October 14-18, 1985

Symposium on
Nonlinear Multivariate Data Analysis

Jan de Leeuw
Department of Data Theory
University of Leiden
Leiden, The Netherlands

Presented at
The University of North Carolina at Chapel Hill

October 14-18, 1985

Symposium on
Nonlinear Multivariate Data Analysis

Jan de Leeuw
Department of Data Theory
University of Leiden
Leiden, The Netherlands

Presented at
The University of North Carolina at Chapel Hill

(c) Robust and computer-intensive statistics

- replace models by super models
- replace assumptions by computations
- create your own stochastic environment

⇝ Robust statistics proper.

⇝ Jackknife, Bootstrap, Subsampling.

⇝ Permutation and Randomization tests.

① Gifi's principles

1 There is no clear-cut distinction between exploratory and confirmatory

deductivism	⟷	inductivism
rationalism	⟷	empiricism
optimism	⟷	pessimism
inference	⟷	data analysis
model	⟷	technique
confirmation	⟷	exploration
Popper	⟷	Feyerabend

2 In most of the situations we are familiar with the standard statistical prejudice (Kalmens) does not make sense.

3 We must explore the many-many between models and techniques.

(5)

© tools for data analysis

1 Gauging
2 Analysis of stability

ad 1: A data set with known properties is a gauge. We want to find out how a particular technique represents these known properties.

a Probabilistic gauges [multinormal]
b Statistical gauges [samples from multinormal]
c Monte Carlo gauges [idem, without formulas]
d Algebraic gauges [simplex, Guttman Scale, ...]
e Empirical gauges [data with well-known structure].

ad 2 A small and/or unimportant change in data, model, or technique should lead to a small and unimportant change in the results. This is stability.

a Replication stability
b Statistical stability [replication without replication]
c Stability under data selection [resampling]
d Stability under model selection
e Numerical stability
f Analytical stability [differentiability]
g Algebraic stability [... with algebraic tools]
h Stability under selection of technique

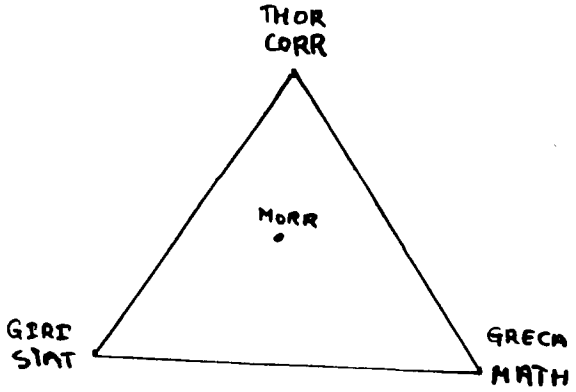
Both tools are technique-oriented But otherwise all of classical statistics fits in painlessly. Gauging is used to study the correspondence models $\rightleftarrows$ techniques

II) Multivariate analysis

- inventory
- problems
- prospects
- definition

(a) Content analysis of 20 MVA Books

[graphs & tables]



b) Current approaches

1 The multivariate normal [Muirhead, Eaton]

- usually a good description (?)
- the central limit effect (?)
- simple formulas (?)

- regressions are linear
- independence ≡ orthogonality
- equiprobability ≡ Euclidean distance
- closed under linear combinations

} (!)

- only one problem : usually a poor description [both the normality, the independence, and the identical distributions]

2 The loglinear model

- more realistic

but

- poor as data description (interpretation)
- only usable if $m \lesssim 4$ [empty cells]
- only usable if n very large [at least for the statistical part].
- model selection problem

3 Tabellary analysis

- close to the data

but

- very heavy output
- problems with integration, reporting, table selection, interpretation

4 The packages

PCA
GLM
etc] by SAS, SPSS, ...

convenient

but

limited to numerical variables.

So

We need techniques for data sets which are

- mixed level
- large m
- nonrandom
- analyzed joint-bivariately

Definition of MVA

Content analysis of books shows confusion

Is MDS a form of MVA?

Must we have a random sample in order to use MVA?

We answer both questions with a firm NO.

Multivariate analysis studies the structure of multivariables

$\phi_j : \Omega \to \nabla_j$ is a variable with domain Ω and range ∇_j.

We can have

$\nabla_j = \mathbb{R} = (-\infty, +\infty)$

$\nabla_j = \{0, 1, 2, \dots\} = \mathbb{N}$

$\nabla_j =$ {protestant, catholic, buddhist}

$\nabla_j =$ {agree, don't know, disagree}

etc

We can have

Ω a finite set $\{\omega_1, \dots, \omega_n\}$

Ω a probability space, and ϕ_j measurable [i.e. ϕ_j a random variable]

$(\phi_1, \dots, \phi_m)$ is a multivariable with domain Ω and range $\nabla_1 \times \dots \times \nabla_m$.

Observe that rows and columns enter asymmetrically [conditionality]

TUESDAY OCT 15 13e

Yesterday we have seen that MVA techniques are needed which can be used

- on large data sets (large m)
- on mixed measurement level data
- on nonrandom data

Today we shall start building a system of MVA techniques that

- has the existing linear techniques as special cases
- containg the necessary extensions

The system is build around the notion of homogeneity and around a particular least squares loss function measuring departure from homogeneity

The techniques will be discussed from the geometrical [MDS], algebraic [MVA], and statistical point of view.

Quantification and coding

Suppose that $\Omega = \{\omega_1, \ldots, \omega_m\}$ is finite and suppose $D_j = \{1, \ldots, k\}$ for all $j = 1, \ldots, m$.

Quantification

A **quantification** of the objects (individuals, ...) is a mapping of Ω into $\mathbb{R}^p$. The number p is the **dimensionality** of the quantification

A **quantification** of (the categories of a) variable is a mapping of D_j into $\mathbb{R}^p$. In some contexts we do not use quantifications, but we use **transformation**

Perfect quantification

A quantification of the objects X, with $x_1, \dots, x_n \in \mathbb{R}$, is perfect if, for all i, j, k

$$\phi_j(\omega_i) = \phi_j(\omega_k) \implies x_i = x_k.$$

i.e. if objects which are in the same category of a variable have the same **score**.

Quantifications $Y_1, \dots, Y_m$ of the variables, with Y_j matrices of order $k \times p$, are perfect if

$$\phi_j(\omega_i) = a \wedge \phi_l(\omega_i) = b \implies (Y_j)_a = (Y_l)_b$$

i.e. if the quantifications corresponding with the values of each object are the same.

1	a	p	u
2	b	q	v
3	a	r	v
4	a	p	u
5	b	p	v
6	c	p	v
7	a	p	u
8	a	p	v
9	c	p	v
10	a	p	v

perfect object scores

from column 1

$$\begin{cases} x_1 = x_3 = x_4 = x_7 = x_8 = x_9 \\ x_2 = x_5 \\ x_6 = x_{10} \end{cases}$$

from column 2

$$x_1 = x_4 = x_5 = x_6 = x_7 = x_8 = x_9 = x_{10}$$

etc

perfect category quantifications

from row 1

$$y_a = y_p = y_u$$

from row 2

$$y_b = y_q = y_v$$

etc

Matrix notation

We need the concept of an indicator matrix

For variable j this is defined as the

$n \times k$ matrix G_j with

$$(G_j)_{il} \begin{cases} 1 & \text{if } \phi_j(\omega_i) = l \\ 0 & \text{otherwise} \end{cases}$$

	abc	pqr	uvw
1	100	100	100
2	010	010	010
3	100	001	010
4	100	100	100
5	010	100	010
6	001	010	010
7	100	100	100
8	100	100	010
9	001	100	010
10	100	100	010
	622	811	370

$D_j = G_j' G_j$
is diagonal with univariate marginals

$C_{jl} = G_j' G_l$ contains bivariate marginals (i.e. cross tables).

This is a complete indicator matrix because
$G_j u = u$ for all j.

Table 2.1

a	p	u
b	q	v
a	r	v
a	p	u
b	p	v
c	p	v
a	p	u
a	p	v
c	p	v
a	p	v

Table 2.1.

Example of data matrix H.

Table 2.2

a	p	u	3
a	p	v	2
a	p	w	0
a	q	u	0
a	q	v	0
a	r	u	0
a	r	v	1
a	r	w	0
b	p	u	0
b	p	v	1
b	p	w	0
b	q	u	0
b	q	v	1
b	q	w	0
b	r	u	0
b	r	v	0
b	r	w	0
c	p	u	0
c	p	v	2
c	p	w	0
c	q	u	0
c	q	v	0
c	q	w	0
c	r	u	0
c	r	v	0
c	r	w	0

Table 2.2.

Profile frequency matrix.

Table 2.3

a	p	u	3
a	p	v	2
a	r	v	1
b	p	v	1
b	q	v	1
c	p	v	2

Table 2.3.

Reduced profile frequency matrix.

Table 2.4

	p	q	r
a	3	0	0
b	0	0	0
c	0	0	0

u

	p	q	r
a	2	0	1
b	1	1	0
c	2	0	0

v

	p	q	r
a	0	0	0
b	0	0	0
c	0	0	0

w

Table 2.4. Higher dimensional cross tabulation.

Table 2.5

a b c	p q r	u v w
1 0 0	1 0 0	1 0 0
0 1 0	0 1 0	0 1 0
1 0 0	0 0 1	0 1 0
1 0 0	1 0 0	1 0 0
0 1 0	1 0 0	0 1 0
0 0 1	1 0 0	0 1 0
1 0 0	1 0 0	1 0 0
1 0 0	1 0 0	0 1 0
0 0 1	1 0 0	0 1 0
1 0 0	1 0 0	0 1 0

Table 2.5 Indicator matrix G for data matrix H of table 2.1

Table 2.6

	a b c	p q r	u v w
a	6 0 0	5 0 1	3 3 0
b	0 2 0	1 1 0	0 2 0
c	0 0 2	2 0 0	0 2 0
p	5 1 2	8 0 0	3 5 0
q	0 1 0	0 1 0	0 1 0
r	1 0 0	0 0 1	0 1 0
u	3 0 0	3 0 0	3 0 0
v	3 2 2	5 1 1	0 7 0
w	0 0 0	0 0 0	0 0 0

Table 2.6 Matrix C of bivariate marginals.

Table 2.7

	a b c	p q r	u v w
a	6 0 0	0 0 0	0 0 0
b	0 2 0	0 0 0	0 0 0
c	0 0 2	0 0 0	0 0 0
p	0 0 0	8 0 0	0 0 0
q	0 0 0	0 1 0	0 0 0
r	0 0 0	0 0 1	0 0 0
u	0 0 0	0 0 0	3 0 0
v	0 0 0	0 0 0	0 7 0
w	0 0 0	0 0 0	0 0 0

Table 2.7 Matrix D of univariate marginals.

Table 2.8

	a b c d e f
old	1 0 0 0 0 0
	1 1 0 0 0 0
	0 1 1 1 0 0
	0 0 1 1 1 0
new	0 0 0 1 1 1

Table 2.8A. Incomplete indicator matrix.

	+a-	+b-	+c-	+d-	+e-	+f-
old	1 0	0 1	0 1	0 1	0 1	0 1
	1 0	1 0	0 1	0 1	0 1	0 1
	0 1	1 0	1 0	1 0	0 1	0 1
	0 1	0 1	1 0	1 0	1 0	0 1
new	0 1	0 1	0 1	1 0	1 0	1 0

Table 2.8B. Completed indicator matrix

Perfect quantification in matrix notation

$(Y_1, \ldots, Y_m)$ is perfect iff $\exists X \Rightarrow$

$G_1 Y_1 = \ldots = G_m Y_m = X$

(X) is perfect iff $\exists Y_1, \ldots, Y_m \ni$

$G_1 Y_1 = \ldots = G_m Y_m = X$

In summary $(X; Y_1, \ldots, Y_m)$ is perfect iff

$X = G_1 Y_1 = \ldots = G_m Y_m.$

Theorem

$(Y_1, \ldots, Y_m)$ is perfect iff

$G_1 Y_1 = \ldots = G_m Y_m = \frac{1}{m} \sum_{j=1}^{m} G_j Y_j$ iff

$$\sum_{j=1}^{m} \sum_{e=1}^{m} Y_j' C_{je} Y_e = m \sum_{j=1}^{m} Y_j' D_j Y_j$$

between variables SSQ

total SSQ.

Theorem X is perfect iff

$$X = P_1 X = P_2 X = \dots = P_m X \quad (\text{with } P_j = G_j (G_j'G_j)^{-1} G_j')$$

iff

$$X'X = X' P_* X \quad (\text{with } P_* = \tfrac{1}{m}\textstyle\sum P_j)$$

↑ total SSQ ↑ between object SSQ.

Theorem X is perfect iff the p largest eigenvalues of P_* are equal to +1.

$Y = (Y_1, \dots, Y_m)$ is perfect iff the p largest generalized eigenvalues of $(C, \mathrm{diag}\, C)$ are +1.

Perfectness cannot be attained for real data sets. Thus we look for optimality. This defines homogeneity analysis.

Loss function measuring departure from perfect homogeneity (of given scores X and quantifications Y).

$$\sigma(X;Y) = \frac{1}{m} \sum_{j=1}^{m} \mathrm{tr}\, [X - G_j Y_j]' [X - G_j Y_j]$$

In homogeneity analysis (also known as

- multiple correspondence analysis [BENZÉCRI]
- GUTTMANS PRINCIPAL COMPONENTS OF SCALE ANALYSIS [G 1941, L 1960]
- FACTORIAL ANALYSIS OF QUALITATIVE DATA [BURT, 1950]
- HAYASHI's a^{th} METHOD OF QUANTIFICATION [H, 1956]

we want to find X and Y such that $\sigma(X;Y)$ is minimized.

BUT

we do not want the trivial solution $X=0 \wedge Y=0$.

NOR

do we want the trivial dimension $X=\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}$ $Y=\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}$

Thus we must impose some normalization restriction.

$$N_1 : \begin{cases} u'X=0 \\ X'X=I. \end{cases}$$

$$N_2 : \begin{cases} u'DY=0 \\ Y'DY=I \end{cases}$$

We only need to normalize either X or Y.

Let us see what happens if we normalize X.

$$b(X;A) \triangleq \min \{ b(X;Y) \mid Y \text{ free} \}$$

minimum attained for $Y_j = D_j^{-1} G_j' X$ [or $G_j^+ X$]

$$b(X;A) = \operatorname{tr} X'[I - P_A]X$$

now define

$$b(A;u) \triangleq \min \{ b(X;A) \mid X \text{ norm} \}$$

then

$$b(A;u) = p - \sum_{s=1}^{p} \lambda_s(P_A) \qquad \boxed{\lambda_c(P_A)=1}$$

Interpretation: X are eigenvectors of the average projector P_*, and Y_j is the centroid of the corresponding X.

Now proceed dually

$$\zeta(\alpha,\gamma) \triangleq \min\{G(X;\gamma) \mid X \text{ free}\}$$

attained for $X = \frac{1}{m}\sum G_j \gamma_j$

$$\zeta(\alpha,\gamma) = \frac{1}{m}\,\mathrm{tr}\,\gamma'[D - \tfrac{1}{m}C]\gamma$$

then [theorem, not definition]

$$\zeta(\alpha,\alpha) = \min\{\zeta(\alpha,\gamma) \mid \gamma \text{ norm}\} =$$

$$= p - \sum_{s=1}^{p} \lambda_s\left[\tfrac{1}{m} D^{-1/2} C D^{-1/2}\right]$$

Proof. Let $H' = m^{-1/2} D^{-1/2} G'$. Then

$P_\alpha = HH'$ and $m^{-1} D^{-1/2} C D^{-1/2} = H'H$.

The nonzero eigenvalues of HH' and $H'H$ are the same. □

$$\begin{array}{ccc} \zeta(X,\gamma) & \xrightarrow{X \text{ free}} & \zeta(\alpha,\gamma) \\ \gamma \text{ free} \downarrow & & \downarrow \gamma \text{ norm} \\ G(X,\gamma) & \xrightarrow{X \text{ norm}} & G(\alpha,\alpha) \end{array}$$

Thus homogeneity amounts, computationally, to finding the p largest eigenvalues of either

- the Burt table C [corrected for marginals]
- the average projector P_α

Pictures of loss

[Example from Beyond]

[Also illustrating the ALS-algorithm]

	Price	Gas	Weight
Chevette	5.6	6.9	9.7
Dodge Colt	5.7	5.1	8.8
Plymouth Horizon	6.3	5.5	9.9
Fort Mustang	7.6	6.7	12.0
Pontiac Phoenix	8.6	6.9	12.1
Dodge Diplomat	9.4	10.2	15.5
Chevrolet Impala	10.1	7.5	16.9
Buick Regal	10.5	7.8	15.0
AMC Eagle	10.7	11.7	15.7
Oldsmobile 98	13.3	8.7	18.3

table 1: Car data.

Chevette	1	1	1
Dodge Colt	1	1	1
Plymouth Horizon	1	1	1
Fort Mustang	2	1	2
Pontiac Phoenix	2	1	2
Dodge Diplomat	2	3	2
Chevrolet Impala	3	2	3
Buick Regal	3	2	2
AMC Eagle	3	3	2
Oldsmobile 98	4	2	3

table 2: Car data, discrete.

Chevette	1	4	2
Dodge Colt	2	1	1
Plymouth Horizon	3	2	3
Fort Mustang	4	3	4
Pontiac Phoenix	5	5	5
Dodge Diplomat	6	9	7
Chevrolet Impala	7	6	9
Buick Regal	8	7	6
AMC Eagle	9	10	8
Oldsmobile 98	10	8	10

table 3: Car data, ranked.

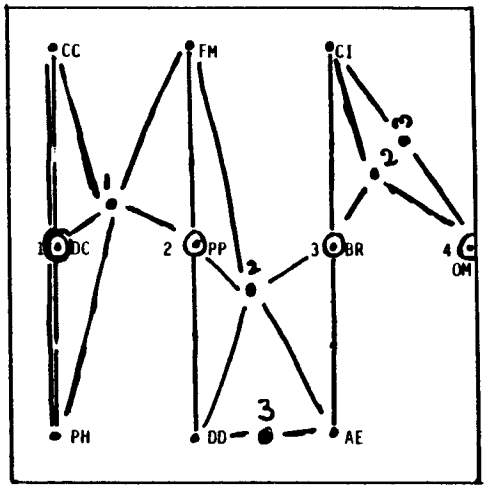


figure 1: loss variable 1, arbitrary solution

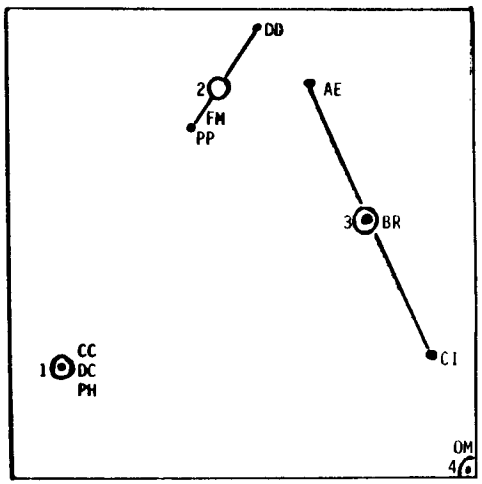
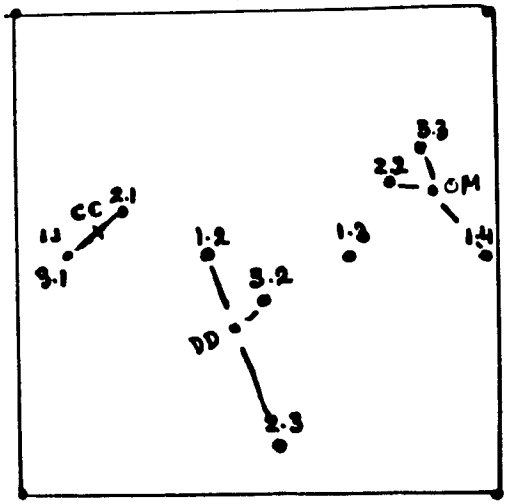


figure 2: loss variable 1, optimal solution



- it shrinks
- it is simple

Alg A X_0 such that $u'X_0 = 0$ $X_0'X_0 = I$

→ step 1: $Y_j \leftarrow B_j^{-1} G_j' X$

step 2: $X \leftarrow \frac{1}{m} \sum G_j Y_j$

step 3: $X \leftarrow$ GRAM (X)

Alg B: Y_0 such that $u'DY_0 = 0$, $Y_0'DY_0 = mI$.

→ step 1: $X \leftarrow \frac{1}{m} \sum_{j=1}^{n} G_j Y_j$

step 2: $Y_j \leftarrow D_j^{-1} G_j' X$

step 3: $Y \leftarrow$ WGRAM (Y)

The algorithms both work by applying
- reciprocal averaging [Horst, 1935; Richardson, Kuder]
- le principe barycentrique
  [the centroid principle]

Unfortunately one cannot have **both** Y as centroids of X and X as centroids of Y.
One has to choose.

HOMALS uses Alg A. Thus X is normalized, and category quantifications are centroids of scores of the objects in the categories.

Variables	Categories & codes
Thread	Yes = Y No = N.
Head	Flat = F Cup = U Cone = O Round = R Cylinder = Y.
Head indentation	None = N Star = T Slit = L.
Bottom	Sharp = S Flat = F.
Length	(in half inches).
Brass	Yes = Y No = N.

Object	1	2	3	4	5	6
TACK	N	F	N	S	1	N
NAIL1	N	F	N	S	4	N
NAIL2	N	F	N	S	2	N
NAIL3	N	F	N	S	2	N
NAIL4	N	F	N	S	2	N
NAIL5	N	F	N	S	2	N
NAIL6	N	U	N	S	5	N
NAIL7	N	U	N	S	3	N
NAIL8	N	U	N	S	3	N
SCREW1	Y	O	T	S	5	N
SCREW2	Y	R	L	S	4	N
SCREW3	Y	Y	L	S	4	N
SCREW4	Y	R	L	S	2	N
SCREW5	Y	Y	L	S	2	N
BOLT1	Y	R	L	F	4	N
BOLT2	Y	O	L	F	1	N
BOLT3	Y	Y	L	F	1	N
BOLT4	Y	Y	L	F	1	N
BOLT5	Y	Y	L	F	1	N
BOLT6	Y	Y	L	F	1	N
TACK1	N	F	N	S	1	Y
TACK2	N	F	N	S	1	Y
NAILB	N	F	N	S	1	Y
SCREWB	Y	O	L	S	1	Y

Table 3.13 Hartigan's hardware

object	dim1	dim2
TACK	0.75	0.46
NAIL1	0.68	0.47
NAIL2	0.96	0.52
NAIL3	0.96	0.52
NAIL4	0.96	0.52
NAIL5	0.96	0.52
NAIL6	1.00	-1.69
NAIL7	1.25	-0.74
NAIL8	1.25	-0.74
SCREW1	-0.38	-3.96
SCREW2	-0.85	0.23
SCREW3	-0.91	0.26
SCREW4	-0.57	0.28
SCREW5	-0.63	0.31
BOLT1	-1.31	0.38
BOLT2	-1.18	-0.51
BOLT3	-1.30	0.40
BOLT4	-1.30	0.40
BOLT5	-1.30	0.40
BOLT6	-1.30	0.40
TACK1	0.93	0.67
TACK2	0.93	0.67
NAILB	0.93	0.67
SCREWB	-0.54	-0.44

Table 3.14A Hartigan's hardware object scores

category	dim1	dim2
1Y	-0.96	-0.15
1N	0.96	0.15
2F	0.90	0.56
2U	1.16	-1.06
2O	-0.70	-1.64
2R	-0.91	0.30
2Y	-1.12	0.36
3N	0.96	0.15
3T	-0.38	-3.96
3L	-1.02	0.19
4S	0.43	-0.08
4F	-1.28	0.25
5/1	-0.34	0.31
5/2	0.44	0.44
5/3	1.25	-0.74
5/4	-0.60	0.24
5/5	0.31	-2.82
6Y	0.56	0.39
6N	-0.11	-0.08

Table 3.14B Hartigan's hardware category quantifications

Variables	dim1	dim2
Thread	0.930	0.024
Head	0.951	0.635
Head indentation	0.945	0.681
Bottom	0.546	0.020
Length	0.292	0.819
Brass	0.064	0.038
Eigenvalues	0.621	0.368

Table 3.14C Hartigan's hardware discrimination measures

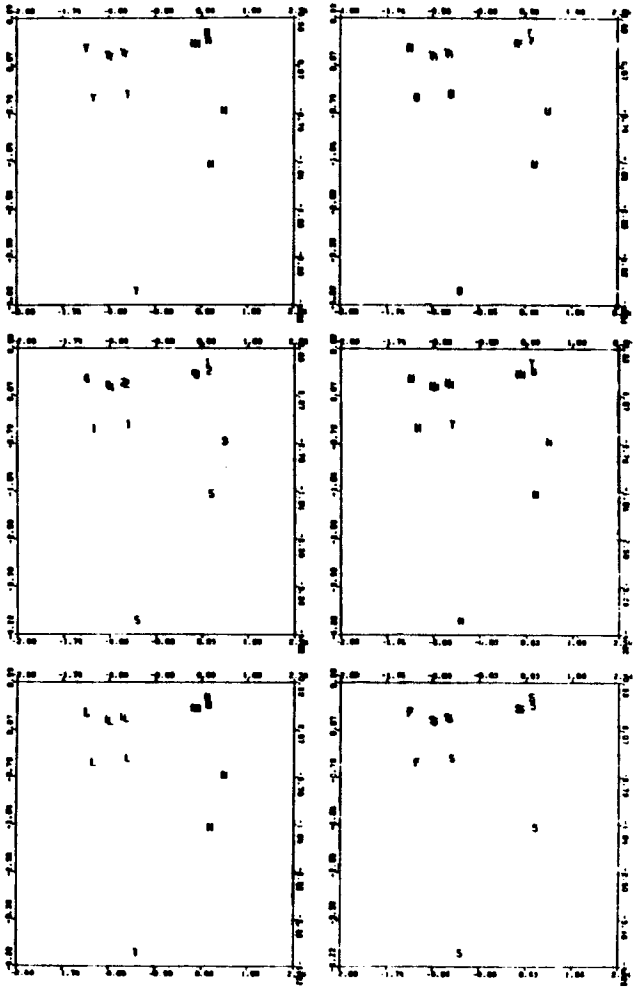


Figure 3.13. Hartigan's hardware
object scores labeled by variables

imensions. This gives the dis-
iables we get the total discri-

(40)

dness-of-fit of a HOMALS-repre-
tion:

(41)

ntation of the data. Classifi-
nfiguration of object points
nfiguration is an undifferen-
consecutive objects in it.
clusters and we need to de-
entify clusters or regions is
entation according to the cat-
with high discrimination
ns of the configuration. An-
by using an outside criterion.
partition can be used to label
this we can compare the old
is used in our example which

iles, meaning the measurements
cription we refer to Vescia

s a fit of .508, which is not
Nevertheless the solution is
up numbers given in the data
ir clusters: (5,6,7,8,9),(4),
1,2) are nearer to the finback
ension divides the river dol-
-dimensional representation
,6), however not far apart.
it of .442 which is lower than
the fit is the mean of the
ons. However the sum of the
mensions increases. The first
ution are similar to the two-
imension adds new information
ked and sperm whales (6,4) from
HOMALS-representation is given
. The end points of the rods on
ution of Fig.4.1. The HOMALS
n of Slijper |10|, which is
baleen, grey and finback whales
The dolphins and porpoises (5,6)
e white whales (7). These rela-
solution. The sperm whales (4)
ijper's classification.

