LEAST SQUARES AND MAXIMUM LIKELIHOOD FOR CAUSAL MODELS WITH DISCRETE VARIABLES

by

Jan de Leeuw
Professor of Data theory
University of Leiden
Holland

Discussion Paper

on

Camilo Dagum's

ON THE PRINCIPLES AND HISTORY OF SCIENTIFIC MODEL-BUILDING

The Thirteenth International Conference on the Unity of the Sciences
Washington, D.C. September 2-5, 1984

1: CAUSAL MODELS

Suppose $\underline{x}_1,\ldots,\underline{x}_m$ are centered random variables with finite variances, not necessarily distinct. We use the convention of underlining random variables (Hemelrijk, 1966). A causal model for our random variables is a directed graph, with the variables as edges. If $\underline{x}_j$ is adjacent to $\underline{x}_1$, then we say that $\underline{x}_j$ is a direct cause of $\underline{x}_1$. If $\underline{x}_j$ is adjacent from $\underline{x}_1$, then $\underline{x}_j$ is a direct effect of $\underline{x}_1$.

Now define $\underline{x}_j$ to be a cause of $\underline{x}_1$ if $j \neq 1$ and if there is a directed path from $\underline{x}_j$ to $\underline{x}_1$. Conversely $\underline{x}_j$ is an effect of $\underline{x}_1$ if $j \neq 1$ and there is a directed path from $\underline{x}_1$ to $\underline{x}_j$. The cause-effect relation defines another digraph, which is the transitive closure of the original causal model. We call it the causal closure of the model.

A causal model is recursive if the graph contains no cycles. Recursive models are obviously asymmetric. All causal models are irreflexive. All causal closures are transitive. Causal closures of recursive models are partial orders.

In recursive models we also define a convenient level assignment. Exogeneous variables (i.e. transmitters, edges with indegree 0, variables without causes) have level 0. The level of any other variable is equal to one plus the maximum of the levels of its direct causes. If $\underline{x}_j$ has a lower level than $\underline{x}_1$, then $\underline{x}_j$ is a predecessor of $\underline{x}_1$ and $\underline{x}_1$ is a successor of $\underline{x}_j$. This defines another partial order, which is an extension of the cause-effect order. All causes are predecessors, but not all predecessors need be causes.

This ends our brief graph-theoretical treatment of causal models. It only defines the qualitative components of the model. Our treatment is by no means new. It is implicit in most older path analysis literature,

and it is already quite explicit in Dörfel (1972), Gordesch (1974). Recent interest in similar, although generally much deeper, uses of graph theory has been stimulated by work of Darroch, Lauritzen, Speed, Wermuth, and others. Compare Kiiveri and Speed (1982) or Hodapp and Wermuth (1983) for reviews.

From now on we restrict our attention to recursive causal models, and to the predecessor-successor relation defined by the level assignment in such models. We incorporate a quantitative component, in the form of two testable assumptions. Assumption A states that the projection of a variable on the space spanned by its predecessors is equal to the projection on the space of its direct causes. If we define the residual of a variable to be the anti-projection on the space of its direct causes, then A states that the residual of a variable is orthogonal to its predecessors. Assumption B states that the residuals of different variables are orthogonal.

Assumptions A and B are much weaker than the assumptions used by Wermuth, Speed, and others. They assume either independence in stead of orthogonality, or they directly assume multivariate normality. In his PLS work Wold assumes predictor specifications, which assume linear conditional expectations. We think that for the least squares analysis of causal models the wide-sense concepts or projection and orthogonality are more natural than the strict-sense concepts of conditional expectation and independence. Compare Doob (1953) for further discussion of this distinction.

2: THE LEAST SQUARES LOSS FUNCTION

The loss function that is minimized in least squares (LS, from now on) analysis of causal models is very simple. It is

$$\sigma(B) = \sum_{j=1}^{m} ||\underline{x}_j - \sum_{l=1}^{m} \beta_{jl}\underline{x}_l||^2. \qquad (1)$$

The norm in (1) is the standard deviation. We restrict $B = \{\beta_{jl}\}$ by requiring that $\beta_{jl} = 0$ if $\underline{x}_l$ is not a direct cause of $\underline{x}_j$. Without loss of generality we can assume that B is lower-triangular. Minimization of (1) is quite trivial. We solve at most m separate linear regression problems, i.e. we project at most m variables on the space spanned by their direct causes. For the exogeneous variables there is no regression problem. Parameters are fitted by single equation ordinary least squares.

