RR 88-03

FITTING LONGITUDINAL REDUCED RANK REGRESSION MODELS BY ALTERNATING LEAST SQUARES

Jan de Leeuw*
Catrien Bijleveld**

*Departments of Psychology and Mathematics,
University of California at Los Angeles

**Department of Data Theory FSW
University of Leiden

our models can deal with situations in which the observations are ordered in some clearly defined way (time and space are merely the most obvious examples), and in which we have reason to suppose that close observations influence each other.

As we shall point out below, our models are quite close to the models studied in mathematical systems theory (Kalman, Falb and Arbib, 1969). The connections of such **linear dynamical systems** with the theory of covariance structure modelling have been discussed recently by Oud, Van den Bercken and Essers (1986), Otter (1986), MacCallum and Ashby (1986). We shall discuss, in a later section, the similarities and differences of the two fields from a different perspective, mainly because we have a somewhat different approach to models of this kind. It seems to us that the differences between these two approaches to data analysis are much more important than their similarities. But first we need to mention some general principles of data analysis that are relevant here.

In fitting models to data there 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). In this paper we will not model sampling errors and measurement errors explicitly. Not because this is not feasible, in fact in subsequent publications we intend to include both types of errors in our modelling process. It seems preferable, however, to start with the relatively simple case in which we merely approximate complicated multivariate data structures by simpler ones. This is, in fact, the usual way in which regression and component analysis are applied in multivariate data analysis.

State space models in probabilistic terms

For vectors we write $\mathbf{q}_i$, for matrices $\mathbf{Q}$. We study the conditional distribution of the output $\mathbf{y}_{1;T} = \{y_1,...,y_T\}$ up to time T, given the input $\mathbf{x}_{1;T} = \{x_1,...,x_T\}$ up to time T. We use a somewhat informal notation, 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. We use a type of simplification that can be introduced by using concepts borrowed from factor analysis. In factor analysis we observe m variables $\mathbf{y}$, and these variables are correlated. We assume that there exist p unobserved variables or

factors z which 'explain' the association between the observed variables, in the sense that the observed variables are independent given the factors. Compare Figure 1.

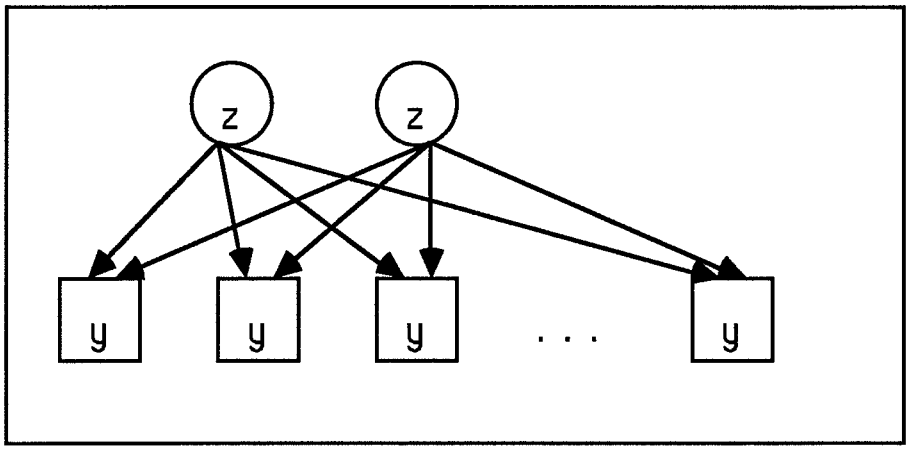

Figure1. Geometric representation of the factor analysis model.

In our informal notation we assume that

$$p(\mathbf{y}|\mathbf{z}) = \Pi_{j=1}^{m} p(y_j|\mathbf{z}), \tag{1}$$

and thus

$$p(\mathbf{y}) = \int \Pi_{j=1}^{m} p(y_j|\mathbf{z})\, p(\mathbf{z})d\mathbf{z}. \tag{2}$$

If we translate this to the (cross-sectional) regression context, following De Leeuw and Bijleveld (1987), the model becomes

$$p(\mathbf{y}|\mathbf{x}) = \int \Pi_{j=1}^{m} p(y_j|\mathbf{z})\, p(\mathbf{z}|\mathbf{x})d\mathbf{z}. \tag{3}$$

Writing the conditional distribution of **y** in this way, the dependency of the output **y** on the input **x** is decomposed into dependency of the output **y** on the latent factor **z**, and dependency of the latent factor **z** on the input **x**. The dependency of **y** on **x** is thus 'explained' through the latent variables **z**. A diagram of this situation is drawn in Figure 2.

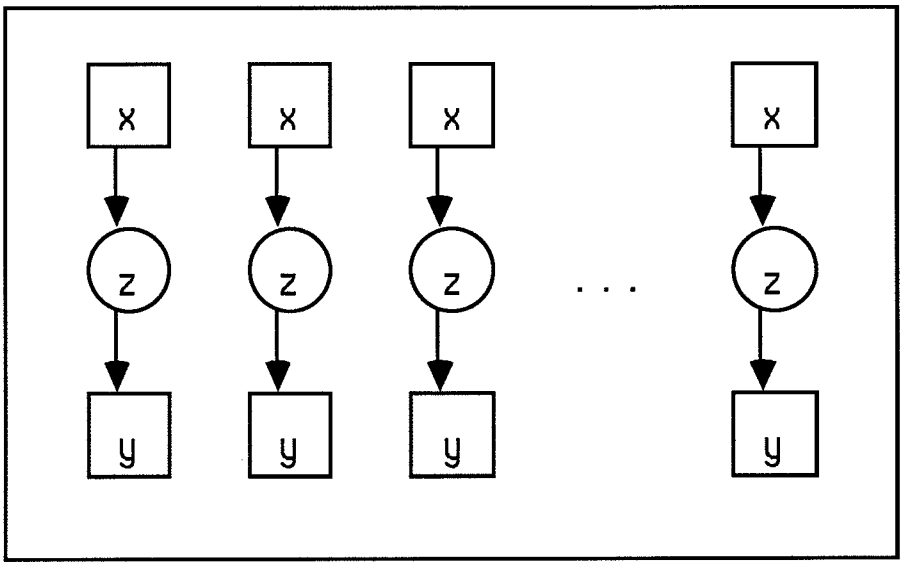

Figure 2. Geometric representation of the model (3).

