UCLA Statistics Series #35 April, 1989

Fitting Reduced Rank Regression Models by Alternating Maximum Likelihood

Jan de Leeuw
Departments of Psychology and Mathematics

UCLA Classification: Theory and methods. Social Sciences-Education

Introduction

Consider the following empirical situation. For a number of objects or individuals, indexed by i = 1, ... , n, we observe two vector variables x_i and y_i. The basic idea behind this partitioning of the variables into two sets is that we have the idea that the y_i are influenced by the x_i, i.e. that the x_i are the *causes* of the y_i. The x_i can be thought of as *input* variables, the y_i as *output* variables. In econometrics the x_i are called *exogenous*, and the y_i *endogenous*. In psychometrics, and in variou other areas of applied statistics, the x_i are called *independent* variables, and the y_i are *dependent*. Thus we have two sets of variables, and the two sets play a different and asymmetric role in our thinking.

In multivariate analysis the individuals are often considered to be replications of the same basic structure. The data can be considered to be a random sample from some well-defined population. Another way of saying this is that there is no causal connection between variables with different indices. Thus x_1 influences y_1, x_2 influences y_2, and so on, but there is no influence of x_1 on x_2 or on y_2. This is called the *independence* assumption. Another important aspect of the usual models in this class is *stationarity*. This means that the influence of x_1 of y_1 is supposed to be the same as that of x_2 on y_2, and so on. Models for which the independence assumption is violated are be discussed in another publication (Bijleveld and De Leeuw, in press).

Independent and stationary models are at the basis of regression analysis, and of linear models in general. More recently a slightly more complicated class of these models has been discussed, which goes under various names. They are called *reduced rank regression* models, *redundancy analysis* models, *growth curve* models, *MIMIC* models, or *errors-in-variables* models. Their basic common idea is that the influence of x on y is mediated by unobserved latent variables z, with x determining z, and z determining y. In general the dimensionality of the z is lower than that of the x, and in this sense z filters the relationships between the two sets of variables. We call the space of the z the latent space, and we use p for its dimensionality. For various versions and applications of reduced rank regression we refer to Anderson (1951, 1984), Izenman (1965),

Keller and Wansbeek (1983), Jøreskog and Goldberger (1976). The basic properties of such models will be discussed in general terms below.

In addition we consider techniques for fitting models of this kind. Some general considerations must be kept in mind here. In fitting models to data there usually are three kinds of errors that we have to take into account. The first error is *approximation error*. This occurs because models are never true, and are at best approximations. The second kind of error is *replication error* or *sampling error*, this is the kind of error studied in statistics. It occurs because we sample from a population. It is often expedient also to discuss *measurement error*, which occurs because of limited precision or other disturbing circumstances. In survey research the measurement errors are often discussed as *non-sampling errors*. Observe that we assume that even if there are no sampling errors and no measurement errors, then there will still be approximation errors. This is because models are not exactly true, by definition. For further discussion of these points we refer to Guttman (1985), Kalman (1983), De Leeuw (1984, 1988a).

Regression with latent variables

In the usual regression situation we study the conditional distribution of y given x. This conditional distribution is studied through conditional expectations and/or conditional variances. Suppose p(y|x) is this conditional distribution. We use a somewhat informal notation here, which can either refer to discrete probability distribution or to densities. The purpose of statistical analysis in this context is to see if we can describe this conditional distribution in simple terms. Often this is done by assuming that the conditional expectations are linear in x, and the variances do not depend on x (i.e. are homoscedastic). But this type of simplification of the models is perhaps a little bit drastic in many circumstances.

