Structural covariance models with categorical variables

Jan de Leeuw
Departments of Psychology and Mathematics
University of California Los Angeles

Peter van der Heijden
Department of Methodology and Psychometrics
University of Leiden

Introduction

Suppose $\underset{\sim}{S}_n$ is an observed covariance matrix based on n observations. A parametric model for such a matrix postulates that the elements of $\Sigma = \mathbf{E}(\underset{\sim}{S}_n)$, the expected value of the matrix, can be written in the form $\sigma_{ij}(\theta_1,...,\theta_p)$, with the σ_{ij} given functions of the vector of parameters θ. Techniques for fitting parametric models to covariance matrices are very popular in psychometrics, econometrics, and many other areas of applied statistics. Reviews are in Joreskog (1978), Bentler and Weeks (1982), Brown (1982), Aigner, Hsiao, Kapteyn, and Wansbeek (1984).

These parametric models are often derived from more simple considerations, for instance from notions of causality or conditional independence of variables. We shall not be concerned in this paper with the more basic methodological problems connected with the usefulness of such models. There is an enormous literature on this. We mention Wilson (1928), Guttman (1978), Kalman (1983), Cliff (1983), Freedman (1987), De Leeuw (1985) for a discussion of some of these problems.

One common criticism of covariance structure models is that their subsequent statistical analysis is usually based on the additional assumption of multivariate normality. Thus we assume that the n observations are n independent drawings from the same multivariate normal distribution. In many situations this assumption is not very realistic. A great deal of research has been done recently on various ways to relax the assumption of multivariate normality, while preserving most of the attractive properties of an analysis based on multinormal assumptions. Brown (..), Bentler (..), Wesselman (..) study the case in which we merely assume ellipticity. Multinormal maximum likelihood estimates remain first order efficient in this case, but their asymptotic distribution (a.d.) varies with the kurtosis parameter of the elliptical distribution. Brown (..) also indicates that multinormal maximum likelihood is efficient, and has the same a.d. as in the multinormal case, for all multivariate distributions with zero kurtosis. Brown (), Bentler(), and De Leeuw (1983) studied the so-called ADF (asymptotic distribution free) methods, which are applicable whenever fourth moments exists, and which improve on the multinormal maximum likelihoods estimates if the multinormality assumption is not satisfied. Many additional results on ADF have been derived in the mean time.

Using ADF methods, however, is definitely not satisfactory in some cases. If the bivariate regressions are distinctly nonlinear, then the methods are applicable, but in these situations there is very little reason to study covariances at all. There is an even more important problem if we are not sure if the parrticular transformation or quantification of the variable we use is the correct one. We use income in our linear model, but perhaps we should use log-income. We use IQ, but perhaps the square root of IQ is better behaved. Or, even more extremely, we use catholic = 1, protestant = 2, buddhist = 3, but we realize that this choice is highly arbitrary. For such situations assuming multinormality does not make much sense, and ADF does not help us at all, because we can only apply it on the condition that we know the correct scoring of the variable.

There are several adaptations of the usual techniques available which can help us to deal with nonlinear regressions and uncertain scoring. Most of them are in the tradition of optimal scaling (Gifi, 1981, Young, 1981). Optimal scaling methods, in this context, choose a reasonable loss function that measures closeness of the observed covariance matrix $\underline{S}_n$ to a fitted covariance matrix $\Sigma(\theta)$. The observed covariance matrix is partially unknown. It is a function of the transformations of the variables that are chosen, each transformation from its appropriate permissible class of transformations (for instance all monotone transformations, or all smooth transformations). We minimize the loss function not only over the structural parameters θ, but also over the permissible quantifications. In the usual case the criterion chosen is of the least squares type. The algorithms alternate the solution of two usually nonlinear least squares problems. The first least squares problem finds the optimal transformation for given structural parameters, the second subproblem finds the optimum parameters for given transformations. The FACTALS algorithm of Takane, Young, and De Leeuw (1979) and the PATHALS algorithm of Coolen and De Leeuw (1987) are clear examples of such methods.

