Reduced Rank Regression Models with Structured Errors

Ab Mooijaart [1)]
Rien van der Leeden [1)]
Jan de Leeuw [2)]

2911
-112

[1)]Department of Psycholgy, University of Leiden, Hooigracht 15, 2312 KM, Leiden, The Netherlands.
[2)]Departments of Psychology and Mathematics, University of California Los Angeles, 405 Hilgard Avenue, Los Angeles CA 90024, USA.

In reduced rank regression models it is assuned that the matrix of regression weights is not of full rank. In multivariate regression models the covariance matrix of the error variables is commonly restricted in some way. In this paper a reduced rank regression model is presented in which a very large class of covariance structures can be formulated for the error covariance matrix. A number of analysis models are shown to be special cases of this general approach, for instance, canonical correlation analysis, principal components analysis and redundancy anlysis. One interesting interpretation is the case in which the sample is divided into different groups and the same covariance structure is fitted for each group. At the same time these groups are found optimally. A simple algorithm is presented which is in fact an alternating least squares algorithm Data from a study on epidemiologics are used to illustrate the model.

Key words: reduced rank regression, covariance structure models, functional relationships, alternating least squares, LISREL, MANOVA.

1. General Model

In this section we will consider the following general multivariate regression model.

$$\mathbf{y}_i = \mathbf{M}'\mathbf{x}_i + \mathbf{e}_i . \tag{1.1}$$

For i = 1,...,N, the $\mathbf{y}_i$ and $\mathbf{e}_i$ are vectors of dimension t, the $\mathbf{x}_i$ are vectors of dimension r. It is assumed that the $\mathbf{x}_i$ vectors contain fixed or non-stochastic elements, while the $\mathbf{y}_i$ vectors contain random or stochastic elements. The matrix $\mathbf{M}$ is of the order r x t and contains regression coefficients. The disturbances $\mathbf{e}_i$ are independent, normally distributed random vectors with zero expectation and dispersion matrix S. If we let N be the sample size, the $\mathbf{y}_i$ and $\mathbf{x}_i$ can be called multi-response vectors, which contain N observations on t and r variables. In matrix notation the model can be written as

$$\mathbf{Y} = \mathbf{XM} + \mathbf{E}. \tag{1.2}$$

In one terminology, this equation specifies a *functional* model (see Kendall & Stuart, 1979, vol.2).This means that the observations are not conceived as random replications, as is the case with *structural* models.

In multivariate regression models, the dispersion matrix S is commonly restricted in some way. The most specific assumption is that S is known, or that $S = sS_0$, with s unknown and S_0 being a known matrix. Other assumptions that have been considered (Rao, 1967; Gleser & Olkin, 1970a, 1970b) are, for instance, that S is an unknown, arbitrary positive definite matrix, or that S is unknown but has some special structure. One can think of, for instance, S known to be diagonal, or to have intraclass correlation structure.

In this paper we will consider multivariate regression models in which a structural equation model is fitted on the dispersion matrix S. This means that a large class of covariance structures is possible for this matrix. This model can be formulated as follows,

$$\begin{aligned} \mathbf{e}_i &= \mathsf{L}h_i + d_i \\ \mathbf{0} &= \mathbf{B}_0 h_i + x_i , \end{aligned} \tag{1.3b}$$

in which $\mathbf{B}_0 = \mathbf{I} - \mathbf{B}$. In a reduced form, the model can be written as

$$\mathbf{e}_i = \mathsf{L}\mathbf{B}^{-1}x_i + d_i . \tag{1.4}$$

This yields the following covariance equation for S,

$$\mathsf{S} = \mathsf{L}\mathbf{B}^{-1}\mathsf{F}(\mathbf{B}^{-1})'\mathsf{L}' + \mathsf{Y},$$

in which $\mathsf{E}(xx') = \mathsf{F}$ and $\mathsf{E}(dd') = \mathsf{Y}$.
Usually, as in the well known LISREL framework (Jöreskog & Sörbom, 1984), this model is specified with three equations, using eight parameter matrices. The formulation chosen here is a shorthand, which requires less parameter matrices. It does not, however, limit the possibilities for specifying models.
The model that will be considered is further generalized by restrictions on the matrix **M**. In this paper it is assumed that rank (**M**) $\leq s \leq$ min (r,t). This rank restriction can be expressed, writing **M** as a product matrix **AB**, where **A** is of the order r x s and B is of the order s x t. The rank of **M** is restricted by choosing s.
In that way the so-called reduced rank regression model is defined (Anderson, 1951, 1984; Izenman, 1975; Tso, 1981; Davies & Tso, 1982). In a later section, some additional remarks are made on this topic.
Having the possibilities of imposing the restrictions on the matrices S and **M**, described above, yields a very general model. A great many of multivariate analysis techniques can be interpreted as special cases of it.
The model can be summarized as follows.

$$\mathbf{Y} = \mathbf{XAB} + \mathbf{E}, \tag{1.6a}$$
the columns of **E** are i.i.d. N (0,S), (1.6b)
$\mathsf{S} = \mathsf{v}(q)$, q unknown, (1.6c)
rank (**M**) $\leq s$. (1.6d)

For convenience it is assumed that the columns of both **X** and **Y** have zero mean.
Another way of formulating the rank restriction on **M** is also useful. Rank (**M**) $\leq s$ can be written as $\mathbf{V}'\mathbf{M} = 0$, with **V** being an unknown matrix of order r x (r -s), or as $\mathbf{MW} = 0$, with **W** being an unknown matrix of order t x (t -s) . Matrices **V** and **W** are supposed to be of full column rank. This is a natural method of formulating the model if our prior knowledge consists of linear restrictions on the regression coefficients. Keller and Wansbeek (1983) call this the principal relations specification of the model, while (1.6) is called the principal factors specification.