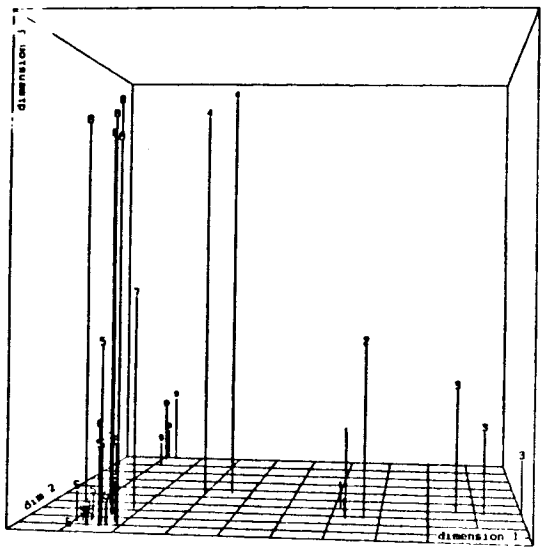


Figure 5

Three-dimensional HOMALS-representation. Objects are labeled by their group numbers.

Whales - example

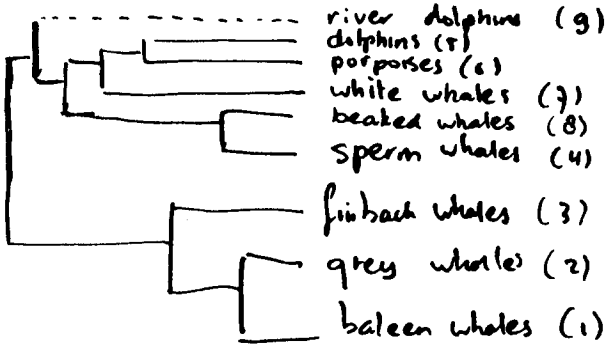
E. van der Burg

HOMALS classification of whales, porpoises, and dolphins

In J.F. Marcotorchino, J.M. Proth, and J. Janssen (eds)

Data analysis in Real Life Environment

Elsevier Science Publishers, 1988



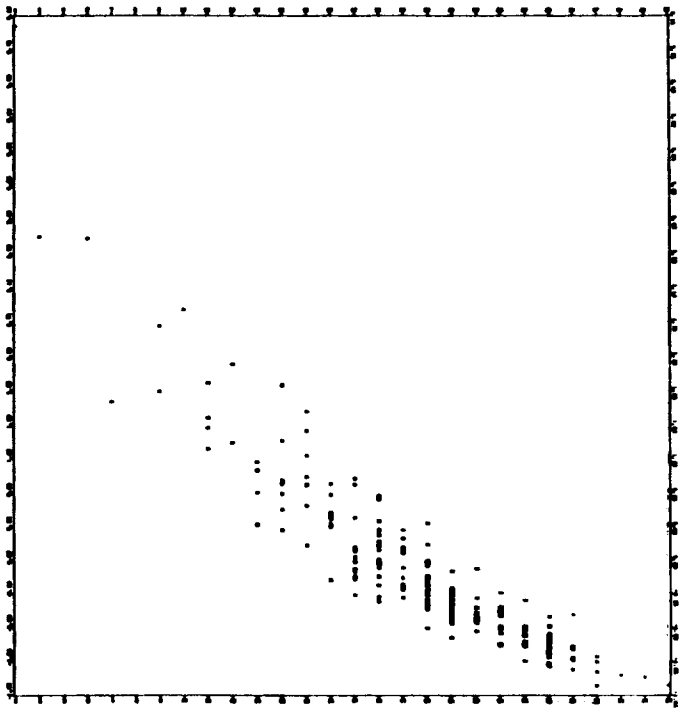


Figure 13.2.2 Individual HOMALS scores (4 categories) plotted against number of correct answers

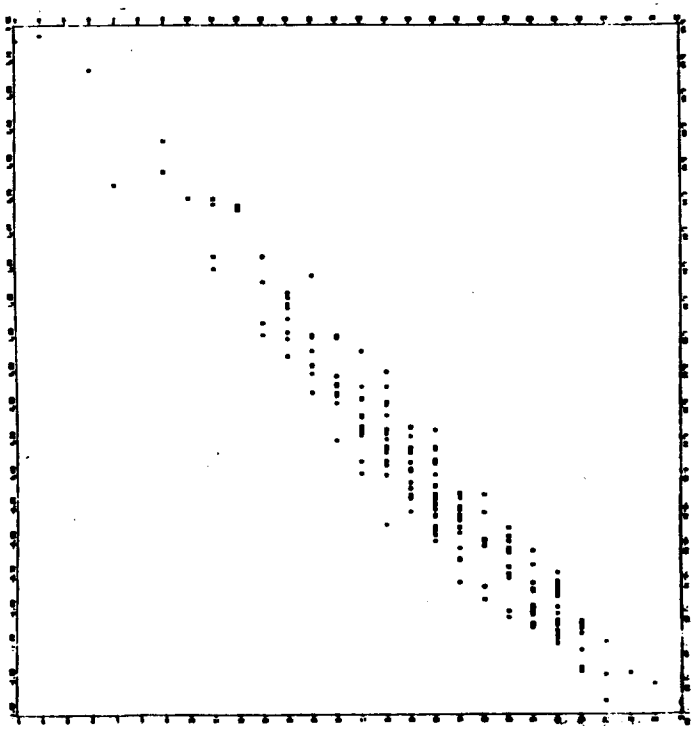


Figure 13.2.3 Individual HOMALS scores (2 categories) plotted against number of correct answers

Our discussion of homogeneity analysis so far has been largely geometrical. What is the purpose of the technique? If $p > 1$ it is drawing a picture in which objects with similar answer patterns are close together, and categories containing the same objects are close together etc. If $p = 1$ the "picture" is really a "scale". How do you choose p? That depends on the type of picture you want. Moreover the dimensions are "nested". Are we "estimating" something? Not necessarily. Neither are we testing anything.

But

There are connections with the chi-square test for independence.

$$\frac{1}{2} n \sum_s \left[m\lambda_s - 1 \right]^2 = \sum_{j<e} \sum \chi^2_{je}.$$

(variance of eigenvalues)

Thus, if all variables are independent; (and if we have a random sample)

$$\frac{1}{2} n \sum_s \left[m\hat{\lambda}_s - 1 \right]^2 \xrightarrow{\mathcal{L}} \chi^2_{\sum_{j<e} (k_j - 1)(k_e - 1)}$$

Moreover, if variable j is independent of the others, then

$$n \left(\frac{\hat{\lambda}_s}{1 - \hat{\lambda}_s} \right)^2 \hat{y}'_{js} D_j \hat{y}_{js} \xrightarrow{\mathcal{L}} \chi^2_{k_j - 1}.$$

($\hat{y}_{js}$: discrimination measure)

A less geometrical approach [$p=1$, already used in MC-example].

We have seen that homogeneity analysis (if $p=1$) can be formulated as

$$y'Cy \quad \max! \qquad y'Dy = m \cdot \quad u'Dy = 0$$

Here $y' = (y'_1 | \ldots | y'_m)$. Now decompose $y_j = \alpha_j z_j$ with $z'_j D_j z_j = 1$ and $u'_j D_j z_j = 0$.

Then

$$y'Dy = \sum_j y'_j D_j y_j = \sum_j \alpha_j^2 z'_j D_j z_j = \sum \alpha_j^2$$

and

$$y'Cy = \sum_j \sum_\ell \alpha_j \alpha_\ell z'_j C_{j\ell} z_\ell \overset{\Delta}{=} \sum_j \sum_e \alpha_j \alpha_e r_{je}$$

with $R = \{r_{je}\}$ the correlation matrix induced by the scores y_j (or z_j).

The homogeneity analysis problem can now be written as

$$\max_{\alpha} \max_{z} \{ \alpha' R \alpha \mid \alpha'\alpha = 1 \wedge z_0' D_j z_j = 1 \}$$

or

$$\max_{z} \{ \lambda_{max}(R) \mid z_0' D_j z_0 = 1 \}.$$

In words: we want to find quantifications of the categories that maximize the largest eigenvalue of the induced correlation matrix.

Data list is [CRIME AND FEAR]

CORRELATION MATRIX BEFORE TRANSFORMATION

	1	2	3	4	5	6	7	8	9	10
1 =	1.000	0.101	0.227	0.410	0.047	0.316	-0.024	-0.029	-0.035	-0.019
2 =	0.101	1.000	0.488	0.166	0.145	0.170	0.042	0.006	-0.020	-0.128
3 =	0.227	0.488	1.000	0.265	0.294	0.217	0.091	-0.023	-0.080	-0.156
4 =	0.410	0.166	0.265	1.000	0.103	0.438	-0.007	-0.065	0.019	-0.142
5 =	0.047	0.145	0.294	0.103	1.000	0.061	0.115	-0.055	-0.051	-0.172
6 =	0.316	0.170	0.217	0.438	0.061	1.000	-0.021	-0.037	0.013	-0.084
7 =	-0.024	0.042	0.091	-0.007	0.115	-0.021	1.000	-0.072	-0.048	-0.084
8 =	-0.029	0.006	-0.023	-0.065	-0.055	-0.037	-0.072	1.000	-0.016	-0.164
9 =	-0.035	-0.020	-0.080	0.019	-0.051	0.013	-0.048	-0.016	1.000	0.044
10 =	-0.019	-0.128	-0.156	-0.142	-0.172	-0.084	-0.084	-0.164	0.044	1.000

EIGENVALUES

2.274	1.365	1.139	1.003	0.974	0.853	0.747	0.663	0.520	0.464

COMPONENT LOADINGS

	1	2
1 =	0.565	-0.450
2 =	0.570	0.315
3 =	0.710	0.311
4 =	0.684	-0.391
5 =	0.394	0.469
6 =	0.612	-0.409
7 =	0.108	0.415
8 =	-0.055	0.174
9 =	-0.083	-0.253
10 =	-0.325	-0.400

DISCRIMINATION MEASURES

VARIABLES

	1	2	3	4	5	6	7	8	9	10
1 =	0.360	0.345	0.505	0.491	0.141	0.407	0.021	0.038	0.038	0.019

CATEGORY QUANTIFICATIONS

CATEGORIES

VARIABLES	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1 =	-1.305	-0.614	-0.203	0.285	0.942									
2 =	-0.627	-0.001	0.841	0.766	1.408									
3 =	-0.522	0.329	0.998	1.225	1.805									
4 =	-1.894	-0.916	-0.704	0.038	0.869									
5 =	-0.416	0.171	0.063	0.572	0.822									
6 =	-1.997	-0.774	-0.575	0.052	0.726									
7 =	-0.050	-0.501	0.128	0.171	0.019									
8 =	0.161	-0.161	-0.054	-0.098	-0.470	0.617	0.284	-0.040	-0.031	-0.147				
9 =	0.888	0.178	-0.726	-0.041	0.185	-0.387	0.125	0.510	0.324	-0.040	0.079	0.032	0.234	-0.044
10 =	0.324	0.307	0.090	-0.075	-0.281	-0.188								

CORRELATION MATRIX AFTER TRANSFORMATION

	1	2	3	4	5	6	7	8	9	10
1 =	1.000	0.126	0.238	0.407	0.054	0.315	0.126	0.054	0.068	0.020
2 =	0.126	1.000	0.498	0.187	0.158	0.180	0.084	0.137	0.065	0.135
3 =	0.238	0.498	1.000	0.279	0.311	0.222	0.080	0.207	0.179	0.163
4 =	0.407	0.187	0.279	1.000	0.103	0.448	0.105	0.098	0.146	0.161
5 =	0.054	0.158	0.311	0.103	1.000	0.083	0.035	0.139	0.120	0.161
6 =	0.315	0.180	0.222	0.448	0.083	1.000	0.098	0.095	0.117	0.098
7 =	0.126	0.084	0.080	0.105	0.035	0.098	1.000	0.134	0.037	0.073
8 =	0.054	0.137	0.207	0.098	0.139	0.095	0.134	1.000	0.103	0.123
9 =	0.068	0.065	0.179	0.146	0.120	0.117	0.037	0.103	1.000	0.131
10 =	0.020	0.135	0.163	0.161	0.161	0.098	0.073	0.123	0.131	1.000

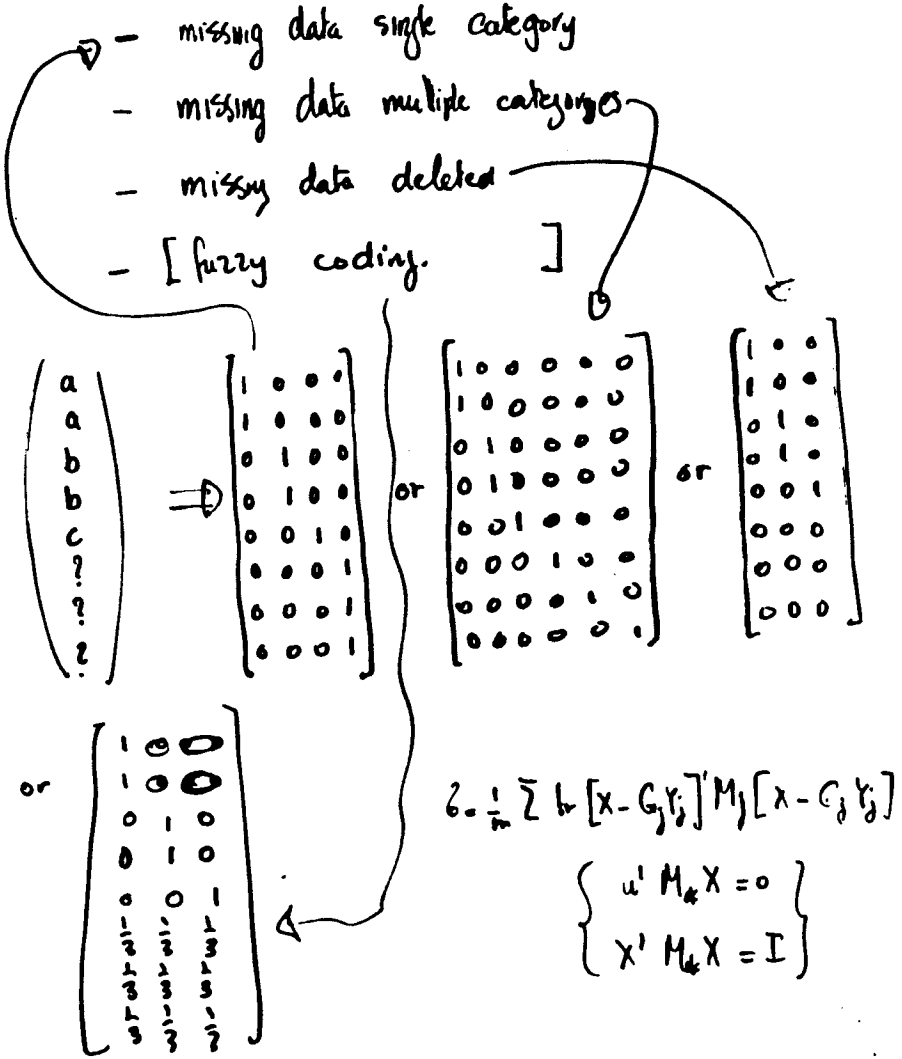
EIGENVALUES

2.510	1.294	1.030	0.999	0.875	0.827	0.820	0.676	0.519	0.450

COMPONENT LOADINGS

	1	2
1 =	0.535	0.526
2 =	0.567	-0.292
3 =	0.707	-0.295
4 =	0.662	0.457
5 =	0.408	-0.471
6 =	0.592	0.443
7 =	0.265	0.060
8 =	0.362	-0.334
9 =	0.344	-0.138
10 =	0.358	-0.319

Missing data [refer to Meulman, in list].



It is clear that these choices can have important consequences.

- single category suggests more homogeneity than there usually is.
- multiple will lead to dimensions determined by missing values.
- deleted is standard [in HOMALS and PRINCALS]

The options can also be used in a much more interesting and creative way.

Example Munsingen - Rain
[Japanese]
[Skiing Resorts]

The example also illustrates the reordering [parallellogram analysis] used of homogeneity analysis. Also multinormal gauge.

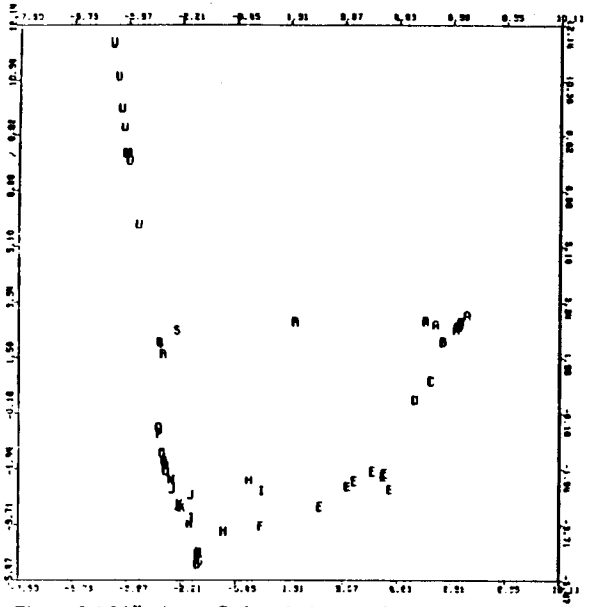


Figure 9.1 Münsingen-Rain missing passive

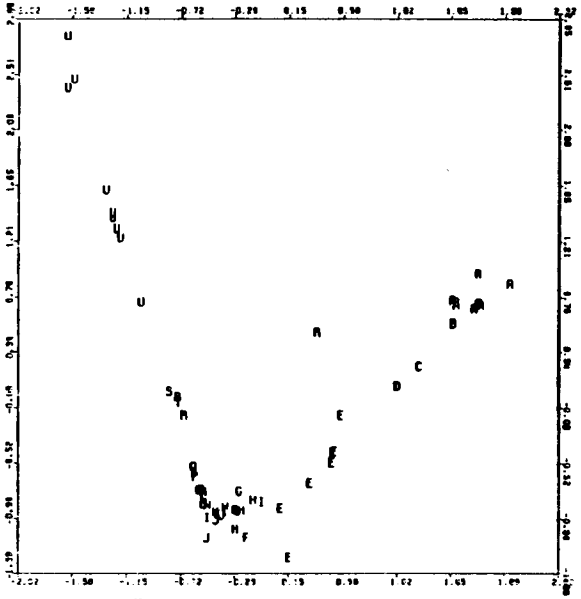


Figure 9.2 Münsingen-Rain missing multiple

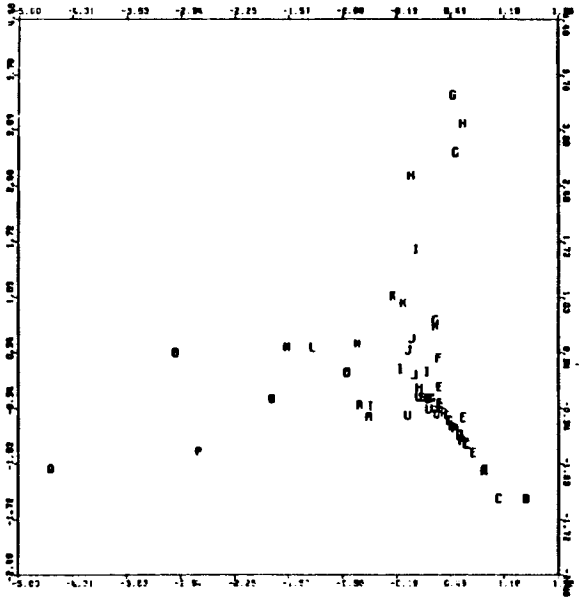


Figure 9.3 Münsingen-Rain missing single

Table 9.1 Permuted data matrix Münsingen-Rain (missing multiple)

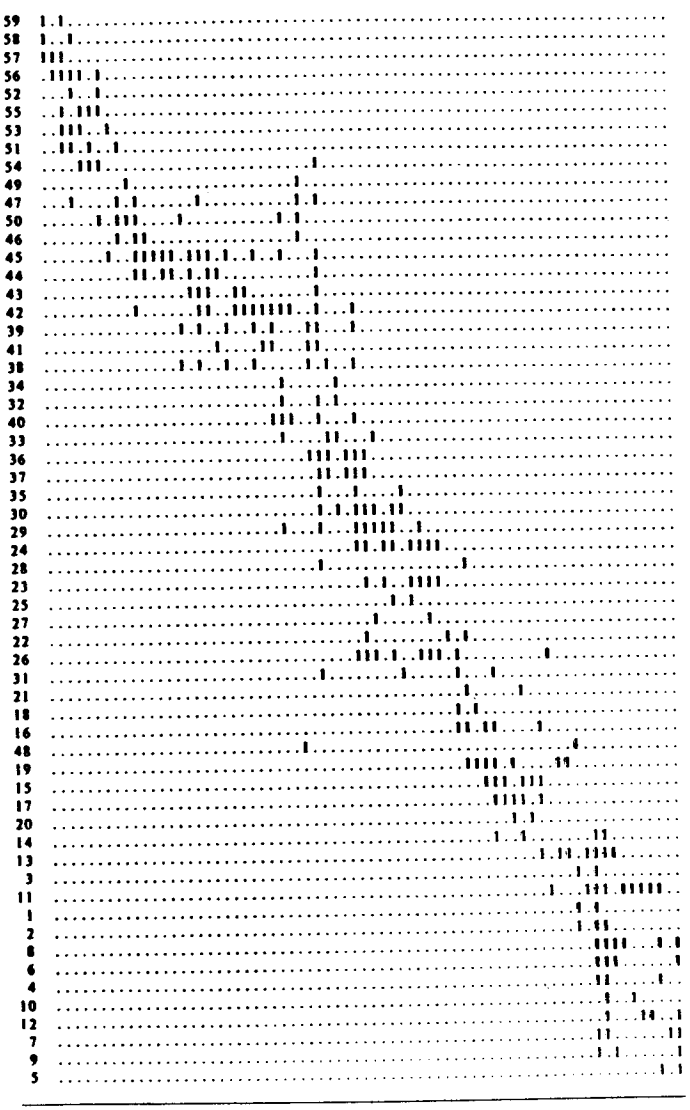


Table 9.2 Permuted data matrix Münsingen-Rain (missing passive)