Using alternating least squares in the analysis of covariance structures has some disadvantages. It has been shown by Dijkstra (198.) that the method does not give consistent estimates of the structural parameters in the usual case in which everything is normally distributed and some variables are completely unknown (are latent variables). This is basically a consequence of the large number of incidental parameters that are estimated. Each observation on a latent variable acts as an incidental parameter (compare Little and Rubin, 198.). In the second place one could say that no explicit statistical model is postulated by these techniques. Thus nothing can be falsified and consequently we cannot really augment our knowledge. This is based on a popular, although perhaps somewhat old-fashioned, philosophy of science. We shall ignore this argument, because it is merely philosophical. Thirdly one can argue that computation of stability information, in the form of confidence intervals for instance, is more complicated in least squares techniques than for instance in maximum likelihood methods. This is usually true, although computation of confidence intervals does not always make sense. And finally it is often argued that using unweighted least squares implies a particular weighting of errors which is often very hard to defend. This is only true if the notion of error makes sense, and if some models for errors are really more sensible than others. Nevertheless the objections, certainly if taken together, are serious enough it consider alternatives to least squares optimal scaling, at least in some cases in which specific models are plausible.

Maximum likelihood optimal scaling

The first such alternative is readily available. The polychoric model was proposed by Pearson in 1900, and many elaborations and extensions of it have been proposed recently by Christofferson, Muthen, Olsson, Lee and Poon. Compare Van der Pol and De Leeuw (1984) for a recent overview. Its is easy to explain the polychoric model. The cells of the multivariate table correspond with blocks in the multivariate space. The probability of a cell is the integral of a multinormal distribution over the corresponding block. Thus we can say that each discrete observed variable is the discretization of an unobserved continuous variable, and the unobserved variables are jointly multivariate normal.

The polychoric model has one major disadvantage. Computing estimates of the parameters is computationally very demanding. In LISREL (Joreskog and Sorbom, 19..) the correlations between the variables are first estimated by polychoric methods, together with their dispersion. These results are then used in a generalized least squares estimation step to find estimates of the structural parameters. In LISCOMP (Muthen, ...) the two steps are combined in a single step, which makes the method more efficient statistically, but less efficient computationally. Still, all these methods are limited information methods, because they ignore the information in the higher order frequencies. Full information methods for the polychoric model are only possible in cases with a small number of variables (say three, compare Lee, ..) or in cases with a simple covariance structure (such as a one-factor model, compare Takane and De Leeuw, 1987).

The relationship with alternating least square methods becomes clear if we develop the EM-algorithm for the block multinormal model. It turns out that in each step we have to minimize a function of the familiar form $\not{f} = \ln |\Sigma| + \mathrm{tr}\ \Sigma^{-1}S$ over the structural parameters in Σ. The matrix S is not constant but depends on the current estimates of the structural parameters, and has element $s_{jl} = E(x_j x_l)$, where the expectation is computed with respect to a probability distribution which is $(p_i/\pi_i)\pi(x)$ in block i, with $\pi(.)$ the current estimate of the multinormal, and π_i its integral over block i. Although this may seem simple, the actual computation of S involves multiple integration, and must usually be carried out by Monte Carlo methods or by Gauss-Hermite approximation. Thus we must solve an optimization problem in each step which has the same form as the problem solved by LISREL or EQS, but we also must compute a very complicated substitute for the observed covariance matrix in each step. This actually makes the problem computationally infeasible, at least at the present time.