One could argue that the modelling of error-covariances with a structural equation model may be somewhat far-fetched. However, in our opinion, we are not 'just modelling error'. Instead of minimizing the difference (**Y** - **XAB**) and thus predicting as much as possible of the variables contained in **Y**, we are 'filtering out' some chosen effects of **Y**. This is done by both specifying **X** in the right way and chosing the proper restrictions on **M**. On the remaining covariances the structural equation model (1.4) is fitted. We will clarify this interpretation in a later section, also showing the relationship of this approach with other techniques.

2. History of Reduced Rank Regression

The history of reduced rank regression is rather complicated. In his thesis of 1945, Anderson already obtained many of the basic results. These were published in Anderson (1951). He considers the general multivariate linear model with linear restrictions on the matrix **M**, and no restrictions on Σ. The maximum likelihood estimates are obtained and the likelihood ratio tests for the rank restrictions are derived. Anderson observed that the 'errors-in-variables' models from econometrics and the discriminant analysis models of Fisher (1938), were special cases of the general set-up.. He also indicated the relationship of the likelihood equations with canonical correlation analysis.

Davies and Tso (1982) recently reintroduced the reduced rank regression model with known Σ. They showed it to be related to principal component analysis. Tso (1981) studied the case with unknown Σ and rederived the relationship with canonical correlation analysis.

Izenman (1975) studies the reduced rank model (1.6) in which both **X** and **Y** are stochastic. The parameters are estimated by minimizing a criterion defined in terms of the latent roots of the dispersion matrix of the residuals. In this stochastic version of the model, the asymptotic theory is much simpler. Izenman also points out the relationship with canonical analysis. In a recent paper Bagozzi et al. (1981) show, independently, that this means that canonical analysis is a special case of LISREL and that the model described in equation (1.6), with both **X** and **Y** stochastic, can be considered as a LISREL interpretation of canonical correlation analysis.

Thus we see that reduced rank regression has been studied in the case in which Σ is known and in the case in which Σ is unknown, but not really in intermediate cases in which there is partial knowledge about Σ. We also see that both the structural case (random regressors) and the functional case (fixed regressors) have been studied.

3. Relationships with Other Models

As we already mentioned in the short historical section, special cases of the model have been around for a long time. The errors-in-variables model , for example, dates back to Adcock (1878). It is, in our notation, and in the functional form, $\mathbf{Y} = \mathbf{M} + \mathbf{E}$, with **Y** stochastic, with the rows of **E** i.i.d. N (0,Σ) and with rank (**M**) $\leq s$. This is the special case of (1.2), in which the matrix **X** equals the identity (or any other nonsingular matrix). In orthogonal regression ,

dating back to Pearson (1901), it is usually assumed that $\Sigma = s\mathbf{I}$. In the model studied by Frisch (1934) and Koopmans (1937), Σ is an unknown diagonal matrix. In the fixed factor analysis model the same assumptions are made, although factor analysis usually concentrates on data with small s (the number of factors) and the Frisch model is mainly concerned with data with large s. The algebraic results connecting the Frisch model with the factor analysis model, have been reviewed recently by Bekker and De Leeuw (1987). The statistical theory for errors-in-variables models is discussed in considerable detail by Gleser (1981) and Anderson (1984). This paper by Anderson also discusses the corresponding structural models.
For the sake of clearness we will now first present an algorithm for the estimation of the parameters of our general multivariate reduced rank regression model. Some of the relationships described above, may then be clarified, while other relationships may become more clear when they are treated as special cases of the algorithm. These special cases, which result in a simplified estimation procedure, directly relate the model to other techniques.

4. Algorithm

In this section an algorithm is presented for the estimation of the parameters for the general multivariate reduced rank regression model formulated in equation (1.6).
In fact, it is a simple, straightforward algorithm. Under the assumption that the observed variables $\mathbf{Y}$ are normally distributed, the likelihood function is optimized. However, as we have already mentioned, our model describes *functional* relationships. Anderson and Rubin (1956) showed, in the framework of factor analysis, that in these cases the likelihood function is unbounded and maximum likelihood estimates do not exist (see also Anderson, 1984). In general, this problem can be solved imposing restrictions on the parameters. It can be shown that the restrictions $\mathbf{XAB}$, imposed in our model, are sufficient restrictions to cause the likelihood function to be bounded. Thus, we can derive maximum likelihood estimates. In a later section we will make some comments on possible choices of linear restrictions and their interpretation.
The estimation procedure proposed here, consists of several steps. One could call it an alternating maximum likelihood procedure (because of the strong ressemblance with alternating least squares techniques).
Note that it is assumed that $\mathbf{X}$ and $\mathbf{Y}$ both have zero column means. Also, without restrictions, $\mathbf{A}$ and $\mathbf{B}$ are not uniquely determined. One could write the model as $\mathbf{Y} = \mathbf{XAZZ}^{-1}\mathbf{B} + \mathbf{E}$, where $\mathbf{Z}$ can be any square, nonsingular matrix of proper order. So $\mathbf{A}^* = \mathbf{AZ}$ and $\mathbf{B}^* = \mathbf{Z}^{-1}\mathbf{B}$ are proper solutions too. If the matrix $\mathbf{XA}$ is required to be orthogonal, i.e. $\mathbf{A'X'XA} = \mathbf{I}$, then it holds that $\mathbf{A}^{*\prime}\mathbf{X'XA}^* = \mathbf{Z'A'X'XAZ} = \mathbf{Z'Z} = \mathbf{I}$. Thus $\mathbf{Z}$ is orthogonal. This means that $\mathbf{A}$ and $\mathbf{B}$ are determined up to an orthogonal rotation.
The function to be maximized can be written as