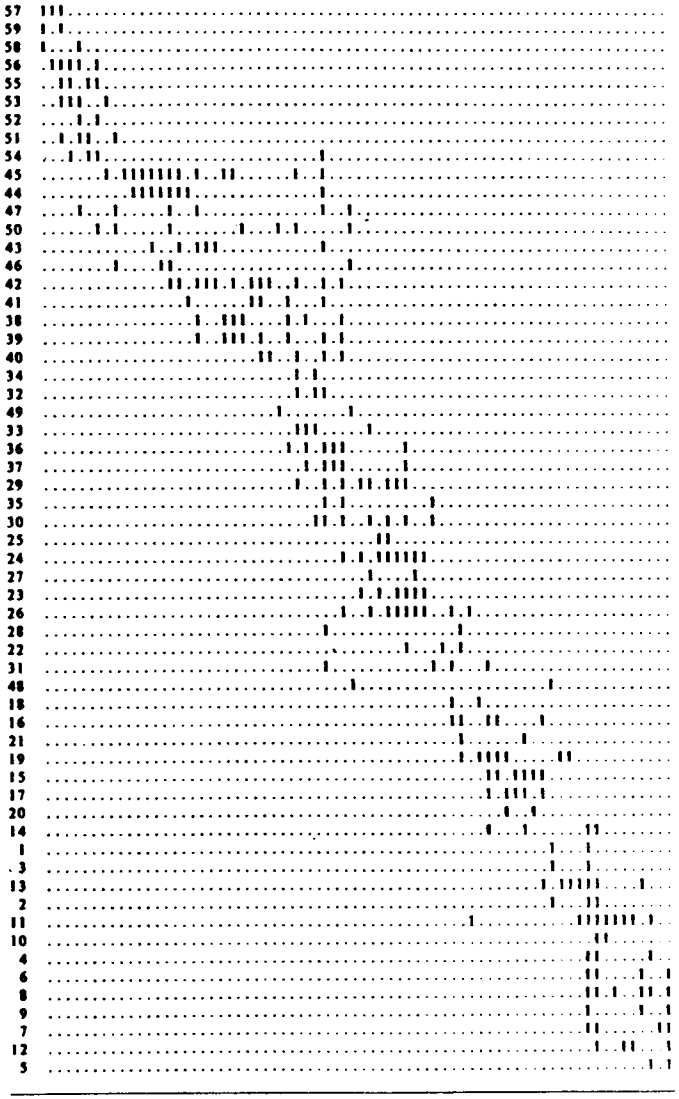
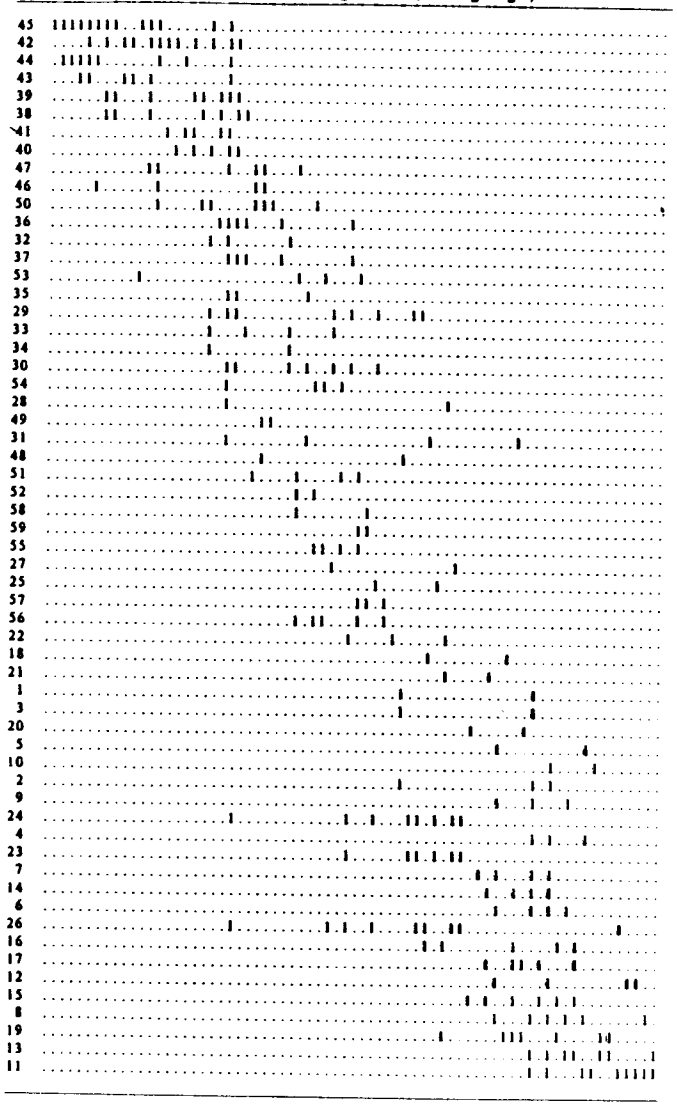


Table 9.3 Permuted data matrix Münsingen-Rain (missing single)



Suppose, however, that we did not know the right rankorder. In that case, perhaps we would have been quite content with the seriation, seeing a structure that approximates a parallelogram and explaining the dots between the one's by low eigenvalues. Similar pictures are in fact obtained when we permute the rows and columns of the gauges II and III. This illustrates how important two-dimensional representations are, since we saw immediately by not finding a horse-shoe that something unexpected is going on. We also give the results for the artifacts.

Table 9.4 Recovery of the rankorder of the artifacts compared to Hodson's rankorder

H	70	69	68	67	66	65	64	63	62	61	60	59	58	57	56	55	54	53	52	51
S	59	56	58	55	51	57	44	46	52	53	60	45	54	49	48	50	47	43	37	42
P	70	69	66	68	63	67	65	60	61	59	56	58	55	57	54	51	50	45	52	53
M	70	69	66	63	68	67	65	60	61	64	54	57	59	56	58	43	55	45	51	50
H	50	49	48	47	46	45	44	43	42	41	40	39	38	37	36	35	34	33	32	31
S	36	34	40	61	62	64	41	38	63	39	65	68	35	67	27	66	70	33	69	26
P	46	44	43	49	47	48	64	42	37	40	38	36	41	34	62	32	35	28	33	31
M	46	52	53	44	49	47	37	48	62	42	36	40	38	41	34	27	35	32	33	39
H	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11
S	1	32	31	23	28	22	30	29	21	7	19	4	25	17	24	2	20	3	18	6
P	27	30	29	39	26	23	22	11	25	17	18	24	20	19	21	14	1	15	16	10
M	28	31	30	29	26	23	22	25	18	17	20	24	19	21	14	11	15	16	1	10
H	10	9	8	7	6	5	4	3	2	1										
S	14	5	9	15	16	11	12	13	8	10										
P	2	3	9	8	12	13	6	5	7	4										
M	2	3	6	8	9	12	13	5	7	4										

9.2 Overall conclusions for gauge IV

This gauge was especially meant to draw attention to the problem of deciding if we could 'missing' let define homogeneous groups. It is an extreme example, but it serves our purpose well: it is worth the trouble of paying substantial attention to newly created categories. In this gauge the missing single strategy is strongly not recommended. Although the results for both passive and multiple are satisfying, we do recommend the passive strategy. From a theoretical point of view this strategy appeals most; from a practical point of view it is computationally far more efficient than the missing multiple strategy with its enormous amount of additional categories (the number of dots in the table).

Yesterday we have discussed **HOMOGENEITY ANALYSIS** (or Multiple Correspondence Analysis). We first introduced the (logical, qualitative) concept of perfect quantification (perfect homogeneity). This was translated into matrix algebra, using the concept of an **INDICATOR MATRIX**, and made quantitative using ANOVA-like concepts. This allowed us to define **OPTIMAL** homogeneity [illegible], and the technique associated with this. Computations could be formulated as eigen-problems, but we used a **LEAST SQUARES LOSS FUNCTION** in order to derive an **ALTERNATING LEAST SQUARES** algorithm, which worked by **RECIPROCAL AVERAGING**. Both geometrical and algebraic aspects were discussed.

ve choose ρ
lo gauge by
e too small.
der certain
has to be.
ie quadratic
one.

d variables
bution with
l with all
ing we also
categorical
s. This can
1980)); we
s minimize
tions or, in
d as well as
and for our
l. For our

alues after

transformation which will be presented in chapter 7 and 8.

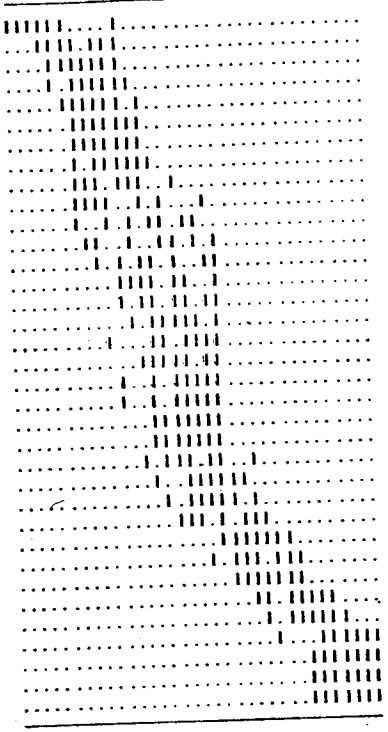
It might be worthwhile to wonder what type of gauge these samples represent in terms of the qualifications we used in describing our first gauge. Imagine that we would have sampled from variables with an overall intercorrelation of 1. Then we would have obtained a data matrix with all columns equal. In that case there would be no work at all for homogeneity analysis, since its ultimate aim — obtaining induced scores which are equal over all variables — is already attained in the data. Every solution will give eigenvalues of 1; the tracelines for the items will all coincide. Sampling from distributions with lower correlations will give data where the pattern described above is disturbed to a certain degree (the smaller the correlation the larger the disturbance). As an illustration we sampled from variables with an overall correlation of .95 (table 5.5), performed a homogeneity analysis and inspected the permuted indicator supermatrix (table 5.6).

Table 5.5 Multinormal
ρ = .95

1	1	1	1	1	2	1
1	2	2	1	1	2	2
1	2	2	1	2	2	2
2	2	2	2	2	2	2
1	2	2	2	2	2	2
2	2	2	2	2	2	2
2	2	2	2	2	2	2
2	2	2	2	3	2	2
2	2	2	2	2	2	3
3	2	2	2	2	3	2
2	3	2	3	3	3	2
2	2	3	3	2	3	3
3	2	3	2	3	2	3
2	3	3	2	3	2	3
3	3	3	2	3	2	3
2	3	3	3	3	3	3
3	3	3	3	3	3	2
3	3	3	2	3	3	3
3	3	3	3	3	2	3
3	3	3	3	3	2	3
3	3	3	3	3	3	3
3	3	3	3	3	3	3
3	3	3	2	4	3	3
3	3	3	3	3	4	4
3	3	3	3	4	3	4
4	3	3	3	4	4	3
4	4	3	4	4	4	4
3	4	4	4	4	4	4
4	4	4	4	4	4	4
4	4	4	5	4	5	5
4	4	4	5	5	5	5
5	5	5	4	5	5	5
5	5	5	5	5	5	5
5	5	5	5	5	5	5
5	5	5	5	5	5	5

65

Table 5.6 Permuted indicator supermatrix multinormal $\rho = .95$



It is clear that homogeneity analysis tries to quantify objects and categories in such a way that a parallelogram structure is approximated as well as possible. The consecutive 1's property applies (imperfectly) to both rows and columns and so we can interpret this gauge as a imperfect version of items with orderly interlocking, or a multicategory Guttman scale. The gauges we shall actually apply would reveal a even less perfect parallelogram, since the intercorrelations between the variables are not that high.

5.2.3 The Münsingen-Rain data: an empirical gauge

The Münsingen-Rain data consist of binary variables. The rows refer to 59 graves from the La Tène cemetery at Münsingen-Rain, Switzerland, and the columns represent 70 varieties of objects found in the graves. The data indicate by means of ones and zeros whether or not an

BIRNBAUM.

Table 1

Item Weights from Homogeneity Analysis

Item	a	w Starting Value	w Final Value
1	0.5	.69	.68
2	0.5	.56	.57
3	0.5	.62	.62
4	0.5	.60	.62
5	0.5	.52	.52
6	1.0	.99	.99
7	1.0	.99	.97
8	1.0	.95	.95
9	1.0	1.00	1.01
10	1.0	1.07	1.08
11	1.5	1.53	1.45
12	1.5	1.30	1.26
13	1.5	1.24	1.21
14	1.5	1.26	1.22
15	1.5	1.37	1.31

right or as a provisional analysis before an analysis with a model based on strong assumptions, like the multicategory rating Rasch model.

References

Bunch, J. R., Nielsen, C. P., & Sorensen, D. C. (1978). Rank-one modification of the symmetric eigenproblem. *Numerische Mathematik, 31*, 31–48.

de Gruijter, D. N. M., & van der Kamp, L. J. Th. (1984). *Statistical models in psychological and educational testing*. Lisse, The Netherlands: Swets & Zeitlinger.

de Leeuw, J. (1973). *Canonical analysis of categorical data*. Unpublished doctoral dissertation, University of Leyden, The Netherlands.

Gifi, A. (1980). *Niet-lineaire multivariate analyse*. Leyden, The Netherlands: University of Leyden, Department of Data Theory.

Guttman, L. (1950). The principal components of scale analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen, *Measurement and prediction* (pp. 312–361). Princeton NJ: Princeton University Press.

Henrysson, S. (1963). Correction of item-total correlations in item analysis. *Psychometrika, 28*, 211–218.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109–133.

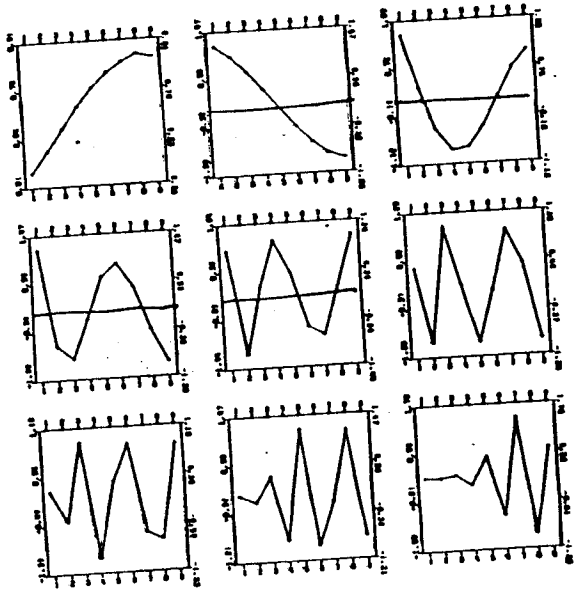
Lord, F. M. (1958). Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika, 23*, 291–296.

McDonald, R. P. (1968). A unified treatment of the weighting problem. *Psychometrika, 33*, 351–381.

McDonald, R. P. (1982). Some alternative approaches to the improvement of measurement in education and psychology: Fitting latent trait models. In D. Spearritt (Ed.), *The improvement of measurement in education and psychology* (pp. 213–232). Hawthorn, Victoria: The Australian Council for Educational Research.

McDonald, R. P. (1983). Alternative weights and invariant parameters in optimal scaling. *Psychometrika, 48*, 377–391.

Guttman 1950



figuur 2.2eigenvektoren Guttman gespreid

At some later point we shall study application of HOMALS to binary data & time series. But now .... CORRESPONDENCE ANALYSIS.

[HISTORICAL REMARKS].

What is it? We can be brief. CA is MCA if $m=2$. This is a bit too brief, perhaps. CA is often [in France, for instance) presented in different ways, not as a special case of MCA. Infact .... the other way around.

(a) Benzécri – distances [χ^2 – distance].

Suppose $F = \{f_{ij}\}$ is an $n \times m$ table with (relative) frequencies adding up to one. Let $e_i = \sum_j f_{ij}$ and $d_j = \sum_i f_{ij}$.

Define the Benzecri-distances between rows by

$$\delta_{ik}^2 = N \sum_{j=1}^{m} \frac{1}{d_j} \left[\frac{f_{ij}}{c_i} - \frac{f_{kj}}{e_k} \right]^2$$

or

$$\delta_{ik}^2 = N (u_i - u_k)' E^{-1} F D^{-1} F' E^{-1} (u_i - u_k).$$

The idea is to approximate this by ordinary Euclidean distances. We use the singular value decomposition

$$N^{1/2} E^{-1/2} F D^{-1/2} = K \Psi L' \qquad \left\{ \begin{array}{l} K'K = I \\ L'L = I \\ \Psi \text{ diag} \end{array} \right\}$$

and define

$$X = E^{-1/2} K \Psi \qquad (\text{thus } X'EX = \Psi K'K\Psi = \Psi^2)$$

Then

$$\delta_{ik}^2 = (u_i - u_k)' X X' (u_i - u_k) = d_y^2(X).$$

If we truncate the SVD at p dimensions, then

$$d_y^2(X_1) \leq d_y^2(X_2) \leq \cdots \leq d_y^2(X_r) = \delta_y^2$$

where $r = \text{rank}(F)$.

[Example Leiden (nº 16)

Again, dually, we can define B-distances between columns and approximate them. This gives $Y = D^{-1/2} L \Psi$.

Transition equations [SVD in disguise]

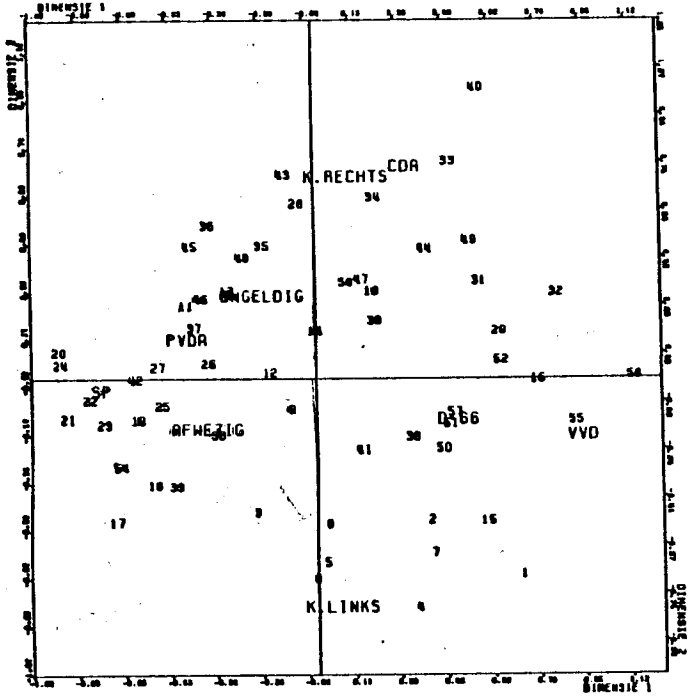
$$X = E^{-1} F Y \Psi^{\alpha}$$

$$Y = D^{-1} F' X \Psi^{\beta}$$

$\alpha = 0 \quad \beta = 1 \rightarrow$ first centroid principle
$\alpha = 1 \quad \beta = 0 \rightarrow$ second ——
$\alpha = 1/2 \quad \beta = 1/2 \rightarrow$ biplot scaling [Gabriel]
$\alpha = 1 \quad \beta = 1 \rightarrow$ French scaling.

Figuur 2.2: Analyse van marginale matrix van stemdistricten x partijen

$\lambda_1 = .273$ ($\tau_1 = .64$), $\lambda_2 = .155$ ($\tau_2 = .21$)

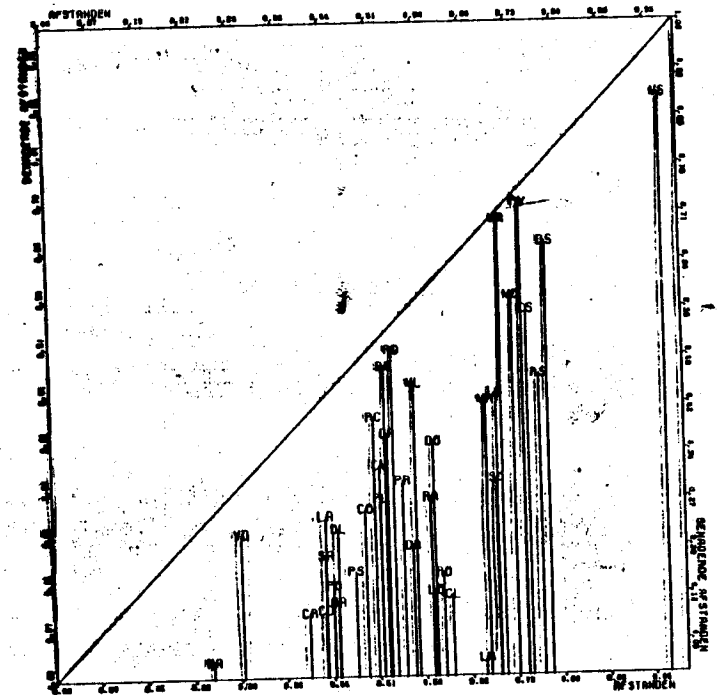


relatief veel VVD en D'66 stemt. Met "relatief" bedoel ik hier "t.o.v. het gemiddelde profiel van Leiden". Het gemiddelde Leidse partijprofiel en stemdistrictprofiel liggen in de oorsprong. Ik roep hier in herinnering dat een partijpunt i dicht ligt bij een stemdistrict j als $x_{ij} > e_{ij}$, en dat een partijpunt i ver weg ligt van stemdistrict j' als $x_{ij'} < e_{ij'}$. Door X en E te vergelijken, zou men eventueel kunnen nagaan of de oplossing wel juist geïnterpreteerd is. Het is immers mogelijk dat twee punten op de eerste twee dimensies dicht bij elkaar liggen, maar een grote afstand hebben op een hogere dimensie. De tweede dimensie maakt een onderscheid tussen stemdistricten uit de binnenstad, waar men veel klein links stemt (een verklaring hiervoor kan zijn dat hier veel studenten wonen) en stemdistricten waar men veel stemt op christelijke partijen. Men kan zien dat CA een helder

inzicht verschaft in de 9x58 matrix.

Met behulp van formule (2.14) worden de χ^2-afstanden nader bestudeerd. Met het linkerlid van (2.14) zijn de χ^2-afstanden tussen de 9 partijen berekend, en deze zijn vergeleken met de in de eerste twee dimensies benaderde χ^2-afstanden, berekend met het rechterlid van (2.14). In figuur 2.3 zijn voor de 36 paren rijpunten de ongekwadrateerde χ^2-afstanden (horizontaal) afgezet tegen de in de eerste dimensie benaderde χ^2-afstanden (verticaal). Als labels zijn de eerste letters van de partijen genomen (L voor klein links, en R voor klein rechts). De grootste afstanden zijn die tussen de SP en resp. de VVD, D'66, klein rechts en het CDA. Verder is de afstand groot tussen de VVD en resp. de PvdA, ONGELDIG en AFWEZIG. In figuur 2.2 worden deze afstanden voor een groot deel op de eerste dimensie benaderd: de verticale lijnen

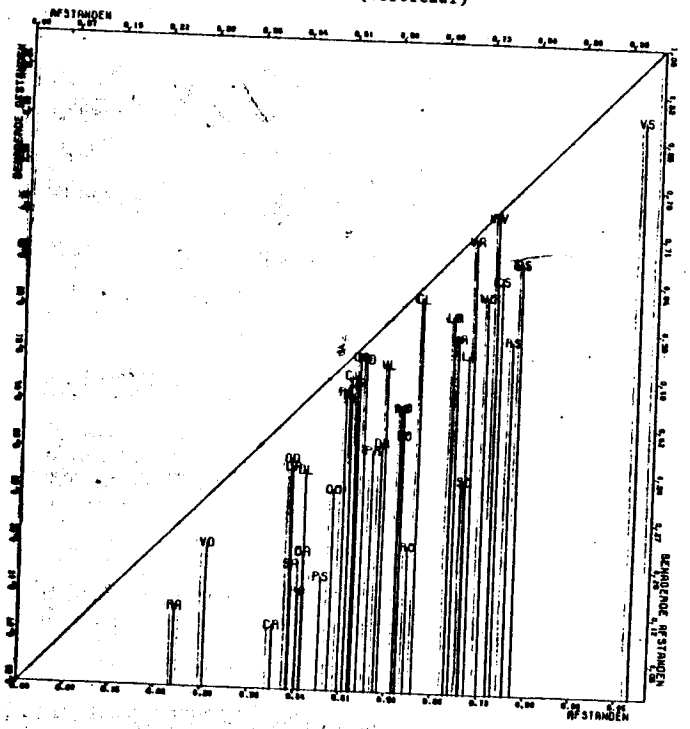
Figuur 2.3: χ^2-afstanden (horizontaal) vs. benaderde afstanden voor eerste dimensie (verticaal).