In the dynamic case there are unobserved state variables $\mathbf{z}_{1;T}$ to mediate the influence of the input **x** onto the time-dependent **y**. This dependency of the output variables is accomodated by assuming that all influence of the past on the present is mediated by the present state. This first and basic assumption renders the model Markovian. Figure 3 illustrates this clearly.

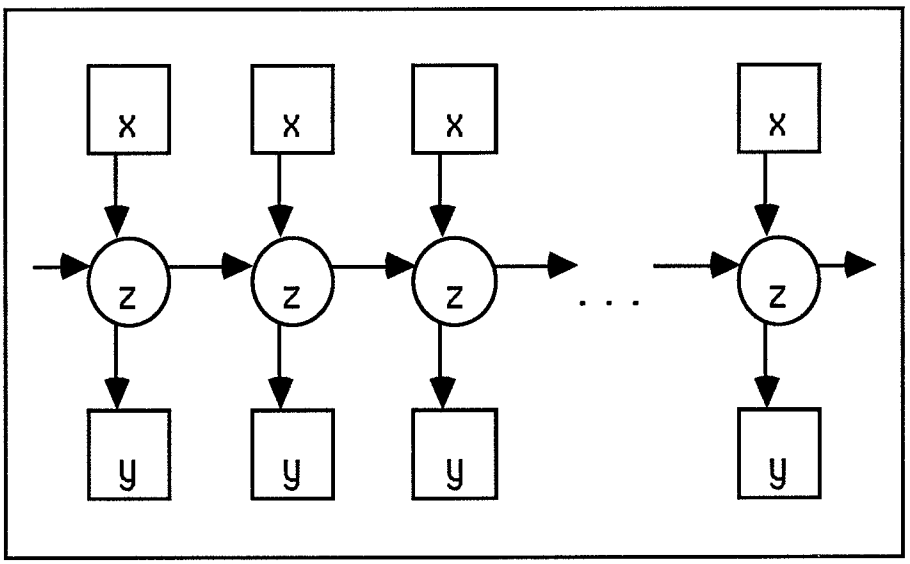

Figure 3. Geometric representation of the dynamic version of model (3).

This means that

$$p(\mathbf{y}_{1;T}|\mathbf{x}_{1;T} \wedge \mathbf{z}_{0;T}) = p(\mathbf{y}_T|\mathbf{y}_{1;T-1} \wedge \mathbf{x}_{1;T} \wedge \mathbf{z}_{0;T})p(\mathbf{y}_{1;T-1}|\mathbf{x}_{1;T} \wedge \mathbf{z}_{0;T}). \quad (4)$$

But according to the model $\mathbf{y}_T$ depends on the past only through $\mathbf{z}_T$, so that

$$p(\mathbf{y}_{1;T}|\mathbf{x}_{1;T} \wedge \mathbf{z}_{0;T}) = p(\mathbf{y}_T|\mathbf{z}_T)p(\mathbf{y}_{1;T-1}|\mathbf{x}_{1;T} \wedge \mathbf{z}_{0;T}). \quad (5)$$

Repeating this for all $\mathbf{y}_1$ to $\mathbf{y}_{\_T}$ we find

$$p(\mathbf{y}_{1;T}|\mathbf{x}_{1;T} \wedge \mathbf{z}_{0;T}) = \{\Pi_{t=1}^{T} p(\mathbf{y}_t|\mathbf{z}_t)\}p(\mathbf{x}_{1;T} \wedge \mathbf{z}_{0;T}). \quad (6)$$

We now reduce the second factor on the right hand side of (6). This gives first

$$p(\mathbf{x}_{1;T} \wedge \mathbf{z}_{0;T}) = p(\mathbf{z}_T|\mathbf{z}_{0;T-1} \wedge \mathbf{x}_{1;T})p(\mathbf{z}_{0;T-1} \wedge \mathbf{x}_{1;T}). \quad (7)$$

But $\mathbf{z}_T$ only depends on the input at time T and the state at time T-1.Thus

$$p(\mathbf{x}_{1;T} \wedge \mathbf{z}_{0;T}) = p(\mathbf{z}_T|\mathbf{z}_{T-1} \wedge \mathbf{x}_T)p(\mathbf{z}_{0;T-1} \wedge \mathbf{x}_{1;T}). \quad (8)$$

If we repeat this we obtain

$$p(\mathbf{x}_{1;T} \wedge \mathbf{z}_{0;T}) = \{\Pi_{t=1}^{T} p(\mathbf{z}_t|\mathbf{z}_{t-1} \wedge \mathbf{x}_t)\}p(\mathbf{z}_0 \wedge \mathbf{x}_{1;T}). \quad (9)$$

The second assumption is that the state at time 0 is independent of the input. If we combine (6) and (9) we have the basic result

$$p(\mathbf{y}_{1;T}|\mathbf{x}_{1;T} \wedge \mathbf{z}_{0;T}) = \{\Pi_{t=1}^{T} p(\mathbf{y}_t|\mathbf{z}_t)\}\{\Pi_{t=1}^{T} p(\mathbf{z}_t|\mathbf{z}_{t-1} \wedge \mathbf{x}_t)\}p(\mathbf{z}_0)p(\mathbf{x}_{1;T}). \quad (10)$$

This implies

$$p(\mathbf{y}_{1;T}|\mathbf{x}_{1;T}) = \int \{\Pi_{t=1}^{T} p(\mathbf{y}_t|\mathbf{z}_t)\}\{\Pi_{t=1}^{T} p(\mathbf{z}_t|\mathbf{z}_{t-1} \wedge \mathbf{x}_t)\}p(\mathbf{z}_0)\, \Pi_{t=0}^{T} d\mathbf{z}_t. \quad (11)$$

Two important special cases of (11) have been studied in detail. If both output and input assume only a finite number of values, and if the state variable is discrete as well, model (11) defines a dynamic version of the latent class model. In particular, if there is not input, we recover the latent Markov chain model of Lazarsfeld and Henry (1968), which was studied recently by Van

der Pol and De Leeuw (1986). They used the EM algorithm to compute maximum likelihood estimates. It is useful to remember the interpretation of (11) as a latent Markov chain with input. The other special case assumes that all distributions involved are multivariate normal. It is then again possible to apply the EM-algorithm to compute maximum likelihood estimates. This is not what we shall do, however. The maximum likelihood methods, both in the discrete and in the continuous (multinormal) case, approximate the complete distribution of the series as closely as possible. In this paper we shall try to approximate merely the expected value structure, assuming linear regressions.