It is very well known that in recursive causal models, in which we assume in addition to A and B that variables are multivariate normal, the negative logarithm of the likelihood can be written as

$$\sigma(B,\Delta) = \sum_{j=r+1}^{m} \ln \delta_j^2 + \sum_{j=r+1}^{m} (\delta_j^2)^{-1} ||\underline{x}_j - \sum_{l=1}^{m} \beta_{jl}\underline{x}_l||^2. \qquad (2)$$

In formule (2) we have ignored irrelevant constants, and terms which only depend on the distribution of the exogeneous variables. There are r exogeneous variables, and the variances of the residuals of the remaining m - r endogeneous random variables are the δ_j^2. Observe that in (1) we could also have numbered the equations starting with j = r + 1.

It follows directly from (2) that the maximum likelihood estimates of the regression weights are the same as the least squares estimates given above. Thus ML = LS. The ML-estimates of the residual variances are the variances of the observed residuals. This result is given explicitly in Wold (1954). It depends critically on the recursivity of the model, because the level assigment can be used to factor the likelihood. It also depends critically on the assumption of multivariate normality, of course.

In general we can say that LS is more widely applicable than ML. The geometrical notion of projection makes sense without using any probability model. The vanishing of regression coefficients, or the orthogonality of residuals, can be tested by asymptotic methods under much more general assumptions than multivariate normality.

3: CAUSAL MODELS WITH LATENT VARIABLES

The 'partial' in partial least squares (PLS, from now on) becomes operative if the causal model contains latent variables. Models of this type have been discussed by Wold in his numerous publications dealing with NIPALS and PLS, and by Jöreskog in his numerous publications dealing with the LISREL-system. Compare Jöreskog and Wold (1982). In econometrics corresponding errors-in-variables models are reviewed by Aigner, Hsiao, Kapteyn, and Wansbeek (1983). In this section we start with the least squares approach to models of this form. Our emphasis is somewhat different from that of Wold, however, because our starting point is the overall LS loss function (1). The model in our approach is completely defined by the digraph. The translation of the model into the loss function (1) is immediate, and the computational problem is to minimize this loss function. Thus there is a single or total least squares criterion, while in Wold's PLS approach the model consists of various submodels which have their own separate loss function. We follow the work of De Leeuw, Young, and Takane (1976, and many later publications) and that of Gifi (1981, and many later publications). Although the minimization algorithm we use is very similar to the algorithms used in PLS, it is designed explicitly to minimize the total least squares criterion (1). We prefer to call these algorithms alternating least squares

(from now on ALS). As far as algorithms are concerned we often have ALS = PLS, but the total LS criterion (1) is not necessarily the same as the PLS criteria for the submodels of Wold.

To explain the basic idea of latent variables in a convenient way, we first observe that random variables on a fixed probability space, with zero mean and finite variance, define a separable Hilbert space H. Let S be the unit sphere in this space, i.e. the variables with unit variance. Suppose J is a subset of the index set $\{1,\ldots,m\}$ that indicates those variables that are unobserved or latent. The problem now is to minimize (1), which we now write as $\sigma(B,\{\underline{x}_j\}_{j\in J})$ over B and over $\underline{x}_j$ in S, for all j in J.

There are several aspects to this extension which are worth discussing at this point. A latent variable which occurs in only one equation is not very useful. This equation can always be fitted perfectly, and the corresponding term simply drops out of the loss function. Thus latent variables usually occur in more than one equation. They can be both endogeneous (as in the Hauser-Goldberger path model) or exogeneous (as in the common factor model). Because the same latent variable occurs in more than one equation we cannot use single equation LS techniques any more, the latent variables link the equations. In some models additional restrictions on the latent variables are needed. If we introduce more than one latent variable in the common factor model or the Hauser-Goldberger model, then we are forced to impose orthogonality conditions on the latent variables to avoid uninteresting duplications of results. In some models, for example the canonical analysis models, we have to require that different latent variables are represented by the same element of S. Such tricks make it possible to introduce various special forms of non-recursivity into the models.

There is another aspect of latent variables that is worth discussing. It is related to the results by Gifi, De Leeuw, Young, Takane and others in what is commonly known as nonlinear multivariate analysis with optimal scaling using alternating least squares. Suppose K_j, with $j=1,\ldots,m$, are convex cones in the space H. A much more general problem than the one we have discussed above is the minimization of $\sigma(B,\underline{x}_j)$ over B and over all $\underline{x}_j$, with the restriction that $\underline{x}_j$ must be in the intersection of K_j and S. We have already discussed two important special cases. If K_j is a single ray then $\underline{x}_j$ is an observed variable, if K_j is the whole space H then $\underline{x}_j$ is a latent variable. There are, however, many interesting intermediate cases. Discrete random variables, for instance, assuming k_j possible values, can be quantified. The quantifications define a $k_j - 1$ dimensional subspace of H. Numerical variables can be transformed. Transformations also define a subspace of H, which can be approximated by using polynomials or splines or other convenient finite dimensional families. If we require monotonicity the quantification or transformation subspaces are replaced by pointed cones. There are many examples in the book by Gifi (1981, 1984). Important special cases have also been discussed separately. Principal component analysis is treated by De Leeuw (1982), two-set canonical analysis by Van der Burg and De Leeuw (1983), and N-set canonical analysis by Van der Burg, De Leeuw, and Verdegaal (1984). Analysis of causal models, based on loss function (1), has only been discussed very briefly and very incompletely in the Gifi or ALSOS system. Our systematic treatment of the latent variable as just another measurement level in the sequence numerical, ordinal, nominal, latent is new, and seems very convenient.