58

komen bijna tot de schuine lijn, waarvoor geldt dat de χ^2-afstand gelijk is aan de benaderde χ^2-afstand. Kleine afstanden zijn die tussen de paren PvdA en AFWEZIG, VVD en D'66, CDA en klein rechts. Deze paren halen relatief veel van hun stemmen uit dezelfde stemdistricten. Grote afstanden, die nauwelijks in de eerste dimensie worden benaderd, zijn bijvoorbeeld de afstanden tussen klein links en resp. klein rechts, CDA en ONGELDIG. Deze benadering vindt plaats in de tweede dimensie. Figuur 2.4 laat zien in hoeverre de χ^2-afstanden benaderd worden in twee dimensies. Alle verticale lijnen zijn hier langer dan in figuur 2.3, omdat een benadering in twee dimensies altijd beter is dan een benadering in één dimensie. Vooral de lijnen LR, CL, LO zijn een stuk langer geworden. Veel afstanden tot ONGELDIG worden nog steeds slecht gerepresenteerd. Deze worden in hogere dimensies afgebeeld. Hogere

Figuur 2.4: χ^2-afstanden (horizontaal) vs. benaderde afstanden voor eerste twee dimensies (verticaal)



14

54

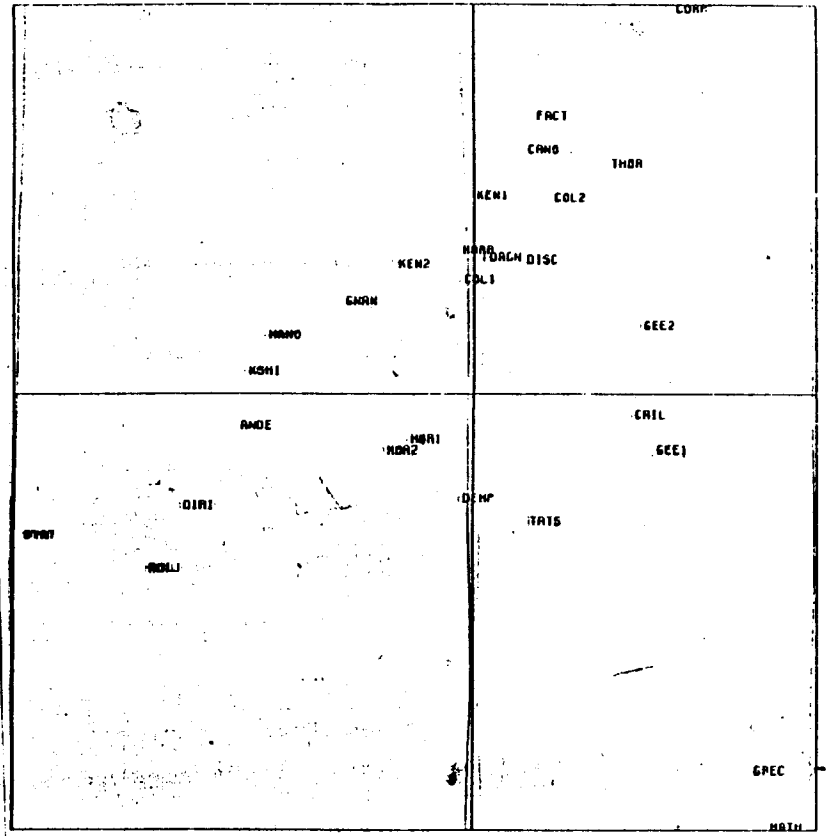
		A	B	C	D	E	F	G
1	Roy (1957)	31	0	0	0	0	164	11
2	Kendall (1957)	0	86	58	18	27	13	14
3	Kendall (1975)	0	40	82	10	42	60	0
4	Anderson (1958)	19	0	35	19	28	163	52
5	Cooley & Lohnes (1962)	14	7	35	22	17	0	56
6	Cooley & Lohnes (1971)	20	69	72	38	55	0	32
7	Morrison (1967)	74	0	86	14	0	184	48
8	Morrison (1976)	78	0	80	5	17	105	60
9	Van de Geer (1967)	74	19	33	12	26	0	0
10	Van de Geer (1971)	80	68	67	15	29	0	0
11	Dempster (1969)	108	48	4	10	46	108	0
12	Tatsuoka (1971)	109	13	5	17	39	32	46
13	Harris (1975)	16	35	69	24	0	26	41
14	Dagnelie (1975)	26	86	60	6	48	48	28
15	Green & Carroll (1976)	290	10	6	0	8	0	2
16	Caillez & Pages (1976)	184	48	82	42	134	0	0
17	Giri (1977)	29	0	0	0	41	211	32
18	Gnanadesikan (1977)	0	19	56	0	39	75	0
19	Kshirsagar (1978)	0	22	45	42	60	230	59
20	Thorndike (1978)	30	128	90	28	68	0	0

table 1.1: number of pages of MVA books devoted to several subjects.

1	ROIJ	-1.1086	-0.6144	-0.3390
2	KEN1	0.0740	0.7025	0.2526
3	KEN2	-0.2115	0.4605	-0.4923
4	ANDE	-0.7779	-0.1107	0.1556
5	COL1	0.0278	0.4065	1.0513
6	COL2	0.3578	0.6960	0.0928
7	MOR1	-0.1641	-0.1572	0.4635
8	MOR2	-0.2502	-0.1963	0.3900
9	GEE1	0.7279	-0.1945	-0.0475
10	GEE2	0.6840	0.2434	-0.1796
11	DEMP	0.0273	-0.3665	-0.4430
12	TATS	0.2680	-0.4475	0.2829
13	HARR	0.0219	0.5089	0.5172
14	DAGN	0.1205	0.4846	-0.1948
15	GREC	1.0831	-0.0321	0.0828
16	CAPA	0.6496	-0.0708	-0.1827
17	GIRI	-0.9835	-0.3927	-0.2502
18	GNAN	-0.4001	0.3292	-0.3383
19	KSHI	-0.7473	0.0810	-0.0051
20	THOR	0.5655	0.8145	-0.3526
A	MATH	1.1440	-1.5268	0.0692
B	CORR	0.8046	1.3602	-1.2504
C	FACT	0.2928	0.9827	0.6468
D	CANO	0.2605	0.8653	0.8399
E	DISC	0.2555	0.4777	-0.7914
F	STAT	-1.5635	-0.4972	-0.6071
G	MANO	-0.6745	0.2086	2.5069
LAMDA		0.6038	0.5183	0.3364

table 1.2:
profections of books, profections of subjects and singular values from correspondence analysis on table 1.1.

CORRESPONDENCE ANALYSIS ON MVA-BOOKS
COMPLETE



SINGULAR VALUES: $\lambda_{.1} = 0.6038$ $\lambda_{.2} = 0.5133$

figure 1.1

Relations with MCA

$$\begin{bmatrix} E & \vdots & F \\ \cdots & \cdots & \cdots \\ F' & \vdots & D \end{bmatrix} \begin{bmatrix} x \\ \cdots \\ y \end{bmatrix} = 2\lambda \begin{bmatrix} E & \vdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \vdots & D \end{bmatrix} \begin{bmatrix} x \\ \cdots \\ y \end{bmatrix}$$

$$Fy = [2\lambda - 1] E x$$

$$F'x = [2\lambda - 1] D y$$

$$\Leftrightarrow \begin{cases} \lambda_s = \frac{1}{2}(1 + \psi_s) & s = 1, \ldots, r \\ \lambda_s = \frac{1}{2} & s = r+1, \ldots, (n+m)-r \\ \lambda_s = \frac{1}{2}(1 - \psi_s) & s = (n+m)-r+1, \ldots, n+m \end{cases}$$

CA is special case of MCA of $m = 2$

On the other hand.

$$\begin{bmatrix} G_1 & \vdots & G_2 & \vdots & \cdots & \vdots & G_m \end{bmatrix} \begin{bmatrix} \\ \\ \end{bmatrix}$$

$$\begin{bmatrix} d_1 & \vdots & d_2 & \vdots & \cdots & \vdots & d_m \end{bmatrix}$$

MCA is CA applied to G. [Benzécri]

This remains true if G is incomplete

(remember Münsingen - Rain etc.)

Carry over from MCA

(LL) $N \Sigma \psi_s^2 = X^2$ (decomposition of departure from independence, analyses of residuals). Van der Heyden ------

CA find scores for rows and columns that maximize the correlation. This makes it [perhaps] interesting as a scaling technique. Example Guilford. Bock, 1960 Guttman, 1946.

CA linearizes bee regressions. Example Raven. Cf transition formulae. MCA.

Fisher decomposition $[[\eta\mu]]$ MCA

(LL) $F = E [[uu' + X \psi \gamma']] D$

$f_{ij} = d_j e_i [[1 + \sum_s \psi_s x_{is} y_{js}]]$

cards in the same groups and equal category quantifications for all ten replications. The homogeneity approach has a natural interpretation in psychophysical contexts, in which we very commonly suppose that there 'is' a one-dimensional scale and that the different variables are merely replications.

Guilford's data matrix is reproduced in table 10.2.1. The first singular value of an ANACOR analysis on this table was .93, the remaining singular values conformed closely to $\lambda_s = (.93)^s$, which shows that we only have to pay attention to a single dimension. Table 10.2.2 contains optimal quantifications of spot patterns and of piles ('intervals'). Both transformations are plotted in figures 10.1 and 10.2. It is clear from the transformation of the intervals that correspondence analysis does not follow the instruction to keep interpile distances equal, the intervals in the middle are larger, near the endpoints the distances are smaller. The transformation of the stimuli is fairly linear, deviations from linearity are in the direction of concavity. In this sense correspondence analysis, which produces the 'best' scale in a specific least squares sense, confirms Fechner's law, but this is obviously a very weak confirmation. It could very well be that a similar technique with a least absolute deviation loss function disconfirms Fechner's law.

S/R	1	2	3	4	5	6	7	8	9
15:	14	18	7	1					
16:	16	19	3	2					
17:	7	18	11	4					
19:	8	18	9	3	2				
20:	3	12	14	3	6	2			
22:	1	11	14	2	2				
24:		3	12	4	5	2			
26:		2	9	18	9	2			
28:			2	20	17	1			
30:				26	11	3			
32:			2	10	16	9	3		
35:				8	17	14	1		
37:				8	18	10	4		
40:				2	14	14	10		
43:					12	19	9		
46:				2	6	18	14		
48:					2	14	23	1	
50:						10	25	5	
56:						12	22	6	
60:						5	22	11	2
64:							14	20	6
69:							7	17	16
74:							6	20	14

Table 10.2.1 Guilford's Spot Pattern data.

64

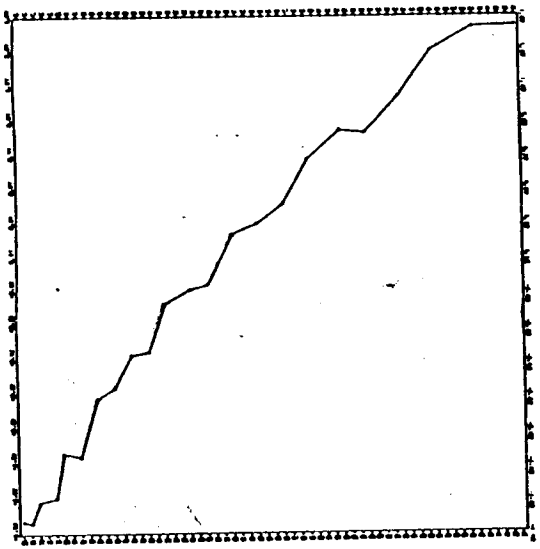


Figure 10.1 Guilford's spot patterns: transformations of the stimuli

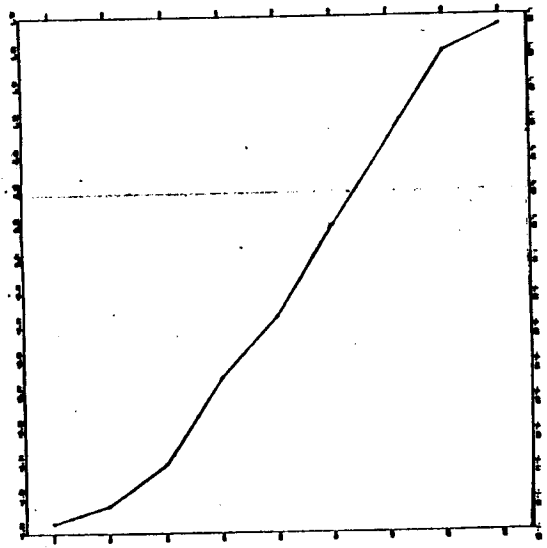
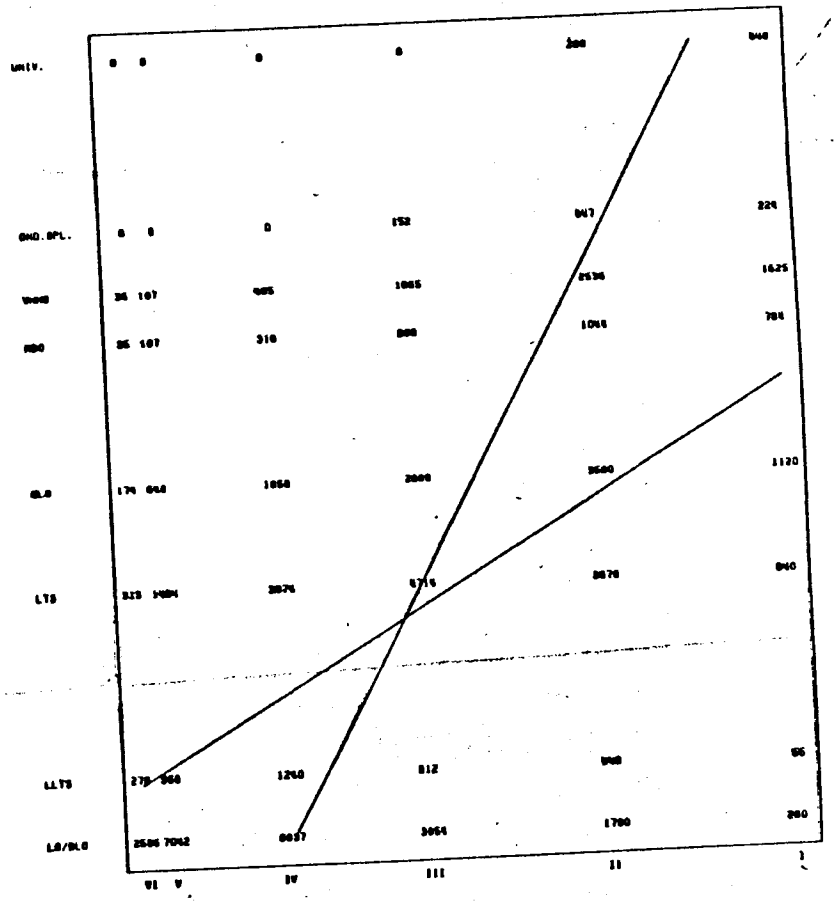


Figure 10.2 Guilford's spot patterns: transformations of the intervals

65

Figuur 3.4: Intelligentie (Raven-scores) en genoten schoolopleiding in 1952



Raven-scores langs de x-as in 6 klassen
(I zijn de hoogste scores)

If $m \geq 3$ then CA can still be applied in various ways.

- on the indicator matrices. (MCA)
- on the BURT table (MCA)
- on compound tables

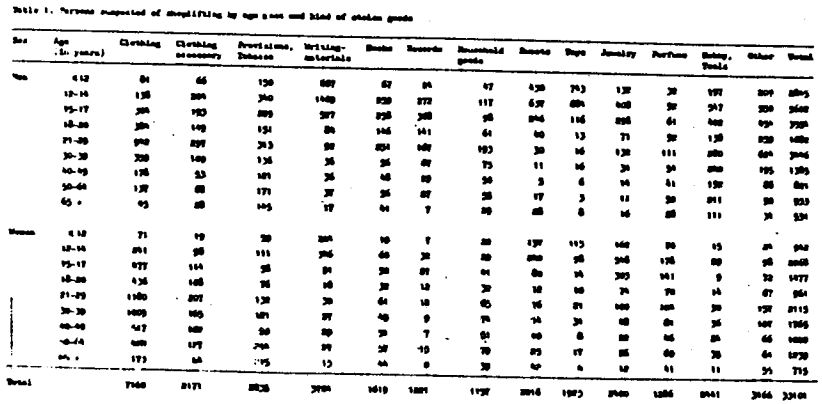
example → shoplifting

→ suicide

LLA: briefly. We have seen that CA decomposes the departure from independence. In the same way it is possible to define versions of CA which decompose departure from

- quasi - independence
- quasi - symmetry
- a hierarchical model.

This is discussed in the more specialized literature.



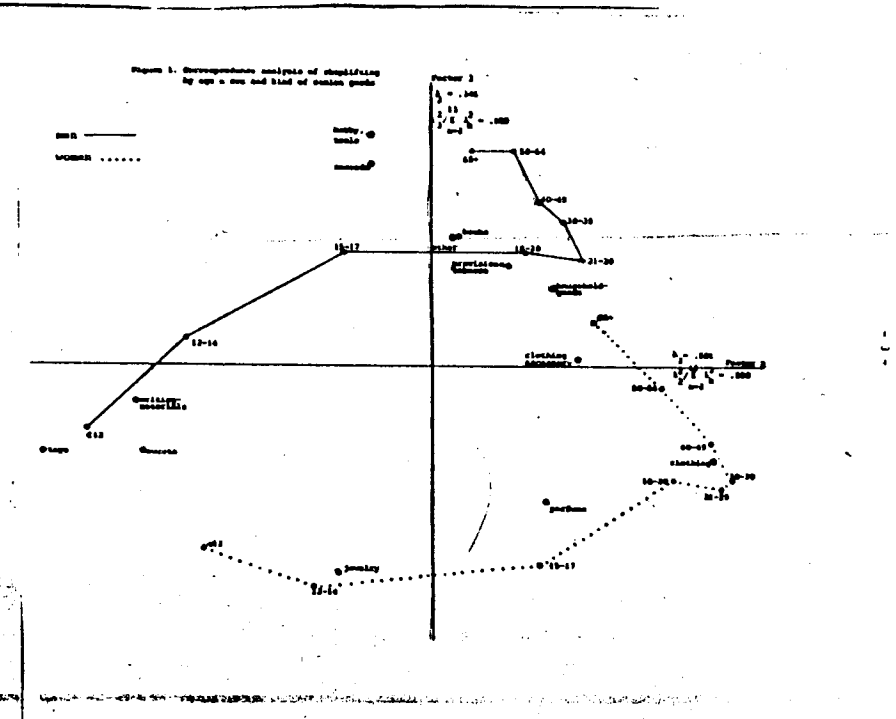


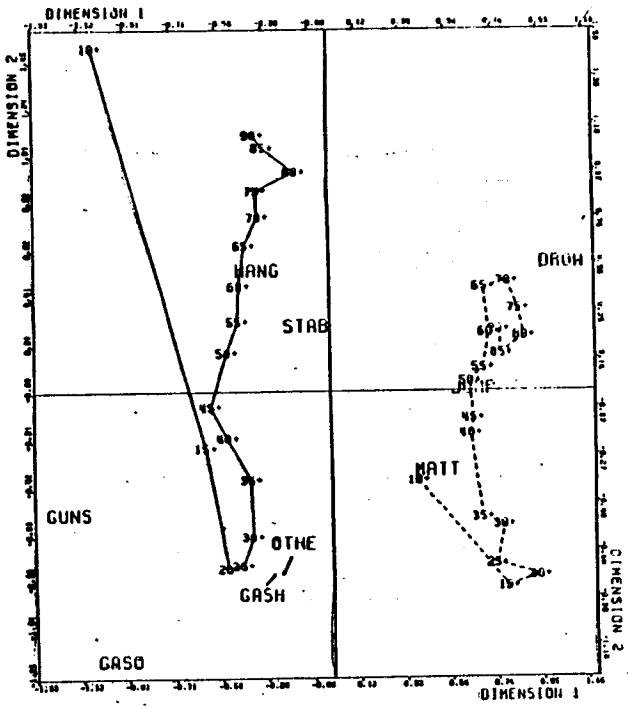
Table 1: Suicide behavior: age by sex by cause of death

Labels for cause of death-categories:

1. Suicide by solid or liquid matter (MATT)
2. Suicide by toxification of gas at home (GASH)
3. Suicide by toxification with other gas (GASO)
4. Suicide by hanging, strangling, suffocating (HANG)
5. Suicide by drowning (DROW)
6. Suicide with guns and explosives (GUNS)
7. Suicide with knifes etc. (STAB)
8. Suicide by jumping (JUMP)
9. Suicide with other methods (OTHE)

	MATT	GASH	GASO	HANG	DROW	GUNS	STAB	JUMP	OTHE	TOTAL
Men										
10-15	4	0	0	247	1	17	1	6	9	285
15-20	348	7	67	578	22	179	11	74	175	1461
20-25	808	32	229	699	44	316	35	109	289	2561
25-30	789	26	243	648	52	268	38	109	226	2399
30-35	916	17	257	825	74	291	52	123	281	2836
35-40	1118	27	313	1278	87	293	49	134	268	3567
40-45	926	13	250	1273	89	299	53	78	198	3179
45-50	855	9	203	1381	71	347	68	103	190	3227
50-55	684	14	136	1282	87	229	62	63	146	2703
55-60	502	6	77	972	49	151	46	66	77	1946
60-65	516	5	74	1249	83	162	52	92	122	2355
65-70	513	8	31	1360	75	164	56	115	95	2417
70-75	425	5	21	1268	90	121	44	119	82	2175
75-80	266	4	9	866	63	78	30	79	34	1429
80-85	159	2	2	479	39	18	18	46	19	782
85-90	70	1	0	259	16	10	9	18	10	393
90+	18	0	1	76	4	2	4	6	2	113
TOT	8917	176	1913	14740	946	2945	628	1340	2223	33828
Women										
10-15	28	0	3	20	0	1	0	10	6	68
15-20	353	2	11	81	6	15	2	43	47	560
20-25	540	4	20	111	24	9	9	78	67	862
25-30	454	6	27	125	33	26	7	86	75	839
30-35	530	2	29	178	42	14	20	92	78	985
35-40	688	5	44	272	64	24	14	98	110	1319
40-45	566	4	24	348	76	18	22	103	86	1242
45-50	716	6	24	447	84	13	21	95	88	1504
50-55	942	7	26	691	184	21	37	129	131	2168
55-60	723	3	14	527	163	14	30	92	92	1658
60-65	820	8	8	702	245	11	35	140	114	2083
65-70	740	8	4	785	271	4	38	156	90	2096
70-75	624	6	4	610	244	1	27	129	46	1691
75-80	495	8	1	420	161	2	29	129	35	1279
80-85	292	3	2	223	78	0	10	84	23	715
85-90	113	4	0	83	14	0	6	34	2	256
90+	24	1	0	19	4	0	2	7	0	57
TOT	8648	77	241	5637	1703	172	309	1505	1090	19382

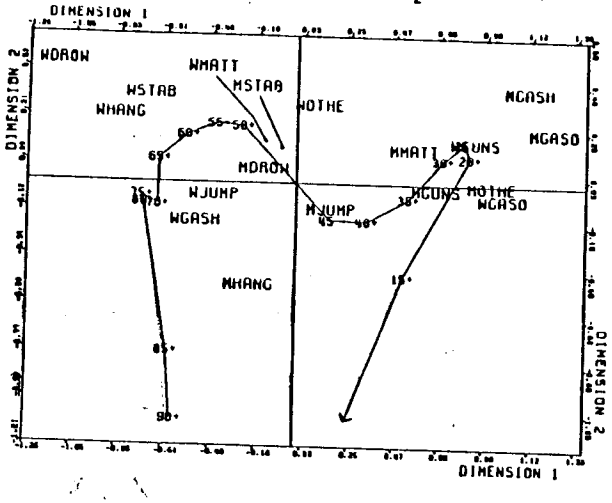
Figure 1: Analysis of $F^{Mx(SxA)}$, first two dimensions. The line for men is solid. Singular values with relative contributions: λ_1 = .312 (.519), λ_2 = .268 (.381).



[M] [SA]

Psychometrika | ID NUMBER Van der Heijden | DATE 1/00
DATE | FIGURE 1 | w.with

Figure 2: Analysis of $F^{Ax(MxS)}$, first two dimensions. Singular values with relative contributions: λ_1 = .315 (.808), λ_2 = .100 (.082)



[A][MS]

JOB	ID NUMBER Van der Heijden	
DATE	FIGURE 2	

Applications to Finnie

Three distinct applications are shown when illustrate

(a) homogeneity analysis (in this context also known as qualitative harmonic analysis). BABY DATA.

