Jan de Leeuw

Departments of Mathematics and Psychology
University of California
Los Angeles - U.S.A.

Graphical Models, etc.

The theory of hierarchical models, decomposable models, graphical models, chain models, or path models, has changed, and will continue to change, the face of multivariate statistical analysis. The accomplishments so far have been impressive. There are important and convincing systematizations of early work by Wright, Goodman, and Dempster in the writings of Wermuth, Speed, Lauritzen, and Cox. The contributions of Pearl, Shafer, Glymour and others provide rich connections with the field of artificial intelligence, and with computational philosophy of science.

Of course such a glowing introduction must necessarily be followed by words of criticism and/or caution. We concentrate on the practical usefulness of graphical model techniques in actual data analysis.

Weak Spots

The theory developed so far is not complete. First, the connections with econometric simultaneous equation theory, and with the related LISREL/EQS systems in psychometrics and sociometrics, have not been worked out in detail. Graphical chain models correspond with (block-)recursive path models with uncorrelated errors. The errors are the only latent (unobserved) variables. In the simultaneous equations models in the Haavelmo tradition there is no longer simple recursiveness. Wermuth's paper, the one we are discussing here, makes an attempt to close this first gap. Her point of view is that block recursive regression equations are the key to understanding the relationship between the approaches. She only discusses "errors-in-equations" models, and we shall argue that these models are not the most interesting generalization of simple path models. It also seems to me that quite a few of the attractive properties of block models get lost in the generalization. This means that the theory is not very robust.

Second, one of the satisfactory aspects of these models is that categorical and numerical variables can be dealt with in one and the same formulation. But this can only be done by assuming multivariate normality for the numerical variables, and then translating conditional independence into vanishing partial correlations; Although parts of the theory can be based on partial correlations only, we lose the connection with maximum likelihood, which requires a full specification of the model. Unfortunately, real variables are not either multinomial or multinormal. They are most often, in my experience, somewhere in between these extremes. Multinomial allows for too many parameters, multinormal for not enough parameters. We shall come back to this below.

Thirdly, there is a gap between theory and practice. In quantitative genetics, in econometrics, in educational science, and in sociology, path models and simultaneous equation models have been, let's say, disappointing. The collective works of David Freedman illustrate this more forcefully than I ever could. Having graphical models is technologically a big step ahead, but we should be careful not to fall into the LISREL trap. A generation of social scientists has been misguided by an appealing metaphor, wrapped in an authoritarian black box, decorated with forbidding equations. As far as I can see, the outcome has been, let's say, disappointing.

Linear Structural Equations

The material in the first four parts of Wermuth's paper is an excellent and useful review of the various facts and results that are known about block recursive path models. The fifth part is an original comparison of block recursive models with general simultaneous or structural equation models. This part we shall discuss somewhat more in detail.

In the full model, not all parameters can be identified from the first and second order moments. Wermuth (p.18) seems to equate "not identifiable" with "not defining a statistical model". But this is misleading, because the model defines a unique manifold in the space of covariance matrices, even if this manifold is not described with a minimal number of parameters. Here, as in the linear case, it helps to think of the parametrization as just incidental, and of the subset of the covariance cone as the model we are trying to estimate. This also takes away the fear that identification could lead to misspecification. If it does, it is by definition not just identification. Identifying a model means describing it in terms of the smallest number of parameters possible. Also, giving the impression that identification problems come from overparametrization (p.22) is somewhat misleading. They tend to come from underparametrization, of course.

I am not an econometrician, but were I one, I would be somewhat unhappy about Wermuth's account of simultaneous equation modeling. The modern econometric literature on identification is not even mentioned. Hsiao (1983) gives a fairly recent overview, while even more recent material is reviewed in Bekker and Pollock (1986) or Bekker and Dijkstra (1990). There are now symbolic computation methods to investigate identification and equivalence of simultaneous equation models.

