On least squares optimization of linear dynamic systems

Frits D. Bijleveld *

Catrien C.J.H. Bijleveld †

Jan de Leeuw ‡

December 15, 1993

Requests for reprints should be addressed to C. Bijleveld, Department of Psychometrics and Research Methodology, Faculty of Social Sciences, P.O. Box 9555, 2300 RB Leiden, the Netherlands.

*SWOV Institute for Road Safety Research Leidschendam, the Netherlands

†Dept. of Psychometrics and Research Methodology Faculty of Social Sciences, Leiden University, the Netherlands

‡Depts. of Psychology and Mathematics University of California at Los Angeles, U.S.A.

Abstract

We discuss two estimation methods for fitting linear dynamic systems. The first is the existing DYNAMALS algorithm, that uses Alternating Least Squares and a majorization substep. With this method, it is difficult to ensure that the latent states are completely in the space of the predictor variables. We propose an alternative method that uses a single step algorithm. After direct implementation, the latent states are in the space of the predictor variables. The proposed method can also estimate intercepts in the system and measurement equations. The porposed method is compared with the existing DYNAMALS method using a real-life example.

Keywords: Longitudinal reduced rank regression, state space modelling, optimization methods.

1 Introduction.

Recently, Bijleveld and De Leeuw proposed an algorithm for fitting the longitudinal reduced rank regression or state space model (Bijleveld & De Leeuw 1991) The algorithm used least squares and majorization substeps. Its main advantage over existing methods is that it can easily incorporate optimal scaling of non numerical variables, by adding a third substep to the iteration. The drawbacks of the algorithm, however, are the slow majorization procedure, standardization of the variables which makes it impossible to model e.g. exponential developments, and the fact that the latent state variables are usually not completely in the space of the predictor variables, for reasons we will detail below. In the following we will propose a different method that directly estimates the unknowns using a quasi-Newton type of optimization procedure. The solutions for the latent variables are completely in the space of the predictor variables. Moreover, the proposed algorithm appears to be simple, efficient and reliable. The two methods will be compared using an example from psychotherapy research.

2 The state space model.

In the following vectors will be denoted by lower case characters, matrices by upper case characters. Suppose we have observed k input variables x and m output variables y at T consecutive occasions. We suppose that there is a time dependence in the measurements, which is modeled by supposing that the x influence the y through p-dimensional latent variables z. The z accommodate the time dependence in the measurements by following a Markov type of dependency; the latent states z thereby serve as the memory of the system. Usually, the dimensionality of the z is lower than that of the smallest of the dimensionalities of x and y. In that sense the z also filter the dependence of the output on the input. A visual representation of this model is in Figure 1. In formula this model can be written as follows:

$$z_t = Fz_{t-1} + Gx_t \qquad \text{(system equation)} \tag{1}$$

$$y_t = Hz_t \qquad \text{(measurement equation)} \tag{2}$$

where the subscript t indicates the timepoints t, F is the $p \times p$ state transition matrix, G is the $p \times k$ control matrix, H is the $m \times p$ measurement matrix, and z_t, x_t, and y_t are vectors of dimensionality $p \times 1$, $k \times 1$, and $m \times 1$

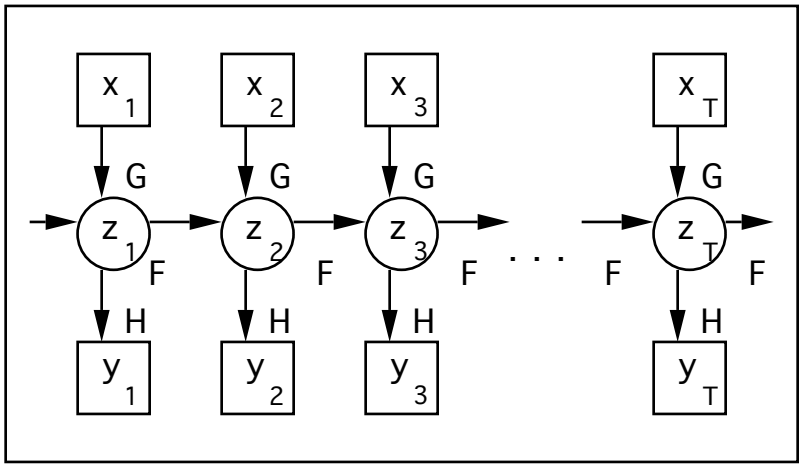

Figure 1: Visual representation of the state space model

respectively. Model (1) (2) is known under various names: linear dynamic system, state space model, longitudinal reduced rank regression model, linear dynamic model. For a discussion of the state space model see Ljung (1987), Willems (1990), Bijleveld & De Leeuw (1991) and Harvey (1989).