Expected values for state space models

Now assume that the regressions are linear. If we translate the general idea that the present only depends on the past through the current state to expected values, we find the model

$$E(\mathbf{z}_t|\mathbf{z}_{0;t-1} \wedge \mathbf{x}_{1;t}) = \mathbf{F}\mathbf{z}_{t-1} + \mathbf{G}\mathbf{x}_t, \tag{12a}$$

$$E(\mathbf{y}_t|\mathbf{z}_{0;t} \wedge \mathbf{x}_{1;t}) = \mathbf{H}\mathbf{z}_t. \tag{12b}$$

This means that we best predict $\mathbf{z}_t$ by using (12a) and $\mathbf{y}_t$ by using (12b). Looking at all predictions simultaneously gives the equations

$$\mathbf{Z} = \mathbf{B}\mathbf{Z}\mathbf{F}' + \mathbf{X}\mathbf{G}', \tag{13a}$$
$$\mathbf{Y} = \mathbf{Z}\mathbf{H}', \tag{13b}$$

with $\mathbf{B}\mathbf{Z} = (\mathbf{z}_0', \mathbf{z}_1', \ldots, \mathbf{z}_{T-1}')$.

There is a special case of (13) which occurs quite often. If there is no input the model becomes

$$\mathbf{Z} = \mathbf{B}\mathbf{Z}\mathbf{F}', \tag{14a}$$
$$\mathbf{Y} = \mathbf{Z}\mathbf{H}'. \tag{14b}$$

Models without measured input are sometimes called **dynamic factor analysis** models (Molenaar, 1981, Immink, 1986). Because we do not explicitly model errors, and consequently do not distinguish common and unique factors, it is more appropropriate to call (14) a **dynamic component model**. Also the special case of (13) with $\mathbf{F} = 0$ is the (cross-sectional)

reduced-rank regression models studied earlier with similar techniques by De Leeuw and Bijleveld (1987).

Defining the loss function

The techniques presented in this paper choose the unknowns $\mathbf{Z}$ and $(\mathbf{F},\mathbf{G},\mathbf{H})$ in such a way that the sum of the squares of the prediction errors is as small as possible. In later sections we shall also consider the case in which the input $\mathbf{X}$ and the output $\mathbf{Y}$ are partially unknown (for instance, known only up to monotone transformations). The computational problem we consequently discuss is the minimization of

$$\sigma_\omega(\mathbf{F},\mathbf{G},\mathbf{H},\mathbf{Z}) = \omega^2\mathbf{SSQ}(\mathbf{Z} - \mathbf{BZF}' - \mathbf{XG}') + \mathbf{SSQ}(\mathbf{Y} - \mathbf{ZH}') \tag{15}$$

over all its arguments.

The weight ω can be used to adjust for the relative importance of predicting the output. If $\omega = 0$ then the first term in (15) becomes irrelevant, and minimizing (15) degenerates to the principal component analysis of the output. The limiting case with $\omega \rightarrow \infty$ is more interesting. In order to study it properly we observe that the first part of the loss function can always be made equal to zero (even if F and G are fixed at known values). We merely need to choose z_0 arbitrarily, and then recursively compute $z_t = \mathbf{F}z_{t-1} + \mathbf{G}x_t$. Thus $z_1 = \mathbf{F}z_0 + \mathbf{G}x_1$, $z_2 = \mathbf{F}^2z_0 + \mathbf{FG}x_1 + \mathbf{G}x_2$, and so on. Let us fix z_0, to make things simple. This makes $\mathbf{Z}$ a function of $\mathbf{F}$ and $\mathbf{G}$, which we write as $\mathbf{Z}(\mathbf{F},\mathbf{G})$. Define

$$\sigma_\infty(\mathbf{F},\mathbf{G},\mathbf{H}) = \mathbf{SSQ}(\mathbf{Y} - \mathbf{Z}(\mathbf{F},\mathbf{G})\mathbf{H}'), \tag{16}$$

and $\sigma_\infty(*,*,*)$ is the minimum of (16). Write $\mathbf{F}_\infty$, $\mathbf{G}_\infty$, $\mathbf{H}_\infty$ for the mininizers. We see that minimizing (16) amounts to a principal component analysis of the output, with restrictions $z_t = \mathbf{F}z_{t-1} + \mathbf{G}x_t$ on the component scores. We now invoke the general theory of penalty functions (Fiacco and MacCormick, 1968), which immediately gives us the following result, where $\sigma(\omega)$ is the minimum and $\mathbf{F}(\omega)$, $\mathbf{G}(\omega)$, $\mathbf{H}(\omega)$, and $\mathbf{Z}(\omega)$ are the minimizers of (15).

Theorem 1. If $\omega \rightarrow \infty$ then $\sigma(\omega) \rightarrow \sigma_\infty(*,*,*)$, $\mathbf{F}(\omega) \rightarrow \mathbf{F}_\infty$, $\mathbf{G}(\omega) \rightarrow \mathbf{G}_\infty$, $\mathbf{H}(\omega) \rightarrow \mathbf{H}_\infty$ and $\mathbf{Z}(\omega) \rightarrow \mathbf{Z}(\mathbf{F}_\infty,\mathbf{G}_\infty)$.