Another type of simplification can be introduced by using concepts borrowed from factor analysis. In factor analysis we observe variables $y_1,...,y_m$, and these variables are correlated. We assume that there exist unobserved variables or factors $z_1,...,z_p$ which 'explain' the association between the observed variables, in the sense that the observed variables are independent given the factors. In our informal notation we assume that

$$p(y_1,\ldots,y_m|z) = \Pi_{j=1}^{m}\ p(y_j|z), \tag{1}$$

and thus

$$p(y_1,\ldots,y_m) = \int \Pi_{j=1}^{m}\ p(y_j|z)\ p(z)dz. \tag{2}$$

Now let us translate this to the regression context. The first possibility is to assume that there are p latent variables $z_1,\ldots,z_p$ such that y and x are independent given z. In formula this is $p(x,y|z) = p(x|z)p(y|z)$, or, equivalently, $p(y|x,z) = p(y|z)$. This means that

$$p(y|x) = \int\ p(y|z)\ p(z|x)dz. \tag{3}$$

But the conditional independence assumption is also equivalent to $p(x|y,z) = p(x|z)$, and consequently the role played by x and y is perfectly symmetric. This is not precisely what we had in mind. We get the necessary asymmetry by assuming in addition that (1) is true. Then (3) becomes

$$p(y|x) = \int \Pi_{j=1}^{m}\ p(y_j|z)\ p(z|x)dz. \tag{4}$$

Model (4) is called a reduced rank regression model, because z has fewer components than x. If the regressions of y on z and of z on x are linear, the name becomes even more clear. Suppose $E(y|z) = Hz$ and $E(z|x) = G'x$ then

$$E(y|x) = \int Hz\ p(z|x)dz = HG'x. \tag{5}$$

Thus the regression coefficients B satisfy $B = HG'$, i.e. B is of reduced rank p. Observe that (5) is also true for the more general model (3). If the conditional dispersions satisfy $V(y|z) = \Theta$ and $V(z|x) = \Omega$, then

$$E(yy'|x) = \int (\Theta + Hzz'H')\, p(z|x)dz = \Theta + H(\Omega + G'xx'G)H', \qquad (6)$$

and thus

$$V(y|x) = \Theta + H\Omega H'. \qquad (7)$$

Again (7) is true for model (3). For model (4) in addition we know that Θ is diagonal.

If y given z and z given x are both multivariate normal, then we get from (3) and (7) for the conditional density p(y|x) of y given x

$$(2\pi)^{-(m+p)/2}|\Theta|^{-1/2}|\Omega|^{-1/2} \int \exp\{-\tfrac{1}{2}[(y - Hz)'\Theta^{-1}(y - Hz) + (z - G'x)'\Omega^{-1}(z - G'x)]\}\, dz =$$

$$= (2\pi)^{-m/2}|\Theta + H\Omega H'|^{-1/2} \exp\{-\tfrac{1}{2}(y - HG'x)'(\Theta + H\Omega H')^{-1}(y - HG'x)\}. \qquad (8)$$

Again model (4) is the special case in which Θ is diagonal.

Thus we have introduced the basic model in various levels of generality. In the nonparametric case we have models (3) and (4), with (4) the restricted asymmetric model. In the case of linear regression we have (5), and homoscedasticity adds (6). In the strongest version of the model, which assumes multivariate normality, we have model (8). This again has a version with diagonal Θ and one with full Θ. A graph picturing the reduced rank regression model is given in Figure 1. It shows clearly how the effect of x on y is *mediated* by, or *filtered* by, z.

INSERT FIGURE 1 ABOUT HERE

We can also introduce (8) in a slightly different way, which connects our approach with multilevel analysis (Mason, Wong, and Entwisle, 1983, Aitkin and Longford, 1985, Goldstein, 1987, Goldstein and McDonald, 1988). Multilevel analysis is also known as random coefficient regression or empirical Bayes regression (see De Leeuw and Kreft, 1986, for references).

Suppose $y = H\beta + \varepsilon$, with ε normal with mean zero and dispersion Θ, while $\beta = G'x + \delta$, with δ again normal, independent of ε, with mean zero and covariance Ω. It follows that $y = HG'x + H\delta + \varepsilon$, which is exactly identical to (8). The regression coefficients β now play the role of the latent variables z. This shows that the random coefficient regression models can be interpreted as reduced rank regression models, or as regression models with latent variables. In this situation both G and H are usually known matrices, while it is commonly assumed that Θ is the scalar matrix $\sigma^2 I$.