Another disadvantage of the polychoric model, already pointed out by Yule in his early comments on Pearson's work, is that usually their are no valid reasons to assume that the underlying distribution is indeed multivariate normal. Why would it be ? And in Yule's own research the observed variable was 'dying from tuberculosis or not dying'. Yule could not very well imagine un underlying continuum here, one either died or one did not die. This gave rise to heated discussions at the time (Pearson and Heron, 1913), which were mainly about the value of probability models versus the construction of simple association coefficients with some desirable properties, and about Pearson's desire to achieve a unity of the sciences on the basis of the coefficient of correlation. We shall not enter into this debate here, but merely point out that Yule's pragmatic approach to the analysis of categorical data has developed into the enormously popular loglinear model, while Pearson's polychoric approach has gained some popularity in psychometrics. Compare Norton (198.), Fienberg (19.) for more dicussion of these isssues.

In the last decade there have been various attempts at unification, inspired on the one hand by correspondence analysis (i.e. optimal scaling) and on the other hand by log-linear models for ordinal variables (Goodman, 198., Agresti, 198.). The model we shall propose in this paper is another such attempt at unification. It was first proposed, in a more limited form, by De Leeuw (1983).

Definition of the point-multinormal model

Suppose $g_1, \ldots, g_N$ are indicators for the profiles of a multidimensional contingency table. Thus if variable 1 has k_1 values, ... , variable m has k_m values, then $N = k_1 \times \ldots \times k_m$, and the g_i have $k_1 + \ldots + k_m$ elements, of which exactly m are equal to one, while the others are zero.

Write $\underline{n}_i$ for the observed frequency of profile i in a multinomial sample of size n, $\underline{p}_i = n_i/n$, and $\pi_i = E(\underline{p}_i)$. We assume that there exist m vectors of scores $y_1, \ldots, y_m$ and a matrix Σ, such that

$$\pi_i = C_1 \exp\{-\tfrac{1}{2}(z_i - \mu)'\Sigma^{-1}(z_i - \mu)\}. \qquad (1)$$

Here $z_i = Y'g_i$, where Y is the direct sum of the y_j, a $(\Sigma\, k_j) \times m$ matrix having m blocks of size $k_j \times 1$ along the diagonal. C_1 is a factor which makes all π_i add up to one. For computational purposes it is also convenient to collect the y_j in a large vector y, and to define z_i by selecting from y. If H_i is the binary matrix that does the selection, of dimension m x N, then $z_i = H_i y$.

Expression (1) looks very familiar indeed. In the continuous multinormal $\mathcal{N}(\mu,\Sigma)$ the density at z_i is

$$\pi(z_i) = C_2 \exp\{-\tfrac{1}{2}(z_i - \mu)'\Sigma^{-1}(z_i - \mu)\}, \qquad (2)$$

with C_2 a factor that makes the integral of (2) over all of $\Re^m$ equal to one. Our formula (1) for $\pi_i = \pi(x_i)$ is a discrete version of this.

We can also say that in the block multinormal or polychoric model the probability of a cell is the integral over the block in space corresponding with the cell. If we approximate the value of the integral by a constant times the value of the density at some point in the cell, then we find the point multinormal model. Thus the point multinormal model approximates the continuous multinormal model in a somewhat different way as the polychoric model, and gets rid of the multiple integration.

Properties of the point multinormal model

There is one obvious property that the point multinormal model does not have, while this property is true for the block multinormal model. If the model is true for a multivariate discrete distribution, then tthe model will in general not be true if the distribution is redefined by grouping categories. And, perhaps even more dramatically, if the model is true for a discrete multivariate distribution then it will generally not be true for its lower-dimensional marginals. This may seem a very serious disadvantage to some persons, but if one considers the role of models a little bit more in detail the seriousness seems to disappear. Models are never exactly true, they are only approximations. And it will be the case that if the point multinormal model is approximately true for a distribution, then it will also be approximately true for its lower dimensional projections. In fact exactly the same objection could be raised against loglinear models in general (Darroch, 196.), the absence of interactions in such models is also not invariant under grouping and marginalizing.