(b) use of passive variables (éléments supplementaires).

	stepping	standing	w w help	w alone
martin	15	?	21	50
carol	15	19	37	50
max	14	?	25	54
virginia ruth	?	21	40	54
sibyl	?	22	37	58
david	19	27	34	60
james d.	19	30	54	60
harvey	14	27	42	62
winifred	15	30	41	62
quentin	15	23	38	64
maurice	18	23	45	66
judy	18	29	45	66
irene may	19	34	45	66
peter	15	29	49	66
walley	18	33	54	68
fred	15	32	46	70
donovan	?	23	50	70
patricia	15	30	45	70
torey	?	21	72	74
larry	13	41	54	76
doris	?	23	44	?

table 1: transitions in weeks

martin	1111112222224444444444444444444444444444445555555555555555555555555555555
carol	1111112222333333333333333333444444444444445555555555555555555555555555555
max	1111122222222224444444444444444444444444444445555555555555555555555555555
virginia	6666666666663333333333333333333444444444444455555555555555555555555555555
sibyl	6666666666666333333333333333444444444444444455555555555555555555555555555
david	1111111111222222223333344444444444444444444444444455555555555555555555555
james d	1111111111222222222233333333333333444444444444445555555555555555555555555
harvey	1111122222222222223333333333333344444444444444444455555555555555555555555
winifred	1111112222222222222233333333333444444444444444444455555555555555555555555
quentin	1111112222222233333333333333444444444444444444444455555555555555555555555
maurice	1111111112222233333333333333334444444444444444444455555555555555555555555
judy	1111111111222222222233333333333334444444444444444444555555555555555555555
irene	1111111111222222222222233333333333444444444444444444455555555555555555555
peter	1111112222222222222333333333333333444444444444444444555555555555555555555
walley	1111111112222222222222233333333333333333344444444444445555555555555555555
fred	1111112222222222222222333333333333344444444444444444444555555555555555555
donovan	6666666666666633333333333333333333333344444444444444444455555555555555555
patricia	1111112222222222222233333333333333444444444444444444445555555555555555555
torey	6666666666663333333333333333333333333333333333333333333333333344555555555
larry	1111222222222222222222222222233333333333344444444444444444444445555555555

table 2: 71 time-points mapping 20 babies into 6 states

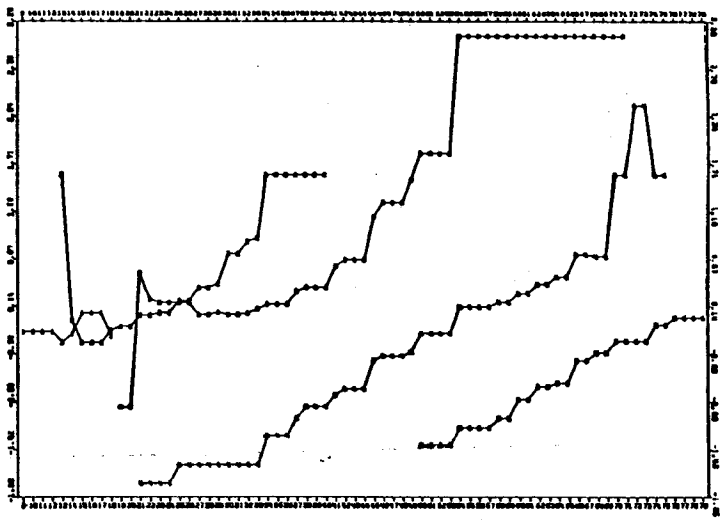


figure 1: category quantifications baby data

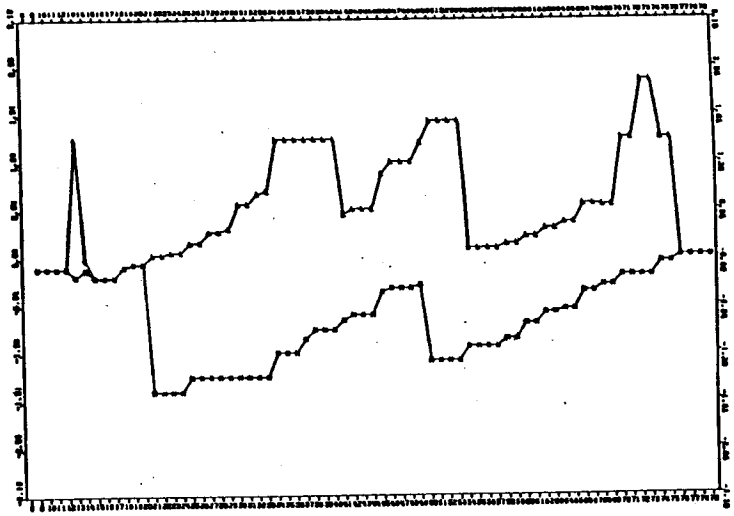


figure 2: induced curves for martin and larry

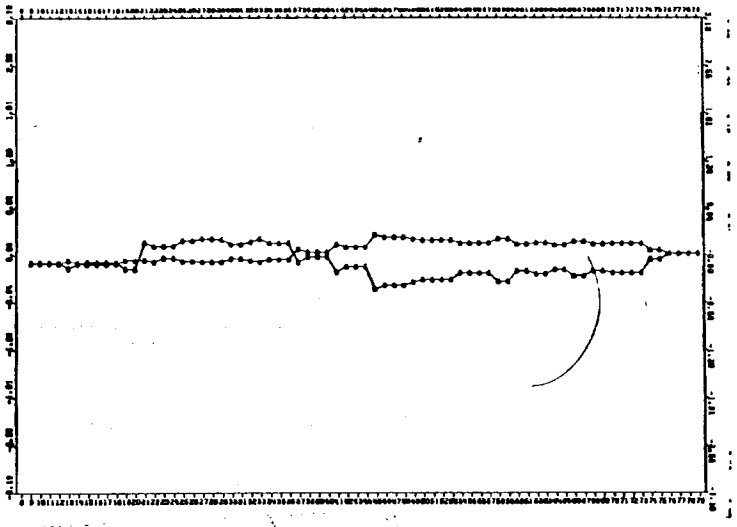


figure 3: average induced curves for boys and girls

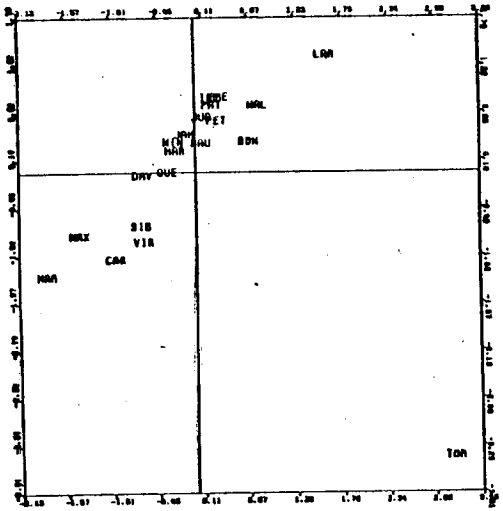


figure 4: baby object scores in two dimensions

	1978	1979	1980	1981	1982
lbo-	384	408	279	95	132
lbo+			115	132	
mavo 1	384	2			
mavo 2	1695	725	64	2	
mavo 3-		1345	829	133	8
mavo 3+		11	28	11	
mavo 4-			377	262	276
mavo 4+			589	460	
havo 1	107				1
havo 2	1329	289	8		
havo 3		1154	469	26	1
havo 4			970	1011	472
havo 5-				233	928
havo 5+				336	
vwo 1	34				
vwo 2	1449	106			
vwo 3		1276	175	5	
vwo 4			1071	215	22
vwo 5			1	952	462
vwo 6				1	792
kmbo				1	22
mbo 1			12	426	608
mbo 2				16	426
mbo 3					18
hbo 1				5	154
uitval	82	148	477	1142	1142
totaal	5464	5464	5464	5464	5464

tabel 2

schoolloopbaan kategorieen, verdeling over de meetpunten.

	leerjaar 2	leerjaar 3	leerjaar 4	leerjaar 5	leerjaar 6
lbo-	-1.598	-1.530	-1.440	-1.332	-1.319
lbo+			-1.393	-1.338	
mavo 1	-1.146				
mavo 2	-0.832	-1.085	-1.158		
mavo 3-		-0.763	-0.968	-1.051	-1.185
mavo 3+		-0.793	-1.021	-1.045	
mavo 4-			-0.691	-0.903	-0.972
mavo 4+			-0.698	-0.861	
havo 1	-0.334				
havo 2	+0.019	-0.255	-0.257		
havo 3		+0.124	-0.118	-0.139	
havo 4			+0.246	-0.145	-0.495
havo 5-				+0.298	-0.017
havo 5+				+0.296	
vwo 1	+0.452				
vwo 2	+1.396	+0.718			
vwo 3		+1.520	+1.008		
vwo 4			+1.633	+1.271	+0.696
vwo 5				+1.665	+1.167
vwo 6					+1.686
kmbo					-1.080
mbo 1			-0.071	-0.863	-0.830
mbo 2				-0.136	-0.697
mbo 3					+0.035
hbo 1					+0.341

tabel

PRIMALS kwantifikaties van de schoolloopbaan kategorieen.

	II	III	IV	V	VI
Groningen	-.247	-.199	-.098	.018	.015
Friesland	-.074	-.032	-.097	-.037	-.035
Drente	.016	.040	.090	.220	.177
Overijssel	-.177	-.219	-.156	-.062	-.062
Gelderland	.006	-.016	.007	.064	.055
Utrecht	.065	.006	.002	.169	.206
Noord Holland	-.085	-.052	.052	.189	.176
Zuid Holland	-.099	-.075	-.008	.088	.083
Zeeland	.192	.236	.307	.407	.468
Noord Brabant	.114	-.095	-.047	.057	.085
Limburg	-.217	-.180	-.072	-.032	-.029
IJsselmeerpolders	-.177	-.141	-.096	.367	.053
eta^2 provincie	.0137	.0115	.0068	.0175	.0197
jongens	-.150	-.142	-.075	.023	.028
meisjes	-.017	.000	.058	.168	.163
eta^2 sexe	.0044	.0047	.0042	.0051	.0047
vwo advies	.728	.769	.819	.868	.817
havo advies	.064	.065	.111	.171	.152
mavo advies	-.673	-.680	-.631	-.534	-.475
lbo advies	-1.054	-1.053	-.991	-.782	-.747
eta^2 advies	.2950	.2998	.2971	.2891	.2656
hogere employes	.249	.271	.315	.378	.365
middel employes	.103	.119	.162	.220	.217
lagere employes	-.114	-.101	-.066	-.002	.015
zelfst met pers	-.231	-.193	-.123	-.035	-.059
zelfst zon pers	-.168	-.153	-.098	-.032	-.033
gesch arbeiders	-.196	-.200	-.144	-.017	-.014
onges arbeiders	-.424	-.445	-.360	-.183	-.222
overigen	-.244	-.216	-.118	.021	.045
eta^2 SES	.0466	.0453	.0359	.0346	.0371
intercept	-2.919	-2.968	-2.892	-2.733	-2.543
beta taal	.064	.065	.064	.061	.057
beta rekenen	.055	.058	.058	.057	.053
R^2 CITO	.3428	.3458	.3375	.3214	.3028

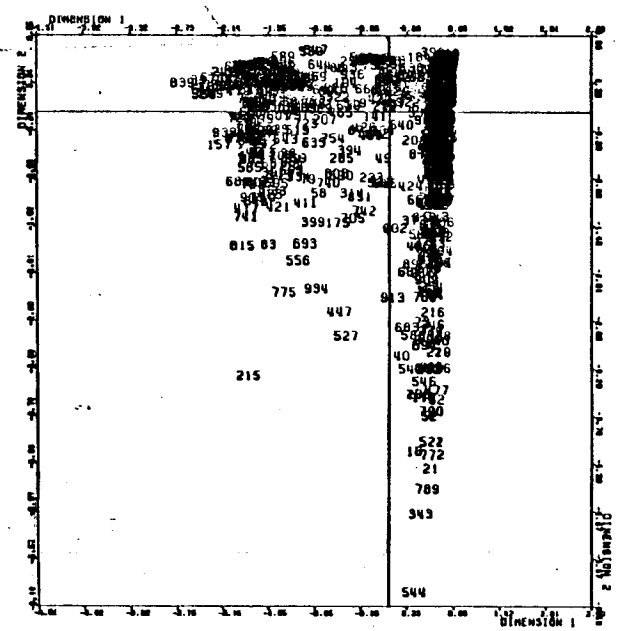
tabel 2.

Relaties van schoolloopbaan variabelen met achtergronds-variabelen.

78

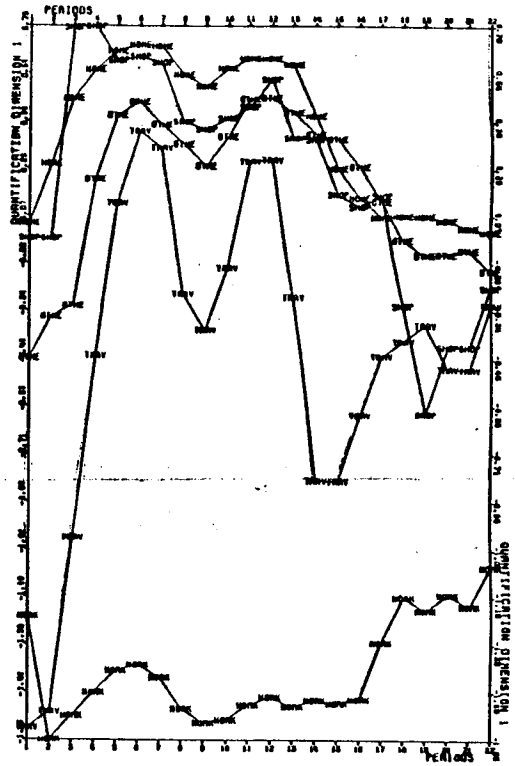
(a) the 'elbow'-criterion indicates that the remaining singular values are approximately equal, (b) the remaining dimensions are not really 'common' factors but contrast either one time period or one activity with the rest, (c) three-dimensional plots are less attractive. Figure 1 shows the projections of the 940 individuals on the first two singular vectors. The very large cluster in the top right-hand section are the individuals who are mainly at home. The second large cluster, top left, are people who work full-time. Thus dimension one, the horizontal dimension, contrasts 'work' with 'being-at-home'. The second, vertical, dimension contrasts 'working' and 'being-at-home', at the top, with other activities outside the house, at the bottom. These alternative activities are mainly in the category 'other', but also in 'shopping' and 'traveling'. This interpretation of the dimensions becomes beautifully clear if we plot the 5 x 22 = 110 category quantifications. In this analysis we have not plotted them in two dimensions, but we have plotted each dimension separately against time-period. Figure 2a is the plot for the first dimension. We have seen that these category quantifications make maximum discrimination of the individuals possible. They are related in a simple way to figure 1: the score for 'travel' in period 6 is the average score of all individuals who travel in period 6 (on dimension 1, and weighted with the number of minutes they travel). Individuals who work a lot are low on the dimension, individuals who are at home are high. During working hours the

Figure 1: analysis 1, 940 object scores in two dimensions



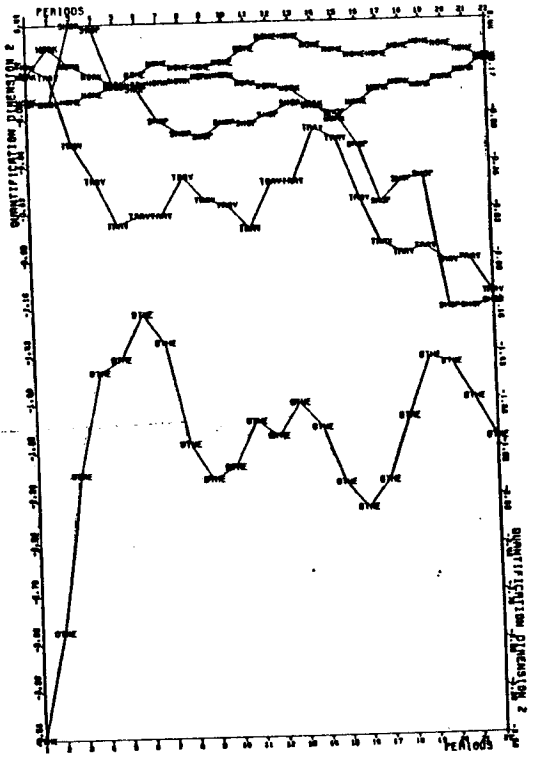
79

Figure 2a: analysis 1, category quantifications, dimension one against time



80

Figure 2b: analysis 1, category quantifications, dimension two against time



average for travelling persons is close to the average for persons at home, during the lunch break. During the early morning and late evening hours it is much closer to the average of the persons who work. The same thing is true for shopping, although working people do not shop a great deal during lunch time, and do not even shop much during late afternoon. Shopping is done by those who stay at home. Figure 2b shows a similar plot for the second dimension. The social and cultural activities are concentrated in the morning, in the early afternoon, and in the early evening. The morning and early

81

afternoon shopping behaves like 'other', in the evening it is quite different. Travelling behaves in the opposite way. If you are going to visit somebody or something, then you have to travel before and after this visit. Thus travelling hours are just before and just after visiting hours. On the first dimension we can best discriminate people during working hours, on the second dimension we can best discriminate them outside working hours. More detailed interpretations of these plots are in Kreft and Mulder (1984).

Another way to interpret the dimensions of the homogeneity analysis is by using passive variables (or supplementary variables). They play no role in the analysis, only in the interpretation. They are used afterwards to label the plots, and to compute centroids of groups of individuals. In our analysis we have used two passive variables: sex combined with work-situation of the family. Work situation has nine possible values: the head of the family can be employed full-time, part-time, or no-time, and the same thing is true for his/her partner. If we combine this with sex we have a new interactive passive variable

Figure 3: analysis 1, centroids for passive sex x work variable

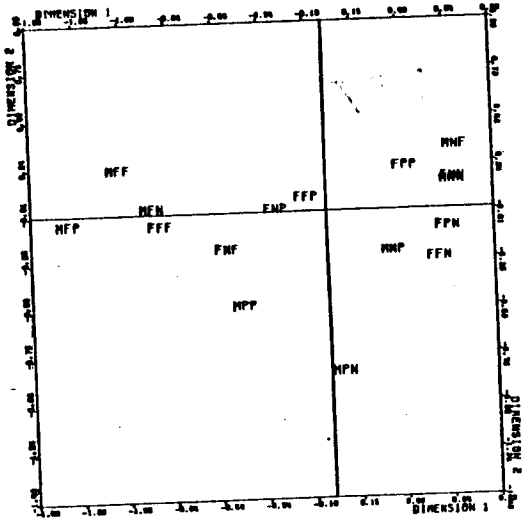


Figure 5a: analysis 2, 20680 x 5 table, average object scores for men and women, dimension one against time

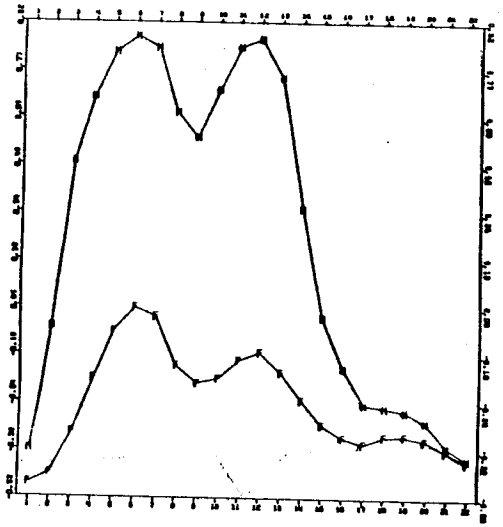
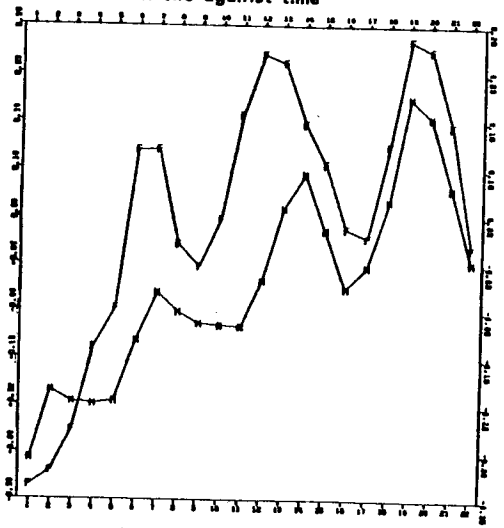


Figure 5b: analysis 2, 2068 x 5 table, average object scores for men and women, dimension two against time



Back to Theory

Nonlinear PCA [program PRINCALS]

MCA is related to PCA, but quite different.

If $p=1$ then our analytical approach [$\max_z \lambda_{max}(R)$] shows that the relationship is quite close. But

if $p>1$ then MCA finds a different transformation of the variables (and thus, implicitly, a new induced correlation matrix) for each dimension.

Since there are $\sum_{j=1}^{m} (k_j - 1)$ nontrivial MCA dimensions, this would mean as many correlation matrices [each of which could be used in a PCA !].

GIFI calls this DATA PRODUCTION.

In order to create a link with PCA which is stronger, we shall now introduce various restrictions into the Gifi systems. They are

(a) rank restrictions
(b) cone restrictions

For ease of reference we repeat our basic loss function

$$\sigma(X;Y) = \frac{1}{m}\sum_{j=1}^{m} \mathrm{tr}\,[X - G_j Y_j]'[X - G_j Y_j]$$

and our normalization

$$u'Y = 0,$$
$$X'X = I.$$

(a) Rank restrictions

IDENTIFICATION

$$\underset{k_j \times p}{Y_j} = \underset{k_j \times p}{Z_j A_j'} \qquad \left\{ \begin{array}{l} u_j' D_j Z_j = 0 \\ Z_j' D_j Z_j = I \end{array} \right\}$$

Geometrical

[Figure CARS]

a_j defines a direction [through the origin] and z_j indicates the location on that direction. Thus we require

(a) perfect homogeneity
(b) all objects must be on parallel straight lines

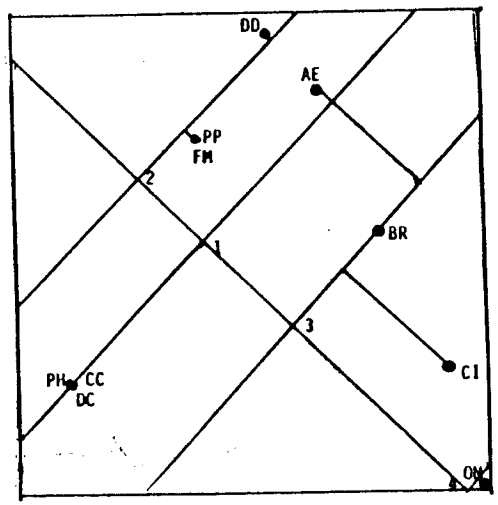


figure 3: single nominal loss, variable 1, arbitrary solution

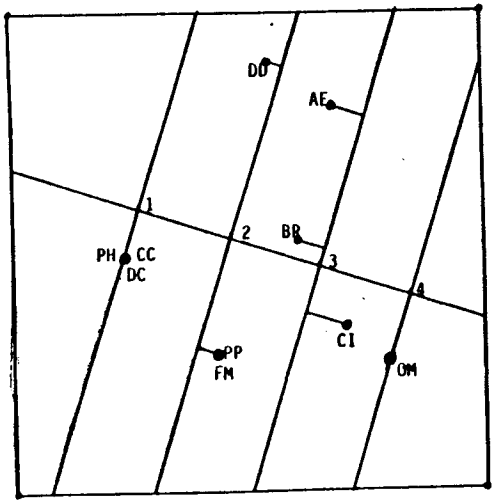


figure 4: single numerical loss, variable 1, optimal solution

ANALYTICAL

For the analytical interpretation we start from the dual formulation

$$\max \operatorname{tr} Y'CY \qquad \text{with} \quad Y'DY \leq mI.$$

If all variables are single (i.e. not multiple, i.e. rank-restricted) then

$$\operatorname{tr} Y'CY = \sum_j \sum_e \operatorname{tr} Y_j'C_{je}Y_e = \sum_j \sum_e \operatorname{tr} a_j z_j' C_{je} z_e a_e' =$$

$$= \sum_j \sum_e r_{je}\, a_j' a_e = \operatorname{tr} RAA' = \operatorname{tr} A'RA.$$

$$\blacklozenge\ Y'DY = \sum_j \blacklozenge\ Y_j' D_j Y_j = \sum_j \blacklozenge\ a_j z_j' D_j z_j a_j =$$

$$= \blacklozenge\ A'A.$$

It follows (in the same way as before) that homogeneity analysis, when all variables single [nominal] is equivalent to

$$\max_{z_1,\dots,z_m} \lambda_1[R] + \cdots + \lambda_p[R]$$

where R depends on the z_j ($r_{je} = z_j' C_{je} z_e$).

It also follows that if we impose the further restriction that the z_j are known vectors, then the only unknown are in A_j and the technique becomes ordinary PCA. [or linear PCA].

Also: if $p=1$ then rank-one is no restriction.