If $0 < \omega < \infty$ the situation becomes a bit more complicated. The main reason for these complications is that unconstrained minimization of (15) over $\mathbf{G}$, $\mathbf{H}$, and $\mathbf{Z}$ is not useful. This

follows from Theorem 2 below. We first discuss an auxilary result. Define $\sigma_* = \min \mathbf{SSQ}(\mathbf{Y} - \mathbf{ZH}')$. We can find σ_* from the singular value decomposition of the output.

Theorem 2. $\inf \sigma(\mathbf{F},\mathbf{G},\mathbf{H},\mathbf{Z}) = \sigma^*$, and the infimum is only attained in very special cases.

Proof. It is clear that $\sigma(\mathbf{F},\mathbf{G},\mathbf{H},\mathbf{Z}) \geq \sigma^*$. Now take $\mathbf{F}_0$ and $\mathbf{G}_0$ arbitrary, $\mathbf{Z}_0$ and $\mathbf{H}_0$ from the singular value decomposition of the output, and define $(\mathbf{G},\mathbf{H},\mathbf{Z}) = (\omega\mathbf{G}_0,\omega^{-1}\mathbf{H}_0,\omega\mathbf{Z}_0)$. Then $\sigma(\mathbf{F},\mathbf{G},\mathbf{H},\mathbf{Z}) = \omega^2\mathbf{SSQ}(\mathbf{Z}_0 - \mathbf{B}\mathbf{Z}_0\mathbf{F}_0' - \mathbf{X}\mathbf{G}_0') + \sigma^*$, and letting $\omega \rightarrow 0$ makes $\sigma(\mathbf{F},\mathbf{G},\mathbf{H},\mathbf{Z}) \rightarrow \sigma^*$. The minimum is attained if and only if we can choose $\mathbf{F}$ and $\mathbf{G}$ such that $\mathbf{SSQ}(\mathbf{Z}_0 - \mathbf{B}\mathbf{Z}_0\mathbf{F}' - \mathbf{X}\mathbf{G}') = 0$, which is possible if and only if $\mathbf{Z}_0$ is in the space spanned by the columns of $\mathbf{B}\mathbf{Z}_0$ and $\mathbf{X}$. **QED.**

Thus unrestricted minimization of (15) is not a good idea, because iterative procedures will produce a trivial solution with a very large $\mathbf{H}$ proportional to $\mathbf{H}_0$, a very small $\mathbf{Z}$ proportional to $\mathbf{Z}_0$, and an arbitrary, but also very small, value of $\mathbf{G}$. Thus we impose the normalization restrictions $\mathbf{Z}'\mathbf{Z} = \mathbf{I}$. Minimization of (15) will be carried out by alternating least squares, as usual. Thus we alternate the solution to two types of problems: first we minimize (15) with respect to $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{H}$ for given fixed $\mathbf{Z}$, then we minimize over $\mathbf{Z}$ for fixed current $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{H}$, under the restriction that $\mathbf{Z}'\mathbf{Z} = \mathbf{I}$. Then we go back to the first type of problem, and so on. The general theory of alternating least squares shows that this process is convergent. It is clear that the subproblem of the first type, solving for $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{H}$ for given $\mathbf{Z}$, is a linear problem which is easy to solve. The subproblem of the second type is much more complicated, however, and we shall discuss it in a separate section.

Majorization

Consider the problem of minimizing (15) over $\mathbf{Z}$, with $\mathbf{Z}'\mathbf{Z} = \mathbf{I}$, and with the parameters $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{H}$ (temporarily) regarded as known constants. Write $\mathbf{Z} = \mathbf{Z}_{\text{old}} + \Delta$, with $\mathbf{Z}_{\text{old}}$ the current best solution, and define $\Delta = \mathbf{Z} - \mathbf{Z}_{\text{old}}$. Now $\sigma(\mathbf{F},\mathbf{G},\mathbf{H},\mathbf{Z})$ equals

$$\mathbf{SSQ}\ \omega\{(\mathbf{Z}_{\text{old}} - \mathbf{B}\mathbf{Z}_{\text{old}}\mathbf{F}' - \mathbf{X}\mathbf{G}') + (\Delta - \mathbf{B}\Delta\mathbf{F}')\} + \mathbf{SSQ}\{(\mathbf{Y} - \mathbf{Z}_{\text{old}}\mathbf{H}') - \Delta\mathbf{H}'\}. \qquad (17)$$

Let $\mathbf{P}_1 = \mathbf{Z}_{\text{old}} - \mathbf{B}\mathbf{Z}_{\text{old}}\mathbf{F}' - \mathbf{X}\mathbf{G}'$ and $\mathbf{P}_2 = \mathbf{Y} - \mathbf{Z}_{\text{old}}\mathbf{H}'$ be the two matrices of residuals for the previous solution. Then

$$\sigma(\mathbf{F},\mathbf{G},\mathbf{H},\mathbf{Z}) = \sigma(\mathbf{F},\mathbf{G},\mathbf{H},\mathbf{Z}_{old}) - 2\omega^2 \text{ tr } \Delta'(\mathbf{B}'\mathbf{P}_1\mathbf{F} - \mathbf{P}_1) - 2 \text{ tr } \Delta'\mathbf{P}_2\mathbf{H} +$$

$$+ \text{SSQ } \omega(\Delta - \mathbf{B}\Delta\mathbf{F}') + \text{SSQ}(\Delta\mathbf{H}'). \tag{18}$$

Now suppose we have a bound of the form

$$\begin{aligned} &\text{SSQ } \omega(\Delta - \mathbf{B}\Delta\mathbf{F}') + \text{SSQ}(\Delta\mathbf{H}') = \\ &\text{SSQ } \Delta \left(\omega(\mathbf{I} - \mathbf{B}\otimes\mathbf{F}') \,//\, (\mathbf{I}\otimes\mathbf{H}) \right) \leq \gamma \, \text{SSQ}(\Delta), \end{aligned} \tag{19}$$

where // stands for vertical concatenation and γ depends on **B**, **F** and **H**. Also define

$$\mathbf{S} = \gamma^{-1}(\omega^2\mathbf{B}'\mathbf{P}_1\mathbf{F} + \mathbf{P}_2\mathbf{H} - \omega^2\mathbf{P}_1). \tag{20}$$