Another important property is shared by the point multinormal and the continuous multinormal model, but not by the block multinormal model. The negative log-likelihood for the block model is

$$\mathcal{f} = -\Sigma_{i=1}^{N} \underline{p}_i \ln \pi_i = \text{tr}\, \Sigma^{-1}S + \ln \Sigma_{i=1}^{N} \exp(-\tfrac{1}{2}(z_i - \mu)'\Sigma^{-1}(z_i - \mu)). \qquad (3)$$

Here S is the observed covariance matrix of the z_i. If the columns of Y are in deviations from the mean then S = YCY', where C is the Burt Matrix of the multivariate table, i.e. the stacked matrix of bivariate cross-tables. This can be compared with the negative log-likelihood for the continuous multinormal model, which is, except for irrelevant constants,

$$\mathcal{L} = \text{tr}\, \Sigma^{-1}S + \ln |\Sigma|. \tag{4}$$

Formulation (3) shows immediately that the maximum likelihood estimates of all parameters in the point multinormal model are functions of the bivariate marginals only, just like in the bivariate normal case. This makes the problem relatively well conditioned (the empty cell problem is not too serious), it also makes the theory comparable to multivariate normal theory on one side and multiple correspondence analysis theory on the other side. Comparison of (3) and (4) shows again that the multiple integral that normalizes the multinormnal density, and that happens to be equal to $|\Sigma|$, is approximated by a multiple sum over the grid defined by the z_i.

For statistical theory, and also for computational purposes, it is useful to look at the first and second derivatives of the likelihood function. We do this for the somewhat more general case in which

$$\pi_i(\theta) = \exp(-f_i(\theta))/\Sigma_{k=1}^{n} \exp(-f_k(\theta)). \tag{5}$$

where θ is a general vector of parameters. In the point multinormal model $f_i(\theta) = \frac{1}{2}(z_i - \mu)'\Sigma^{-1}(z_i - \mu)$, and the parameters are y, μ, and Σ. Now (5) gives

$$\mathcal{L} = \Sigma_{i=1}^{N} p_i f_i(\theta) + \ln \Sigma_{i=1}^{N} \exp(-f_i(\theta)). \tag{6}$$

To derive a compact expression we the derivatives we use

$$\pi_{is} = -\pi_i(f_{is} - f_s[\pi]), \tag{7}$$

where $\pi_{is} = \partial\pi_i/\partial\theta_s$, $f_{is} = \partial f_i/\partial\theta_s$, and $f_s[\pi]$ is the weighted average $\Sigma_{i=1}^{n} \pi_i f_{is}$. Because

$$\mathcal{L}_s = -\Sigma_{i=1}^{n} (p_i/\pi_i)\, \pi_{is}, \tag{8}$$

we find

$$\mathcal{L}_s = \Sigma_{i=1}^{n} p_i(f_{is} - f_s[\pi]) = (f_s[p] - f_s[\pi]). \tag{9}$$

We now use a similar notation for second derivatives. Thus

$$\ell_{st} = (\ell_s)_t = (f_s[p] - f_s[\pi])_t = (f_{st}[p] - (\Sigma_{k=1}^n \pi_k\, f_{ks})_t) =$$

$$= (f_{st}[p] \ - \Sigma_{k=1}^n \pi_{kt}\, f_{ks} - \Sigma_{k=1}^n \pi_k\, f_{kst}) = (f_{st}[p] - f_{st}[\pi]) + w_{st}[\pi]. \qquad (10)$$

Here

$$w_{st}[\pi] = \Sigma_{i=1}^n \pi_i(f_{is} - f_s[\pi])(f_{it} - f_t[\pi]). \qquad (11)$$

It follows directly that $E(\ell_{st}) = w_{st}[\pi]$, which is positive semidefinite. Clearly these are convenient expressions to have around. They do show that ℓ is not necessarily convex unless the f_i are linear. If p and π are sufficiently different, because the model does not fit or because we do not have good estimates, then the negative log likelihood could be very irregular, and very difficult to minimize.

