RR 87-16

PROGRESS REPORT ON DYNAMALS

Catrien Bijleveld

November 1987

Preface

This Progress Report describes the DYNAMALS computer program in the version which is operational at the time of writing.

This version of DYNAMALS has been written in SAS-version 5. It is planned to have a Fortran version operational by mid 1988. Future versions of DYNAMALS will be more user friendly and produce more elaborate output.

November 1987

Introduction

DYNAMALS is an acronym of Linear DYNAMical Systems analysis (or state space analysis) by Alternating Least Squares. Linear dynamical or state space models are a very general class of models of which a number of specific models are special cases; amongst the best known of these are the models for redundancy analysis and dynamic factor analysis. These specific models can be specified by using the appropriate options.

In Chapter I a theoretical framework for the program is sketched. In Chapter II guidelines for running the program are given. In Chapter III several examples are discussed that together give an overview of the possible applications of DYNAMALS.

I. THEORY OF THE DYNAMALS PROGRAM

I.1 The data matrix

DYNAMALS analyzes a *T* by *m* data matrix, where the number of rows *T* equals the number of observation points and the number of columns *m* equals the number of variables.

The observation points that serve as input for the DYNAMALS program have a different meaning than their counterparts used in the other -ALS programs. In those cases, the observations are imperfectly measured, replicated, and independent measurements of one phenomenon. DYNAMALS, however, analyzes observations that are dependent, whether that may be through time or space or through any other criterion. In the two dimensional datamatrix, no replication of observations over individuals occurs; DYNAMALS essentially analyzes n=1 designs. We will in the following paragraphs always use the time axis for illustration.

The variables can be divided into two groups: one set of input variables and one set of output variables. We will call the number of input variables d_i and the number of output variables d_o. The two sets play an asymmetric role in the model.

I.2 The model

For measurements of some observation unit at *m* $(= d_i + d_o)$ variables and at *T* time points, the state space model essentially states that:

- the output at time *t* is influenced by the input at time *t*, mediated by a latent factor z;
- the state at time *t* is influenced by the state at time *t-1*.

If $\mathbf{x}_t$ is the input vector at time t, $\mathbf{y}_t$ the output vector at time t, and $\mathbf{z}_t$ the latent state vector at time t the model can be formulated as follows:

$$\mathbf{z}_t = \mathbf{F}\,\mathbf{z}_{t-1} + \mathbf{G}\,\mathbf{x}_t \qquad \text{(system equation)}$$

$$\mathbf{y}_t = \mathbf{H}\,\mathbf{z}_t \qquad \text{(measurement equation)}$$

A diagram of this model is presented in Figure 1.

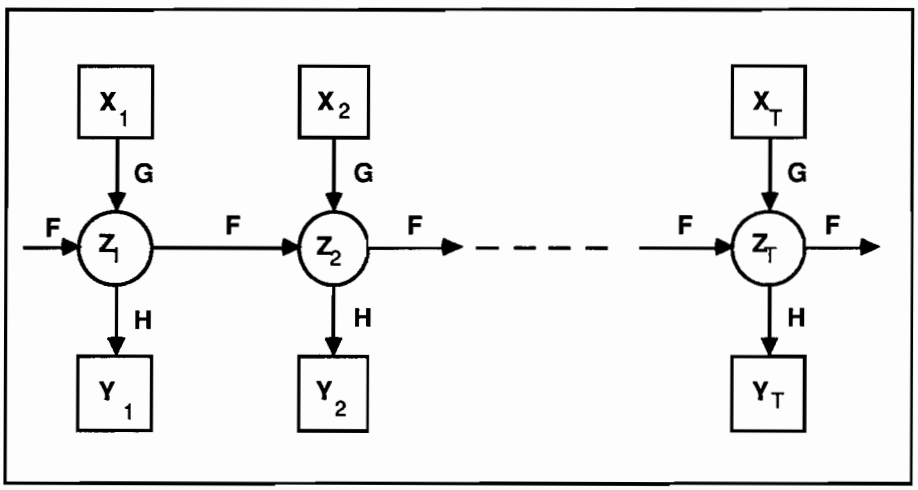


Figure 1. Geometrical representation of the state space model

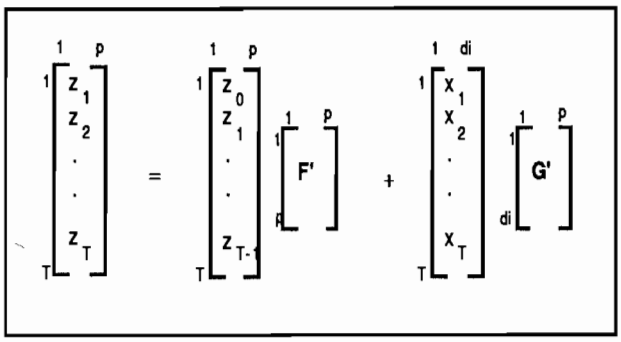
In matrix notation the model can be written as:

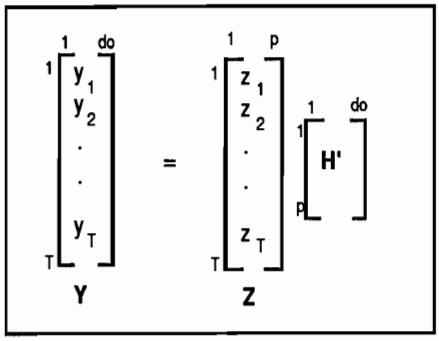
$$\mathbf{Z} = \mathbf{BZF}' + \mathbf{XG}' \qquad \text{(system equation)}$$

$$\mathbf{Y} = \mathbf{ZH}' \qquad \text{(measurement equation)}$$

with $\mathbf{BZ} = (\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_{T-1})$. .

Diagrams of the system and measurement equations are presented in Figures 2 and 3.





Figures 2 (left) and 3 (right): the matrix equations of the state space model

I.3 The algorithm

Fitting the state space model using least squares amounts to minimizing

$$\sigma(\mathbf{Z},\mathbf{F},\mathbf{G},\mathbf{H}) = \omega^2\sigma_1(\mathbf{Z},\mathbf{F},\mathbf{G}) + \sigma_2(\mathbf{Z},\mathbf{H})$$

$$= \omega^2\mathrm{SSQ}(\mathbf{Z} - \mathbf{BZF}' - \mathbf{XG}') + \mathrm{SSQ}(\mathbf{Y} - \mathbf{ZH}') \qquad (1)$$

The ω is a weight that takes into account the relative importance of the input. Varying values of ω produce different types of solutions. At the one extreme where $\omega = 0$, principal components analysis of the output is obtained; redundancy analysis appears for $\mathbf{F} = 0$ when $\omega = \infty$. In the normal case $\omega = 1$ (see De Leeuw and Bijleveld, 1987).