Maximum likelihood estimation

If we assume multivariate normality, then we can use the result (8) to compute maximum likelihood estimates. The negative log likelihood is, except for irrelevant constants, equal to

$$\mathfrak{L}(G,H,\Theta,\Omega) =$$

$$= \ln \det \Sigma + n^{-1} \operatorname{tr} \Sigma^{-1}(Y-XGH')'(Y - XGH'), \tag{9}$$

with

$$\Sigma = H\Omega H' + \Theta. \tag{10}$$

Compare Jøreskog and Goldberger (1971). As a general point (Keller and Wansbeek, 1983, De Leeuw and Kreft, 1986) it is quite possible to interprete loss function (9) without actually referring back to the multivariate normal distribution or the principle of maximum likelihood. If $\Sigma = \sigma^2 I$, then (9) reduces to the ordinary least squares loss function $\operatorname{tr} (Y - XGH')'(Y - XGH')$, and in general weighted least squares can still be used if Σ is proportional to any known matrix. But if we do not want to make that assumption, we have to estimate Σ as well as the mean structure $Y = XGH'$. Loss function (9) measure the distance between Σ and $S = n^{-1}(Y - XGH')'(Y - XGH')$, in fact it defines an eminently reasonable metric on the cone of positive definite matrices. Thus, by minimizing (9) over its parameters, we obtain two goals at the same time. In the first place the dispersion of the residuals S is made as small as possible, and in the second place S and Σ

are made as close as possible. This generalizes the least squares idea, which merely concentrates on the dispersion of the residuals.

Minimizing (9) can be quite complicated in general, but there are some special cases in which the problem simplifies. If Θ is unrestricted, then Σ is unrestricted as well. It follows that the partial minimum of (9) is, except for constants,

$$L(G,H,*,*) = \ln \det (Y-XGH')'(Y - XGH'), \tag{11}$$

and this is minimized, under identification conditions $G'X'XG = I$, by solving

$$X'YH = X'XG\Omega, \tag{12a}$$

$$(Y'Y)^{-1}Y'XG = H. \tag{12b}$$

But (12) defines canonical correlation analysis (Bagozzi et al., 1981, Tso, 1981), which is consequently a special case of our general framework. Observe, however, that the role of X and Y is (12) is perfectly symmetric, which is because we have analyzed a model of form (3) and not of form (4). Thus canonical correlation analysis is not really what we want.

If Σ is known (or proportional to a known matrix) then minimizing (9) over G and H , again requiring $G'X'XG = I$, amounts to solving the problem

$$X'Y\Sigma^{-1}H = X'XG\Omega, \tag{13a}$$

$$Y'XG = H. \tag{13b}$$

which is a weighted version of redundancy analysis (Davies and Tso, 1982). De Leeuw, Mooijaart, and Van der Leeden (1985) discuss the case in which $\Omega = 0$ and Θ is restrricted to be diagonal, or simplex-like, or of factor analytic form, or whatever. The important point here is that if $\Omega = 0$, then the problem of estimating Σ and the problem of estimating G and H separate, and

we can consequently use *alternating maximum likelihood* or AML methods. We use this name because it describes the structure of the algorithm: we alternate iteratively over adjusting (G,H) and Σ, and because it emphasizes the analogy with the *alternating least squares* techniques usec in the nonlinear multivariate analysis methods of Gifi (in press).

The step of estimating Σ, under restrictions and with G and H currently considered known, is alternated with the step of estimating G and H with Σ currently known. The first step fits a covariance structure model to the current dispersion matrix of the residuals, the second step solves the weighted redundancy problem (13). This does not work if $\Omega \neq 0$, because in that case H occurs in both subproblems. If H is known, as in the Pothoff-Roy (1964) growth curve models or the random coefficient regression problems, this is no problem, and alternating maximum likelihood can still be used. But if H is (partially) unknown the subproblems are confounded, and other more complicated optimization methods must be used.