Then

$$\sigma(\mathbf{Z},\mathbf{F},\mathbf{G},\mathbf{H}) \leq \sigma(\mathbf{Z}_{old},\mathbf{F},\mathbf{G},\mathbf{H}) + \gamma\text{SSQ}(\Delta - \mathbf{S}) - \gamma\text{SSQ}(\mathbf{S}). \tag{21}$$

But $\text{SSQ}(\Delta - \mathbf{S}) = \text{SSQ}(\mathbf{Z} - (\mathbf{Z}_{old} + \mathbf{S}))$. An iteration step of our majorization algorithm consists of minimizing $\text{SSQ}(\mathbf{Z} - (\mathbf{Z}_{old} + \mathbf{S}))$ over **Z** satisfying $\mathbf{Z}'\mathbf{Z} = \mathbf{I}$. This is a simple (weighted) Procrustus problem (Cliff, 1969), whose solution is well known. If $\mathbf{Z}_{old} + \mathbf{S} = \mathbf{K}\Lambda\mathbf{L}'$ is a singular value decomposition, then $\mathbf{Z}_{new} = \mathbf{K}\mathbf{L}'$ is the solution of the minimization problem. After computing $\mathbf{Z}_{new}$ we set $\mathbf{Z}_{old} = \mathbf{Z}_{new}$, and we repeat the computations.

Theorem 3. The algorithm $\mathbf{Z}_{new} = \mathbf{K}\mathbf{L}'$, with $(\mathbf{Z}_{old} + \mathbf{S}) = \mathbf{K}\Lambda\mathbf{L}'$ and **S** given by (20), converges to a stationary point, i.e. to a point satisfying $\mathbf{Z}_{new} = \mathbf{Z}_{old}$.

Proof. The convergence proof of the procedure is based on the chain

$$\sigma(\mathbf{Z}_{new},\mathbf{F},\mathbf{G},\mathbf{H}) = \min\{\sigma(\mathbf{Z}_{old},\mathbf{F},\mathbf{G},\mathbf{H}) + \gamma\text{SSQ}(\mathbf{Z} - (\mathbf{Z}_{old} + \mathbf{S})) - \gamma\text{SSQ}(\mathbf{S} \mid \mathbf{Z}'\mathbf{Z} = \mathbf{I}\} \leq$$

$$\leq \sigma(\mathbf{Z}_{old},\mathbf{F},\mathbf{G},\mathbf{H}) + \gamma\text{SSQ}(\mathbf{Z}_{old} - (\mathbf{Z}_{old} + \mathbf{S}) - \gamma\text{SSQ}(\mathbf{S}) = \sigma(\mathbf{Z}_{old},\mathbf{F},\mathbf{G},\mathbf{H}). \tag{22}$$

Thus the transformation $\mathbf{Z}_{old} \rightarrow \mathbf{Z}_{new}$ decreases the loss function. Because the transformation is generally continuous (excluding the degenerate case of zero singular values) it follows from Zangwill (1967, chapter 4) that we have convergence to at least a stationary point. **QED.**

After convergence of the iterative procedure for computing the optimal $\mathbf{Z}$ we compute a new $\mathbf{F},\mathbf{G},\mathbf{H}$ by simple least squares. An alternative (which may be better in terms of overall speed of convergence) is to alternate a single $\mathbf{Z}_{old} \rightarrow \mathbf{Z}_{new}$ step with a single $(\mathbf{F},\mathbf{G},\mathbf{H})$ step. Now consider what happens if we do not use $\mathbf{Z}_{new}$, but $\mathbf{Z}_{new}\mathbf{M}$, with $\mathbf{M}$ an arbitrary rotation matrix. Denoting arguments over which we have minimized by stars, it follows that $\sigma(\mathbf{Z}_{new},*,*,*) = \sigma(\mathbf{Z}_{new}\mathbf{M},*,*,*)$. Thus the decrease of the loss function as a result of the two substeps taken together will be the same, and is independent of $\mathbf{M}$. It follows that we can also compute an update (much more cheaply) by setting $\mathbf{Z}_{new} = \mathbf{GRAM}(\mathbf{Z}_{old} + \mathbf{S})$, with $\mathbf{GRAM}(.)$ the Gram-Schmidt orthogonalization. This situation is analogous to the situation in other alternating least squares methods (Gifi, 1981, chapter 3).

There is one step in the actual implementation of the algorithm which is still unclear. This the choice of γ in (19). Write $\lambda_{max}(\mathbf{A})$ for the largest singular value of a matrix $\mathbf{A}$.

Theorem 4. If $\gamma \geq \lambda^2_{max}\{(\omega(\mathbf{I}-\mathbf{B}\otimes\mathbf{F}'))//(\mathbf{I}\otimes\mathbf{H})\}$ then (19) is true.

Proof. Define $\underline{\delta} = \text{vec}(\underline{\Delta})$ and $\mathbf{A}$ as the matrix $(\omega(\mathbf{I}-\mathbf{B}\otimes\mathbf{F}'))//(\mathbf{I}\otimes\mathbf{H})$. Then (19) can be written as $\underline{\delta}'\mathbf{A}'\mathbf{A}\underline{\delta}$, wherewith $\mathbf{SSQ}\ (\underline{\delta}'\mathbf{A}'\mathbf{A}\underline{\delta}) \leq \mathbf{SSQ}\ (\underline{\delta}'\underline{\delta})\ \lambda^2_{max}(\mathbf{A})$. **QED.**

By using the results on Theorem 4, in combination with the earlier results, we obtain a convergent algorithm to minimize (15) over $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{H}$, and all $\mathbf{Z}$ such that $\mathbf{Z}'\mathbf{Z} = \mathbf{I}$. This does not guarantee, of course, that convergence is fast enough for practical purposes, and ceratinly not that the solutions found by the algorithm will be satisfactory. This will have to be studied by extensive numerical studies, and by the analysis of practical examples.

Example 1: Eigen analysis of the American states data.