It is easy to show, by elementary algebraic manipulations, that in the

case of the common factor model, or the Hauser-Goldberger model, or the canonical correlation model, minimization of the loss function (1) over weights and latent variables gives the same solution as some of Wold's PLS techniques for these models. Other solutions can be derived by modifying the model (changing the direction of arrows, duplicating variables at different places in the model). We shall not show this in detail, because it does not belong in a discussion paper. Nevertheless we re-emphasize that our formulation of PLS (or ALS) makes it easy to extend all causal models to nominal, ordinal, 'splinical', and other types of variables. The algorithm alternates the fitting of B (a linear projection problem) with the fitting of the $\underline{x}_j$ in their cones (cone projection problems). The computer program that will do all this is called PATHALS. It only exists in preliminary APL versions.

There is another obvious generalization of the ALS approach that we discuss here. Suppose we do not minimize $\sigma(B)$ of (1) but $\sigma(B,\Delta)$ of (2) by our alternating least squares algorithms. Several different special cases of this could be considered. If the δ_j^2 are known, then we simply apply weighted least squares to fit the latent variables. The regression weights are still computed by single equation OLS. If the δ_j^2 are known up to a proportionality constant, the same thing is true. If the δ_j^2 are known to be equal, we are in fact back in the situation of minimizing (1). If the δ_j^2 are unknown, we consider them as additional structural parameters that must be estimated. We insert an additional step in our minimization cycles, in which $\sigma(B,\Delta)$ is minimized over δ_j^2. This is very easy, the estimate is simply the variance of the observed current residual. We have to be very careful in this most general of cases, however. If we can choose latent

variables and regression weights in such a way that the residual for one of the $\underline{x}_j$ vanishes, then loss function (2) is unbounded below, and the minimum we are looking for does not exist. This was already discussed for the special case of the common factor model by Anderson and Rubin (1956). Of course if an equation contains a completely unrestricted latent variable, then the residual can always be made equal to zero.

Procedures that minimize (2) over latent variables and structural parameters are still called least squares methods. This is true if the residual variances are known, partially known, or completely unknown.

4: MAXIMUM LIKELIHOOD WITH LATENT VARIABLES

If there are no latent variables, then ML estimation can be done by single equation OLS. This makes it interesting to look at algorithms which first complete the likelihood, by computing an estimate or proxy for the latent variable, and then perform ML = LS to estimate the structural parameters. We have already seen that ALS, our version of PLS, is such an algorithm. It computes proxies by minimization. Or, to put it differently, the missing information is treated as a set of additional parameters over which we minimize. In a very interesting recent paper Little and Rubin (1983) discuss this technique of estimating missing information in general terms. They find that it can lead to estimates that are badly biased, also of the structural parameters. We briefly explain the problems in terms of the difference between structural and functional models.

Suppose $\underline{x}_i$ are the n independent m vectors corresponding with the n observations. We suppose that $\underline{x}_i = \mu_i + \underline{\chi}_i$, where the $\underline{\chi}_i$ are independent and identically distributed centered m-vectors. For an observed variable

$\underline{x}_{ij}$ we assume that $\mu_{ij} = 0$ for all i. But for a latent variable the μ_{ij} can all be different. The negative logarithm of the likelihood is

$$\sigma(B,\Delta,M) = \sum_{j=r+1}^{m} \ln \delta_j^2 + \sum_{j=r+1}^{m} (\delta_j^2)^{-1} n^{-1} \sum_{i=1}^{n} ((\underline{x}_{ij} - \mu_{ij}) - \sum_{l=1}^{m} \beta_{jl}(\underline{x}_{il} - \mu_{il}))^2.$$

Mimimizing this over the unknown parameters is the same thing as PLS or ALS. Thus we have shown that PLS can be interpreted as ML fitting of a partial functional and partial structural multinormal model with incidental parameters. Inconsistencies arise, because the number of parameters tends to infinity with the number of observations.

Our emphasis dictates, that PLS should not be treated as a somewhat aberrant and deficient method of estimation in structural models, but as the ML method in models with functional latent variables. This makes our results different from, and perhaps more enlightening than, those of Dijkstra (1983), who compares PLS and LISREL as techniques to fit a common structural model. Of course we can continue to interprete PLS, as we have done from the beginning, as a very natural method to fit causal models that are defined only in terms of projection.