In matrix notation for all time points simultaneously the model becomes:

$$Z = BZF' + XG' \tag{3}$$

$$Y = ZH' \tag{4}$$

with Z the $T \times p$ matrix containing the latent states from time 1 until time T, X the $T \times k$ matrix containing the input variables from time 1 until time T, Y the $T \times m$ matrix containing the output variables from time 1 until time T, and B the $T \times T$ shift matrix such that $B(z_1', \ldots, z_T') = (z_0', \ldots, z_{T-1}')$.

Note that z_0 is an unknown entity, in fact $z_0', z_1', \ldots$ is unknown, and several options are available for estimating it from the available data. Bijleveld & De Leeuw (1991) proposed to set $z_0 \equiv z_1$ by default but this may in a number of cases be suboptimal.

In fitting this model the following loss function was proposed by Bijleveld & De Leeuw (1991):

$$\sigma_\omega = \omega^2 \, \mathbf{SSQ}(Z - BZF' - XG') + \mathbf{SSQ}(Y - ZH') \tag{5}$$

where '**SSQ**' stands for the sum of squares over all the arguments. The weight ω specifies the relative importance of the input and output. There are only two choices for ω which are non-arbitrary: $\omega = 0$ and $\omega \to \infty$. All other choices for ω are arbitrary, even the choice $\omega = 1$ gives a solution that is strictly speaking dependent upon the scaling of the variables. As ω becomes smaller, the last(output)-part of the loss function (5) becomes more and more important, which has the consequence that the latent states Z are more and more in the space of the output variables; in the limiting case with $\omega = 0$, (5) yields principal components analysis. As ω becomes larger, the input becomes more and more important, and in the limiting case where $\omega \to \infty$, the latent states are situated in the space of the input variables.

In state space analysis, algebraic methods such as the one advanced by Ho & Kalman (1966), solve (1) and (2) for Z, F, G and H, using a minimal dimension of the state space; these algebraic methods also produce solutions in which the latent state variables are completely in the space of the predictor variables.

As the option $\omega = 0$ is not relevant in a dynamic context, the option $\omega \to \infty$ is the only interesting option here, being the straightforward dynamic analogue of (reduced rank) regression analysis. Thinking of the reduced rank regression model as a special case of the state space model, namely the special case of (3), (4) where $F = 0$:

$$Z = XG' \tag{6}$$

$$Y = ZH' \tag{7}$$

the choice for $\omega \to \infty$ is then also analogous to the situation in reduced rank regression where $Z = XG'$ (Davies & Tso 1982). A third way to think of the model (1), (2) is as a principal components analysis of the output, with restrictions $z_t = Fz_{t-1} + Gx_t$ on the component scores Z.

3 Fitting the state space model using alternating least squares with majorization.

Bijleveld & De Leeuw (1991) showed how it is necessary in their case to restrict the parameters to obtain a meaningful solution. When F, G, H and Z are unrestricted, the algorithm can trivially decrease the loss to its lower bound by producing a solution with a very small z, a very small trivial G and a very large H. This can be seen as follows. Define:

$$\sigma_\omega^\alpha = \omega^2 \, \mathbf{SSQ}(Z_\alpha - BZ_\alpha F' - XG_\alpha') + \mathbf{SSQ}(Y - Z_\alpha H_\alpha'), \tag{8}$$

then by letting $Z_\alpha \to 0$, $G_\alpha \to 0$, and $H_\alpha \to \infty$, σ_ω^α becomes minimal. Bijleveld & De Leeuw (1991) chose to restrict Z to $Z'Z = I$ to avoid this problem. It is also possible to restrict either G or H; the latter is sometimes encountered in state space analysis (Harvey 1989), the former resembles the restriction encountered in reduced rank regression. Because of the restriction $Z'Z = I$, estimates for Z cannot be obtained anymore by ordinary least squares, and a (cumbersome) majorization procedure was needed. Bijleveld and De Leeuw's algorithm is named DYNAMALS; for details see Bijleveld & De Leeuw (1991) and Bijleveld (1989).