(b) Cone restrictions

$$z_j \in K_j$$

- K_j is $\mathbb{R}^{k_j}$ [no restriction, single nominal]
- K_j is a subspace of $\mathbb{R}^{k_j}$ [for instance of polynomials]
- K_j is a ray [known vector, single numerical]
- K_j is the cone of monotone transformations [single ordinal]

In PRINCALS variables can be mixed

- multiple nominal [MCA]
- single nominal
- – ordinal
- – numerical [PCA]

TABLE 1. Table of social indicator statistics taken from statistical abstracts of the US (1977). U.S. Department of Commerce: Bureau of the census.

Popul : 1975 population in thousands
Incom : Per capita income in dollars
Illit : Illiteracy rate in percent of population
Life : Life expectancy in years
Homic : 1976 homicide and non-negligent manslaughter rate (per 1000)
School: Percent of the population over age 25 who are high school graduates
Freeze: Average numbers of days of the year in which temperature falls below freezing.

State		Popul	Income	Illit	Life	Homic	School	Freeze
Alabama	AB	3615	3624	2.1	69.05	15.1	41.3	20
Alaska	AK	365	6315	1.5	69.31	11.3	66.7	152
Arizona	AZ	2212	4530	1.8	70.55	7.8	58.1	15
Arkansas	AR	2110	3378	1.9	70.66	10.1	39.9	65
California	CA	21198	5114	1.1	71.71	10.3	62.6	20
Colorado	CO	2541	4884	0.7	72.06	6.8	63.9	166
Connecticut	CT	3100	5348	1.1	72.48	3.1	56.0	139
Delaware	DW	579	4809	0.9	70.06	6.2	54.6	103
Florida	FL	8277	4815	1.3	70.66	10.7	52.6	11
Georgia	GA	4931	4091	2.0	68.54	13.9	40.6	60
Hawaii	MI	868	4963	1.9	73.60	6.2	61.9	0
Idaho	ID	813	4119	0.6	71.87	5.3	59.5	126
Illinois	IL	11197	5107	0.9	70.14	10.3	52.6	127
Indiana	IN	5313	4458	0.7	70.88	7.1	52.9	122
Iowa	IA	2861	4628	0.5	72.56	2.3	59.0	140
Kansas	KS	2280	4669	0.6	72.58	4.5	59.9	114
Kentucky	KY	3387	3712	1.6	70.10	10.6	38.5	95
Louisiana	LA	3806	3545	2.8	68.76	13.2	42.2	12
Maine	ME	1058	3694	0.7	70.39	2.7	54.7	161
Maryland	MD	4122	5299	0.9	70.22	8.5	52.3	101
Massachus.	MA	5814	4755	1.1	71.83	3.3	58.5	103
Michican	MI	9111	4751	0.9	70.63	11.1	52.8	125
Minnesota	MN	3921	4675	0.6	72.96	2.3	57.6	160
Mississippi	MS	2341	3098	2.4	68.09	12.5	41.0	50
Missouri	MO	4767	4254	0.8	70.69	9.3	48.8	108
Montana	MT	746	4347	0.6	70.56	5.0	59.2	155
Nebraska	NB	1544	4508	0.6	72.60	2.9	59.3	139
Nevada	NV	590	5149	0.5	69.03	11.5	65.2	188
New Hampsh.	NH	812	4281	0.7	71.23	3.3	57.6	174
New Jersey	NJ	7333	5237	1.1	70.93	5.2	52.5	115
New Mexico	NM	1144	3601	2.2	70.32	9.7	55.2	120
New York	NY	18076	4903	1.4	70.55	10.9	52.7	82
N. Carolina	NC	5441	3875	1.8	69.21	11.1	38.5	80
N. Dakota	ND	637	5087	0.8	72.78	1.4	50.3	186
Ohio	OH	10735	4561	0.8	70.82	7.4	53.2	124
Oklahoma	OK	2715	3983	1.1	71.42	6.4	51.6	82
Oregon	OR	2284	4660	0.6	72.13	4.2	60.0	44
Pennsylv.	PA	11860	4449	1.0	70.43	6.1	50.2	126
Rh. Island	RI	931	4558	1.3	71.90	2.4	46.4	127
S. Carolina	SC	2816	3635	2.3	67.96	11.6	37.8	65
S. Dakota	SD	681	4167	0.5	72.08	1.7	53.3	172
Tennessee	TN	4173	3821	1.7	70.11	11.0	41.8	70
Texas	TX	12237	4188	2.2	70.90	12.2	47.4	35
Utah	UT	1203	4022	0.6	71.72	4.3	67.3	137
Vermont	VT	472	3907	0.6	71.64	5.5	57.1	168
Virginia	VA	4981	4701	1.4	70.08	9.5	47.8	85
Washington	WA	3559	4864	0.6	71.72	4.3	63.5	32
W. Virginia	WV	1799	3617	1.4	69.48	6.7	41.6	100
Wisconsin	WI	4589	4468	0.7	72.48	3.0	54.5	149
Wyoming	WY	376	4566	0.6	70.29	6.9	62.9	173

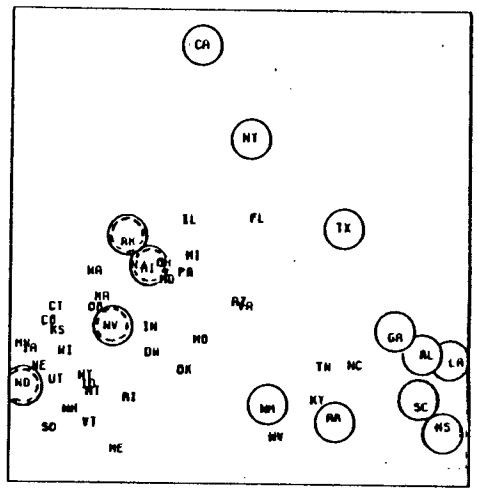


Figure 1. PCA solution for 50 states. Encircled points have dissimilarities larger than average. Dotted circles indicate more than average stress im addition.

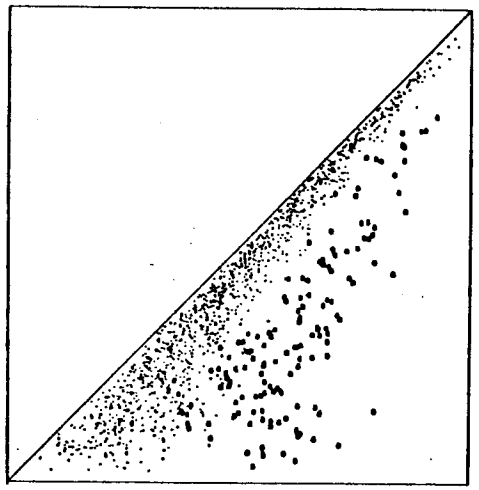


Figure 2. PCA solution for 50 states. δ(Z) (borizontal axis) versus d(X) (vertical axis). Approximation from below. Ellipses refer to all pairs including AK, HI and NV.

92

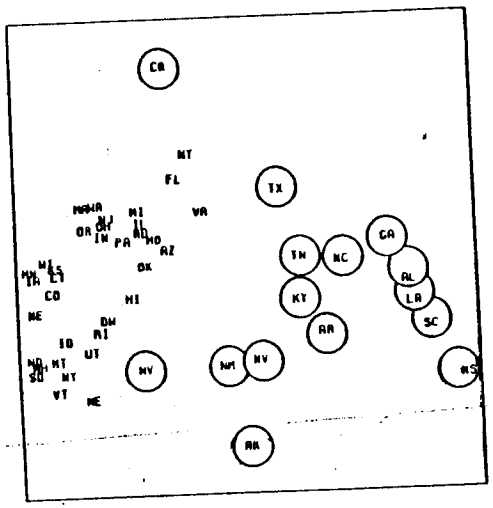


Figure 5. Nonlinear PCA solution for 50 states. Encircled points have dissimilarities larger than average.

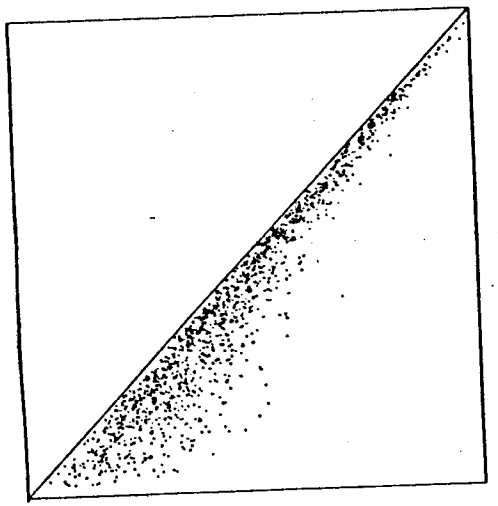


Figure 6 Nonlinear PCA solution for 50 states. $\delta(\underline{Z})$ (horizontal axis) versus $d(\underline{X})$ (vertical axis). Approximation from below.

93

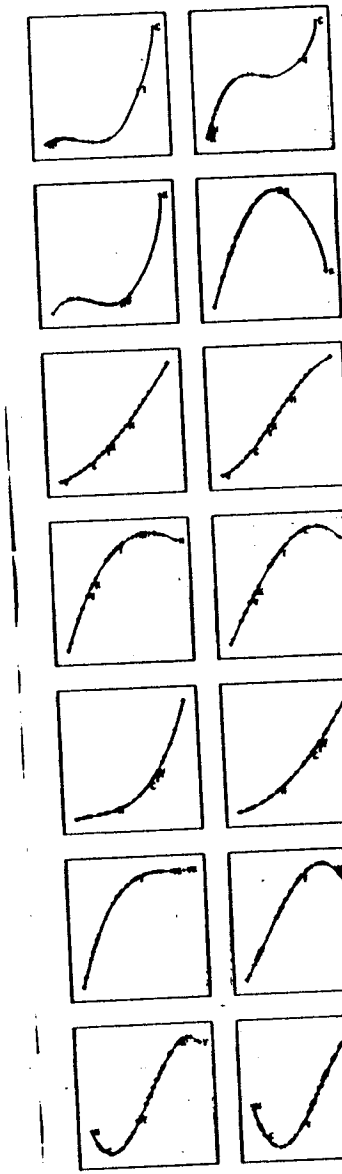


Figure 9. Transformation. MDS on the left, PCA on the right.

Variable 1: Population

Variable 2: Income

Variable 3: Illiteracy

Variable 4: Life expectancy

Variable 5: Homicide

Variable 6: High school

Variable 7: Freezing

1:	7	4	1	8	10	9	5	2	3	6	(S)
2:	7	6	2	9	3	8	10	1	4	5	(S)
3:	10	5	1	7	4	6	8	2	3	9	(S)
4:	6	5	3	7	4	8	9	2	1	10	(S)
5:	6	3	5	10	4	2	9	7	8	1	(D)
6:	8	7	4	9	2	5	10	6	3	1	(D)
7:	5	9	4	8	6	2	10	7	3	1	(D)
8:	6	7	4	9	5	3	10	8	2	1	(D)
9:	2	3	6	4	5	8	9	7	10	1	(D)
10:	5	8	2	9	1	7	10	6	4	3	(D)
11:	7	2	6	10	5	1	9	8	4	3	(D)
12:	8	7	2	9	1	6	10	5	3	4	(C)
13:	10	7	1	9	4	6	8	2	3	5	(C)
14:	5	2	3	4	1	8	7	9	6	10	(C)
15:	6	5	2	7	1	10	9	8	4	3	(C)
16:	4	7	5	2	8	9	1	6	3	10	(M)
17:	4	7	5	3	9	8	1	6	2	10	(M)
18:	5	4	7	3	9	8	1	10	2	6	(M)
19:	1	5	6	7	10	9	3	8	2	4	(E)
20:	1	5	8	7	9	3	6	10	2	4	(E)
21:	3	7	6	2	8	4	5	9	1	10	(E)
22:	1	3	8	6	9	7	4	10	2	5	(E)
23:	1	4	6	5	9	10	2	8	3	7	(E)
24:	1	7	5	4	10	9	3	8	2	6	(E)
25:	1	8	6	5	9	4	3	10	2	7	(E)
26:	1	2	5	6	10	4	7	9	3	8	(E)
27:	1	5	6	4	8	7	2	9	3	10	(E)
28:	4	6	5	1	7	10	3	8	2	9	(E)
29:	8	7	1	2	9	10	6	3	4	5	(R)
30:	7	4	1	2	9	10	8	6	3	5	(R)
31:	9	8	2	7	1	4	10	5	6	3	(R)
32:	7	1	5	8	2	6	3	9	4	10	(T)
33:	2	3	7	8	10	9	1	6	4	5	(T)
34:	10	4	2	9	3	5	6	8	1	7	(T)
35:	3	2	10	6	8	4	7	9	1	5	(T)
36:	6	1	3	9	4	7	10	2	5	8	(T)
37:	2	1	6	4	10	9	5	7	3	8	(T)
38:	2	3	6	5	7	8	4	9	1	10	(A)
39:	2	6	7	3	10	8	4	9	1	5	(A)

table 5.6: Roskam's journal data
preference rank orders

III.3. Roskam's journal preference data.

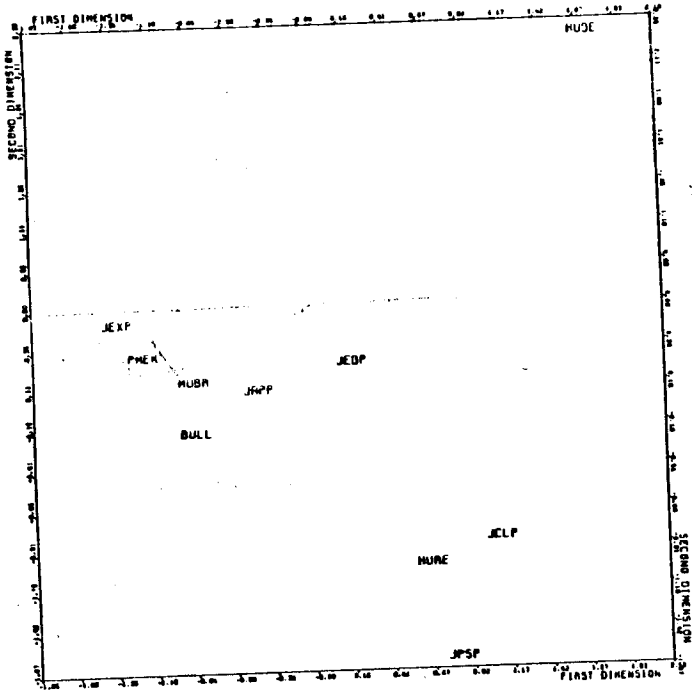
III.3.1.

Table III.3.1. gives preference rank orders of 39 psychologists for ten psychological journals (from Roskam, 1968).
Columns of the table refer to the following journals:
1: JEXP: Journal of experimental psychology,
2: JAPP: Journal of applied psychology,
3: JPSP: Journal of personality and social psychology,
4: MUBR: Multivariate behavioral research,
5: JCLP: Journal of consulting psychology,
6: JEDP: Journal of educational psychology,
7: PMEK: Psychometrika,
8: HURE: Human relations,
9: BULL: Psychological bulletin,
10: Hude: Human development.
In addition, table III.3.1. contains a final column that identifies each psychologist with respect to the department he or she is affiliated. The codes are:
S: social psychology (4),
D: educational and developmental psychology (7),
C: clinical psychology (4)
M: mathematical psychology and psychological statistics (3),
E: experimental psychology (10),
R: cultural psychology and psychology of religion (3),
T: industrial psychology (6),
A: physiological and animal psychology (2),
Numbers between parentheses in the list above give the number of psychologists from the department.
Table III.3.1. gives preferences in the usual way, from 1 (most preferred journal) to 10 (least preferred journal).

III.3.2.

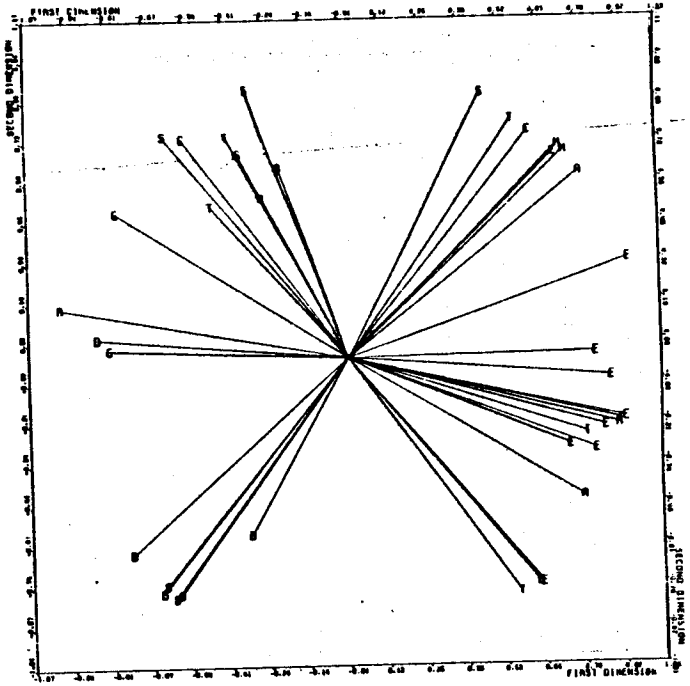
A matrix of ranking such as given in table III.3.1. can be analyzed in two ways.
(a) columns (journals) as variables, rows (psychologists) as objects,
(b) rows (psychologists) as variables, columns (journals) as objects.

Figure III.3.1: ordinal solution, object scores (journals)



"developmental" journals (6,10) to the NE, and the more "social" and "clinical" journals in the SE. Note that in figure III.3.1. the "hard" psychologists are on the right, whereas in figure III.3.2 the "hard" journals are on the left - the reason is that a high score reflects low preference, a low score reflects high preference.
The solution given here is somewhat different from that given in Gifi (1981a, p.191-192). This is probably due to

Figure III.3.2: ordinal solution, component loadings (psychologists) labeled by department



the fact that the solution in Gifi was not computed on the basis of the standard PRINCALS program, and differences in initial configuration may have had their effect on the final solutions. Nevertheless, the sum of the eigenvalues in both solutions is the same. It is quite probable with data of this sort, that there are different quantifications with about the same total fit; the choice of initial configuration then becomes decisive with respect to which of such equivalent solutions will turn up.

The PRINCALS algorithm

(a) Start with X_0, where $u'X_0 = 0$ and $X_0'X_0 = I$. (set $j=1$)
and z_0, where $u'D_j z_0 = 0$ and $z_0'D_j z_0 = 1$.

step 1 : $Y_j \leftarrow D_j G_j' X$

step 2 : if variable j is multiple go to step 10

step 3 : $a_j \leftarrow Y_j' D_j z_j$

step 4 : if variable j is single numerical nominal go to step 9

step 5 : $z_j \leftarrow Y_j a_j$

step 6 : if variable j is single nominal go to step 8

step 7 : $z_j \leftarrow Proj(z_j)$

step 8 : $z_j \leftarrow z_j / (z_j' z_j)^{1/2}$

step 9 : $Y_j \leftarrow z_j a_j'$

step 10 : if $j = m$ go to step 12

step 11 : $j \leftarrow j+1$, go to step 1

step 12 : $X = GRAM\left[\sum_{j=1}^{m} G_j Y_j\right]$

step 13 : $j=1$, go to step 1.

PRINCALS is more general than comparable programs in some respects, but less general in others. All transformations are discrete, ties remain tied. But multiple and single can be mixed.
We do not have favourable experience with the continuous nominal / ordinal / numerical options. They are (a) slow and (b) unstable. In order to remedy this we use fuzzy coding again, this time in the form of splines.

Definition, example —
Cars
Public spending
Cylinders

Program SPLINALS

= Yesterday we have discussed correspondence analysis as a special case of homogeneity analysis [with $m=2$], and we have shown that Benzécri presents a geometrical derivation of the technique which makes multiple correspondence analysis a special case of ordinary correspondence analysis. We have also shown that CA maximizes correlation and linearizes regression.

= Homogeneity analysis gives multiple quantifications and thus multiple induced correlation matrices. This makes it somewhat dissimilar from PCA (except when $p=1$). We make it more similar by introducing rank-one restrictions, in fact we make it identical by combining this with linear cone restrictions. The combination of homogeneity analysis, rank restrictions, and cone restrictions defines NLPCA, program PRINCALS.

Of course this is not the usual way in which PCA is introduced. Other derivations

- the inner product approximation [MDS]
- the linear approximation with unknown approximands [Reed Rank Regr]
- the French derivation [$\delta \ll d$]
- the construction of typical variables (index numbers).

The one we have chosen is PCA to MCA, and thus to the centroid principle. We do not like "inner products as a generic "model". Some of the other approaches are perhaps covered by our "dual" approach via the Burt-table.

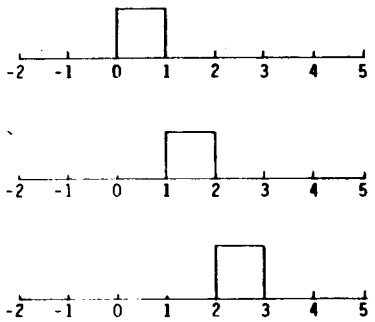


Figure 11.1.a
zero degree B-splines

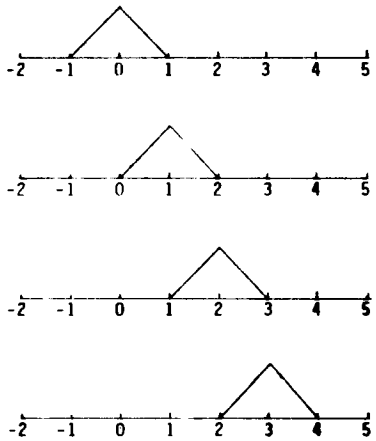


Figure 11.1.b
first degree B-splines

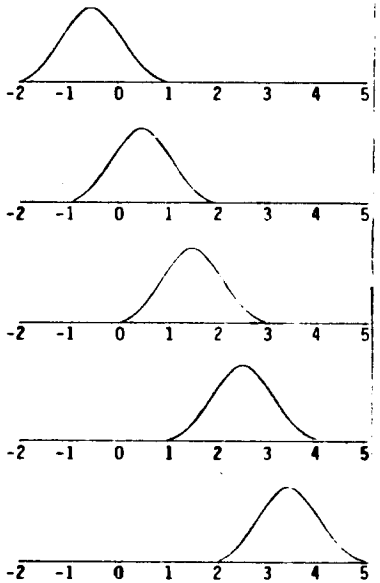


Figure 11.1.c
second degree B-splines

0.50	1	0	0	0.50	0.50	0.00	0.00	0.12500	0.75000	0.12500	0.00000	0.00000
0.75	1	0	0	0.25	0.75	0.00	0.00	0.03125	0.68750	0.28125	0.00000	0.00000
0.90	1	0	0	0.10	0.90	0.00	0.00	0.00500	0.59000	0.40500	0.00000	0.00000
2.30	0	0	1	0.00	0.00	0.70	0.30	0.00000	0.00000	0.24500	0.71000	0.04500

table 11.1: B-splines of degree zero, one, two for four values of x.

var 1	$= a_i$
var 2	$= b_i$
var 3	$= 2(\pi b_i)^{\frac{1}{2}}$
var 4	$= 2a_i(\pi b_i)^{\frac{1}{2}}$
var 5	$= a_i b_i$
var 6	$= (2\pi)^{-1} a_i b_i^2$
var 7	$= (2\pi)^{-\frac{1}{2}} a_i b_i^{-\frac{1}{2}}$
var 8	$= a_i b_i^{-1}$
var 9	$= a_i^{-1} b_i$
var 10	$= 2a_i b_i^{-2}$

Table 1. Thurstone's cylinder data

The first two variables are a sample of size 20 from two independent uniform random variables on the unit interval. The meaning of these variables (in terms of cylinders) is not relevant in our case. We simply use the fact that a centered log transformation of all variables gives a matrix of exact rank two. The aim of the techniques discussed here is to recover these transformations. The idea is that non linear principal components analysis will fit this matrix perfectly in two dimensions, and that correspondence analysis will not accomplish this result because the second eigenvalue is not the second root of the correlation matrix $R(\phi)$. The fit in table 2 is defined as the sum of the first two eigenvalues of $\frac{1}{m}R(\phi)$. The maximum fit is thus equal to 1.

Technique used	Type of transformations approximating ϕ	Total fit for p = 2	eigenvalues	
metric PCA	single linear	.87	.60	.27
ordinal PCA	single monotone	.94	.66	.28
splines-PCA k=1	single LS splines k=1	.89	.61	.28
splines-PCA k=2	single LS splines k=2	.98	.68	.30
splines-PCA k=3	single LS splines k=3	.99	.69	.29
Correspondence-analysis+splines	multiple splines k=2	.77	.77	.66

Table 2. Fit, eigenvalues and transformation-types for several techniques.