The EM algorithm

It seems as if using maximum likelihood methods does not give estimates of the scores on the latent variables, but this is only apparently so. In the first place we can simple set $Z = XG$, and use this as the estimate of the scores. In the second place we can use (8) to derive a different form of the loglikelihood function, which is also very useful for computational purposes. This amounts to a new derivation of the EM-algorithm (Dempster, Laird, and Rubin, 1973), which has been applied earlier in this context by Chen (1981). We present the relevant argument here, because the EM-algorithm is often discussed in purely statistical terms, which makes its simple computational structure somewhat mysterious. Our derivation uses the concavity of the logarithm (Jensen's inequality), together with the general idea of majorization, which is a very useful methodology to extend the scope of linear optimization techniques (compare De Leeuw, 1988b). A careful observer will note that we tried, in the previous paragraph, to deemphasize the statistical interpretation of the likelihood function, and its role in so-called inference. In this paragraph we try to deemphasize the statistical interpretation of the EM algorithm, by reformulating in perfectly general algorithmic terms. The reason for this shift of emphasis is that we think that fitting mean structures and covariances structures to observed data in this way

makes geometrical and computational sense, even if the assumption of multivariate normality does not.

Ignoring irrelevant constants we find, using $\pi_i(G,H,\Theta,\Omega)$ for the density (8) at (x_i,y_i,z_i),

$$\mathfrak{L} = \Sigma_{i=1}^{n} \ln \int \pi_i(G,H,\Theta,\Omega)\, dz_i = \tfrac{1}{2}n \ln \det(\Theta) + \tfrac{1}{2}n \ln \det(\Omega) +$$

$$- \Sigma_{i=1}^{n} \ln \int \exp\{-\tfrac{1}{2}[(y_i - Hz_i)'\Theta^{-1}(y_i - Hz_i) + (z_i - G'x_i)'\Omega^{-1}(z_i - G'x_i)]\}dz_i. \qquad (14)$$

Suppose we use underlining for current, tentative estimates of the parameters, and $\pi_i(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega})$ for the density with these current estimates. Then, by concavity of the logarithm,

$$\ln \int \pi_i(G,H,\Theta,\Omega)\, dz_i - \ln \int \pi_i(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega})\, dz_i =$$

$$\geq [\int \pi_i(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega}) \ln \{\pi_i(G,H,\Theta,\Omega)/\pi_i(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega})\}\, dz_i]/[\int \pi_i(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega})\, dz_i]. \qquad (15)$$

Now let $\eta_i(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega}) = \pi_i(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega})/[\int \pi_i(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega})\, dz_i]$. Then

$$\ln \int \pi_i(G,H,\Theta,\Omega)\, dz_i \geq \ln \int \pi_i(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega})\, dz_i +$$

$$\int \eta_i(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega}) \ln \pi_i(G,H,\Theta,\Omega)\, dz_i - \int \eta_i(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega}) \ln \pi_i(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega})\, dz_i. \qquad (16)$$

We have equality in (16) if $(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega}) = (G,H,\Theta,\Omega)$. In fact, because the logarithm is strictly concave, this condition is actually necessary and sufficient for equality.

If we sum both sides of (16) over the observations we find an inequality of the form

$$\mathfrak{L}(G,H,\Theta,\Omega) \geq \mathfrak{L}(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega}) +$$

$$+ \{\Delta(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega};G,H,\Theta,\Omega) - \Delta(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega};\underline{G},\underline{H},\underline{\Theta},\underline{\Omega})\}. \qquad (17)$$

Now suppose that we find improvements of $(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega})$ by maximizing $\Delta(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega};G,H,\Theta,\Omega)$ over (G,H,Θ,Ω). This gives updates $(G^+,H^+,\Theta^+,\Omega^+)$. By definition

$$\Delta(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega};G^+,H^+,\Theta^+,\Omega^+) \geq \Delta(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega};\underline{G},\underline{H},\underline{\Theta},\underline{\Omega}), \tag{18}$$

and thus, by (17),

$$\mathfrak{L}(G^+,H^+,\Theta^+,\Omega^+) \geq \mathfrak{L}(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega}), \tag{19}$$

with equality if and only if $(G^+,H^+,\Theta^+,\Omega^+) = (\underline{G},\underline{H},\underline{\Theta},\underline{\Omega})$. This, together with the continuity of the update mapping, proves convergence to a stationary point of the likelihood function (Zangwill, 1969).