$$f(\mathbf{A},\mathbf{B},\Sigma) = c + N \ln|\Sigma^{-1}| - \mathrm{tr}[(\mathbf{Y} - \mathbf{XAB})\Sigma^{-1}(\mathbf{Y} - \mathbf{XAB})'], \tag{4.1}$$

where $c = -.5N \ln 2\pi$. This function f equals twice the ln likelihood function.The matrix Σ is a

function of the unknown parameters in q, say $S = v(q)$. Note that maximization of f is equivalent to minimizing the sum of squared differences $(\mathbf{Y} - \mathbf{XAB})$, weighted by S^{-1}. So in First we will find an expression for $\hat{\mathbf{B}}$ in terms of $\mathbf{A}$ and q. To find the maximum of the function f,we will minimize the trace term in f. Let this term be the function f^*. This function can be rewritten as

$$f^*(\mathbf{A},\mathbf{B},q) = \text{tr}[\mathbf{Y}S^{-1}\mathbf{Y}' - 2\mathbf{Y}S^{-1}\mathbf{B}'\mathbf{A}'\mathbf{X}' + \mathbf{XAB}S^{-1}\mathbf{B}'\mathbf{A}'\mathbf{X}']. \tag{4.2}$$

To find the values for $\mathbf{B}$ that minimize f^*, the partial derivative of f^* w.r.t. $\mathbf{B}$ should be zero. This yields

$$\partial f^*/\partial \mathbf{B} = 2S^{-1}(\mathbf{A}'\mathbf{X}'\mathbf{Y} - \mathbf{A}'\mathbf{X}'\mathbf{XAB}) = 0.$$

So we find as an estimator for $\mathbf{B}$

$$\hat{\mathbf{B}} = \mathbf{A}'\mathbf{X}'\mathbf{Y}, \tag{4.4}$$

under the condition $\mathbf{A}'\mathbf{X}'\mathbf{XA} = \mathbf{I}$. Note that $\hat{\mathbf{B}}$ consists of the regression coefficients of $\mathbf{Y}$ on $\mathbf{XA}$. Although $\hat{\mathbf{B}}$ is numerically still unknown, we have found an expression for it by means of which the optimization of f is simplified. Instead of maximizing $f(\mathbf{A},\mathbf{B},q)$, the function $g(\mathbf{A},q)$ will be maximized. This function can be written as

$$g(\mathbf{A},q) = \max_{\mathbf{B}} f(\mathbf{A},\mathbf{B},q)$$

$$= c + N \ln|S^{-1}| - \text{tr}[(\mathbf{Y} - \mathbf{XAA}'\mathbf{X}'\mathbf{Y})S^{-1}(\mathbf{Y} - \mathbf{XAA}'\mathbf{X}'\mathbf{Y})'], \tag{4.5}$$

which simplifies into

The optimization of $g(\mathbf{A},q)$ is done stepwise. In step 1, $\mathbf{A}$ is estimated for given q. Then, in step 2, q is estimated for given $\mathbf{A}$. These steps are repeated and in each step, the parameters being held fixed, are updated with the result of the previous step. The estimation procedure stops when a certain convergence criterium has been reached.

Step 1. Find $\hat{\mathbf{A}}$ for given q. This means that the function $g(\mathbf{A}|q)$ has to be maximized. This is equal to minimizing the function

$$g^*(\mathbf{A}|q) = \text{tr}[\mathbf{Y}'(\mathbf{I} - \mathbf{XAA}'\mathbf{X}')\mathbf{Y}S^{-1}], \tag{4.7}$$

under the condition $\mathbf{A}'\mathbf{X}'\mathbf{XA} = \mathbf{I}$. The function g^* can be rewritten in such way that minimizing g^* is equal to maximizing the function

$$g^{**}(\mathbf{A}|\mathsf{q}) = \mathrm{tr}[\mathbf{Y}'\mathbf{X}\mathbf{A}\mathbf{A}'\mathbf{X}'\mathbf{Y}\mathsf{S}^{-1}]$$