Implementation of the option $\omega \to \infty$ often induces numerical problems in DYNAMALS, as it may lead to an ill-conditioned Hessian. In practice, this means slow minimization and inaccurate solutions, a problem well known from penalty methods in minimization theory (Murray 1969). In the following we will present an alternative optimization method that easily implements the option $\omega \to \infty$ by writing the latent states as a function of the relevant parameters, and in addition does not need the constraint $Z'Z = I$. We extend the model with additional intercepts in the next section.

4 Extensions of the state space model.

In contrast to common mutivariate techniques, the state space model is sensitive to the manner in which the input and output variables are standarized before the model is estimated. This can be clarified by a simple example. Suppose $z_0 = 1$, $F = \frac{1}{2}$, $G = 0$ and $H = 1$. When data is simulated, a sequence $\frac{1}{2}, \frac{1}{4}, \ldots$ of values of y is generated. Obviously, if this sequence is standarized to having mean value 0 first, the simple structure of the model cannot be recovered anymore. In general terms it means that the fit of a model on any series of points y_t is influenced by the mean value at which it is

entered in the estimation procedure. The same applies to the input variables.

As this mean value is generally unknown, we circumvent this problem by estimating the mean value along with the other parameters. The formulas (1) and (2) then become:

$$z_t = Fz_{t-1} + Gx_t + u \tag{9}$$

$$y_t = Hz_t + v, \tag{10}$$

where the vectors u and v are called intercept-vectors that have properties similar to the properties of intercepts in common regression models.

5 Fitting the extended state space model using direct least squares implementation.

Suppose the latent states *are* completely in the space of the predictor variables. Then, the latent state values $z_1, \ldots, z_T$ are completely defined by the input x_t and the matrices F and G, z_0 and the shift u. This does not mean that Z is determined by the input only, as F, G, z_0 and u depend also on the output. This can be formulated in the following theorem:

Theorem 1 *The state z_t at time t can be written as:*

$$z_t = F^t z_0 + \sum_{k=0}^{t-1} F^k \left(G x_{t-k} + u\right) \tag{11}$$

proof 1 *Clearly, $z_1 = F z_0 + G x_1 + u$, If (11) holds, then:*

$$\begin{aligned} z_{t+1} =& F z_t + G x_{t+1} + u \\ =& F\left(F^t z_0 + \sum_{k=0}^{t-1} F^k \left(G x_{t-k} + u\right)\right) + G x_{t+1} + u \\ =& F^{t+1} z_0 + \sum_{k=1}^{t+1-1} F^k \left(G x_{t+1-k} + u\right) + G x_{t+1} + u \\ =& F^{t+1} z_0 + \sum_{k=0}^{t} F^k \left(G x_{t+1-k} + u\right), \end{aligned}$$

completing the proof by induction.

Corollary 2 *If $\|F\| < 1$, then if $k \to \infty$, $\|F^k z_0\| \to 0$, implying that the influence of the initial state of the system can be neglected after a longer period, if $\|F\| < 1$.*

Corollary 3 *If $\|F\| < 1$, then if $k \to \infty$, $\sum_{t=0}^{k} F^t u \to (I - F)^{-1} u$, implying that the state is, in the limiting case, a linear combination of the input.*

In practice, the terms z_t are best computed recursively, using $z_t = F z_{t-1} + G x_t + u$. This also holds for its partial derivatives: (with δ_{ij} the Kronecker

delta)

$$\frac{\partial z_{it}}{\partial u_k} = \delta_{ik} + \sum_{s=1}^{p} F_{is} \frac{\partial z_{st-1}}{\partial u_k}, \tag{12}$$

similary:

$$\frac{\partial z_{it}}{\partial G_{kr}} = \delta_{ik} x_{rt} + \sum_{s=1}^{p} F_{is} \frac{\partial z_{st-1}}{\partial G_{kr}}, \tag{13}$$

and, finally:

$$\frac{\partial z_{it}}{\partial F_{kr}} = \delta_{ik} z_{rt} + \sum_{s=1}^{p} F_{is} \frac{\partial z_{st-1}}{\partial F_{kr}}. \tag{14}$$

Given the definition of z_t outlined above, the *loss* function can be defined as:

$$\text{loss} = \sum_{t=1}^{T} \sum_{j=1}^{m} \left(y_{jt} - \sum_{k=1}^{p} H_{jk} z_{kt} - v_j \right)^2 \tag{15}$$

The matrix H and the vector v are determined by F, G, z_0, u and the input variables. For each combination of these, H and v can be determined.