It remains to show that maximizing $\Delta(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega};G,H,\Theta,\Omega)$ is fairly simple. In the first place observe that $\eta_i(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega})$ is the conditional density of z_i, given x_i and y_i. It is thus normal, with mean vector of the form $\underline{m}_i = \underline{A}x_i + \underline{B}y_i$ and dispersion $\underline{W}$. Straightforward computation gives

$$\underline{A} = (I - \underline{\Omega H}'\underline{\Lambda}^{-1}\underline{H})\underline{G}', \tag{20a}$$

$$\underline{B} = \underline{\Omega H}'\underline{\Lambda}^{-1}, \tag{20b}$$

$$\underline{W} = \underline{\Omega} - \underline{\Omega H}'\underline{\Lambda}^{-1}\underline{H\Omega}, \tag{20c}$$

with

$$\underline{\Lambda} = \underline{H}'\underline{\Omega H} + \underline{\Theta}. \tag{21}$$

Also

$$\ln \pi_i(G,H,\Theta,\Omega) = -\tfrac{1}{2} \ln \det(\Theta) - \tfrac{1}{2} \ln \det(\Omega) +$$

$$- \tfrac{1}{2}[(y_i - Hz_i)'\Theta^{-1}(y_i - Hz_i) + (z_i - G'x_i)'\Omega^{-1}(z_i - G'x_i)]. \qquad (22)$$

Thus, taking conditional expectations $\underline{E}$ with respect to z_i, for given x_i and y_i, at the current parameter values $(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega})$,

$$-2\Delta(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega};G,H,\Theta,\Omega) = \ln\det(\Theta) + \ln\det(\Omega) +$$

$$+ \operatorname{tr}\Theta^{-1}\underline{E}\{(y_i - Hz_i)(y_i - Hz_i)'\} + \operatorname{tr}\Omega^{-1}\underline{E}\{(z_i - G'x_i)(z_i - G'x_i)'\}. \qquad (23)$$

It is now straightforward, although somewhat tedious, to develop the algorithm from here by summation over the observations, and by collecting terms. We use the conditional means $\underline{m}_i$ and the conditional dispersions $\underline{W}$.

We find that we now can separate the estimation of G and H from that of Θ and Ω, even in the more complicated models, at the price of using approximations which will undoubtedly slow down the convergence. In order to find a new Θ, for temporarily fixed G and H, we have to minimize $\ln\det(\Theta) + \operatorname{tr}\Theta^{-1}\underline{S}$, where

$$\underline{S} = H\underline{W}H' + n^{-1}\Sigma_{i=1}^{n}(y_i - H\underline{m}_i)(y_i - H\underline{m}_i)', \qquad (24a)$$

and in order to find a new Ω we have to minimize $\ln\det(\Omega) + \operatorname{tr}\Omega^{-1}\underline{T}$, where

$$\underline{T} = \underline{W} + n^{-1}\Sigma_{i=1}^{n}(\underline{m}_i - G'x_i)(\underline{m}_i - G'x_i)'. \qquad (24b)$$

This will be simple if the restrictions on Ω and Θ are simple, but in any case we know how to solve subproblems like these. In order to find the optimal G and H for temporarily fixed Ω and Θ we have to minimize $\operatorname{tr}\Theta^{-1}\underline{S}$ and $\operatorname{tr}\Omega^{-1}\underline{T}$, which are linear regression problems.