We analyze an example, merely for illustrative purposes, which shows what the effect in practice of our theorems is. For this purpose we take data on the fifty states of the USA, analyzed earlier by many people. We have used a version of these data taken from Meulman (1986, p. 48-54), in which there is a total of twelve variables. The first seven variables are to be considered as input variables. They are, respectively, percentage of blacks, percentage of hispanos, ratio of urban to rural, per capita income in dollars, life expectancy in years, homicide rate, and unemployment rate. The last five variables are output variables, having to do with educational achievement in the fifty states. They are: percentage high school graduates, percentage public school enrollment, pupil to teacher ratio, illiteracy rate, and failure rate on selective service mental ability test.

In order to illustrate the theorems we have first standardized all variables (sum equal to zero, sum of squares equal to one). Next we set $\underline{\mathbf{X}} = \mathbf{GRAM}(\mathrm{X})$, and we performed the eigen-analysis of $\omega^2\underline{\mathbf{X}}\underline{\mathbf{X}}' + \mathbf{YY}'$ for ω equal to 0, 1, and 10. Table 1 shows the ordered eigenvalues, with the first seven eigenvalues corrected by subtracting ω^2. Theorem 1 tells us that the first seven eigenvectors, with ω^2 substracted, converge to $\mathbf{Q}$, the eigenvalues of $\underline{\mathbf{X}}'\mathbf{YY}'\underline{\mathbf{X}}$. This happens fairly rapidly. Table 1 gives the eigenvalues for various values of ω; the last five eigenvalues converge to the largest eigenvalues of $\mathbf{X}_{\_}'\mathbf{YY}'\mathbf{X}_{\_}$, with $\mathbf{X}_{\_}$ a basis for the orthogonal complement of the column space of $\mathbf{X}$.

Table 1: Eigenvalues for various values of ω

ω :	0	1	10	100
01	2.674	2.543	2.245	2.237
02	1.331	.685	.267	.265
03	.526	.223	.166	.165
04	.323	.076	.050	.050
05	.146	.036	.028	.028
06	.000	.000	.000	.000
07	.000	.000	.000	.000
08	.000	.703	1.166	1.169
09	.000	.385	.528	.533
10	.000	.182	.269	.272
11	.000	.108	.209	.210
12	.000	.058	.071	.071

Table 2a: Loadings for $\omega = 0$

black	.852	.041	-.132	.010	-.281
hispa	.065	.219	.177	.462	.369
urban	-.057	-.176	.299	.111	-.153
incom	.520	.211	.354	.214	-.242
lifex	-.704	-.219	.079	-.012	.241
homic	.723	.257	.045	.172	-.042
unemp	.269	-.047	.301	-.076	-.156
highs	-.894	.154	.295	.400	.086
publi	-.259	.844	-.460	.086	-.047
pupil	.411	.769	.458	-.175	.002
illit	.917	-.028	.059	.320	-.230
failu	.936	.063	-.118	.150	.289

Table 2b: Loadings for $\omega = 1$

black	.892	-.116	-.049	-.056	-.201
hispa	.074	.403	.547	-.538	.284
urban	-.068	-.305	.537	.158	.073
incom	-.552	-.270	.608	-.013	-.359
lifex	-.742	-.231	.007	-.153	.531
homic	.763	.307	.279	.026	-.179
unemp	.277	-.152	.471	.682	.109
highs	-.839	.193	.352	-.104	-.071
publi	-.186	.802	-.281	-.118	-.042
pupil	.438	.655	.218	.233	-.014
illit	.900	-.002	.186	-.165	.118
failu	.947	-.041	.006	-.060	-.160

Table 2c: Loadings for $\omega = 10$

black	.933	-.267	.029	-.087	-.140
hispa	.085	.642	.519	-.414	.242
urban	-.086	-.333	.586	.170	.145
incom	-.591	-.250	.655	-.013	-.303
lifex	-.786	-.156	.002	-.165	.561
homic	.806	.325	.274	.093	-.210
unemp	.279	-.221	.470	.719	.135
highs	-.768	.139	.284	-.064	-.053
publi	-.050	.376	-.224	-.079	-.038
pupil	.433	.278	.105	.167	-.020
illit	.806	.136	.150	-.098	.094
failu	.908	-.116	.045	-.051	-.120

Tables 2a, 2b, and 2c show the correlations between $\mathbf{Z}_1$ to $\mathbf{Z}_5$, the eigenvectors corresponding with the five largest eigenvalues, and the input and output variables $\mathbf{X}$ and $\mathbf{Y}$. For $\omega = 0$ these eigenvectors are the principal components of the output, if ω increases they become related more and more to the input, and for large ω they are in the space of the input variables. We see that especially the first one is still strongly related with the output, but the second component is no longer strongly related to anything. On the basis of this analysis it seems that the first dimension can be interpreted as a general poverty dimension, and moreover it can be chosen almost completely in the space of the input variables. A more detailed interpretation of these results will be given below.

Example 2: The American states data analyzed with ALS.

The same data were analyzed with the alternating least squares algorithm described above. For performing this analysis we wrote a program we named DYNAMALS, which is an acronym of linear DYNAMical systems analysis by Alternating Least Squares.

We performed the analysis for $\omega = 0$, 1 and 10. The fit of the respective ALS-solutions was: 1, .856 and .996. In each case the results conformed closely to the eigen solution described above; the ALS-computed correlations of the input and output variables with the state variables $\mathbf{Z}_1$ to $\mathbf{Z}_5$ were approximately the same as those computed by the eigen-analysis. For the first dimension of the state $\mathbf{Z}_1$ no differences were found between the eigen- and ALS-solutions; from $\mathbf{Z}_2$ towards higher and less important dimensions of the state differences appeared, with the largest absolute difference found .005. Thus, the interpretations of the ALS solutions are identical to those of the eigen solutions. To summarize what we have said about the influence of the weight ω, we have drawn a picture of the development of the correlations of the input and output variables with the states for ω is 0, 1 and 10.