Now let $\mathbf{X}$ being decomposed by a Gram-Schmidt orthogonalization as $\mathbf{X} = \mathbf{X}^*\mathbf{T}$, where the columns of $\mathbf{X}^*$ are orthogonal and $\mathbf{T}$ is an upper-triangular matrix. The function g^{**} can then be written as

$$g^{**}(\mathbf{A}|\mathsf{q}) = \mathrm{tr}[\mathbf{A}'\mathbf{T}'\mathbf{X}^{*\prime}\mathbf{Y}\mathsf{S}^{-1}\mathbf{Y}'\mathbf{X}^*\mathbf{T}\mathbf{A}],$$

(4.9)

where $\mathbf{A}'\mathbf{T}'\mathbf{T}\mathbf{A} = \mathbf{I}$. If we define $\mathbf{P} = \mathbf{X}^{*\prime}\mathbf{Y}\mathsf{S}^{-1}\mathbf{Y}'\mathbf{X}^*$, the function g^{**} simplifies into

This function has to be maximized for orthogonal $\mathbf{TA}$. Because the matrix $\mathbf{A}$ has s columns, g^{**} is obviously maximized, if we let $\mathbf{TA}$ be the eigenvectors corresponding to the s largest eigenvalues of $\mathbf{P}$. Let $\mathbf{F}$ be these eigenvectors, then we can write as an estimator for $\mathbf{A}$,

$$\hat{\mathbf{A}} = \mathbf{T}^{-1}\mathbf{F} \tag{4.11}$$

Step 2. Find $\hat{\mathsf{q}}$ for given $\mathbf{A}$. This means the function $g(\mathbf{A},\mathsf{q})$ has to be maximized for given $\mathbf{A}$. We can write this function as

$$g(\mathsf{q}|\mathbf{A}) = c + N \ln|\mathsf{S}^{-1}| - N \,\mathrm{tr}[1/N\ \mathbf{Y}'(\mathbf{I} - \mathbf{X}\mathbf{A}\mathbf{A}'\mathbf{X}')\mathbf{Y}\mathsf{S}^{-1}], \tag{4.12}$$

where the addition of N is just for convenience. Define $\mathbf{Q} = 1/N\ \mathbf{Y}'(\mathbf{I} - \mathbf{X}\mathbf{A}\mathbf{A}'\mathbf{X}')\mathbf{Y}$. For given $\mathbf{A}$, the matrix $\mathbf{Q}$ is known and can be considered a temporary estimate of S. Now we can rewrite $g(\mathsf{q}|\mathbf{A})$ in such way, that maximizing $g(\mathsf{q}|\mathbf{A})$ is equivalent to minimizing the function

$$g^*(\mathsf{q}|\mathbf{A}) = \ln|\mathsf{S}| + \mathrm{tr}[\mathbf{Q}\mathsf{S}^{-1}]. \tag{4.13}$$

In this function S is a function of the unknown parameters in q. So for minimization of g^*, we have to consider this 'modelling' of S. The minimization of g^* may be done using a general algorithm such as the Fletcher-Powell algorithm (Fletcher and Powell, 1963). However, for some special cases the derivative of g^* can easily be obtained and estimates for q can be found in a simple way.

We can distinguish between the following cases :

Case 1. S is a diagonal matrix with unknown elements. The model then defines *factor analysis*. We have a factor model for $\mathbf{Y}$, with the special provision that the factors are on the space of the variables contained in $\mathbf{X}$ (De Leeuw, Mooijaart & Van der Leeden, 1985).

The function g^* can now be simplified and written as

$$g^{**}(q|\mathbf{A}) = \sum_{i=1}^{t} \ln(s_{ii}) + \sum_{i=1}^{t} \mathbf{q}_{ii} s_{ii}^{-1} \tag{4.14}$$

For the minimization of g^{**}, we consider the derivative of g^{**} with respect to s: $\partial g^{**}/\partial s$. This derivative can be written as

$$\partial g^{**}/\partial s = \sum_{i=1}^{t} s_{ii}^{-1} - \sum_{i=1}^{t} \mathbf{q}_{ii} s_{ii}^{-2} \tag{4.15}$$

and set to zero, it yields as an estimator for s,

$$\hat{s}_{ii} = \mathbf{q}_{ii} \tag{4.16}$$

Case 2 . $S = sS_0$, where S_0 is a known matrix and s is unknown. So S is proportional to a parameter s, which is to be estimated. In other words, S is being structured by S_0 . The function g^* (4.13) can now be written as

$$g^*(q|\mathbf{A}) = \ln|sS_0| + \mathrm{tr}[\mathbf{Q}(sS_0)^{-1}], \tag{4.17}$$

which has to be minimized over s. Writing out the ln and trace term in equation (4.17), yields

$$g^*(q|\mathbf{A}) = \ln|s^t| + \ln|S_0| + s^{-1}\mathrm{tr}[\mathbf{Q}S_0^{-1}]. \tag{4.18}$$