The loss function cannot be minimized in a straightforward way, because both the latent state space **Z** and the transition matrices **F**, **G** and **H** are unknown. Therefore, **Z**, **F**, **G** and **H** are estimated using an alternating least squares algorithm in which during each iteration (in the so called **F G H** step) **F**, **G**, and **H** are first estimated with **Z** fixed, and subsequently (in the **Z** step) **Z** is estimated with **F**, **G**, and **H** fixed.

Before computations can be started, a value for the hypothetical $\mathbf{z}_0$ has to be selected. By letting **B** have the following shape:

$$\mathbf{B} = \begin{matrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ & \dots & & & \dots & \\ 0 & 0 & 0 & & 1 & 0 \end{matrix}$$

$\mathbf{z}_0$ is set equal to $\mathbf{z}_1$. Other choices for $\mathbf{z}_0$ exist, such as $\mathbf{z}_0 = 0$, $\mathbf{z}_0 = \mathbf{z}_T$ or $\mathbf{z}_0 = \text{mean}(\mathbf{z}_t)$; these options will be added to later versions of the program. **Z** is restricted by requiring **Z'Z** = **I**, and to start the first iteration with the estimation of the unknown transition matrices **F**, **G** and **H**, the first *p* columns of the singular value decomposition of the output **Y** are used as initial estimates of **Z**.

Because of the restriction **Z'Z** = **I**, ordinary least squares does not suffice to estimate **Z**. A quadratic function is constructed whose values are always greater than or equal to those of the loss function. If the constructed quadratic function is minimized, the loss function will have decreased too; minimizing the constructed quadratic function in each iteration, the minimum of the loss function will eventually be approached. This numerical method of minimizing an unknown function via the minimization of a known function that has values always greater or equal than the unknown function is called *majorization*; see DeLeeuw & Bijleveld (1987,1988). In this case the constructed quadratic function looks as follows:

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

with $\mathbf{S} = \gamma^{-1}\,(\omega\mathbf{B}'\mathbf{P}_1\mathbf{F} + \mathbf{P}_2\mathbf{H} - \omega\mathbf{P}_1)$,

$\mathbf{P}_1 = \omega\,(\mathbf{Z}_{old} - \mathbf{B}\mathbf{Z}_{old}\mathbf{F}' - \mathbf{X}\mathbf{G}')$,

$\mathbf{P}_2 = \mathbf{Y} - \mathbf{Z}_{old}\mathbf{H}'$,

$\gamma = \lambda^2_{max}\,(\omega(\mathbf{I}-\mathbf{B}\otimes\mathbf{F}'))\;//\;(\mathbf{I}\otimes\mathbf{H})$, where // stands for vertical concatenation and $\lambda_{max}(.)$ is the largest singular value of a matrix.

By minimizing the quadratic loss function at the right hand side of the inequality iteratively and setting $\mathbf{Z}_{new} = \mathbf{Z}$, where $\mathbf{Z}$ is the minimizer, the loss function at the left hand side will decrease. In order to do this, $\mathbf{SSQ(Z - (Z}_{old} + \mathbf{S))}$ has to be minimized under the restriction that $\mathbf{Z'Z = I}$. This is a Procrustus problem and it is solved by setting $\mathbf{Z = PQ'}$ with $\mathbf{SVD(Z}_{old} + \mathbf{S}) = \mathbf{P\phi Q'}$.

I.4 The fit

The minimum of the loss function

$$\sigma(\mathbf{Z,F,G,H}) = \omega^2\mathbf{SSQ(\ Z - BZF' - XG')\ +\ SSQ(\ Y - ZH'),} \qquad (1)$$

is always smaller than the minimum of $\sigma(\mathbf{Z},0,0,0)$. In formula:

$\min\ \sigma(\mathbf{Z,F,G,H}) \leq \min\ \sigma(Z,0,0,0)\ =\ \omega^2 p + d_o$,

so that the normalized fit equals $(\omega^2 p + d_o\ -\ \text{loss}\ )\ /\ (\omega^2 p + d_o)$.

To give an impression of the amount of output variance that has been explained, DYNAMALS also prints all squared correlations between input, state and output variables and the average of these squared correlations.

I.5 Rotations

After minimizing $\omega^2\ \mathbf{SSQ(\ Z - BZF' - XG')} + \mathbf{SSQ(\ Y - ZH')}$, $\mathbf{Z}$ is rotated. This is achieved by multiplying $\mathbf{Z}$ by a rotation matrix $\mathbf{R}$; the other parameters are then rotated along with it. Clearly,

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

$$\omega^2\ \mathbf{SSQ\ (\ ZR - BZRR'F'R - XG'R)} + \mathbf{SSQ\ (Y - ZRR'H')}.$$

Several options exist for the choice of **R**; one can choose to diagonalise either **R'H'HR** or **R'GX'XG'R**. In reduced rank regression (where **F** = 0) **R** is chosen such that **R'**(ω^2**GX'XG'** + **H'H**) **R** is diagonal. If ω^2 is very large then ω^2 **R'GX'XG'R** is diagonal (i.e., the component scores of the input are orthogonal). If ω^2 is very small, then **R'H'HR** will be diagonal, which means that the component scores of the output are orthogonal.

These 3 cases have been implemented as an option of the program:

either **R'**ω^2**GX'XG'R**,

R'H'H R, or

R'(ω^2**GX'XG'** + **H'H**) **R** is diagonalized.

In all three cases **R** is found by computing the eigenvectors of the matrixproduct enclosed by **R'** and **R**. The transition matrices are also rotated as follows:

G	→	**R'G**
H	→	**H R**
F	→	**R'F R**

I.6 Interpreting the results: The correlations

As in other multivariate analyses, the interpretation of the results is given by the correlations of the original variables with the constructed variables. The correlations between the state and output variables reflect the extent to which the respective output variables have been predicted. The correlations between the input and state variables reflect the predictive value of the respective input variables. The transition matrices **F**, **G** and **H** are useful mainly for computational purposes.

I.7 Predictions

When DYNAMALS is used for prediction purposes, the correlations between input, output and state variables are of less interest. DYNAMALS always computes and prints three default predictions; these are the estimates of $\mathbf{y}_{T+1}$ given that $\mathbf{x}_{T+1}$ is equal to zero, one and two standard deviations from the mean. Using the estimated $\mathbf{F}$, $\mathbf{G}$, $\mathbf{H}$ and last $\mathbf{Z}$, $\mathbf{z}_T$, the DYNAMALS user may compute estimates of future output, given other inputs.

I.8 Plots

For $p > 1$, DYNAMALS plots the scores on the first two state variables, as well as the correlations of the input and output variables with those two state variables. This plot serves the same purpose as the plots of object scores and component loadings that are given in other -ALS programs.

I.9 Missing data

Because of the (mostly time-) order in the type of data that DYNAMALS analyzes, missing data having to be treated actively, i.e. they have an effect on the solution. In each new iteration the missing values are substituted with optimal estimates that have been computed by least squares minimization of the loss function for the missing values with $\mathbf{F}$, $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{Z}$ fixed.

I.10 Ordinal and nominal variables

In the same manner in which missing values are substituted with optimal estimates, the categories or order values of nominal and ordinal variables may be transformed to optimal quantifications.