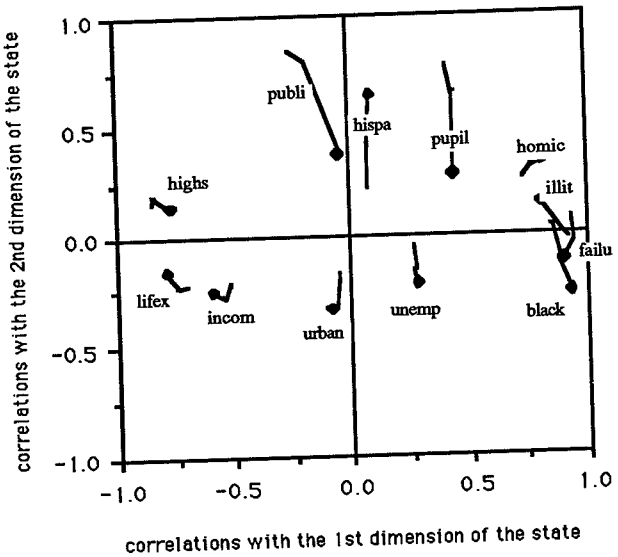

Figure 4. Correlations of the variables with the state for $\omega = 0$, 1, 10.

Figure 4 shows that the correlations of the seven input variables increase for increasing ω; the correlations of the five output variables decrease. The correlations of the variables with the second dimension of the state change most; this second dimension is less stable than the first dimension.

Optimal Scaling

The alternating least squares techniques discussed in this paper can be combined easily with optimal scaling of the variables. This is illustrated, for example, in De Leeuw (1987b). In stead of two substeps in a main iteration, one for updating **G** and **H** for given **Z** and one for updating **Z** for given **G** and **H**, we now have three substeps. In the third substep the scaling of the variables in **X** and **Y** is updated, for given **Z**, **G**, and **H**.

If we take a look at loss function (15) we see that for given **Z**, **G**, and **H** the only part which depends on variable $\mathbf{y}_j$ is of the form $\mathbf{ssq}(\mathbf{y}_j - \tilde{\mathbf{y}}_j)$, where $\tilde{\mathbf{y}}_j = \mathbf{Z}\mathbf{h}_j$. It follows that the update of variable $\mathbf{y}_j$ is of the form $\mathbf{y}_j \leftarrow \mathbf{norm}(\mathbf{proj}(\tilde{\mathbf{y}}_j))$, with **proj** denoting the projection on the cone of admissible transformations. We use **ssq** and **norm** in lower case, because they are now applied to vectors and not to matrices. The admissible transformations can be the cone of monotone transformations, the subspace of nominal transformations, a subspace of spline transformations, and so on. For details we refer to the optimal scaling literature mentioned above.

For updating variable $\mathbf{x}_i$ the situation is a bit more complicated. We can write the relevant part of the loss function as $\mathbf{SSQ}((\mathbf{Z} - \mathbf{X}_i\mathbf{G}_i') - \mathbf{x}_i\mathbf{g}_i)$. Here $\mathbf{X}_i\mathbf{G}_i'$ contains the contributions of the input variables except $\mathbf{x}_i$. Let $\tilde{\mathbf{x}}_i = (\mathbf{Z} - \mathbf{X}_i\mathbf{G}_i')\mathbf{g}_i/\mathbf{ssq}(\mathbf{g}_i)$. Then we have to minimize $\mathbf{ssq}(\mathbf{x}_i - \tilde{\mathbf{x}}_i)$, giving $\mathbf{x}_i \leftarrow \mathbf{norm}(\mathbf{proj}(\tilde{\mathbf{x}}_i))$. Cycling over the variables, changing them one at a time gives the third alternating least squares substep.

Of course there are many variations of this algorithm possible. We can cycle over the scaling of **X** and **Y** various times before we update **Z** and **G** and **H**. We can iterate the updating of **Z** and **G** and **H** until convergence before computing a new scaling of the variables. The general experience so far is that not too much work in each substep leads to simple computations and reasonable overall convergence, but we have no formal proof for this general statement.

Example 3: The American states data treated ordinally in the ALS-analysis.

The American states data were again analyzed with the ALS algorithm; this time the seven input variables were treated ordinally. The fit improved to .948 with ordinal treatment of the variables.

A picture of the American states' scores on the first two state variables, together with the correlations of the input and output variables with these state variables is in Figure 5.

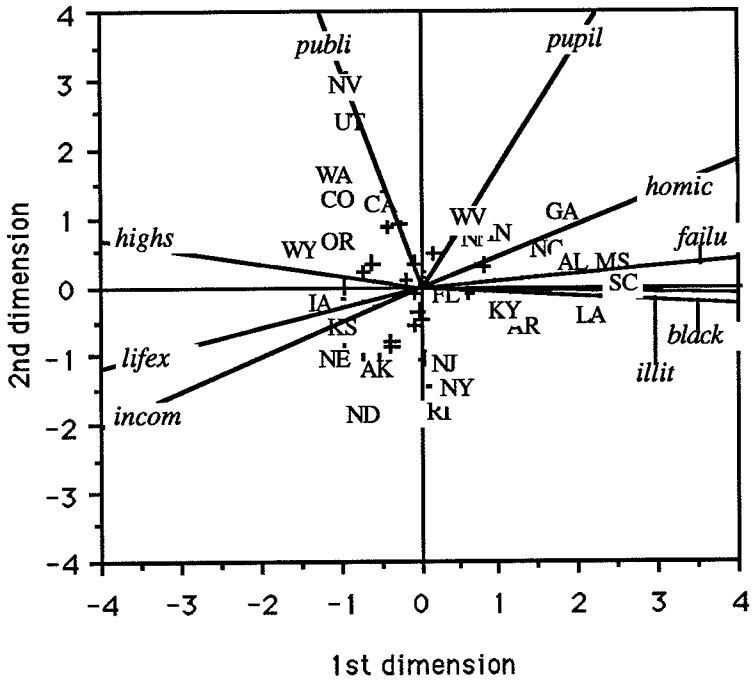

Figure 5. Correlations of the variables and scores of the American states.