In this function, for given $\mathbf{A}$, $\ln|S_0|$ and $\mathrm{tr}[\mathbf{Q}S_0^{-1}]$ are known. Let $\mathrm{tr}[\mathbf{Q}S_0^{-1}] = p$. Then g^* is simplified as

$$g^{**}(q|\mathbf{A}) = c + t\ln|s| + s^{-1}p\,. \tag{4.19}$$

To find the minimum of g^{**} ,the derivative of g^{**} with respect to s,

$$\partial g^{**}/\partial s = t\,s^{-1} - p\,s^{-2}, \tag{4.20}$$

is set to zero. This yields as an estimator for s

$$\hat{s} = pt^{-1} = \mathrm{tr}[\mathbf{Q}S_0^{-1}]t^{-1} \tag{4.21}$$

Case 3 . $S = s\mathbf{I}$. So S is an unknown, diagonal matrix with identical elements. The model then defines *redundancy analysis* (Van den Wollenberg, 1977). Actually this is a special case of Case 2, when $S_0 = \mathbf{I}$. This means that $p = \sum_{i=1}^{t} q_{ii}$ and we find as an estimator for s

$$\hat{S} = t^{-1} \sum_{i=1}^{t} q_{ii} . \tag{4.22}$$

Case 4 . S is a full and unknown matrix and completely free. In this case the model defines *canonical analysis* (Tso, 1981; Izenman, 1975). Because there are no restrictions on S, the estimation procedure now proceeds differently. According to Tso, the trace term can be eliminated out of the likelihood function. For given **A** and **B**, an estimate for S is

$$\hat{S} = N^{-1}(\mathbf{Y} - \mathbf{XAB})'(\mathbf{Y} - \mathbf{XAB}). \tag{4.23}$$

Inserting this result in the likelihood function, causes the trace term in this function to be set to a constant value. Thus maximizing f is equivalent to minimizing

$$f^{*}(\mathbf{A},\mathbf{B}) = \ln|(\mathbf{Y} - \mathbf{XAB})'(\mathbf{Y} - \mathbf{XAB})|. \tag{4.24}$$

Substituting the expression found for $\hat{\mathbf{B}}$ in (4), yields the function

$$f^{**}(\mathbf{A}) = |\mathbf{Y}'(\mathbf{I} - \mathbf{XAA}'\mathbf{X}')\mathbf{Y}|, \tag{4.25}$$

which is to be minimized. Now let $\mathbf{Y} = \mathbf{PWG}' = \mathbf{PT}$ be the singular value decomposition of **Y**, and $\mathbf{X} = \mathbf{KVL}'$ be the singular value decomposition of **X**. Then f^{**} can be written as

$$f^{**}(\mathbf{A}) = |\mathbf{T}^2| \, |\mathbf{I} - \mathbf{A}'\mathbf{LVK}'\mathbf{PP}'\mathbf{KVL}'\mathbf{A}|, \tag{4.26}$$

which is to be minimized under the restriction $\mathbf{A}'\mathbf{X}'\mathbf{XA} = \mathbf{I}$. When we let $\mathbf{F} = \mathbf{VL}'\mathbf{A}$ be the eigenvectors corresponding with the s smallest eigenvalues of the matrix $\mathbf{I} - \mathbf{K}'\mathbf{PP}'\mathbf{K}$, this minimum is attained. Thus we will find as an estimator for **A**,

$$\hat{\mathbf{A}} = \mathbf{LV}^{-1}\mathbf{F}. \tag{4.27}$$

Now S and **A** can be found alternatingly, which completes the estimation provedure. Note that this estimator for **A** strongly ressembles the one given in equation (4.11). The difference is due to a Gram-Schmidt decomposition of **X**, instead of a singular value decomposition. However, having performed the singular value decompositions of both **X** and **Y**, the relationship with canonical analysis can easily be shown. The eigenvectors **F** also correspond with the s largest eigenvalues of the matrix $\mathbf{K}'\mathbf{PP}'\mathbf{K}$. Now the matrix $\mathbf{R} = \mathbf{K}'\mathbf{P}$ can be defined. It is well known that the singular values of **R** are equal to the canonical correlations between the variables contained in **X** and **Y**. So the eigenvalues of $\mathbf{K}'\mathbf{PP}'\mathbf{K}$ are equal to the squares of these canonical correlations.

Case 5 . Finally, we mention the case in which S is completely known. The estimation procedure is less complicated now. The problem is actually reduced to the minimization of the

likelihood function (4.1) over values for **A**, under the restriction $\mathbf{A'X'XA} = \mathbf{I}$. This means that to find **A**, we only have to solve for the eigen-problem described in equation (4.10). Again **B** follows from equation (4.4).

So summarizing the estimation in the general case:

step 0 - take some starting values for Σ.
step 1 - estimate **A** from $\hat{\mathbf{A}} = \mathbf{T}^{-1}\mathbf{F}$ (equation (4.11)).
step 2 - compute $\mathbf{Q} = N^{-1}\mathbf{Y'}(\mathbf{I} - \mathbf{XAA'X'})\mathbf{Y}$ and set $\hat{\Sigma} = \mathbf{Q}$.
repeat step 1 and step 2 untill convergence has been reached.
step 3 - estimate **B** from $\hat{\mathbf{B}} = \mathbf{A'X'Y}$.