I.11 Flow chart

START

$\mathbf{X} \leftarrow$ STAND($\mathbf{X}$)
$\mathbf{Y} \leftarrow$ STAND($\mathbf{Y}$)
$q \leftarrow$ number of input variables
$p \leftarrow$ chosen number of state variables
$\omega \leftarrow$ value of weight ω
$\mathbf{B} \leftarrow {}_T1\ 0\ 0\ \ldots\ 0_T$
$\mathbf{Z} \leftarrow \mathbf{Z}_0 \leftarrow$ SVD($\mathbf{Y}$)(,1:p)
$\mathbf{X/Y}_{(mis,)} \leftarrow$ MEAN($\mathbf{X/Y}$(mis+1,) + $\mathbf{X/Y}$(mis-1,))
$\mathbf{X/Y}_{(nom/ord)} \leftarrow$ ($\mathbf{X/Y}$(nom/ord))
$n \leftarrow 0$

NEW ITERATION

F_G_H_step

$n \leftarrow n+1$

$\mathbf{F} \leftarrow [(\mathbf{BZ}||\mathbf{X})^{+}\mathbf{Z}_{(1:p,)}]'$ → new $\mathbf{F}$
$\mathbf{G} \leftarrow [(\mathbf{BZ}||\mathbf{X})^{+}\mathbf{Z}_{(p+1:p+q,)}]'$ → new $\mathbf{G}$
$\mathbf{H} \leftarrow [\mathbf{Z}^{+}\mathbf{Y}]'$ → new $\mathbf{H}$
$\mathbf{P}_1 \leftarrow \omega\,(\mathbf{Z} - \mathbf{BZF}' - \mathbf{XG}')$
$\mathbf{P}_2 \leftarrow \mathbf{Y} - \mathbf{ZH}'$

Z_step

$\gamma \leftarrow \lambda^2_{max}\omega(\mathbf{I}-\mathbf{B}\otimes\mathbf{F}')///(\mathbf{I}\otimes\mathbf{H})$
$\mathbf{S} \leftarrow \gamma^{-1}(\omega^2\mathbf{B}'\mathbf{P}_1\mathbf{F}-\omega^2\mathbf{P}_1+\mathbf{P}_2\mathbf{H})$
$\mathbf{P},\mathbf{Q} \leftarrow$ SVD($\mathbf{Z}_{old}+\mathbf{S}$)
$\mathbf{Z} \leftarrow \mathbf{PQ}'$ → new $\mathbf{Z}$
$\mathbf{P}_1 \leftarrow \mathbf{Z} - \mathbf{BZF}' - \mathbf{XG}'$

missing values_step

$\mathbf{D}^{(1)}\mathbf{X} \leftarrow$ STAND($\mathbf{D}^{(1)}\mathbf{X}$)
$\mathbf{D}^{(1)}\mathbf{Y} \leftarrow$ STAND($\mathbf{D}^{(1)}\mathbf{Y}$)
$\mathbf{X}_{pred} \leftarrow \mathbf{X} + (\lambda_{max}(\mathbf{G}))^{-1}\,\mathbf{P}_1\mathbf{G})$
$\mathbf{Y}_{pred} \leftarrow (\mathbf{ZH}')$
$\mathbf{D}^{(2)}\mathbf{X} \leftarrow \mathbf{D}^{(2)}\mathbf{X}_{pred}$
$\mathbf{D}^{(2)}\mathbf{Y} \leftarrow \mathbf{D}^{(2)}\mathbf{Y}_{pred}$
$\mathbf{X} \leftarrow$ STAND $((\mathbf{D}^{(1)} + \mathbf{D}^{(2)})\mathbf{X})$ → estimates of missing values $\mathbf{X}$
$\mathbf{Y} \leftarrow$ STAND $((\mathbf{D}^{(1)} + \mathbf{D}^{(2)})\mathbf{Y})$ → estimates of missing values $\mathbf{Y}$

optimal scaling_step

$\mathbf{X}_{(nom)} \leftarrow$ OPSCAL $(1,\mathbf{X}_{pred}, \mathbf{X}_{(nom)})$ → quantifications of nominal input
$\mathbf{X}_{(ord)} \leftarrow$ OPSCAL $(2,\mathbf{X}_{pred}, \mathbf{X}_{(ord)})$ → quantifications of ordinal input
$\mathbf{Y}_{(nom)} \leftarrow$ OPSCAL $(1,\mathbf{Y}_{pred}, \mathbf{X}_{(nom)})$ → quantifications of nominal output
$\mathbf{Y}_{(ord)} \leftarrow$ OPSCAL $(2,\mathbf{Y}_{pred}, \mathbf{X}_{(ord)})$ → quantifications of ordinal output
$\mathbf{P}_1 \leftarrow \omega\,(\mathbf{Z} - \mathbf{BZF}' - \mathbf{XG}')$
$\mathbf{P}_2 \leftarrow \mathbf{Y} - \mathbf{ZH}'$
$\sigma(\mathbf{Z},\mathbf{F},\mathbf{G},\mathbf{H}) \leftarrow \mathbf{SSQ}(\mathbf{P}_1) + \mathbf{SSQ}(\mathbf{P}_2)$

END OF ITERATION

yes ← convergence?

no

yes ← n > 100? → no

___rotation_____

(according to the options)

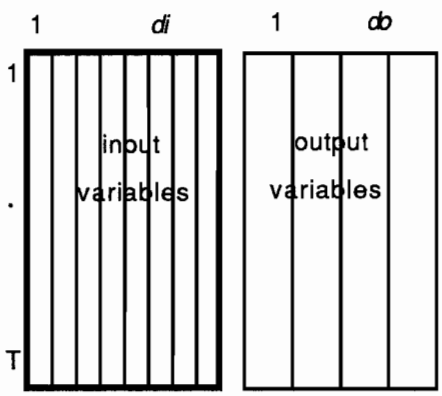
Z ← **Z R** → final **Z**
F ← **R'** **F R** → final **F**
G ← **R'G** → final **G**
H ← **H R** → final **H**

PRINTING

II. USING DYNAMALS

II.1 Input data for the program

The *T* by m data matrix should be presented to DYNAMALS in that form. DYNAMALS cannot upon request leave out certain variables or observations or perform other modifications. The values should have at least one blank column in between them for SAS to read them. The rows must be the time points in their actual order, and the columns must represent the variables. The input variables should come first, the output variables last. The data should look like this:



II.2 Job control

DYNAMALS consists of two parts. The first part reads the input parameters. After the first part the data must come. After the data, the second part that is much larger and contains the algorithm itself must come. Each time when one wants to run DYNAMALS these three components, the first part - the data - the second part, must be joined. The data have been

described in the last paragraph, the second part has been described in Chapter I. An example of the first part is shown below.

```
DATA PARM1;
     INPUT     P T DI DO IDF IDG PRINTLS WEIGHT
          ROTATION CONOUTIT INIT CONINIT ITERMAX;
CARDS;
     2     46   2     2     1     1     1     1
     3     .0005       1     .0005       100
;
DATA MLEV1;
     INPUT MLEV @@ ;
CARDS;
     3     3     3     3
;
DATA LAG1;
     INPUT LAG @@ ;
CARDS;
     2     0
;
TITLE 'inoculation data: measles & whooping cough';
DATA TIME1;
     INPUT DAT @@;
     CARDS;
```

II.3 Input parameters for the program

The input parameters for the program must be specified in three SAS datasets named PARM1 and MLEV1 and LAG1. PARM1 contains respectively:

P dimensionality of the state or number of state variables
T: number of time points
DI: number of input variables
DO: number of output variables
IDF: identification of the transition matrix **F**. If IDF=0, **F** is zero, which implies that no time or other dependency is fitted. If IDF=1, dependency across observations is assumed and the matrix **F** is estimated alongside with **G** and **H**.