The structure of the algorithm is now clear. We start, for current $(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega})$, to compute the conditional means and variances $\underline{m}_i$ and $\underline{W}$ using (20) and (21). Then we compute $\underline{S}$ and $\underline{T}$ by using (24). Construct the auxilary loss function $\Delta_1 = \ln\det(\Theta) + \operatorname{tr}\Theta^{-1}\underline{S}$, and minimize this over

Θ and H (by AML). Also construct the auxilary $\Delta_2 = \ln \det(\Omega) + \text{tr}\ \Omega^{-1}\underline{T}$, and minimize over Ω and G (by AML). It is not necessary to actually carry out the minimizations of the auxilaries completely, in fact this would not be possible as a subset. Usually we only carry out one AML cycle for each auxilary, and then proceed to use the new values as substitutes for $(\underline{G},\underline{H},\underline{\Theta},\underline{\Omega})$ in a new majorization. In the random regression case, in which G and H are known, this reduces to the EM algorithm used, for instance, by Mason, Anderson, and Hayat (1988), and by Raudenbusch, Bryk, Seltzer, and Congdon (1988).

Optimal scaling

This algorithm, complicated as it is, does not yet exhaust the scope of alternating maximum likelihood in this conbtext. By a conceptually very easy extension of these general principles we now incorporate transformation of the variables, a.k.a *optimal scaling* . This uses basically the same principles as the familiar Box-Cox (1964) approach in regression analysis. In the context of regression analysis, path analysis, and structural equations modeling these procedures have been proposed by De Leeuw (1986), Van Wijk (1987), Mooijaart, Meyerink, and De Leeuw (1988).

The assumption we now make is that we have a reduced rank regression model exactly like the one above, except that what used to be observed variables originally now become latent variables as well. We now write η where we first wrote y and ξ where we wrote x. Thus the general model becomes

$$p(\eta|\xi) = \int \Pi_{j=1}^{m}\ p(\eta_j|z)\ p(z|\xi)dz. \tag{25}$$

This is not enough, of course, because we have to introduce observed variables at some point. The observed variables, which are still written as x and y, are in a one-one correspondence with ξ and η, and we assume

$$p(x_j|\xi, \eta, z) = p(x_j|\xi_j), \tag{26a}$$

$$p(y_j|\xi, \eta, z) = p(y_j|\xi_j). \tag{26b}$$

This means that x_j only depends on ξ_j, and y_j only depends on η_j. From this we find

$$p(x,y) = \int\int \Pi_{j=1}^{m} p(y_j|\eta_j) \Pi_{j=1}^{m} p(x_j|\xi_j)\ p(\eta|\xi)p(\xi)\ d\xi d\eta. \tag{27}$$

Equation (27) is the general formulation of the model, which does not assume linearity, homoscedasticity, or normality. If we want to make this much more specific we assume multivariate normality of all latent variables, as before. We also assume $\phi_j(y_j) = \eta_j$ and $\psi_j(x_j) = \xi_j$, with ϕ_j and ψ_j continuously differentiable and increasing, but otherwise unknown. Collect them into vector functions Φ and Ψ. Now suppose F is the cdf of x and y, and U is the cdf of ξ and η. Then

$$F(a,b) = \text{prob}(x < a,\ y < b) = \text{prob}(\Psi(x) < \Psi(a),\ \Phi(y) < \Phi(b)) =$$

$$\text{prob}(\xi < \Psi(a),\ \eta < \Phi(b)) = U(\Psi(a),\ \Phi(b)). \tag{28}$$

Thus we find for the density

$$\partial^2 F(a,b) = \partial^2 U(\Psi(a),\ \Phi(b))\partial\Psi(a)\partial\Phi(b). \tag{29}$$

It follows that the log likelihood for observation i is given by our previous log likelihood (14) evaluated at $(\Phi(y_i),\Psi(x_i),z_i)$, minus one half times the logarithm of the Jacobians ln $\det(\partial\Phi)$ + ln $\det(\partial\Psi)$. We can now maximize the likelihood over the usual parameters (G,H,Θ,Ω), as well as over the transformations Φ and Ψ. Again the problem separates nicely, in the sense for given transformations we are in back in the situation of the previous sections, and we can use the AML methods developed there.