From the picture we see that on the first dimension the vectors of ILLIT, BLACK, FAILU and HOMIC point in approximately the same direction; in the opposite direction point INCOME, LIFEX and HIGHS. On the second dimension PUPIL and PUBLI load positively. The correlations of HISPA and UNEMP with either dimensions were low, so they will not be considered in the interpretation. The first dimension may be interpreted as a *poverty* dimension; states with high scores on this dimension have high percentages of blacks, illiteracy, failure on the Selective Service mental ability test, high homicide rate, amall percentage of high school graduates, low life expectancy and low income. The vectors for HOMIC and INCOM/LIFEX are at almost opposite angles. The second dimension may be interpreted as an *education* dimension; while PUBLI and PUPIL are the variables that load on this dimension, they are at an angle of approximately 50 degrees. States with low scores on this dimension like North Dakota, Nebraska and Arkansas on the left side, and Rhode Island, New Jersey and New York on the right side have low pupil to teacher ratio's and small public school enrollment. States with high scores on this dimension like Nevada and Utah are marked by high public scholl enrollment.

The results may be summarized as follows. Southern states like Missouri, South Carolina, Louisiana, Alabama and Georgia that are situated in the right part of the picture are poor states;

rich states are Wyoming, Iowa, Washington, Nebraska, Kansas, Oregon and Colorado. States with high educational achievements are Nevada, Utah, Washington, Colorado and California; on the opposite end are Rhode Island, New York, North Dakota and New Jersey.

Example 4: Analysis of time-dependent blood pressure data.

We will now analyze the relation between medication and blood pressure from data obtained for a 57-year old white male under medical treatment for hypertension. For 113 days this patient recorded every morning his diastolic and systolic blood pressure. Added to this were two series of data. The first set, which we will call 'medication', marks the phases in the recording period; these phases may be marked by changes in medicines taken or by otherwise important changes. The other series, which we will call 'weekday', consisted of the day of the week on which blood pressure was measured. As blood pressure can be influenced by stress and other factors, we expected that blood pressure might be generally lower in the weekends and higher during the working-week. The mean diastolic and systolic blood pressure data in mm mercury for the various periods are:

	diastolic blood pressure	systolic blood pressure
meto 400 mg	104.19	160.16
meter	101.53	149.65
sota 240 mg	97.18	137.55
+ diureticum	98.75	131.25
sota 160 mg	84.89	125.70
visit	97.20	138.20

First the patient took 400 mg a day of metoprololtartraat, which we abbreviated as 'meto'; tthen he patient switched to sotalolhydrochloride, abbreviated as 'sota', of which 240 mg and later 160 mg a day were taken respectively. 'Meter' refers to the patient starting to use a new blood sphygmomanometer, at '+ diureticum' the patient took one diureticum, and during "visit" the patient visited a relative on another continent. Medication and weekday were thus the input variables; as no ordering is apparent for either of the two, we treated them on a nominal level. The diastolic and systolic blood pressures served as the output variables, they were treated numerically. A number of blood pressure measurements was missing; these were substituted by least squares optimal estimates. As medication does not take effect immediately, and these were morning blood pressure estimates, we used a lag of one day for the medication variable. The algorithm converged in 9 iterations to a fit of .864.

The correlations of input variables and blood pressure data with the one-dimensional state are in Table 3.

Table 3. Correlations of input and output variables with the state

state	z
medication	.885
weekday	.004
diastolic blood pressure	-.927
systolic blood pressure	-.923

Weekday correlates barely with the states, but medication does. To evaluate the effects of the various "medicines", the category quantifications of the categories of medication and the actual average diastolic and systolic blood pressure for those periods are presented below:

	category quantifications
meto 200 mg	-.116
meter	-.008
sota 240 mg	-.039
+ diureticum	.057
sota 160 mg	.129
visit	.038

As the blood pressures have negative correlations and medication has a positive correlation with the state, blood pressure is highest for the medication period with the lowest quantification. Thus, in especially the first period, when metoprololtartraat was used, blood pressure was high. The new sphygmomanometer gives a considerable improvement, but the first sotalolhydrochloride period has a lower quantification again, indicating that this medicine did in fact increase blood pressure. The diureticum constitutes a sudden improvement and blood pressure is comparably improved when the change from 240 mg to 160 mg of sotalolhydrochloride a day is made. The visit to the relative reverses this trend, and blood pressure rises in this period.

References

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

Cliff, N. (1966). Orthogonal rotation to congruence. **Psychometrika**, 31, 33-42.

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

De Leeuw, J., & Bijleveld, C. (1987). **Fitting longitudinal reduced rank regression models by alternating least squares.** Research Report 87-00. Department of Data Theory, University of Leiden.

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.

A. Fiacco & G. MacCormick (1968). **Non-linear programming: Sequential Unconstrained Minimization Techniques.** New York, Wiley.

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

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

Immink, W. (1986). **Parameter estimation in Markov models and dynamic factor analysis.** PhD dissertation, University of Utrecht.

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.

Kalman, R.E., Falb, P.L. and Arbib, M.A. (1969). **Topics in mathematical system theory**, New York, McGrawHill.

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

Lazarsfeld, P.F. and Henry, N.W, (1968). **Latent structure analysis**, Boston: Houghton-Mifflin.

MacCallum, R. and Ashby, F.G. (1986). Relationships between linear systems theory and covariance structure modeling, **Journal of Mathematical Psychology**, 30, 1-27.

Meulman, J. (1986). **A distance approach to nonlinear multivariate analysis**. Leiden, DSWO- Press.

Molenaar, P.C.M. (1981). **Dynamic factor models**. PhD dissertation, University of Utrecht.

Otter, P.W. (1986). Dynamic structural systems under indirect observation: Identifiability and estimation aspects from a system theoretic perspective, **Psychometrika**, 51, 415-428.

Oud, J.H., Van den Bercken, J.H. and Essers, R.J. (1986). Longitudinal factor score estimation using the Kalman filter, **Kwantitatieve Methoden**, 20, 109-129.

Van de Pol, F. and De Leeuw, J. (1986). A latent Markov model to correct for measurement error, **Sociological Methods & Research**, 15, 118-141.

Zangwill, W.I. (1969). **Nonlinear programming, a unified approach**, Englewood Cliffs, Prentice Hall.