Regularization of the problem by majorization

The basic problem we study in this paper is to maximize the likelihood, i.e. to minimize its complement ℓ, over all score-vectors in y and over all μ and Σ that satisfy certain parametric restrictions (must correspond to a given path model, to a given factor analysis model, or to a given analysis of variance model, or something like that). This is done by a three-step algorithm of the following type. First minimize over y for fixed Σ and μ, then over Σ for fixed y and μ, and finally over μ for fixed y and Σ. This constitutes one cycle of the algorithm. We repeat the cycles until they have converged. Many variations on this basic theme are possible. We discuss one that we actually are going to use, which consists of a form of local regularization of ther problem before we start our cycles. We regularize by using majorization, a technique which has been used before in multidimensional scaling (De Leeuw, 1975, De Leeuw and Heiser, 1980) and which is also the basis of the popular EM algorithm.

Suppose f_i is convex and differentiable. Thus

$$f_i(\theta) \geq f_i(\xi) + (\theta - \xi)'g_i(\xi). \qquad (12)$$

where $g_i(\xi)$ is the vector of partials at ξ. This implies

$$\mathscr{f}(\theta) \leq \Sigma_{i=1}^{N} p_i f_i(\theta) + \ln \Sigma_{i=1}^{N} \exp\{-f_i(\xi) - (\theta - \xi)'g_i(\xi)\}. \tag{13}$$

Letting

$$w_i(\xi) = \exp\{-f_i(\xi) + \xi'g_i(\xi)\}, \tag{14}$$

and

$$\mathscr{h}(\theta,\xi) = \Sigma_{i=1}^{N} p_i f_i(\theta) + \ln \Sigma_{i=1}^{N} w_i(\xi)\exp\{-\theta'g_i(\xi)\}, \tag{15}$$

we can also write

$$\mathscr{f}(\theta) \leq \mathscr{h}(\theta,\xi).$$

Two things are also clear from this formulation. If $\theta = \xi$ or if all f_i are linear, then $\mathscr{f}(\theta) = \mathscr{h}(\theta,\xi)$. If one of the f_i is strictly convex and $\theta \neq \xi$, then $\mathscr{f}(\theta) < \mathscr{h}(\theta,\xi)$. Suppose now that our algorithm consists of finding θ^{k+1} by minimizing $\mathscr{h}(\theta,\theta^k)$ over θ. If θ^k actually minimizes $\mathscr{h}(\theta,\theta^k)$, then we stop. Suppose this is not the case. Then from () we have $\mathscr{f}(\theta^{k+1}) \leq \mathscr{h}(\theta^{k+1},\theta^k)$. Because θ^{k+1} minimizes we also have $\mathscr{h}(\theta^{k+1},\theta^k) < \mathscr{h}(\theta^k,\theta^k) = \mathscr{f}(\theta^k)$. Thus $\mathscr{f}(\theta^{k+1}) < \mathscr{f}(\theta^k)$, and we have increased the likelihood. Under fairly general conditions the transformation $\theta^k \rightarrow \theta^{k+1}$ is continuous, and this proves convergence of the procedure, in the sense that each accumulation point θ^* of the sequence satisfies the condition that θ^* minimizes $\mathscr{h}(\theta,\theta^*)$ over θ, which actually means that θ^* satisfies the likelihood equations.

The procedure is illustrated in Figure 1. Here we have used the fact that $\mathscr{h}(\theta,\xi)$ is convex in θ for every ξ, which is exactly what we meant by regularization above. In stead of minimizing the complicated function function $\mathscr{f}(\theta)$, for which we presumably need step-size procedures and various safeguards, we minimize the regular function $\mathscr{h}(\theta,\xi)$, which can be done quickly and reliably. Of course this regularization is combined with the idea that the variables are partitioned in the three sets μ, Σ, and y. This means that there are many possibilities for algorithm construction here which should be exploited. We can use our cycles to minimize $\mathscr{h}(\theta,\xi)$ fairly exactly before we attempt a new regularization, or we can proceed with crude minimizations of $\mathscr{h}(\theta,\xi)$, and update the regularization a bit more often. Also it is not clear if all three substeps in a cycle should be carried out with equal precision.