5. Different Choices of Linear Restrictions

Thus far we have only considered special cases of our model resulting from the modelling of Σ. It is, however, also interesting to take into account some of the different choices of linear restrictions, i.e. different choices for **X**. One distinction is between *discrete* and *continuous* variables for **X**. In the discrete case, **X** could be interpreted as a 'design matrix'. For instance, one can think of **X** being composed of categorical 'background' variables. 'Main effects' and 'interactions' can then be coded in a MANOVA-type way. When numerical variables are contained in **X**, we have the possibility to introduce polynomial or other functions, for instance, by adding powers of these variables.
When the different choices for **X** are combined with the modelling features of Σ, other submodels can be specified. Two examples will illustrate the generality of our model. Firstly, we can imagine an application in *longitudinal* studies. For instance by manipulating **X** we can deal with individuals having scored on the same variables on different occassions. At the same time we could consider correlated error components by restricting Σ to be block-diagonal. Another example is the case in which **X** consists of only one column containing the scores of all individuals on a categorical variable. If that variable refers to a subdivision of the individuals into different groups, and Σ is unconstrained, our model can be interpreted as *discriminant analysis*.

6. Reduced Rank Regression, LISREL and MANOVA

In this section we will concentrate on our interpretation as described in the section on the general model. We will also discuss the relationships with the LISREL model and MANOVA. First, we will show that LISREL with m groups and equal covariance matrices in each group, is equivalent to a *full rank* regression model with m variables contained in **X**, i.e. **X** consists of m dummy variables, denoting to which group each individual belongs. This result is not very surprisingly, because in this case, and with no restrictions imposed on the covariance matrix,

both models are in fact equal to MANOVA. So we will prove that both methods are also equivalent when the covariance matrix is constrained. Also we will discuss what a reduced rank regression model means in LISREL terms. In order to simplify the derivation, we will not assume the varibables to be put in deviation of their means.
The log-likelihood for each group can be written as (see Jöreskog & Sörbom, 1984)

$$\ln L_i = c_i - 1/2N_i(\ln|S| + \mathrm{tr}[S^{-1}\mathbf{T}_i]), \tag{6.1}$$

where $\mathbf{T}_i = \mathbf{S}_i + (\bar{\mathbf{y}}_i - \mu_i)(\bar{\mathbf{y}}_i - \mu_i)'$. In this expression $\bar{\mathbf{y}}_i$ is the sample mean vector of the **y** variables in group *i* and μ_i the corresponding population mean vector. $\mathbf{S}_i$ is the sample covariance matrix for group *i* and $\Sigma = v(\theta)$, a function of some unknown parameters in θ. Note that in MANOVA the matrix Σ is unconstrained.
The log-likelihood for the whole data matrix is

$$\ln L = \sum_i \ln L_i$$

If no restrictions are assumed on the means of the **y** variables, it follows that μ_i can be estimated by $\bar{\mathbf{y}}_i$. This means that $\mathbf{T}_i = \mathbf{S}_i$.
In the full rank regression model the log-likelihood equals

$$\ln L = c - 1/2N \ln|S| - 1/2\mathrm{tr}[(\mathbf{Y} - \mathbf{XM})S^{-1}(\mathbf{Y} - \mathbf{XM})'].$$

It can easily be verified that if **X** consists of *m* dummy variables, denoting the group to which a sample element belongs, LISREL will give the same result as the full rank regression model. The likelihood function (6.3) is optimal for $\mathbf{M} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. So row *i* of matrix **M** consists of the means of the **y** variables in group *i*. This means that for each group the **y** variables are put in deviation of the group means. It follows that combining equations (6.1) and (6.2) yields (6.3) exactly. Thus we can conclude that, in the case when there are no restrictions on the means, full rank regression with *m* dummy variables and equal covariance matrices within the groups, is equivalent to LISREL with *m* groups and equal covariance matrices.
However, in our reduced rank regression model, there do hold restrictions. Instead of $\mathbf{Y} = \mathbf{XM} + \mathbf{E}$, we have $\mathbf{Y} = \mathbf{XAB} + \mathbf{E}$. These restrictions, like rank restrictions, can be interpreted as restrictions on the group means. If **X** divides the sample into *m* groups, the mean of the **y** variables per group is subtracted of **Y**. This is done leaving the covariance structure for each group the same.
An interesting example is a model with equal group means. In that case **A** is a column vector with unit elements and **B** can be estimated as a vector with the means of the **y** variables over all groups. Testing this model against a full rank regression model will be equal to MANOVA when there do not hold restrictions on the covariance matrix.
More generally, reduced rank regression can be interpreted as a model with regressors **XA** instead of **X**. This means that the regressors are not just the different groups, but some weights of these groups. These weights, the elements of matrix **A**, are chosen optimally by maximizing

the likelihood function. In doing so, the algorithm is looking for new groups by combining the original ones in such a way that the covariance matrices of the new groups have the same structure. In other words, our model is looking for *optimal regressors* and these are found by a linear transformation of the original regressors contained in matrix **X**. This feature is not possible neither with LISREL, nor with MANOVA.
Summarizing we can compare MANOVA, LISREL with *m* groups and reduced rank regression as follows: MANOVA does not have any restriction on the covariance matrices, which are equal in each group. Reduced rank regression and LISREL do have the possibility of imposing restrictions on the covariance matrices. In the way we have formulated the reduced rank regression model, all group covariance matrices are equal. Under this restriction, our model finds optimal, new regressors. A possibility with LISREL is the specification of different covariance matrices for different groups.
Thus we have mainly concentrated on the discrete case, in which **X** is conceived as a design matrix, containing only dummy variables. However, a categorical variable could also make up one column of **X**. If such a variable contains a large number of categories, for instance age ranging from 18 to 65 years, one could speak of a pseudo-numerical variable. In such cases, and also for real numerical variables, for the whole group some weighted score is subtracted of the **y** variables. Finally one can imagine combinations of discrete and (pseudo) numerical variables for **X**.