IDG: identification of the transition matrix **G**. If IDG=0, **G** is zero, which implies that no input is modeled and a dynamic factor analysis model is fitted. If IDG=1, input is modeled and the matrix **G** is estimated alongside with **F** and **H**.

PRINTLS: option to print (1) or not to print (0) the history of iterations.

WEIGHT: value for the weight ω. Default is 1.

ROTATION: options for rotations, requiring - $\mathbf{R}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{R}$ (1)
diagonality of - $\mathbf{R}'\mathbf{H}'\mathbf{H}\mathbf{R}$ (2)
- $\mathbf{R}'(\omega^2\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}' + \mathbf{H}'\mathbf{H})\,\mathbf{R}$ (3)

CONOUTIT: convergence criterion for the loss.

INIT: option to use inner iterations in the **Z** step. It may make a- slight- difference in speed, exactness and CPU timeto use (1) this option.

CONINTIT: convergence criterion for the loss.

NOTE. No default values are available. The user must always specify a value for the options; a blank will produce an error!!

MLEV1 contains the measurement levels of the variables. The categories for the respective measurement levels are:
1 = nominal measurement level
2 = ordinal measurement level
3 = interval/numerical measurement level

LAG1 contains the lags for the input variables. If, for example, LAG1 is {3 2 4}, this implies that input variable 1 starts influencing the output only after a lag of three time units. For input variable 2 this lag is two; for input variable 3 it is four. When there is no lagged influence of the input, LAG1 should consist of only zero's.

II.4 Printed Output

1 inoculation data: measles & whooping cough 1
11:38 FRIDAY, JANUARY 29, 1988

D Y N A M A L S EXAMPLE
FITTING THE INFLUENCE OF INOCULATIONS AGAINST MEASLES & WHOOPINGCOUGH (X1,X2) ON INCIDENCE OF MEASLES & WHOOPING COUGH (Y1,Y2)

DATA COVER THE PERIOD DECEMBER 1982 - SEPTEMBER 1986, GATHERED BY CATRIEN BIJLEVELD FROM CENTRE DE SANTE CRETE ZAIRE-NIL, PREFECTURE KIBUYE, RWANDA.

****
FITTING LINEAR DYNAMICAL SYSTEMS BY ALTERNATING LEAST SQUARES
****

D Y N A M A L S ***** VERSION 1.1

SPECIFI dim Z n of T dim X dim Y F G
rotat weight maxit convcr innit conin

CATIONS 2 46 2 2 1 1
3 1 100 500E-6 0 500E-6 0

LINEAR DYNAMICAL SYSTEMS ANALYSIS

ROTATION 3 ROTATES TO DIAGONALITY OF RT W2G XT X GT + HT H R

THE SPECIFIED LAGS FOR THE RESPECTIVE INPUT VARIABLES ARE:

LAGS	COL1	COL2
ROW1	2	0

THE NUMBER OF (TIME) POINTS USED FOR THE ANALYSIS IS:

NT	COL1
ROW1	44

THE EVENTUAL NORMALIZED FIT AMOUNTS TO:

FIT	COL1
ROW1	0.9481

1 inoculation data: measles & whooping cough 2
11:38 FRIDAY, JANUARY 29, 1988

THE DEVELOPMENT OF THE VALUE OF THE LOSS WAS

VALUEOF	loss
ROW1	0.443757
ROW2	0.317422
ROW3	0.262576
ROW4	0.237657

ROW5	0.225278
ROW6	0.218557
ROW7	0.214613
ROW8	0.212148
ROW9	0.210525
ROW10	0.209410
ROW11	0.208617
ROW12	0.208036
ROW13	0.207601

THE NUMBER OF ITERATIONS WAS

NITER	COL1
ROW1	13

THE F-CANONICAL COEFFICIENTS ARE:

FEC	COL1	COL2
ROW1	1.0139	-0.271
ROW2	.60634	1.173

THE LARGEST EIGENVALUE OF F IS:

EIGENF	COL1
ROW1	1.3479

THE G-CANONICAL COEFFICIENTS ARE:

GEC	COL1	COL2
ROW1	-.0114	-0.004
ROW2	-.1605	-.0537

THE H-CANONICAL COEFFICIENTS ARE:

HEC	COL1	COL2
ROW1	0.8678	.44342
ROW2	0.8473	-0.457

1 inoculation data: measles & whooping cough 3
11:38 FRIDAY, JANUARY 29, 1988

CORRELATIONS BETWEEN THE INPUT- AND THE STATE VARIABLES:

CORRIN	COL1	COL2

ROW1	-.1812	-.1559
ROW2	-.0089	.07233

CORRELATIONS BETWEEN THE OUTPUT - AND THE STATE VARIABLES:

CORROUT	COL1	COL2
ROW1	.86794	.44245
ROW2	.84726	-.4566

PROPORTION OF VARIANCE OF THE RESPECTIVE STATE VARIABLES EXPLAINED BY THE INPUT VARIABLES:

PROPSTAT	COL1	COL2
ROW1	.31982	7.3146

PROPORTION OF VARIANCE OF THE RESPECTIVE OUTPUT VARIABLES EXPLAINED BY THE STATE VARIABLES:

PROPOUT	COL1	COL2
ROW1	94.909	92.633

MEAN PROPORTION OF VARIANCE EXPLAINED BY THE STATES

PROPMEAN	COL1	COL2
ROW1	47.614	49.974

MEAN OVERALL PROPORTION OF EXPLAINED OUTPUT VARIATION:

PROPTOT	COL1
ROW1	93.771

COMPUTING THE FUTURE Y FOR XT+1=0 0 . . .

YFUT	COL1	COL2
ROW1	255.84	235.45

COMPUTING THE FUTURE Y FOR XT+1=1 1 . . .

DYNAMALS progress report **using**

YFUT	COL1	COL2
ROW1	223.12	202.74

COMPUTING THE FUTURE Y FOR XT+1=2 2 . . .