Computing optimal transformations for given (G,H,Θ,Ω) is somewhat less simple. In fact we deal with a semi-nonparametric problem here, because the space of all possible transformations has infinite dimensionality. Thus we have to use finite dimensional approximations in this case.

in the literature mentioned above we use monotone B-splines, which are very similar to the M-splines recently discussed by Ramsay (1988). The monotone B-splines are positive linear combinations of a given number of finite number of basic splines, with the coefficients of the linear combinations chosen to be increasing. The same type of representation is true for the derivatives. It can be shown that the problem of minimizing the negative log likelihood over the coefficients of the B-spline representation is a convex minimization problem (with simple linear inequality constraints), and efficient Newton-Raphson type methods are available to solve such problems. This makes it quite feasible to solve the corresponding AML substep efficiently.

Discussion

This paper has various contributions. We offer an algorithmic interpretation and reformulation of the multinormal likelihood function and the EM-algorithm, and we discuss a class of models which have canonical correlation analysis, redundancy analysis, reduced rank regression analysis, and multilevel analysis as special cases. The alternating least squares algorithms of the Gifi system, and of the PLS approach by Wold, are replaced by very similar (although slightly more complicated) alternating maximum likelihood methods. The alternating maximum likelihood methods, which alternate rescaling of the variables with optimization over the regression parameters, seem to be a promising alternative to the techniques based on least squares, in particular because the marginalization implicit in the likelihood function does not lead to some of the problems with incidental parameters that occur in the PLS/ALS approach.

References

Aitkin, M.A., & Longford, N.T. (1985) Statistical modeling issues in school effectiveness studies. Journal of the Royal Statistical Society, A149, 1-43.

Anderson, T.W. (1951). Estimating linear restrictions on regression coefficients for multi-variate normal distributions. Annals of mathematical statistics, 22, 327-351.

Anderson, T.W. (1984). Estimating linear statistical relationships. Annals of Statistics, 12, 1 - 45.

Bagozzi, R.P., Fornell, C., and Larcker, D.F. (1981). Canonical correlation analysis as a special case of a strctural relations model. Multivariate Behavioural Research, 16, 437-454.

Bryk, A.S., Raudenbusch, S.W., Seltzer, M., & Congdon, R.T. (1988). An introduction to HLM: computer program and users' guide. University of Chicago.

Chen, C.F. (1981). The EM approach to the multiple indicator multiple causes model via estimation of the latent variables. Journal of the American Statistical Association, 76, 704-708.

Davies, P.T. and Tso, M.K.-S. (1982) Procedures for reduced rank regression. Applied Statistics, 31, 244-255.

De Leeuw, J. (1984). Models of data. Kwantitatieve Methoden, 13, 17-30.

De Leeuw, J. (1986). Regression with Optimal Scaling of the Dependent Variable. Research Report 86-08. Department of Data Theory, University of Leiden.

De Leeuw, J, & Kreft, G.G. (1986) Random coefficient models for multilevel analysis. Journal of Educational Statistics, 11, 57-88.

De Leeuw, J., & Bijleveld, C. (1987). Fitting reduced rank regression models by alternating least squares. Research Report 87-05. Department of Data Theory, University of Leiden.

De Leeuw, J. (1988a). Multivariate Analysis with Linearization of the Regressions, Psychometrika, 53, 437 - 454.

De Leeuw, J. (1988b). Convergence of the majorization method for multidimensional scaling. Journal of Classification, 5, 163-180.

De Leeuw, J., & Bijleveld, C. (in press). Fitting longitudinal reduced rank regression models by alternating least squares. Psychometrika, in press.

De Leeuw, J., Mooijaart, A, & Van der Leeden, R. (1985). Fixed factor score models with linear restrictions. Research Report 85-06. Department of Data Theory, University of Leiden.

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

Gifi, A. (in press). Nonlinear Multivariate Analysis. New York, Wiley.

Goldstein, H. (1987) Multilevel models in educational and social research. London, Griffin.