7. Testing of Hypotheses

In this sectionn we discuss some aspects of testing hypothesis with our model.
Nested models could be tested against each other using *likelihood-ratio* 6tests. Let λ be the ratio of two optimal function values, corresponding with two nested models. From standard literature it is well known that $-2\ln\lambda$ is chi-square distributed, with degrees of freedom equal to the difference of the number of estimated parameters under both models to be tested. So it is possible to have a so called most powerful test. In this way, it is for instance possible to test if adding columns to the matrix **X**, is really meaningful. So, given a certain covariance structure for S, one can test different groupings against each other.
Another possibility is to fit a certain structural model on the covariance matrix of the **y** variables and then compare the χ^2 overall goodness-of-fit measure with the goodness-of-fit of the same covariance structure fitted on S, i.e. on the covariances of **Y** after subtracting the part **XAB**. Examining the difference between the chi-square values in both cases, makes us decide if the grouping of the sample is meaningful.

8. Illustration : Zutphen Data

In this section we will illustrate our interpretation with a real data example. The data for this illustration come from a large, longitudinal study on epidemiologics. This investigation started in 1960 and was conducted in the town of Zutphen, a small industrial town in the eastern part of the Netherlands. It took place as the Dutch contribution to the 'Seven Countries Study', a research program on coronary heart disease (Keys, Aravanis, Blackburn, et al, 1967). The object of study was the effect of risk factors on morbidity and mortality in middle-aged men. For this purpose a random sample of all men born between 1900 and 1919, who had lived in Zutphen for at least five years, was selected. Every year, between 1960 and 1985, the men from this sample were subject to medical examination and detailed information was gathered on morbidity. Also psycho-social and socio-economic data were recorded (see also Duijkers, Kromhout & Spruit, 19?? of een relevantere verwijzing).
For this illustration, we analyzed data collected in the year of 1985, from a sample of 535 men. In this year, the ages of these men varied from 65 to 85 years. A structural equation model was formulated in which the effect was studied of alcohol use (Alk), cigarette smoking (Sm) and energy intake (En) in relation to Quetelet Index (QI), systolic (Sy) and diastolic (Dia) bloodpressure. The Quetelet Index can be considered a measure of body fat. It is calculated as weight in kilograms divided by height squared in metres. Alkohol use was measured as the number of grams alkohol consumed per day, energy intake was measured as the number of calories consumed per day. The variable cigarette smoking is binary, consisting of the categories smoking versus non-smoking.
Two models with the same variables were tested, using the very popular LISREL program (Jöreskog & Sörbom, 1984). These models are represented graphically in Figure 1 (a) and (b). In these path diagrams the coefficients of the standardized solutions are given.

-- insert Figure 1 about here --

For model (a) the overall goodness-of-fit test yields $\chi^2 = 18.04$ with 6 degrees of freedom (p=.01), while for model (b) the χ^2 is 10.68 with p=.10. Thus it is clear that this model has an acceptable fit and we will continue considering model (b).
The algorithm described in this paper will now be used to study the effect of the variables age and level of activity on the causal system described by model (b). Level of activity was made up by means of a number of variables measuring different kinds of active behaviour. It was derived by taking the componentscores on the first component of a PRINCALS solution (Van Rijckevorsel & De Leeuw, 1979; Gifi, 1981), relating these variables. This resulted in a continuous variable ranging from -4.735 to 1.976, a high score indicating a high level of activity.
We interpreted the effect of age and activity level on the described causal system in terms of *noise*. Taking age and activity level as variables in **X**, the 'disturbing' effects of these variables

on the relations in model (b) are 'filtered out' of the variables constituting this model. But we will first show the results of our algorithm and then continue with this interpretation.
As is described in an earlier section, the estimation proceeds stepwise. In each step the model for the covariance structure of $\mathbb{S}$ is updated. In Table 1 the χ^2 goodness-of-fit measures are given for each step. Also the χ^2 value for this model fitted on the covariances of the **y** variables is given in this table.

-- insert Table 1 about here --

From Table 1 it becomes clear that the goodness-of-fit for the structural model for $\mathbb{S}$ has strongly increased. The subtraction of a certain part of **Y**, that is predicted by the variables 'age' and 'level of activity' causes the covariance structure to fit much better. Table 1 shows that the χ^2 value decreases from 10.72 to 6.76. Although these two χ^2 values cannot be tested against each other while both models have the same number of degrees of freedom, the value of $p > .30$ is more convincing than the $p < .10$ belonging to the χ^2 of 10.72. In Figure 2 the standardized solution is presented, i.e. the parameter values belonging to the final step.