Block recursive models were already used as basis for discussing indentification by Fisher in his classical book (1966). There have also been interesting discussions in the econometric literature (between Wold and Bassman, for instance) about causal interpretation of coefficients in non-recursive models, as compared to block-recursive models. This anticipates some of the distinctions mentioned on page 22-23. While it is true the LISREL framework has been accepted quite uncritically by many social scientists, it is also true there has been a great deal of critical discussion of the algebraic and methodological properties of simultaneous equation models in the econometric research literature. This is not obvious from Wermuth's paper, because she mainly refers to some excellent, but quite ancient, textbooks.

Latent Variables

Wermuth does not mention latent variables, but it is difficult to see how any discussion of linear structural models can be complete without them. In the situations in which such models are typically applied (with some exceptions in theoretical population genetics) there are errors in variables that cannot be ignored. This is even true in engineering and systems theory, where there currently is a lot of interest for latent variable models. I happen to think that factor analysis, with the closely related models of true score theory, errors in variable theory, and latent trait theory, are the most interesting contributions of the social sciences to data analysis. The fact that these techniques have been misused almost to extinction does not take away this basic fact.

Related to the idea of latent variables is the idea of optimal scaling. A latent variable only exists because of its position in the path diagram. In factor analysis, for intance, we say that there exists a variable z such that the observed y_j are independent given z. We do not observe z, it is missing, but we can still test the consequences of our assumptions. If we want to, we can also "estimate" z.

One step further along the road is the basic indicator idea, which is that we never observe the variables we are interested in. The variables in the path model are theoretical constructs, we only observe indicators for these constructs. The indicators are related to the constructs by small factor models. This is the seductive metaphor I referred to earlier. It seems the only sensible way to model errors-in-variables, and potentially it is a great way to link theory and observation. Unfortunately, in many applications the theory component is missing, and theories are "constructed" by heuristic search over path models. This is the nightmarish part of latent variable modeling. It is a reasonable class of techniques, much more reasonable than block recursive modeling of the observed variables, but I have not seen many convincing applications, and I have seen quite a few hair-raising ones.

Nevertheless I agree with Goldberger (1972), Griliches (1974), Wold (1982), and Aigner et al. (1984) that there is no way around errors-in-variables, and that the errors-in-equations models in the Haavelmo tradition are simply too narrow to be of interest in most social science situations. It is perhaps true, that block-recursive models are more basic and "a key to understanting". They are elegant, they are simple, but they have little practical relevance because of the omnipresence of measurement error. In a sense this means that the vanishing tetrad is of more importance than the vanishing partial correlation coefficient.

Between Multinomial and Multinormal

Relationships between variables can be pictured as arrow diagrams. These qualitative diagrams can be translated into quantitative statements about the joint distribution of the variables in various ways. One translation uses conditional independence, another uses vanishing partial correlations.

In the path diagrams some variables are quantitative, some are qualitative, some are latent and some are manifest. In a number of special cases we know how to integrate all four types of variables into a single model, and large steps have been made towards a general approach. I think, however, that in order to build realistic models we need intermediate types of variables. Intermediate both between continuous normal and discrete multinomial, but also intermediate between latent and manifest.

Some of these intermediate types have been studied in probit and logit models, in which binary variables are regressed on continuous latent variables for which they are indicators. If we generalize these ideas, which is done in LISCOMP (1982), then we have a kernel with a path model for the theoretical constructs, and we have nonliner regressions relating the constructs to the indicators. Indicators now can be nominal, ordinal, numerical, binary, truncated, censored, and so on. The basic idea of having conditional independence of the indicators given the construct still applies, and the basic inheritance from factor analysis is that constructs are continuous variables.

Building the likelihood function for such models is not very difficult. A fairly complete review is in the book by Bartholomew (1988). But maximizing the likelihood can be very costly and pratically impossible. New developments (such as Gibbs sampling) continue to make the boundaries of computability wider and wider, but in the end the empty cell problem and the complicated integrals that are bound to appear in the likelihood function often defeat us.

