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

[10] Eichler, M, Dahlhaus R. & Sandkühler J. (2003). *Partial correlation analysis for the identification of synaptic connections*. Biological Cybernetics. **89**, 289-302.

[11] Fried R. & Didelez, V. (2003). Decomposability and selection of graphical models for time series. *Biometrika*. **90**, 251-267.

[12] Gibbs, W. (1902). *Elementary Principles of Statistical Mechanics*. Yale Univ. Press, New Haven.

[13] Green, P.J., Hjort, N.L. & Richardson, S. (2003). *Highly Structured Stochastic Systems*. Oxford: University Press.

[14] Grzebyk M. & Wild, P. & Chouaniére, D. (2003). On identification of multi-factor models with correlated residuals. *Biometrika*. **91**, 141-151.

[15] Koster, J.T.A. (1999). On the validity of the Markov interpretation of path diagrams of Gaussian structural equations systems with correlated errors. *Scand. J. Statist.*, **26**, 413–431.

[16] Koster, J.T.A. (2002). Marginalizing and conditioning in graphical models. *Bernoulli*, **8**, 817–840.

[17] Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press, Oxford.

[18] Lauritzen S.L. & N. A. Sheehan (2003). Graphical models for genetic analyses. *Statistical Science*, 18, 489–514.

[19] Lindley, D.V. (2002). Seeing and doing: the concept of causation. *Int. Statist. Rev.* **70**, 191-214.

[20] Marchetti, G. M. (2004), R functions for computing graphs induced from a DAG after marginalization and conditioning. 2004 Proceedings of the American Statistical Association, Statistical Computing Section [CD-ROM], Alexandria: VA American Statistical Association.

[21] Marchetti, G. M. & Drton, M. (2003). GGM: an R package for Gaussian graphical models. URL: http://cran.r-project.org.

[22] Markov, A.A. (1912). *Wahrscheinlichkeitsrechnung* (German translation of 2nd Russian edition: Markoff, A.A., 1908). Teubner, Leipzig.

[23] Pearl J. (1998). Graph, causality and structural equation models. *Sociological Methods and Research* **27**, 226-284.

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

[25] Roverato, A. & Consonni, G (2004). Compatible prior distributions for DAG models. *J. Roy. Statist. Soc. B.* **66**, 17-61.

[26] Stanghellini, E. (1997). Identification of a single-factor model using graphical Gaussian rules. *Biometrika*, **84**. 241-254.

[27] Stanghellini, E. & Wermuth, N. (2004). On the identification of path analysis models with one hidden variable. *Biometrika.* To appear.

[28] Vicard, P. (2000). On the identification of a single-factor model with correlated residuals. *Biometrika*, **84**. 241-254.

[29] Wermuth, N. (1998). Graphical Markov models. *Encyclopedia of Statistical Sciences.* S. Kotz, C. Read and D. Banks (eds). Wiley, New York, Second Update Volume, 284-300.

[30] Wermuth, N. & Cox, D.R. (2001). Graphical models: overview. In: *International Encyclopedia of the Social and Behavioral Sciences* P.B. Baltes and N.J. Smelser (eds), Elsevier, Amsterdam, **9**, 6379–86.

[31] Wermuth, N. & Cox, D.R. (2004). Joint response graphs and separation induced by triangular systems. *J. Roy. Statist. Soc. B.* **66**, 687-717.

[32] Wermuth, N. & Cox, D.R. (2004a). Correcting for indirect confounding in linear systems. Submitted manuscript.

[33] Wermuth, N., Cox, D.R. & Wiedenbeck, M. (2004) On partial inversion and closing paths in graphs, Submitted manuscript.

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

[35] Wright, S. (1921). Correlation and causation. *J. Agric. Res.* **20**, 162–177.