The linear fit is surprisingly high compared with the ordinal fit, which on the other hand is clearly inferior to the single

TABLE 1

Data on 1980 Automobiles

	Plot Symbol	Price in $	Engine Size	Miles/Gal.		Weight
				City	Highway	
Audi 4000	1	7700	98	19.3	35.8	2190
Chev Chevette	2	5100	98	18.6	33.9	2120
Datsun 210	3	4750	86	25.2	40.7	2020
Datsun 510	4	5950	119	21.9	42.7	2430
Dodge Colt	5	4900	98	24.2	41.2	1880
Ford Mustang	6	5800	141	15.6	29.8	2610
Honda Accord	7	6500	110	19.9	35.7	2290
Honda Civic CVCC	8	4100	91	22.7	39.7	1780
Mazda GLC	9	4150	86	26.9	47.2	1960
Plymouth Horizon	10	5400	105	21.6	38.3	2150
Plymouth Sapparo	11	6500	159	13.8	26.6	2790
Pontiac Sunbird	12	4950	151	14.6	28.8	2700
Toyota Cellica	13	6650	134	14.5	26.6	2410
Toyota Corona	14	5700	134	16.6	34.9	2570
Toyota Tercel	15	4350	92	23.3	39.5	1970
VW Rabbit	16	6100	97	19.7	45.7	1870
AMC Concord	17	6850	152	15.6	29.2	2910
AMC Eagle	18	8200	258	10.2	22.8	3360
Chev Citation	19	7200	173	14.4	31.6	2660
Chev Malibu	20	7200	232	13.9	22.8	3180
Dodge Aspen	21	6700	225	10.4	23.8	3330
Dodge Diplomat	22	7550	318	10.4	22.2	3540
Ford Fairmont	23	6350	140	13.1	29.6	2600
Mercury Monarch	24	7150	302	11.5	20.9	3510
Olds Cutlass	25	7550	231	11.9	25.5	3440
Pontiac LeMans	26	7400	231	11.9	25.5	3320
Pontiac Phoenix	27	7350	151	17.8	37.9	2550
Buick Regal	28	7400	231	9.8	22.4	3380
Chev Impala	29	8150	231	12.1	26.1	3560
Chev Monte Carlo	30	7500	305	10.3	22.2	3360
Dodge St. Regis	31	8100	318	11.2	21.5	3730
Mercury Marquis	32	8550	302	11.2	20.6	3720
Pontiac Catalina	33	8100	31	10.8	22.8	3610

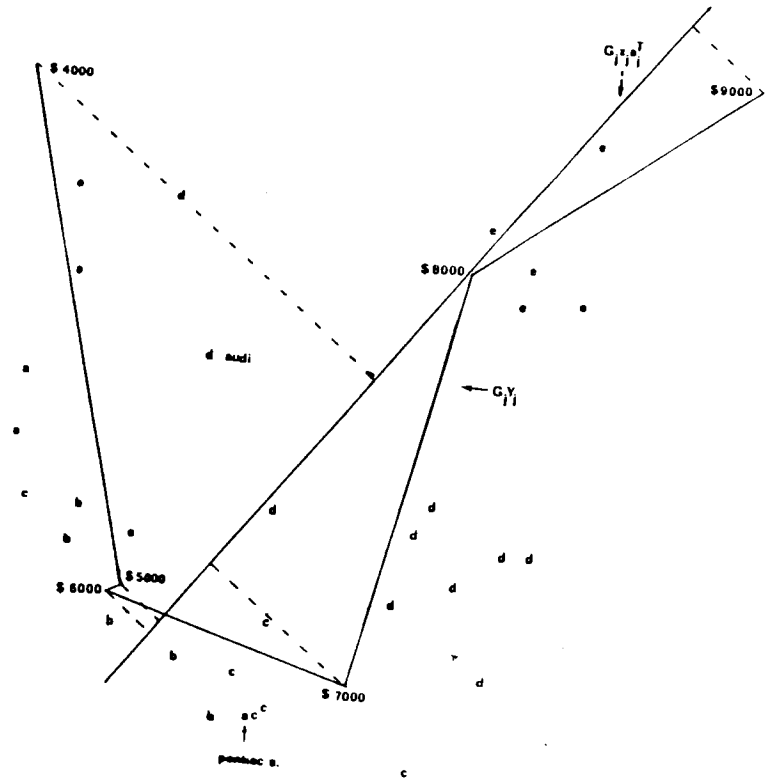


Figure 4 Multiple and single spline functions for price in the object scores plot, cars labeled for price level (a = between ... and $ 5000, etc.). Single hat analysis.

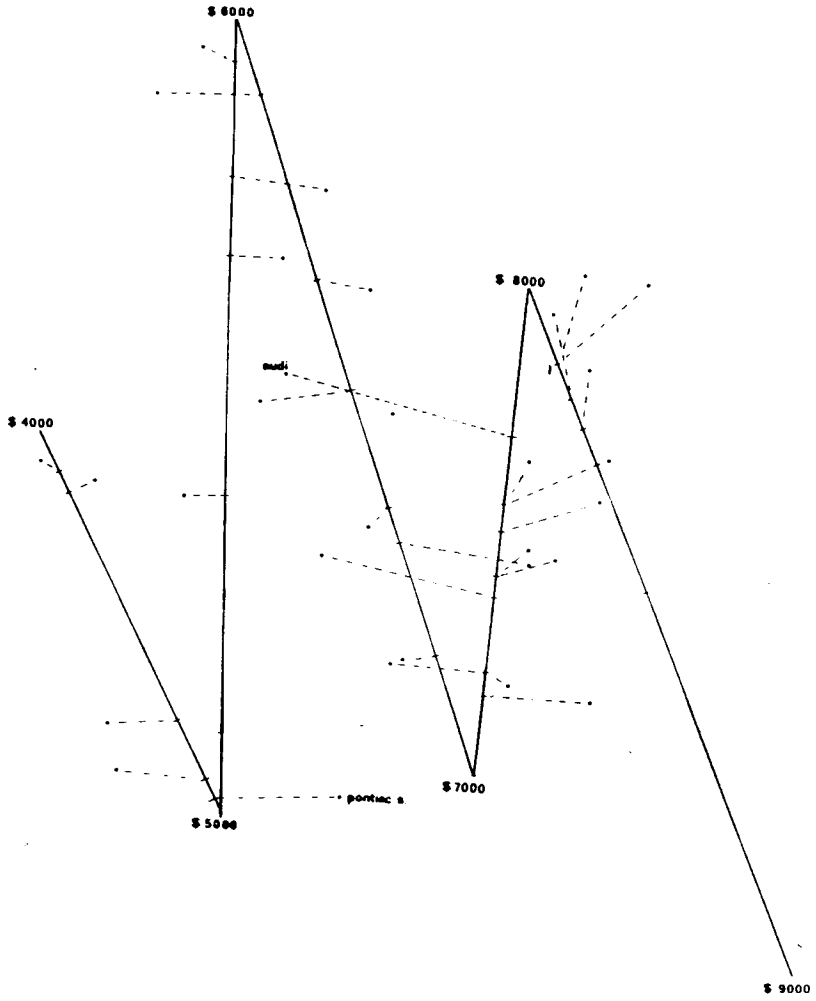
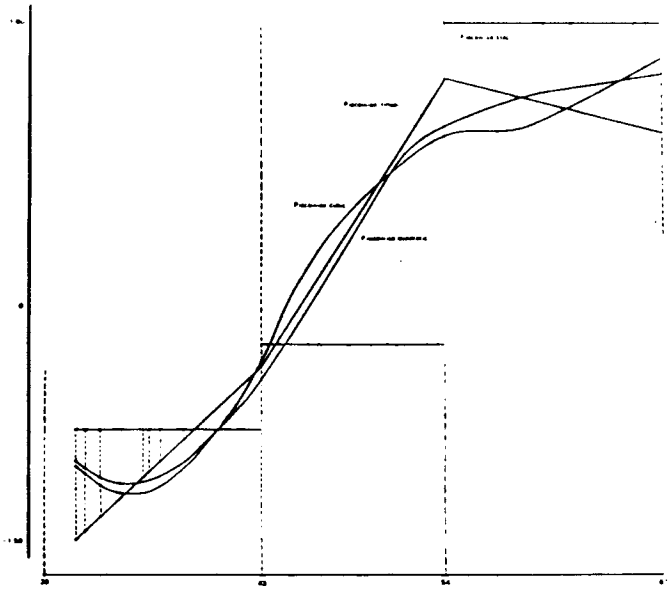


Figure 3 Multiple hat transformation of price with remaining variables linear (omitted) in the object scores plot. The object scores and the corresponding spline values are connected with dotted lines.

components accounted for 96% of the sum of all eigenvalues. From NCA with three components and with less restricted transformations we may expect the equal amount or more.

Figure 2: Transformations of public spending for varying degrees of smoothness.



RESULTS: THE INTERPRETATION

The results are reported concisely. An extensive, conceptually more detailed interpretation with further analysis is VAN KOOTEN & VAN RIJCKEVORSEL (1985).

The eigenvalues are well separated: the first, .68, being three times as large as the second, .22. The third is half as important as the se-

Scaling criteria

We have seen that homogeneity analysis with $p=1$ can be formulated as

$$\max_{Z} \quad \lambda_+\{R(Z)\}$$

and NLPCA can be formulated as

$$\max_{Z_j \in K_j} \quad \sum_{s=1}^{p} \lambda_s \{R(Z)\}.$$

Are there other "PCA-like" criteria that could be optimized.

- criteria should be symmetric in the variables.
- concentrating variance on the largest eigenvalues is good. [shift from $p \geq 2$ to nonlinear transformation]

Suppose

$$\phi(Z) = \Phi\{R(Z)\}$$

is such a criterion. We have a satisfactory algorithm for the case in which

Φ is a convex function of R.

Special cases
- $\sum\sum r_{ij}$ — HORST
- $\sum\sum |r_{ij}|^s$ — KETTENRING ($s=2$)
- $-\det(R)$ — MGV
- $\lambda_1 + \cdots + \lambda_p$ — PRINCALS etc

More satisfactory (in a theoretical sense) if

Φ is a norm on the eigenvalues of R

- $\sum\sum r_{ij}$ not
- $\sum\sum |r_{ij}|^s$ only for $s=2$
- $-\det(R)$
- $\lambda_1 + \cdots + \lambda_p$
- $\sum |\lambda_s|^r$

In the meantime our activities have created two serious problems

- a proliferation of NLPCA techniques
- the fact that homogeneity analysis with $p>1$ [let alone mixed PRINCALS] does not really fit in, and leads to Data Production.

We try to remedy both effects with a theoretical study, which (somewhat unfortunately) leads to two new forms of NLPCA.

COMPSTAT - paper.

We can best study these theoretical results if we remember that CA linearized regressions

Thus, for the cross product of the two variables,

$$C_{12}Y_2 = \rho_{12} D_1 Y_1 \quad \Rightarrow \quad D_1^{-1} C_{12} Y_2 = \rho_{12} Y_1$$

$$C_{21}Y_1 = \rho_{12} D_2 Y_2 \quad \Rightarrow \quad D_2^{-1} C_{21} Y_1 = \rho_{12} Y_2$$

If $m>2$ then homogeneity analysis (or any other technique) in general cannot linearize all $2\binom{m}{2}$ bivariate regressions

Example SMVO

Now let us suppose that the data is such that all regressions can be linearized [a model at last!]. Then there exist $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ and ρ_{jl} such that

$$C_{jl}\mathbf{Y}_l = \rho_{jl} D_j \mathbf{Y}_j \qquad (\forall j,l).$$

Fig. 7.5: Regressies PRE en KEUS (.69)

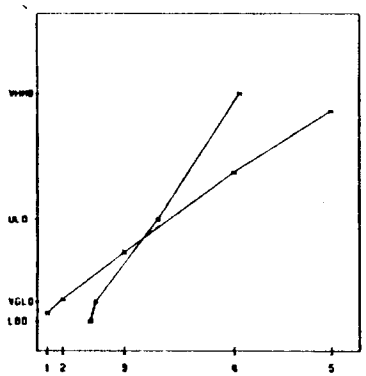


Fig. 7.6: Regressies KEUS en BVA (.43)

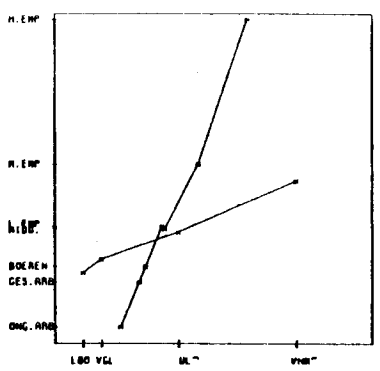


Fig. 7.7: Regressies EIN en BVA (.39)

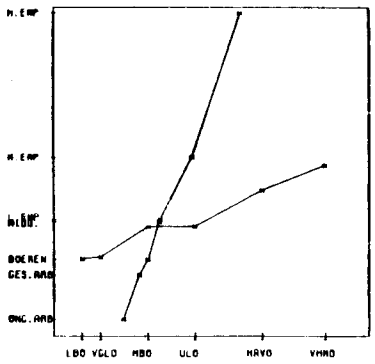
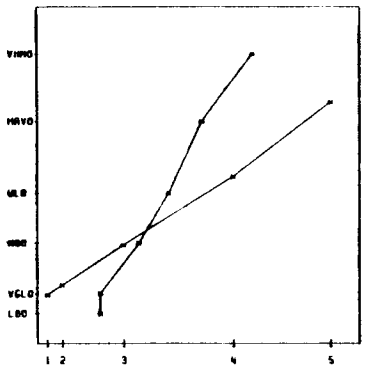


Fig. 7.8: Regressies PRE en EIN (.63)



Theorem: If such quantifications exist, homogeneity analysis will find them.

Proof:

$$\sum_\ell C_{j\ell} x_\ell = \sum_\ell x_\ell \rho_{j\ell} D_j Y_j =$$

$$= \left(\sum_\ell \alpha_\ell \rho_{j\ell}\right) D_j Y_j$$

Now choose α to be an eigenvector of $R = \{\rho_{j\ell}\}$ with eigenvalue λ. Then

$$\sum_\ell C_{j\ell} (\alpha_\ell Y_\ell) = \lambda D_j (\alpha_j Y_j).$$

i.e. $(\alpha_j Y_j, \alpha_2 Y_2, \ldots, \alpha_m Y_m)$ satisfies the equations of homogeneity analysis. □

Actually the construction above provides us with m solutions, because R has m eigenvectors, of the form $\alpha_j s Y_j$. Collecting them in an $k_j \times m$ matrix Y_j gives $Y_j = Z_j a_j$, i.e. there are single quantifications which all induce the same R.

If there is a second system $z_1, \ldots, z_m$ whose linearizes all regressions then it is (variable-wise) orthogonal to the first system. It gives another set of rank-one quantifications inducing a second correlation matrix R_2. And so on.

If there are k systems whose linearize the regressions, i.e. if there exist $z_1, \ldots, z_m$ and Γ_{je} (diagonal) such that

$$C_{jk} z_k = D_j z_j \Gamma_{je} \quad \longleftarrow \text{SVD}$$

then we only have k induced correlation matrices.

[and no more, or far less, data production].

$$\begin{pmatrix} I & D_1^{-1/2} C_{12} D_2^{-1/2} & D_1^{-1/2} C_{13} D_3^{-1/2} \\ D_2^{-1/2} C_{21} D_1^{-1/2} & I & D_2^{-1/2} C_{23} D_3^{-1/2} \\ D_3^{-1/2} C_{31} D_1^{-1/2} & D_3^{-1/2} C_{32} D_2^{-1/2} & I \end{pmatrix}$$

4 cats

orth

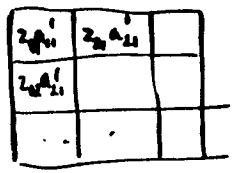
$$\begin{pmatrix} I & \Gamma_{12} & \Gamma_{13} \\ \Gamma_{21} & I & \Gamma_{23} \\ \Gamma_{31} & \Gamma_{32} & I \end{pmatrix}$$

perm →

$\bar{R}_0$ = ave	0	0	0
0	R_1	0	0
0	0	R_2	0
0	0	0	R_3

eval $[C, D]$ = eval $[R_0 + \ldots + R_{k-1}]$

Y : blockwise of rank one.



$\underset{\sim}{\text{ete}}$

Examples

- $m = 2$
- $k = 2$
- multinormal

Practical relevance

COMPSTAT paper [PREHOM]

- - [LINEALS]

Lineals min $\sum_j \sum_\ell [\zeta^2_{j\ell} - r^2_{j\ell}]$

Prehom take $D^{-1/2} C D^{-1/2}$

{ minimize SSQ of "off-diagonal" elements by a generalization of the Jacobi-method

Table 1. Separate classification by seven pathologists of most involved histological lesion of the uterine cervix

slide No.	pathologist 1	2	3	4	5	6	7
1	4	3	4	2	3	3	3
2	1	1	1	1	1	1	1
3	3	3	3	3	3	3	3
4	4	3	3	4	3	3	3
5	3	3	3	3	3	3	3
6	2	1	2	1	1	1	1
7	1	1	1	1	2	1	1
8	3	3	2	3	2	2	3
9	2	2	2	2	3	1	2
10	1	1	1	1	2	1	1
11	5	5	5	4	5	5	5
12	1	1	3	1	2	1	1
13	3	3	3	2	3	3	3
15	2	2	2	1	1	1	2
16	4	3	3	2	3	2	3
17	3	3	2	3	3	3	3
18	2	3	2	2	3	2	3
19	2	1	2	1	2	1	1
22	2	3	2	2	2	1	3
23	1	1	2	1	1	1	1
24	4	3	3	4	3	3	3
25	1	1	2	1	2	1	1
26	1	1	1	1	1	1	1
27	2	1	2	2	2	1	2
28	4	4	4	2	4	3	3
29	3	3	3	2	3	2	3
30	3	3	3	3	3	2	3
31	1	1	1	1	1	1	1
32	4	3	3	3	3	2	3
33	3	3	3	3	3	3	3
34	1	1	1	1	1	1	1
35	3	3	3	2	3	1	2
36	2	2	2	2	3	1	3
37	3	3	2	2	3	1	3
38	5	3	3	3	4	1	3
39	2	1	1	1	2	1	1
40	3	3	2	2	3	1	3
41	3	3	3	3	3	2	3
42	5	5	5	5	5	5	5
43	5	3	3	2	3	2	3
44	3	2	2	2	2	1	2
45	1	1	1	1	2	1	1
46	2	3	1	2	3	1	3
47	4	4	4	3	3	3	3
48	3	3	3	2	3	2	3
49	3	2	2	2	2	1	1
51	2	3	2	2	2	2	2
52	3	3	3	4	3	2	3
53	4	3	3	3	3	5	3
54	3	3	3	2	4	2	3
55	3	3	3	3	3	2	3
56	2	2	2	1	2	2	2
57	2	3	2	1	2	1	1
58	1	1	1	1	3	1	1
59	3	3	3	3	3	3	3
60	1	1	2	1	1	1	1
61	1	3	2	1	2	1	1
62	4	3	3	3	3	2	3
63	1	3	2	2	2	1	2

slide No	pathologist 1	2	3	4	5	6	7
64	2	3	2	2	3	2	3
65	4	3	3	3	3	3	3
66	3	3	3	4	3	2	4
67	1	1	1	1	1	1	1
68	2	3	2	2	3	2	2
69	3	3	2	3	3	1	3
70	1	1	1	1	1	1	1
71	4	3	3	3	3	3	3
72	3	3	3	2	3	1	3
73	3	3	3	3	3	2	3
74	4	3	1	3	3	2	3
76	1	2	1	1	1	1	1
77	2	3	1	2	2	1	2
78	2	3	2	1	3	2	2
79	2	1	1	2	1	1	1
80	4	4	3	2	4	1	3
81	1	1	1	1	1	1	1
82	4	4	3	3	4	3	3
83	5	5	1	4	5	5	4
84	2	3	2	2	2	1	2
85	4	4	4	2	5	1	3
86	3	3	2	3	3	3	3
87	4	3	3	1	3	2	3
88	4	2	3	2	4	1	3
89	2	3	2	2	4	2	3
90	3	3	3	2	4	2	3
91	3	3	2	1	3	2	2
92	4	4	3	2	4	1	3
93	3	3	2	2	3	2	2
94	1	1	2	1	2	1	1
95	3	3	3	2	4	3	3
96	4	3	1	4	4	3	3
98	4	3	2	1	2	1	2
99	1	2	2	1	2	1	2
100	3	3	3	2	4	2	3
101	4	4	3	4	4	3	4
102	3	3	2	2	3	3	3
103	1	1	1	2	1	1	1
104	2	3	2	2	4	1	2
105	3	3	3	3	3	2	3
106	2	3	1	1	3	1	1
107	3	3	2	2	3	2	3
108	3	3	2	2	3	1	3
110	2	2	1	1	2	1	1
111	1	1	1	1	2	1	1
112	3	3	2	2	2	2	3
113	3	3	2	2	2	1	2
114	2	3	1	1	2	1	1
115	3	3	2	2	3	2	3
116	1	1	1	1	2	1	1
117	3	3	3	2	3	2	3
118	3	3	2	2	3	1	3
119	1	1	1	1	2	1	1
120	1	1	1	1	1	1	1
121	2	2	1	1	2	1	2
122	5	3	4	2	3	4	3
123	4	3	4	2	4	1	3
124	1	1	1	1	2	1	1
126	2	3	1	1	2	1	2

PREHOM R	PREHOM COMP		HOM.	
1	1	5.50	1	5.56
2	1	5.24	2	5.19
3	1	2.50	3	2.75
4	1	2.21	4	2.10
4	2	1.53	5	1.68
3	2	1.15	6	1.42
4	3	1.06	7	1.17
3	3	0.86	8	0.89
1	6	0.16	25	0.14
2	6	0.09	26	0.09
1	7	0.09	27	0.05
2	7	0.04	28	0.00

Figure 5.1: Category quantifications of the first dimension

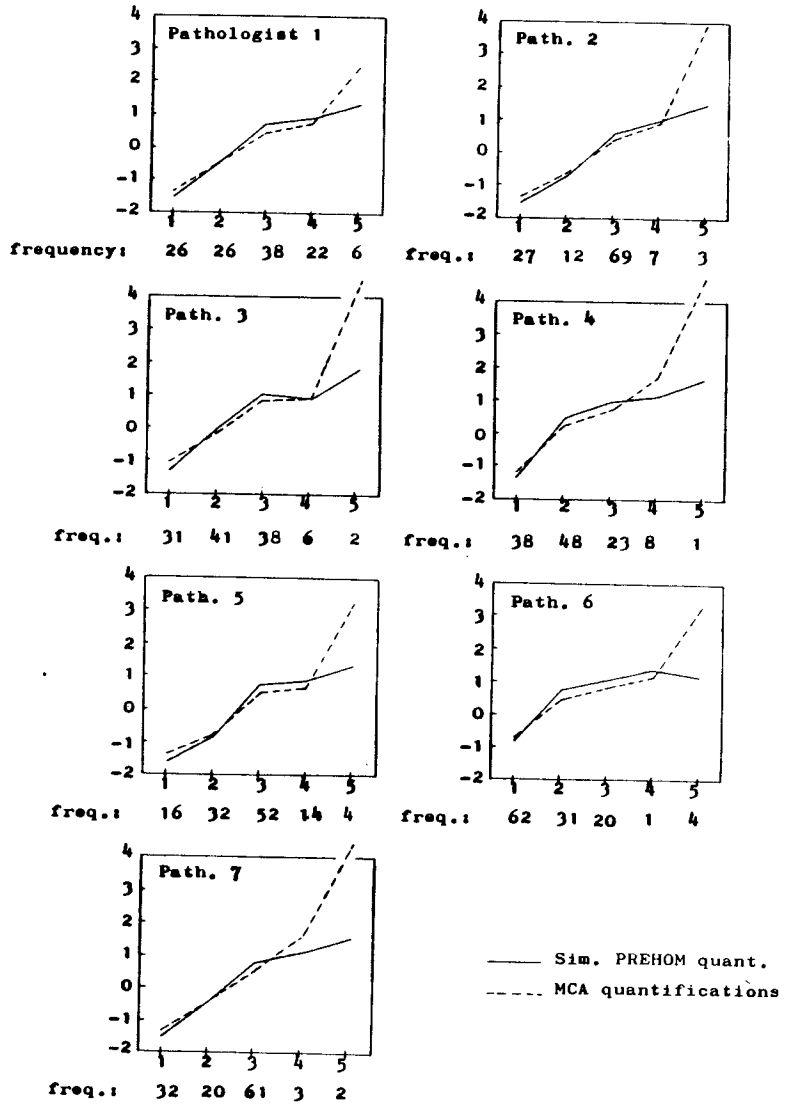
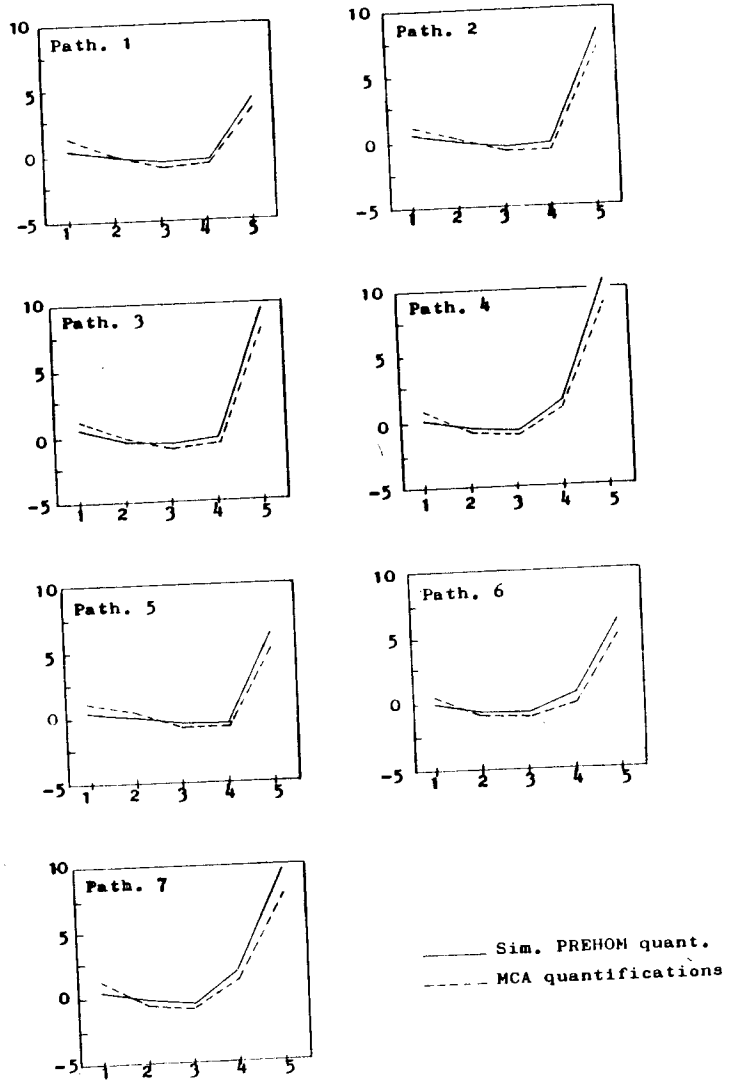