If we want to avoid going to a fully specified model, using likelihood methods, then we can use the alternatives proposed by Gifi (1991) or De Leeuw (1988), (1986). These optimal scaling techniques can perhaps be best understood by using the concept of *linearizable regressions.* Variables can be of mixed type (nominal, ordinal, numerical). We do not make specific distributional assumptions, only that the variables can be transformed in such a way that all bivariate regressions become linear. We then apply an optimal scaling technique (any reasonable one will do), which will recover the linearizing transformations. And we fit the errors-in-variables model to the optimally scaled variables. Statistical theory for this two-step approach is developed in De Leeuw (1988).

Summary

Wermuth's paper limits itself, quite appropriately, to a class of statistical models about which we can say a great deal mathematically. It is unfortunate, but significant, that these models may have considerable normative value but little descriptive value. The example analyzed in Section 6 of the paper is not typical for the types of problems analysed with path analysis. But even for this tiny example the blunt statement that "associations of a linear type are considered to be appropriate descriptions of pairwise relations between the variables" gives rise to many doubts. "By whom ?", for instance. Or, "In what sense ?" In social science applications of this sort there is no royal road around measurement error, around ordinal variables, and around nonlinear regression. If one does not know how to model these (and generally we don't), then descriptive techniques should be used.

Block recursive models have great advantages over LISREL type models. Both are elegant in their formulation, but LISREL rapidly takes you into a swamp of identification, interpretation, and computation problems. Block recursive models are not only elegant in their formulation, they are elegant in the analysis. My thesis is that here, as everywhere, elegance comes at a terrible price. They will only be applicable to small, selected data sets, in which we refuse to look at some of the basic assumptions. As far as I can see, the robustnes properties which have saved the t-test from oblivion, do not apply here. The most surprising result of the paper, for me, is the brevity of Section 5, indicating how few results carry over from the block recursive case.

References

Albert Gifi (1991). *Nonlinear multivariate Analysis.* Wiley, Chichester, Great Britain.

Arthur S. Goldberger (1972). Structural equation models in the social sciences. *Econometrika*, **40**, 979-1001.

Bengt O. Muthém (1982). Some categorical variable models with latent variable. In Karl G. Joreskog and Herman O. Wold, editors, *Systems under indirect observation: causality, structure, prediction*, 65-79. North Holland Publishing Co, Amsterdam, The Netherlands.

Cheng Hsiao (1983). Identification. In Zvi Griliches and Michael D. Intriligator, editors, *Handbook of Econometrics*, volume I, chapter 4, pages 224-283. North Holland Publishing Co, Amsterdam, The Netherlands.

David J. Bartholomew (1987). *Latent variable models and factor analysis.* Griffin, London, Great Britain.

Dennis J. Aigner, Cheng Hsiao, Arie Kapteyn, and Tom Wansbeek (1984). Latent variable models in econometrics. In Zvi Griliches and Michael D. Intriligator, editors, *Handbook of Econometrics*, volume II, chapter 23, pages 1323-1393. North Holland Publishing Co, Amsterdam, The Netherlands.

Franklin M. Fisher (1966). *The identification problem in econometrics.* McGraw-Hill, New York, New York.

Herman O. Wold (1982). Soft modeling: the basic desing and some extensions. In Karl G. Joreskog and Herman O. Wold, editors, *Systems under indirect observation: causality, structure, prediction*, 65-79. North Holland Publishing Co, Amsterdam, The Netherlands.

Jan de Leeuw (1986). Multivariate analysis with optimal scaling. Calcutta, India. Indian Statistical Institute.

Jan de Leeuw (1988). Multivariate analysis with linearizable regressions. *Psychometrika.*

Paul A. Bekker and Theo K. Dijkstra (1990). On the nature and number of contraints on the reduced form as implied by the structural form. *Econometrika*, **58**, 507-514.

Paul A. Bekker and Stephen Pollock (1986). Identification of linear stochastic models wich convariance restrictions. *Journal of Econometrics*, **31**, 179-208.

Zvi Griliches (1974). Errors in variables and other unobservables. *Econometrika*, **42**, 971-988.