Goldstein, H. & MacDonald, R.P. (1988). A general model for the analysis of multilevel data. Psychometrika, 53, 455-467.

Guttman, L. (1985). The illogic of statistical inference for cumulative science, Applied Stochastic Models and Data Analysis, 1, 3-9.

Izenman, A.J. (1965). Reduced-rank regression for the multivariate linear model. Journal of Multivariate Analysis, 5, 248-264.

Jøreskog, K.G., & Goldberger, A.S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable, Journal of the American Statistical Association,70,631-639.

Kalman, R.E. (1983). Identifiability and modeling in econometrics. In: Krishnaiah, (Ed.) Developments in Statistics, vol 4, Amsterdam, North Holland.

Keller, W.J., & Wansbeek, T. (1983). Multivariate methods for quantitative and qualitative data. Journal of Econometrics, 22, 91-111.

Mason, W.M., Wong, G.Y, & Entwisle, B. (1983) Contextual analysis through the multilevel linear model. In S. Leinhardt (ed.), Sociological Methodology 1983. San Francisco, Jossey-Bass.

Mason, W.M., Anderson, A. F., & Hayat, N. (1988), Manual for GENMOD, Population Studies Center, University of Michigan.

Mooijaart, A., Meijerink, F. & De Leeuw, J. (1988). Non-linear Recursive Pathmodels. In preparation.

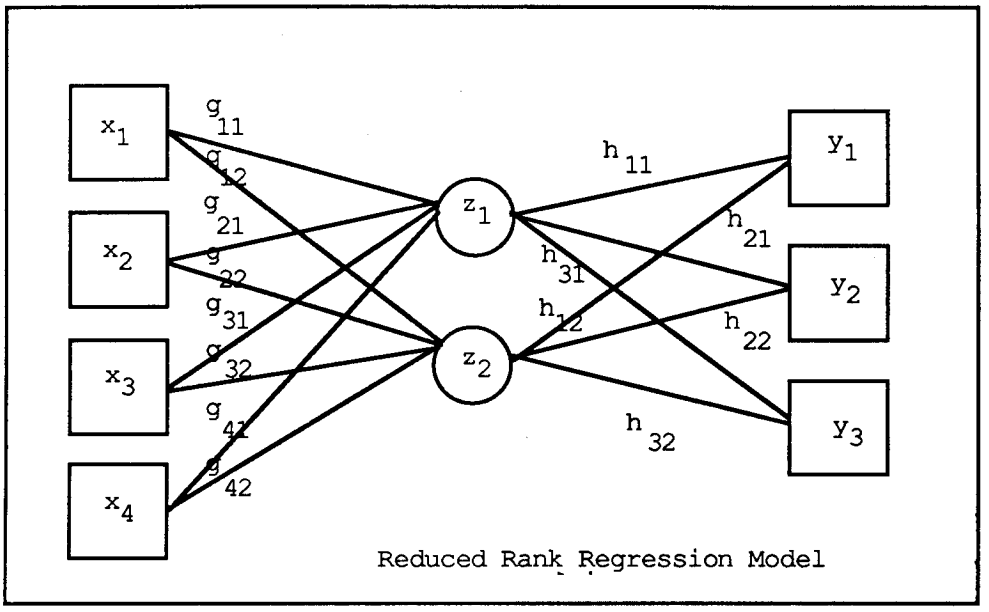
Pothoff, R.F. & Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika, 51, 313-326.

Ramsay, J.O. (1988). Monotone regression splines in action. Statistical Science, 3, 425-460.

Tso, M.K.-S. (1981). Reduced rank regression and canonical analysis. Journal of the Royal Statistical Society, 43, 183-189.

Van Wijk, M. (1987). An ML-method for Non-linear Path models. Leiden: Doctoral Thesis.

Zangwill, W.I. (1969). Nonlinear Programming. Englewood Cliffs, Prentice Hall.


x1
x2
x3
x4
g11
g12
g21
g22
g31
g32
g41
g42
z1
z2
h11
h21
h31
h12
h22
h32
y1
y2
y3
Reduced Rank Regression Model