-- insert Figure 2 about here --

Figure 2 shows that the improved goodness-of-fit results in increased regression weights and a larger amount of variance explained.
As far as the interpretation is concerned, the model shows us that the use of alkohol directly influences bloodpressure, while cigarette smoking and energy intake only have influence through the Quetelet Index, i.e. in an indirect way. The application of our model established a reduction of noise on this causal system. For instance, people that are very active, are obviously compensating for the neagtive effect of alkohol use on bloodpressure, thus obscuring this relationship. Also people that are relatively young are supposed to have a lower bloodpressure in any case, though they are smoking cigarettes or not. People with a very high age tend to weight less, in spite of the number of calories they take. So the removing of the 'age effect' and the 'activity effect' makes the model relations more clear.
In fact, what we do is holding age and activity level constant for a number of groups. This is sometimes called *stratification*. However, stratification differs from our procedure in two ways. Normally stratification is caried out with only a few groups, for instance, five levels of age instead of the original variable, while our procedure deals with very many groups. Also this grouping is fixed, while our groups are found optimally. In practical research, stratification is often used to study so called 'confounding variables'. These are variables that explain causal relationships between other variables. They produce 'spurious relationships' when they are left out of the model. These relationships vanish if such variables ar held constant on different levels. In our illustration, if age and activity would behave as confounders, we could expect a certain increase in χ^2. This reduced goodness-of-fit would indicate the model relationships to be spurious for a part. However, we found the opposite. Therefore, as we have done, our results can be interpreted in terms of noise.

Finally, a remark with respect to LISREL can be made. One could argue that our procedure has a strong resemblance with a multi-sample analysis in LISREL. There are however differences. Using LISREL, in the case that we require the same covariance structure for all groups, these groups are not weighted optimally, as in our procedure. Further, LISREL is not quite suitable in situations with a large number of groups, while our model can handle any number of groups. Finally LISREL provides only one goodness-of-fit measure for all groups together. This test measures the fit of all LISREL models in all groups, including all constraints, to the data from all groups. Our procedure also provides one goodness-of-fit test for the covariance structure fitted on the variables contained in (**Y - XAB**). This test could be compared with the LISREL goodness-of-fit measure in the case that all groups would be analyzed seperately. However, this could not be done imposing the required equality constraints at the same time.

Question : comparison RRR on 20 groups & comparison LISREL on 20 groups useful?

9. References

Adcock, R.J. (1878). A problem in least squares. *The Analyst*, **5**, 53-54.

Anderson, T.W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*, **22**, 327-351.

Anderson, T.W. (1984). The Wald Memorial Lectures. Estimating linear statistical relationships. *Annals of Statistics*, **12**, 1-45.

Anderson, T.W., & Rubin, H. (1956). Statistical inference in factor analysis. In: Jerry Heyman (ed.), *Proceedings of the Third Berkeley Symposium in Statistics and Probability*, **5**, 111-150.

Bagozzi, R.P., Fornell, C., & Larcker, D.F. (1981). Canonical correlation analysis as a special case of a structural relations model. *Multivarite Behavioral Research*, **16**, 437-454.

Bekker, P.A., & De Leeuw, J. (1987). The rank of reduced dispersion matrices. *Psychometrika*, **52**, 125-135.

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

De Leeuw, J., Mooijaart, A. & Van der Leeden, R. (1985). *Fixed factor score models with linear restrictions*. Research Report 87-00. Department of Data Theory, University of Leiden.

Duijkers, T.J., Kromhout, D. & Spruit, I.P. (19??) Of relevantere verwijzijng.

Fisher, R.A. (1938). The statistical utilization of multiple measurements. *Annals of Eugenetics*, **8**, 376-386.

Fletcher, R., & Powell, M.J.D. (1963). A rapidly convergent descent method for minimization. *The Computer Journal*, **2**, 163-168.

Frisch, R. (1934). *Statistical Confluence Analysis by Means of Complete Regression Systems*. Universitetets Økonomiske Institutt, Oslo [Publikasjon Nr. 5. Separate impression *Nordic Statistical Journal*, **5**, 1933].

Gifi, A. (1981). *Nonlinear Multivariate Analysis*. Department of Data Theory, University of Leiden.

Gleser L.J. (1981). Estimation in a multivariate 'errors in variables' regression model: large sample results. *Annals of Statistics*, **9**, 24-44.

Gleser & Olkin (1970a).

Gleser & Olkin (1970b).

Izenman, A.J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, **5**, 248-264.

Jöreskog. K.G., & Sörbom, D. (1984). *LISREL VI Users Guide.* Department of Statistics, University of Uppsala.

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

Keys, Aravanis, Blackburnm, et al. (1967).

Koopmans, T.C. (1937). *Linear Regression Analysis of Economic Time Series.* Haarlem: De Erven F. Bohn NV.

Kendall, M.G., & Stuart, A. (1979). *The Advanced Theory of Statistics, Vol. 2, Inference and Relationship, Fourth Edition.* London: Griffin.

Pearson, K. (1901). On lines and planes of closest fit to points in space. *Philosophical Magazine*, **2**, 559-572.

Rao (1967).

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

Van den Wollenberg, A.L. (1977). Redundancy analysis. An alternative for canonical analysis. *Psychometrika*, **42**, 207-219.

Van Rijckevorsel, J., & De Leeuw, J. (1979). An outline of PRINCALS. Report RB 002-79. Department of Data Theory, University of Leiden.

Table 1

χ^2 values for each model step and p-values; df=6

step	χ^2	p
initial	10.7171	<.10
1	7.1984	-
2	6.7700	-
3	6.7664	>.30