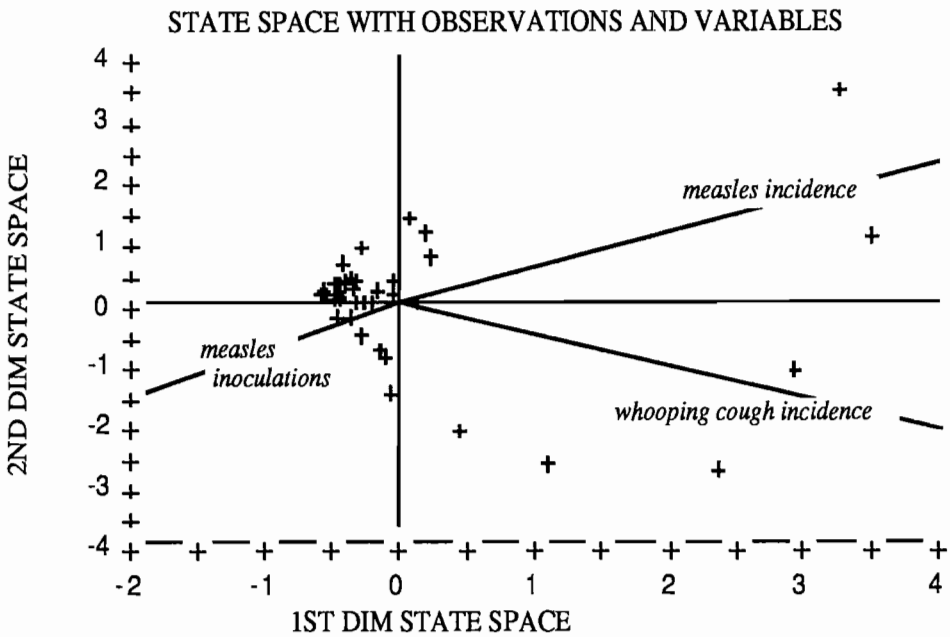
YFUT	COL1	COL2
ROW1	190.41	170.02

1 inoculation data: measles & whooping cough 5
11:38 FRIDAY, JANUARY 29, 1988

THE ESTIMATED STATES ARE:

ZE	COL1	COL2
ROW1	-.0473	.10393
ROW2	.13314	-.0444
ROW3	-.1571	.16849
ROW4	-.3437	.20339
ROW5	-.4004	.26291
ROW6	-.4383	.01582
ROW7	-.2012	-.0188
ROW8	-.0513	.30531
ROW9	.23586	.69011
ROW10	.18134	1.11
ROW11	.08134	1.3541
ROW12	-0.285	.84959
ROW13	-.4261	0.6292
ROW14	-.4096	0.3135
ROW15	-.3688	.35632
ROW16	-.3279	.29834
ROW17	-0.255	-.0444
ROW18	-0.32	-.0277
ROW19	-0.437	.08536
ROW20	-.4789	.24876
ROW21	-.5362	0.1166
ROW22	-.5557	.10305
ROW23	-.5698	.11161
ROW24	-.5635	.16145
ROW25	-.5455	.13324
ROW26	-.4663	.13407
ROW27	-.4394	0.1551
ROW28	-.4551	.26126
ROW29	-.4473	.26886
ROW30	-.4064	.25395
ROW31	-0.487	.01116
ROW32	-.4869	.00894
ROW33	-.4607	-.2685
ROW34	-.3566	-.2792
ROW35	-.2836	-.5654

ROW36	-.1536	-.8238
ROW37	-.0997	-0.932
ROW38	-.0725	-1.55
ROW39	.45112	-2.159
ROW40	1.1002	-2.701
ROW41	2.3669	-2.809
ROW42	2.9371	-1.162
ROW43	3.5038	1.061
ROW44	3.2732	3.4114



III. EXAMPLES

III.1 Exploration: Rwanda inoculation data

Data were gathered from a small missionary hospital in Rwanda, Central Africa (Centre de Santé Crete Zaire-Nil, Kibuye). For 46 consecutive months the number of inoculations and number of cases of measles and whooping-cough in the hospital's district were registered. One is primarily interested in whether and to what extent inoculation influences disease incidence. The variables are presented schematically in Table III.1.

Table III.1: input and output variables of Rwanda inoculation data

measinoc	number of inoculations against measles	INPUT
whooinoc	number of inoculations against whooping-cough	
meascase	incidence of measles	
whoocase	incidence of whooping-cough	OUTPUT

Graphs of the variables are in Figures 4 and 5.

We first ran the program several times for measles and whooping cough seperately. We wanted to assess the optimal lag for inoculating to influence incidence; it is very unlikely that inoculating would take effect immediately. In fact, running the program for lag 0, one finds that higher inoculating levels lead to higher incidence: the more inoculations, the more people fall ill! Clearly, an unwanted relation is fitted in this way, similar to: "the greater the number of firemen, the larger the damage" ; an explanation for this may be that the moment an