There is another way to define proxies, which corresponds with ordinary ML in structural models. Wold has said, on many occasions, that one of the advantages of PLS over LISREL is that PLS gives estimates of the latent variables and LISREL does not. Moreover PLS is the simpler algorithm, because it consists of linear regressions only. This may be true for LISREL as a program, but it is not true for ML as a technique. The work of Dempster, Laird, and Rubin (1977), which has been applied to common factor analysis by Rubin and Thayer (1982) and to the Hauser-Goldberger model by Chen (1981), shows that ML defines it proxies by expectation. The current best proxy

is the conditional expectation of the latent variable, given the observed variables and the current estimates of the structural parameters. Thus proxies are also computed by regression in the ML case. Of course we know that structural ML estimates have statistical optimality properties, among which consistency. It may be true, as it is in factor analysis (Anderson and Rubin, 1956), that structural ML estimates are also consistent if the functional model is true. If the data are multinormally distributed it seems that there is no reason to apply PLS or ALS, we should simply apply structural ML, even if the model is partly functional. Our result that PLS = ML in certain functional models is mostly or only of theoretical interest. If the data are not normal, the choice is not so simple. The EM-algorithm for multinormal ML defines the proxies by expectation, the PLS and ALS algorithms define them by minimization. Both are forms of projection, and are closely related. The two techniques will be especially close if residual variances are equal, but a more detailed comparison is certainly needed.

5: REFERENCES

Aigner, D.J., Hsiao, C., Kapteyn, A., & Wansbeek, T.J. Latent variables in econometrics. In Griliches, Z. and Intriligator, M.D. (eds) Handbook of econometrics, Amsterdam, North Holland Publishing co, 1983.

Anderson, T.W. & Rubin, H. Statistical inference in factor analysis. In Neyman, J. (ed) Proceedings of the third Berkeley symposium, volume V, Berkeley, University of California Press, 1956.

Chen, C.F. The EM approach to the multiple indicators and multiple causes model via the estimation of the latent variable. Journal of the American Statistical Association, 1981, 76, 704-708.

De Leeuw, J. Nonlinear principal component analysis. In Caussinus, H., Ettinger, P., & Tomassone, R. (eds) COMPSTAT 1982. Wien, Physika Verlag, 1982.

De Leeuw, J., Young, F.W., & takane, Y. Additive structure in qualitative data: an alternating least squares method with optimal scaling features. Psychometrika, 1976, 41, 471-503.

Dempster, A.P., Laird, N.M., & Rubin, D.B. Maximum likelihood for incomplete data using the EM algorithm. Journal of the Royal Statistical Society, 1977, B39, 1-38.

Doob, J.L. Stochastic Processes. New York, Wiley, 1953.

Dörfel, H. Pfadkoeffizienten und Strukturmodelle. Biometrische Zeitschrift, 1972, 14, 12-26.

Dijkstra, T.J. Somme comments on maximum likelihood and partial least squares methods. Journal of Econometrics, 1983, 22, 67-90.

Gifi, A. Nonlinear multivariate analysis. Leiden, DSWO Press, 1981, 1984.

Gordesch, J. Causal models. Colloquia Mathematica Societas János Bolyai IX. Proceedings European Meeting of statisticians, Amsterdam, North Holland Publishing Co, 1974.

Hemelrijk, J. Underlining random variables. Statistica Neerlandica, 1966, 20, 1-8.

Hodapp, V. & Wermuth, N. Decomposable models: a new look at interdependence and dependence structures in psychological reserach. Multivariate Behavioural Research, 1983, 18, 361-390.

Jöreskog, K.G. & Wold, H.O.A. Systems under indirect observation: causality, structure, prediction. Amsterdam, North Holland Publishing Co, 1982.

Kiiveri, H. & Speed, T.P. Structural analysis of multivariate data: a review. In Leinhardt, S. (ed) Sociological Methodology. San Francisco, Jossey Bass, 1982.

Little, R.J.A. & Rubin, D.B. On jointly estimating parameters and missing data by maximizing the complete-data likelihood. American Statistician, 1983, 37, 218-220.

Rubin, D.B. & Thayer, D.T. EM algorithms for ML factor analysis. Psychometrika, 1982, 47, 69-76.

Van der Burg, E. & De Leeuw, J. Nonlinear canonical correlation analysis. British Journal of Mathematical and Statistical Psychology, 1983, 36, 54-80.

Van der Burg, E., De Leeuw, J., & Verdegaal, R. Nonlinear canonical correlation for m sets of variables. Submitted for publication, 1984.

Wold, H.O.A. Causality and econometrics. Econometrica, 1954, 22, 162-177.