We estimate the values of the parameters for which the derivative is closest to zero using the BFGS or Broyden-Fletcher-Goldfarb-Shanno algorithm (Fletcher 1981). This quasi-Newton type algorithm uses the gradient and information on the values of the function to determine the search area for the minimum. It incorporates knowledge on the past iterations in its search, which makes it very efficient and stable.

6 Example: the formation of therapeutic alliance.

We have at our disposal data on the alliance formation between therapist and patient. The data that we will analyse were recorded during 28 consecutivepsychotherapy sessions; for one session alliance scores were lacking, these were ignored during our analyses. The male therapist had passed his basic training as a psychotherapist and was attending a course for a certificate as a behavioral therapist. The male patient was an electro-technical specialist, suffering from severe anxiety attacks with somatic symptoms. monodramatic of conflict and failure, connected with The alliance rating system used to record the alliance scores of both therapist and patient was a modified version of the Penn Helping Alliance Scales (Penn-HAS), for details, see Hentschel, Kieß ling, Heck & Willoweit (1992).

The data were first analyzed using DYNAMALS. We used the option $\omega = 1$, which is the default option in DYNAMALS; all other options were default as well. The therapist's alliance scores served as input, the patient's alliance scores served as output. We modelled one dimension for the latent state. The results are in Table 1. Correlations of the input and output

variables with the state are extremely high, the fit is high, and the norm of F is on the low side, indicating a fairly instantaneous reaction of the patient to the therapist, contradicting our intuitive notion of a gradual process of alliance formation building up in the patient (as well as the therapist), steered by the therapist. Next, we ran the same example with the direct method.

The results from the analysis of the alliance scores using the direct implementation resemble the DYNAMALS results, viz Table 2. The correlation of the input with the state has increased (as expected), and the correlation of the output with the state has decreased (as expected). The norm of F is similar. The fit is comparable to the fit values in DYNAMALS, and it is, as expected, somewhat lower. The direct method found as values for the intercepts $u = -.33$ and $v = .38$.

Next, we analyzed the data with higher dimensionalities for the latent state. We found that a two-dimensional solution describes the data succinctly; for all higher dimensions of z the corresponding singular values of F were zero. The results from the two-dimensional analysis are in Table 3.

On the first dimension, the correlations of input and output variable with the state have remained approximately the same. On the second dimension, the correlations of input and output variable with the state are lower. The

Correlation of therapist's alliance scores with the latent state	.96
Correlation of patient's alliance scores with the latent state	.97
Fit	.94
Norm of F	.18

Table 1: DYNAMALS Analysis of Alliance Scores

Correlation of therapist's alliance scores with the latent state	.98
Correlation of patient's alliance scores with the latent state	.89
Fit	.90
Norm of F	.17

Table 2: Analysis of Alliance Ratings with the Direct Method

Correlation of therapist's alliance scores with the latent states	.94	.50
Correlation of patient's alliance scores with the latent state	.92	.39
Fit	.95	
F	$\begin{pmatrix} 0.36 & -0.12 \\ 0.68 & 0.97 \end{pmatrix}$	
Norm of F	1.19	
intercept u	.23	.21
intercept v	-.18	