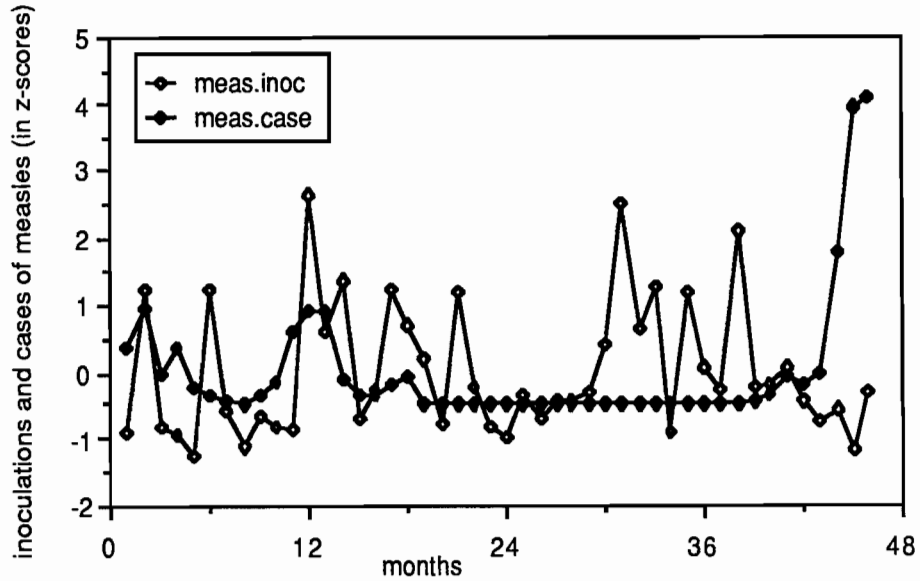


Figure 4. Measles inoculations and incidence data

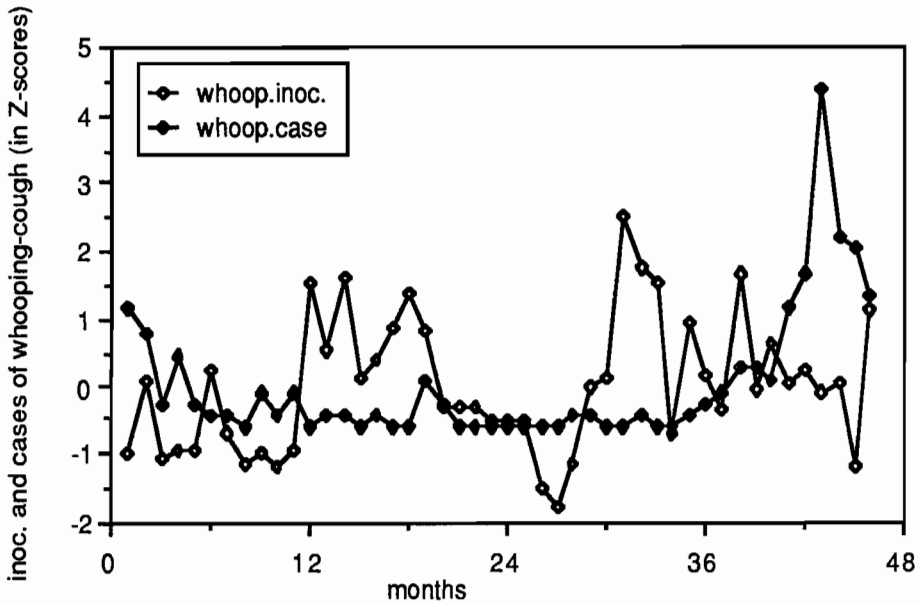


Figure 5. Whooping cough inoculations and incidence data

epidemic rose, inoculating was intensified. We looked up some relevant characteristics of the diseases:

	measles	whooping-cough
incubation period	10-14 days	7-17 days
duration	2 weeks	6 weeks
contagiousness	maximal at 3rd day	1 month

If someone is infected with measles today, he can still infect another person after a maximum period of 4 weeks. If however a person is infected with whooping cough today, he can still infect another person after 6 weeks. Persons who have measles are classified as infected for a period of 2 weeks; those who have whooping cough are classified as infected for 6 weeks. We therefore expect that the time-lag with which whooping-cough inoculations take effect is longer than the time-lag for measles. We ran several analyses for whooping cough and measles seperately to identify the optimal lag (ie. the lag that gives the best fit). The best lag for measles turned out to be 2 months. We were however unable to find a satisfactory solution for whooping cough. For each lag from zero to 9 months, incidence went up for increasing inoculating levels; this may have been caused by the fact that incidence of whooping cough is very low over the 46 months (for illustration purposes a two dimensional solution with lags 2 and 0 for measles and whooping cough was given in section II. 4)

The correlations between the in- and output variable and the state variable for the analysis with lag 2 for measles are given in Table III.2.

Table III.2: correlations of input and output variable with the state variable

state variable	z
measinoc	.223
meascase	-.981

The fit of the solution was .944. The number of inoculations against measles is through the state variable z negatively correlated with the incidence of measles. Predictions of future incidence are thus inversely related to inoculating; for inoculating levels 0, 1 and 2 standard deviations from the mean the predictions of future measles cases are: 310, 262 and 215 respectively.

III.2. Confirmation: Mexico historical data

Historical data about the region of Central Mexico from Mexican and Spanish colonial archives were obtained for the years 1710-1809. Three groups of variables were of interest: one group of variables measuring the agricultural (Mexican) economy, one group of variables measuring the monetary (Spanish) economy and bureaucratic reforms, and the last group of variables measuring the level of consumption. Table III.3 lists the three groups of variables.

Table III.3: input and output variables of Central Mexican historical data

agricultural economy	tributarios	number of tributarios, inhabitants of the villages	INPUT
	crisis	number of crisis years with epidemics and crop failures	
monetary economy	silver1	silver production of the silver mines	
	mining taxes	tax collected on the silver production	INPUT
	silver2	silver production as registered at the mintage	
level of con-sump-tion	land disputes	number of land disputes of Indian villagers against great landowners, reflecting the shortage of land in the Indian villages	
	maize prices	the maize prices	
	revolts	number of peasants' revolts	OUTPUT
	monuments	number of monuments built, reflecting surplus wealth of the Mexican government	

Classical historical theory states that the monetary economy (and especially the silver variables) determines the development of the level of consumption. New dissident theories (see Ouweneel and Bijleveld, 1987) increasingly tend to view the rural and agricultural

economy as a just as decisive factor. In order to test the two theories the two 'determinants' (monetary and agrarian economy) served as input to predict the output level of consumption. Comparison of the respective correlation coefficients of the two types of input variables with the states would indicate the relative importance of each of the two determinants.

Because it was presumed that the variable tributarios, the number of village inhabitants, would start taking effect on the output variables only after approximately 2 years, the tributarios series from 1710-1807 was used; for all other variables the series from 1712-1809 were used. A uni-dimensional state space was chosen, as theory has it that land disputes, maize prices, revolts and monuments are four indices of one concept, the level of consumption. DYNAMALS was run several times, with lags of 0, 1, and 2 years for the output variables. The results are summarized in Table III.4.

Table III.4: correlations of input and output variables with the state variable

lags:	0	1 year	2 years
state variable	z	z	z
tributarios	.822	.843	.850
crisis	.498	.497	.492
silver1	.738	.782	.818
mining taxes	.727	.758	.784
silver2	.691	.740	.778
land disputes	.902	.901	.899
maize prices	.744	.740	.739
revolts	.550	.758	.550
monuments	-.876	-.874	-.873
fit	.683	.680	.679
transition matrix F	.875	.882	.905

The solution for lag zero has the best fit, which indicates that in 18th century Mexico monetary changes and crisis years took effect quickly. The fact that the results are quite stable over the various analyses with different lags, might indicate that the influence of such changes could stretch out over longer periods. This is also reflected in the value of the matrix

F, which is near to one. It is immediately clear that the variables tributarios and crisis play an important role in explaining developments in the output variables that measure the level of consumption; tributarios, in fact, loads the highest of all input variables, but crisis also keeps up a reasonable correlation with the state variable. Of the monetary variables, silver1 has the strongest correlation with the state variable, but differences with the correlations of mining taxes and silver2 are slight. In fact, from other analyses, it appeared that the three monetary variables measure approximately the same thing. Clearly, it is not tenable to maintain that the monetary economy is the sole determinant of the level of consumption; the agrarian variables play a just as important role.

III.3. Prediction: Blue nile data

A dataset was obtained from the Ministry of Irrigation in Khartoum in Sudan, pertaining to the continuous measurements made at and around the barrage in the Blue Nile at El Roseires. During the summer, the usually low levels of water in the Blue Nile swell to immense amounts due to the rainy season upstream in Ethiopia. With available data on inflow of the Blue Nile at the border with Ethiopia, water contents of the lake, and discharge through the barrage's gates, it is possible to predict the total amount of water (contents of the lake + discharge) that would accumulate at a certain time point. With this information one could then predict the amount of water that must be let through at that certain time point to prevent the dam from overflowing or other disasters. A schematic representation of input and output is in Table III.5.

Table III.5: input and output variables of Blue Nile data

INFLOW	mean inflow per day measured with the gauge ruler	INPUT
CONTENTS	contents of the lake measured with the gauge ruler	
DISCHARGE	mean discharge in Mm3 per day	OUTPUT

It should be remarked that the model fitted here is a very minimal and certainly not optimal model to predict the future total amount of water. Many other factors should also be taken into account, such as rainfall over the lake, inflow from downstream tributaries to the lake, evaporation, infiltration, etc. In addition, the variables contents and discharge should be tranformed into one and the same scale, summed and analyzed as one variable. This simple model is, however, useful for illustration purposes.

DYNAMALS was run on the measurements of the first 150 days in 1986, which corresponds with the dry season. One discharge measurement was missing; its least squares optimal estimate was given at the end of the printed output. It had been estimated that the inflow water would take between 0 and 1 days to arrive at the dam; DYNAMALS was thus run twice. Both analyses converged in 11 iterations. The results are given in Table III.6.

Table III.6: transition matrices F, G and H

lag	0 days	1 day
F	.958	.936
G	.052	.077
H	.927	.926
	.713	.708
fit	.776	.773

As the solution with lag 0 has the best fit, that solution was chosen for further discussion. The small coefficient for the **G** matrix is somewhat deceptive; the correlation of inflow with the state variable is actually .814. The one step ahead forecasts for input zero, one and two standard deviations from the mean are given in Table III.7.

Table III.7: one step ahead forecasts of contents and discharge for input 0,1,2 standard deviations (STD's) from the mean

	contents	discharge
STD 0	452.53	21.328
STD 1	457.43	70.374
STD 2	462.34	119.42

III.4 Spatial dependency: Microalgae data

The dataset considered here contains observations on the ecology of the Canadian arctic and has been analyzed and described in Gosselin, Legendre et al. (1986). The distribution of sea-ice microalgae under the ice-cap of the Arctic is postulated to depend on the level of irradiance at the bottom of the ice, which in turn is determined by the snow depth, ice thickness and salinity. The irradiance at the bottom of the ice could not be measured, but the irradiance and the amount of chlorophyll under the ice may serve as indicators of it, "irradiance at the bottom of the ice" is thus clearly a latent variable. All variables are summarized in Table III.8.

Table III.8: input and output variables of microalgae data

SNOW DEPTH	snow depth	
ICE THICKNESS	ice thickness	INPUT
SALINITY	salinity	
LIGHTUNDER	light under the ice	
CHLOROPHYLL	chlorophyll	OUTPUT

Measurements were taken at 100 points on the ice, at intervals of approximately 5 metres each. It was thus assumed that the measurements were dependent, albeit not through time, but through space. Measurements were collected at three periods, 15-17 April 1983, 3-4 May and 6-7 May 1983. Different times of the growth season were thought to establish different dependencies between the variables. DYNAMALS was run with dimensionality 1 for the state vector, which will be called LIGHTB (light at the bottom of the ice). Results for the seperate analyses for the three different sampling periods are given in Table III. 9.

From the analyses it appears that from April to May the relation of chlorophyll to the latent variable LIGHTB reverses. Ice thickness gains in importance and so does salinity; the importance of snow depth decreases however.

Table III.9: correlations of input and output variables with the state variable

	15-17 April 1983	3-4 May 1983	6-7May 1983
state variable	LIGHTB	LIGHTB	LIGHTB
SNOW DEPTH	-.376	-.107	-.228
ICE THICK	-.330	-.346	-.441
SALINITY	-.038	-.288	-.200
LIGHTUNDER	.696	.833	.745
CHLOROPHYLL	.819	-.745	-.671
fit	.668	.692	.589

The correlations for 3-4 and 6-7 May, though differing slightly in absolute magnitude, lead to the same interpretation.

One problem in analyzing data that are related through space is that, unless there is something like a 'current', the direction of the dependency is not defined a priori. In other words: the starting point of the process is unassessed. To check the obtained results, the program was run again, but in the other direction. The results from this analysis are presented in Table III.10.

Table III.10: correlations of input and output variables with the state variable for the backwards analysis

	15-17 April 1983	3-4 May 1983	6-7 May 1983
state variable	LIGHTB	LIGHTB	LIGHTB
SNOW DEPTH	-.358	-.094	-.229
ICE THICKNESS	-.362	-.394	-.410
SALINITY	.027	-.290	-.225
LIGHTUNDER	.697	.838	.723
CHLOROPHYLL	.815	-.753	-.701
fit	.669	.707	.590

Although the correlations from the backwards analysis are not exactly the same as those of the forward analysis, the interpretation of the two solutions is identical.

III.5 Redundancy analysis: American states data

The data analyzed here were collected by Meulman (1986) measuring some properties of the 50 American states. The variables could again be divided into input and output variables. No dependency was assumed between the 50 measurements (the American states) so that the model fitted is a redundancy analysis model. Table III.11 lists the variables that were analyzed.

Table III.11: input and output variables of the American states data

BLACK	% of population of the black race	
HISPA	% of population of Spanish origin	
URBAN	ratio of urban to rural	
INCOME	per capita income in dollars	INPUT
LIFEX	life expectancy in years	
HOMIC	1976 homicide and non-negligent manslaughter rate	
UNEMP	1976 unemployment rate	
HIGHS	% of over 25 yrs population high school graduates	
PUBLIC	% public school enrollment	
PUPIL	public schools pupil to teacher ratio	OUTPUT
ILLIT	illiteracy rate	
FAILU	failure rate on the Selective Service mental ability test	

It was postulated that the education variables in the output could be predicted from the combined input variables. DYNAMALS was run with weight ω equal to 1. (In order to perform reduced rank regression, the weight a should have been set to ∞. For a more detailed discussion of this matter, see De Leeuw and Bijleveld (1987)). The dimensionality of the state was chosen to be 5. All 7 input variables were treated ordinally. After 58 iterations a fit of .947 had been obtained and 93.7% of output variance explained. The proportions of variance of the respective output variables that were explained by the state variables are as follows:

HIGHS	93.5 %
PUBLIC	94.1 %
PUPIL	91.9 %
ILLIT	92.0 %
FAILU	96.8 %

All output variables are predicted nicely, none is really below the mark. FAILU is predicted best, PUPIL least. To interpret the state variables the correlations of input and output variables with the state variables must be used; they are in Table III.12. For interpretation we will use the first two state variables as they correlate highest with the input and output variables.

Table III.12: correlations of input and output variables with the state variables

state variables:	z1	z2	z3	z4	z5
BLACK	.887	-.007	-.094	.055	-.291
HISPA	.058	.215	.223	.597	.663
URBAN	.071	.122	.786	-.070	-.106
INCOM	-.558	-.291	.434	.404	-.260
LIFEX	-.702	-.231	.219	-.269	.523
HOMIC	.721	.330	.150	.150	.100
UNEMP	.405	-.045	.706	-.294	-.170
HIGHS	-.850	.151	.266	.342	-.042
PUBLI	-.260	.837	-.409	.061	.035
PUPIL	.396	.751	.417	-.156	-.010
ILLIT	.905	-.012	.057	.238	.203
FAILU	.940	.071	-.091	.164	-.212

A plot 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 shown in Figure 6.