Table 3: Two-dimensional Analysis of Alliance Ratings with the Direct Method

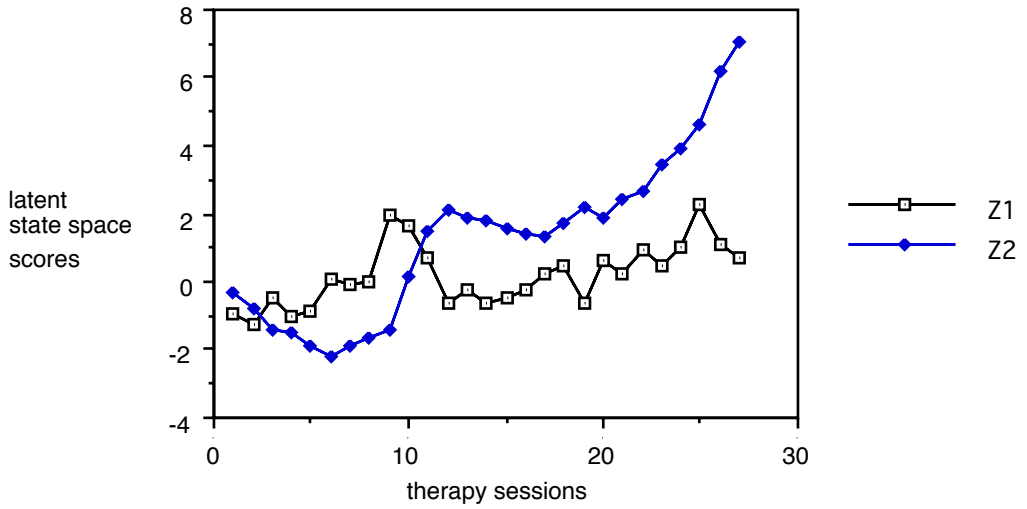

Figure 2: Development of the state space scores over time

largest singular value of F is greater than 1, indicating an ever increasing process. The shape of F indicates that cyclical patterns are captured in this second dimension of the latent state. The latent state space scores over the therapy sessions are in Figure 2, where indeed we see ever increasing state space scores on the second dimension. Summarizing, we can say that the one-dimensional DYNAMALS and direct method analyses gave results that were rather similar. In both cases, the alliance scores of the patient showed an almost instantaneous dependence on the alliance scores of the therapist, with little or no influence of past alliance scores. In the two-dimensional analysis,

the first latent state resembled the latent state from the one-dimensional solutions. On the second dimension however, a latent state was modelled that had ever-increasing scores, strongly governed by the prior latent states, capturing cyclical developments. This second dimension corresponds much more with our idea of alliance formation as a gradual time-dependent process.

7 Discussion.

One of the main advantages of the linear dynamic systems analysis algorithm that was proposed by Bijleveld and De Leeuw was that it can easily incorporate optimal scaling; in fact the DYNAMALS program has been equipped with optimal scaling of nominal and ordinal variables, and can estimate missing values. The direct algorithm proposed here can do the same. In addition, it is far more efficient in that it does not need the majorization step. This is because the constraint $Z'Z = I$ is not needed anymore as a consequence of the direct implementation (Bijleveld & De Leeuw 1991). Non-orthogonality may be an advantage as this makes it possible to model cyclical patterns of development, or it may be a disadvantage as it can complicate the interpretation of solutions.

If the norm of F is much larger than 1, and T is also large, the direct implementation may become very slow due to the necessity of decreasing the step sizes. Alternating least squares using majorization might then be a better option. For investigating latent states that are supposed to follow something that resembles an exponential curve, the direct method is probably best suited. Stability information can be obtained from the Fisher information matrix, an approximation of which is computed in the BFGS algorithm. From practical experience, it appears that both methods converge at acceptable speed.

References

Bijleveld, C. & De Leeuw, J. (1991), 'Fitting longitudinal reduced rank regression models by alternating least squares', *Psychometrika* **56**, 433–447.

Bijleveld, C. C. J. H. (1989), Exploratory linear dynamic systems analysis, PhD thesis, Rijksuniversiteit Leiden.

Davies, P. T. & Tso, M. K. (1982), 'Procedures for reduced-rank regression',

Applied Statistics **31**, 244–255.

Fletcher, R. (1981), *Practical methods of optimization*, Vol. I, John Wiley & Sons, Chicester.

Harvey, A. C. (1989), *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press, Cambridge.

Hentschel, U., Kieß ling, M., Heck, M. & Willoweit, I. (1992), 'Therapeutic alliance: What can be learned from case studies?', *Psychotherapy Research* **2**, 204–223.

Ho, B. L. & Kalman, R. E. (1966), 'Effective construction of linear state-variable models from input/output functions', *Reglungstechnik* **12**, 545–592.

Ljung, L. (1987), *System Identification. Theory for the user*, Prentice-Hall, Englewood Cliffs, NJ.

Murray, W. (1969), *An algorithm for constrained minimization*, Academic Press, London, chapter ?

Willems, J. C. (1990), *From data to model*, Springer Verlag, Heidelberg.