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. The

first dimension may be interpreted as a *poverty* dimension; the second dimension as an *education* dimension.

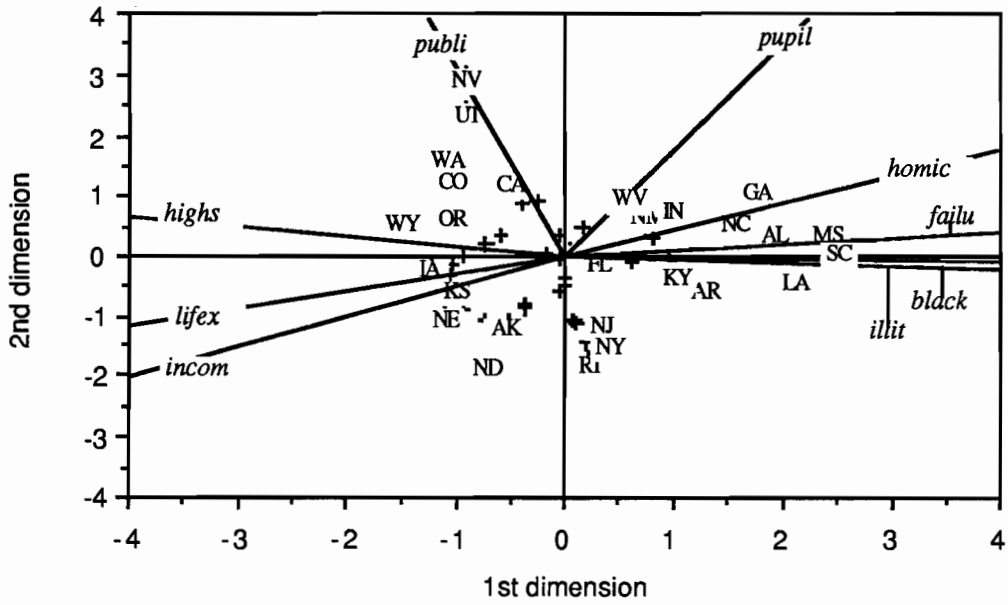


Figure 6. Plot of the states and correlations of the variables with the states

Southern states like Missouri, South Carolina, Louisiana, Alabama, Georgia and North Carolina that are situated in the left part of the picture, are poor states; the 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 and New Jersey. The other states take more moderate positions in the picture.

III.6 Dynamic factor analysis: Menstruation data

The data analyzed here contain one woman's self report on 6 point rating scales of 6 physiological variables related to abdominal pain and water retention (see Bijleveld and Van den Boogaard, 1987); they are presented in Table III.13.

Table III.13: output variables of the menstruation data

abdominal	abdominal pain	
cramps	id	
weight gain	id	
skin disorders	id	OUTPUT
breasts	painful breasts	
swelling	id	

Women's pain and physiological symptoms are presumed to vary cyclically and predictably through the menstrual cycle; a few days before the onset of menstruation water retention increases and subsequently decreases with the onset of menstruation. During especially the first days of menstruation, abdominal pain is reported. To find the hypothetical constructs abdominal pain and water retention, DYNAMALS was run with a two dimensional state variable. As the 6 variables consist of scores on a rating scale, they were treated as ordinal variables. DYNAMALS converged in 33 iterations to a fit of .69046. The results are in Table III.14.

Table III.14: correlations between state variables and output variables

state variables:	z1	z2
abdominal	.783	.478
cramps	.799	.433
weight gain	.454	-.662
skin disorders	.257	-.820
breasts	.348	-.437
swelling	.855	-.067

On the basis of the correlations in Table III.14 the first state variable may be called "abdominal pain" and the second "water retention". The variable swelling assumes an unusual position, as it had been expected to load on "water retention" and not on "abdominal pain". Breasts is less important than had been supposed. To see what happened to this subject during the menstrual cycle, scores on the state variables were plotted against the time axis. The result is shown in Figure 7.

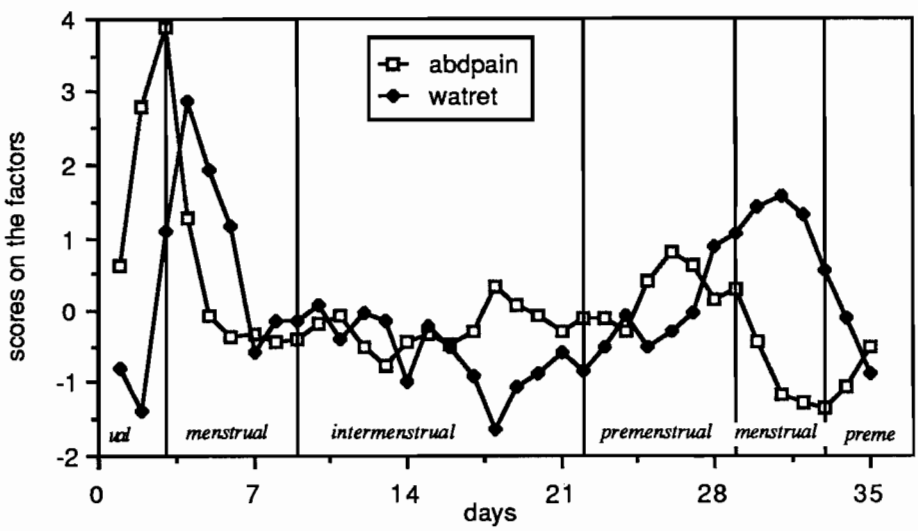


Figure 7. Plot of days against dynamic factor scores.

From the picture it is apparent that abdominal pain and water retention do not always vary as had been expected. For example, water retention does not peak before, but after the onset of menstruation for this particular woman. Abdominal pain peaks just before and at the onset of menstruation. During her first menstruatuon in the experimental period, this subject reported more abdominal pain than during the second menstruation.

References

Bijleveld, C.C.J.H. & Van den Boogaard, T.G.H.M.(1987): Daily menstrual symptoms measures in women and men using an extended version of Moos's instrument. (submitted for publication)

De Leeuw, J. & Bijleveld, C.C.J.H. (1987): **Fitting reduced rank regression models by alternating least squares**. Leiden: Dept. of Datatheory. RR-87-05.

De Leeuw, J. & Bijleveld, C.C.J.H. (1988): Fitting linear dynamical systems by alternating least squares. (in preparation).

Gosselin, M., Legendre, L., Therriault, J-C. et al. (1986): Physical control of the horizontal patchiness of sea-ice microalgae. **Marine Ecology - Progress Series**. Vol. 29: 289-298.

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

Ouweneel, A. & Bijleveld, C.C.J.H.(1987): The agrarian cycle of Bourbon Central-Mexico: a critique of the recaudación del diezmo líqido en pesos. (submitted for publication)