Figure 5.2: Category quantifications of the second dimension



Example list no 9

T2

SOC	1.00							
LBO	0.59	1.00						
WRK	0.26	0.10	1.00					
BIB	-0.15	-0.15	0.24	1.00				
TYP	-0.28	-0.15	0.22	0.54	1.00			
ONG	0.25	0.15	0.82	0.26	0.29	1.00		
JGD	0.30	0.17	0.82	0.20	0.15	0.65	1.00	
ABO	0.47	0.29	0.19	-0.10	-0.17	0.14	0.18	1.00

TABEL 3: Korrelaties ruwe data gediskretiseerd

SOC	0.50	-0.69
LBO	0.35	-0.62
WRK	0.90	0.17
BIB	0.30	0.64
TYP	0.25	0.72
ONG	0.86	0.20
JGD	0.86	0.08
ABO	0.37	-0.54

TABEL 4: Komponent-ladingen uit tabel 3

SOC	1.00	0.62	0.34	0.15	0.32	0.33	0.37	0.51
LBO	0.60	1.00	0.14	0.16	0.17	0.19	0.21	0.33
WRK	0.29	0.16	1.00	0.24	0.23	0.83	0.82	0.24
BIB	0.16	0.20	0.28	1.00	0.55	0.28	0.25	0.14
TYP	0.32	0.22	0.25	0.54	1.00	0.34	0.18	0.21
ONG	0.30	0.21	0.83	0.26	0.29	1.00	0.66	0.22
JGD	0.32	0.21	0.82	0.20	0.16	0.67	1.00	0.30
ABO	0.48	0.30	0.24	0.10	0.18	0.19	0.22	1.00

TABEL 5: Korrelatie-ratios ruwe data gediskretiseerd

SOC	1.00							
LBO	0.62	1.00						
WRK	0.32	0.17	1.00					
BIB	-0.14	-0.07	0.26	1.00				
TYP	-0.15	-0.03	0.20	0.53	1.00			
ONG	0.33	0.23	0.78	0.26	0.29	1.00		
JGD	0.36	0.24	0.81	0.22	0.17	0.68	1.00	
ABO	0.44	0.32	0.26	-0.07	-0.04	0.24	0.32	1.00

TABEL 6: Korrelaties na PRIMALS

SOC	0.60	-0.61
LBO	0.48	-0.53
WRK	0.87	0.20
BIB	0.29	0.71
TYP	0.27	0.69
ONG	0.84	0.21
JGD	0.86	0.11
ABO	0.50	-0.42

TABEL 7: Komponent-ladingen na PRIMALS

SOC	1.00	0.65	0.34	0.14	0.25	0.35	0.37	0.47
LBO	0.65	1.00	0.18	0.07	0.11	0.25	0.27	0.34
WRK	0.33	0.17	1.00	0.26	0.20	0.78	0.82	0.27
BIB	0.16	0.20	0.28	1.00	0.55	0.28	0.25	0.14
TYP	0.24	0.21	0.23	0.52	1.00	0.33	0.18	0.19
ONG	0.33	0.24	0.79	0.26	0.30	1.00	0.68	0.25
JGD	0.37	0.25	0.82	0.22	0.17	0.68	1.00	0.32
ABO	0.48	0.33	0.29	0.07	0.10	0.25	0.32	1.00

TABEL 8: Korrelatie-ratios na PRIMALS

SOC	1.00							
LBO	0.65	1.00						
WRK	0.32	0.15	1.00					
BIB	-0.15	-0.12	0.26	1.00				
TYP	-0.22	-0.12	0.21	0.55	1.00			
ONG	0.32	0.21	0.76	0.26	0.29	1.00		
JGD	0.36	0.23	0.81	0.23	0.17	0.68	1.00	
ABO	0.49	0.34	0.26	-0.09	-0.10	0.23	0.30	1.00

Tabel 9: Korrelaties na LINEALS

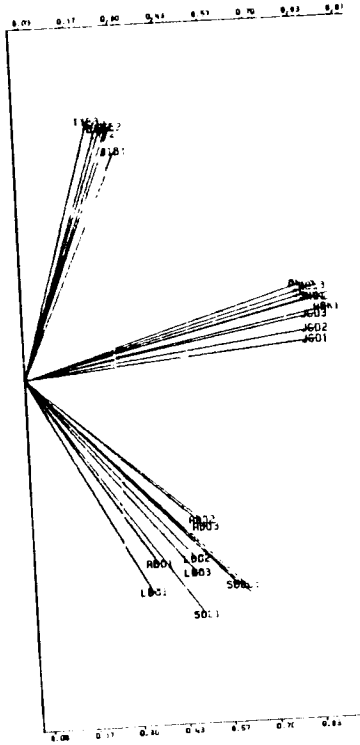
SOC	0.62	-0.61
LBO	0.48	-0.57
WRK	0.86	0.23
BIB	0.26	0.71
TYP	0.22	0.73
ONG	0.83	0.24
JGD	0.86	0.15
ABO	0.51	-0.44

Tabel 10: Komponent-ladingen na LINEALS

SOC	1.00	0.66	0.34	0.15	0.28	0.35	0.37	0.49
LBO	0.66	1.00	0.17	0.12	0.15	0.23	0.26	0.35
WRK	0.33	0.17	1.00	0.26	0.21	0.77	0.82	0.27
BIB	0.16	0.20	0.28	1.00	0.55	0.28	0.25	0.14
TYP	0.30	0.22	0.24	0.55	1.00	0.34	0.19	0.20
ONG	0.33	0.25	0.78	0.26	0.30	1.00	0.68	0.25
JGD	0.37	0.25	0.81	0.23	0.17	0.68	1.00	0.32
ABO	0.50	0.35	0.28	0.09	0.12	0.24	0.30	1.00

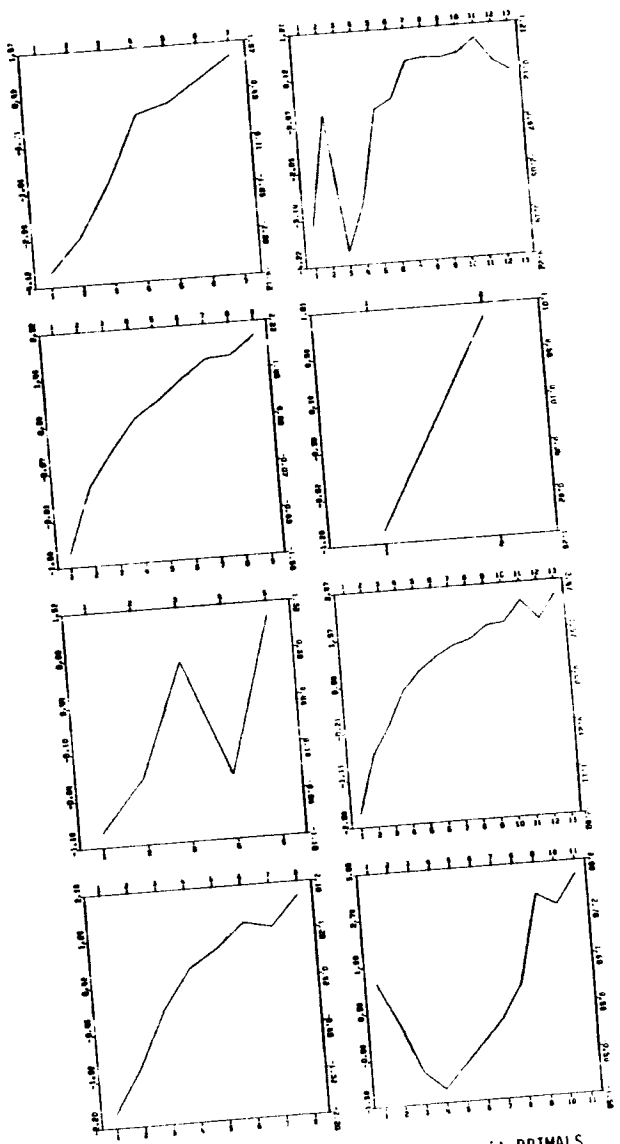
Tabel 11: Korrelatie-ratios na LINEALS

F3



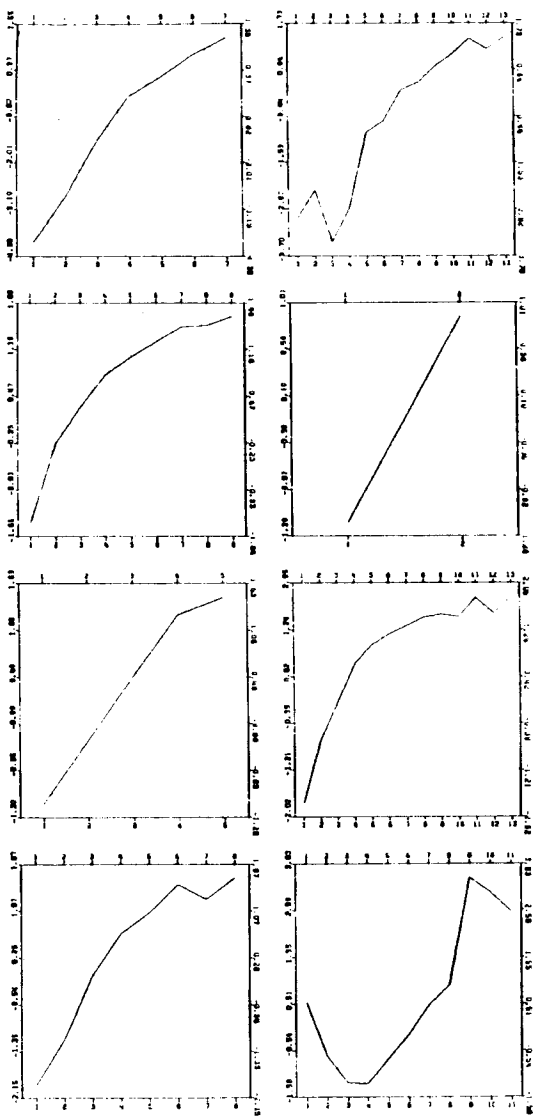
Figuur 3: Komponent-ladingen
1: ruwe data, gediskretiseerd
2: na PRIMALS
3: na LINEALS

F1



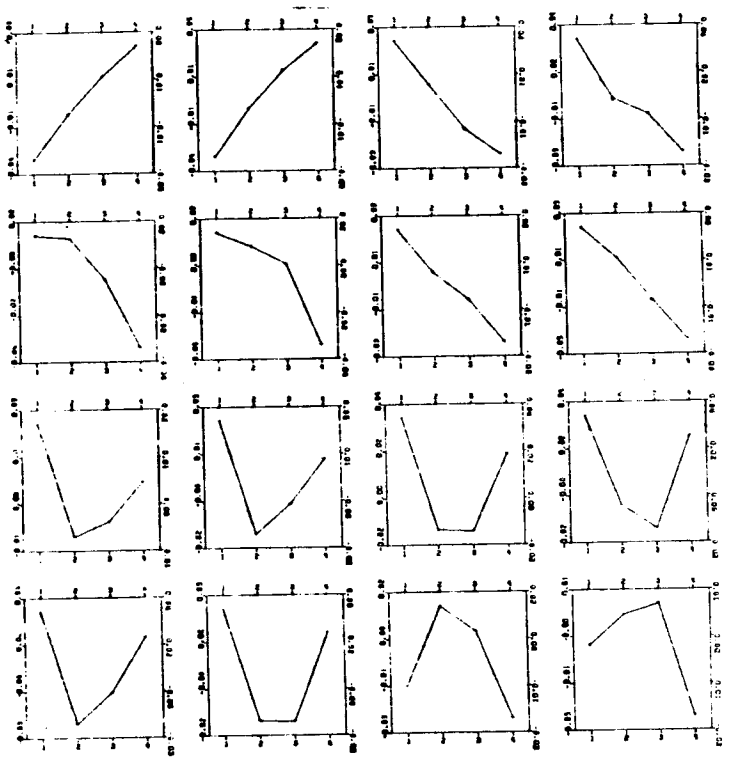
Figuur 1: Transformaties/kwantifikaties uit PRIMALS

F2



Figuur 2: Transformaties/kwantifikaties uit LINEALS

MULTINORMAL GAUGE



four eigenvectors (rows), four variables (columns).

We now introduce another set of restrictions into the system [still using the same loss-function !]. They are **additivity restrictions** and they are used to introduce sets of variables. In order to see this in proper perspective, we have to emphasize that our choice of multivariable is indeed a choice (a form of coding).

- 3 vars 3 cats
- 2 vars (3×3) and 3 cats [in 3 ways]
- 1 var 27 cats

and, of course, we can also group and combine cats.

The next transparency shows that a set of variables can be interpreted as a variable with additivity restrictions on the quantificators.

a	p u
b	q v
a	r v
a	p u
b	p v
c	p v
a	p u
a	p v
c	p v
a	p v

table 6.2.a
data matrix

a	b	c	p	q	r	u	v
1	0	0	1	0	0	1	0
0	1	0	0	1	0	0	1
1	0	0	0	0	1	0	1
1	0	0	1	0	0	1	0
0	1	0	1	0	0	0	1
0	0	1	1	0	0	0	1
1	0	0	1	0	0	1	0
1	0	0	1	0	0	0	1
0	0	1	1	0	0	0	1
1	0	0	1	0	0	0	1

table 6.2.b
indicator supermatrix G_+

apu	apv	aqu	aqv	aru	arv	bpu	bpv	bqu	bqv	bru	brv	cpu	cpv	cqu	cqv	cru	crv
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

table 6.2.c: interactive indicator matrix. G_x

	a	b	c	p	q	r	u	v
apu	1	0	0	1	0	0	1	0
apv	1	0	0	1	0	0	0	1
aqu	1	0	0	0	1	0	1	0
aqv	1	0	0	0	1	0	0	1
aru	1	0	0	0	0	1	1	0
arv	1	0	0	0	0	1	0	1
bpu	0	1	0	1	0	0	1	0
bpv	0	1	0	1	0	0	0	1
bqu	0	1	0	0	1	0	1	0
bqv	0	1	0	0	1	0	0	1
bru	0	1	0	0	0	1	1	0
brv	0	1	0	0	0	1	0	1
cpu	0	0	1	1	0	0	1	0
cpv	0	0	1	1	0	0	0	1
cqu	0	0	1	0	1	0	1	0
cqv	0	0	1	0	1	0	0	1
cru	0	0	1	0	0	1	1	0
crv	0	0	1	0	0	1	0	1

table 6.2.d:
transformation or
design matrix. T

$$G_x T = G_+$$

$$G_x [TY] = G_+ Y$$

$$G_x U = G_+ Y$$

$$U = TY$$

Of course additivity restrictions can be combined with some restrictions and rank restrictions. This combination defines the 3WAYALS algorithm.

Start with

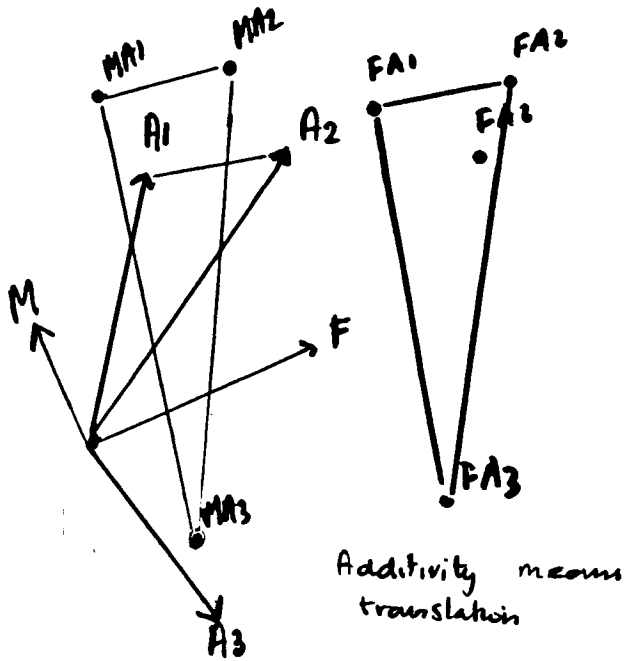
$$f(X;Y) = \frac{1}{m}\sum_{j=1}^{m} \text{tr}\,[X - G_j Y_j]'[X - G_j Y_j]$$

Now suppose there are additivity restrictions on all Y_j. We can write

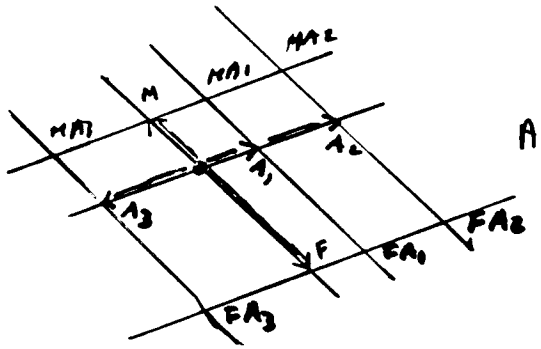
$$f(X;Y) = \frac{1}{m}\sum_{j=1}^{m} \text{tr}\,[X - \sum_{s} G_{js} Y_{js}]'[X - \sum_{s} G_{js} Y_{js}]$$

and then of course we can have $Y_{js} = \Xi_{js} a_{js}'$ for all or some (j,s).

Geometry



Additivity means translation



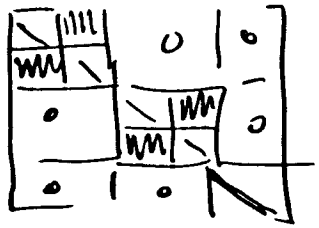
Additivity + rank means grid.

Analytically

Using the same methods as before

to $Y'CY$ max $\qquad Y'DY = mI$

where now

$D =$ 

In the case of single quantifications we can again try to formulate the problem in terms of correlations, and a property of them that is optimized. But what if some variables are multiple?

To integrate multiple variables we need the notion of copies. Suppose G_j is the indicator of a variable, and we use this same G_j more than once in the set, in all case treating it as single. It will occur in the loss function in the form

$$G_j z_{j1} a'_{j1} + \cdots + G_j z_{j\mu} a'_{j\mu} =$$
$$= G_j Y_j \quad \text{with} \quad Y_j = z_j A'_j.$$

But if $\mu \geq \min [p, k_j - 1]$ then any Y_j can be represented in the form $Y_j = z_j A'_j$. Thus multiple nominal $\equiv$ p copies of a single variable in the set.

Actually the notion of copies is more fruitful still.

If we have $q < p$ copies, we have rank-q restrictions (between single and multiple). If we impose ordinality restrictions on the z_j of the copies we have multiple ordinal. We can also make the first copy linear and the second quadratic, etc.

It also means that if we have found a multiple Y_j [from some homogeneity analysis] we could look at decompositions $Y_j = Z A_j'$ which translate it to a set of single variables.

- use orth. pols. in Z_j
- use the SVD
- use $Z_j = D_j^{-1}$

$\Big\}$ $Z_j' D_j Z_j = I$

We can now make a familiar list

	number of sets	number of vars in set
PCA	m	1
MR	2	$m_1 > 1$, $m_2 = 1$
DISCR	2	$m_1 > 1$, $m_2 = 1$ [mult]
CANCOR	2	$m_1 > 1$, $m_2 > 1$
MANOVA	2	$m_1 > 1$, $m_2 > 1$ [Des]
ANOVA	2	[Des] $m_1 > 1$, $m_2 = 1$
MANCOVA	3	

Clearly we can add many "new" special cases to this list.

$\Big\{$ All these techniques (as techniques) as special cases of homogeneity analysis. Of course their statistical properties depend on the additional assumptions one wants to make.

Selected Illustrations

It will become obvious now, why there was so much emphasis on MCA/CA/PCA in this course.

MORALS (example list no 27) } $m_1 > 1$
Wilson, Science, 64, 1926, 47-57. } $m_2 = 1$

[illegible] (example list no 29)

CANALS (example list no 30, in Gifi).
(example in paper —) ($n = 7$)
(more of this)
(political example) ($n = 31$)

REDUNDANCY ANALYSES

ISRAELS, PM 1984, 331-346

	r_{xy}	r_{xz}	r_{yz}	b_x	b_y	R^2
linear	-0.3804	0.8156	0.1534	1.0219	0.5421	0.9166
rational	-0.4008	0.8286	0.1663	1.0666	0.5938	0.9826
step 1	-0.2302	0.7248	0.2189	0.8186	0.4073	0.6825
step 2	-0.3847	0.8475	0.1406	1.0582	0.5477	0.9739
spline 1	-0.3891	0.8394	0.1663	1.0654	0.5809	0.9910
spline 2	-0.3884	0.8413	0.1655	1.0664	0.5797	0.9931

Table 11.2 Regression statistics for ten different analyses of Gibbs data.

a	temperature	16	39	10								
	pressure	34	17	13	1							
	density	28	21	16								
b	temperature	6	10	7	8	9	8	7	2	8		
	pressure	9	12	13	8	7	2	5	4	4	1	
	density	9	12	7	9	6	6	7	4	5		
c	temperature	0.2	16.2	35.0	12.8	0.8						
	pressure	2.9	29.4	19.2	11.0	2.6						
	density	3.4	24.7	21.1	14.0	1.8						
d	temperature	6.0	10.0	7.0	8.0	9.0	8.0	7.0	2.0	8.0		
	pressure	0.2	8.3	12.2	12.6	9.3	5.6	2.5	4.6	4.6	3.1	1.1	0.8
	density	1.5	7.2	11.0	8.4	8.4	6.3	6.3	6.6	4.6	4.0	0.7	

Table 11.3 Marginals for.

a: step 1: step-functions crude.
b: step 2: step-functions fine.
c: spline 1: splines crude.
d: spline 2: splines fine.

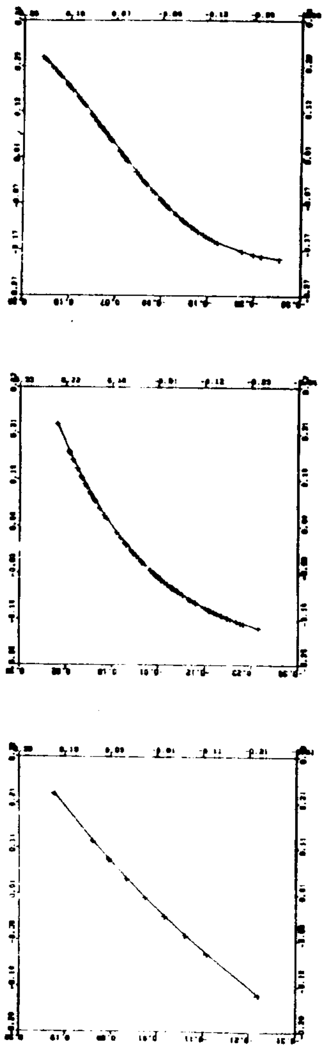


Figure 11.2 Gibbs Wilson example rational transformations.

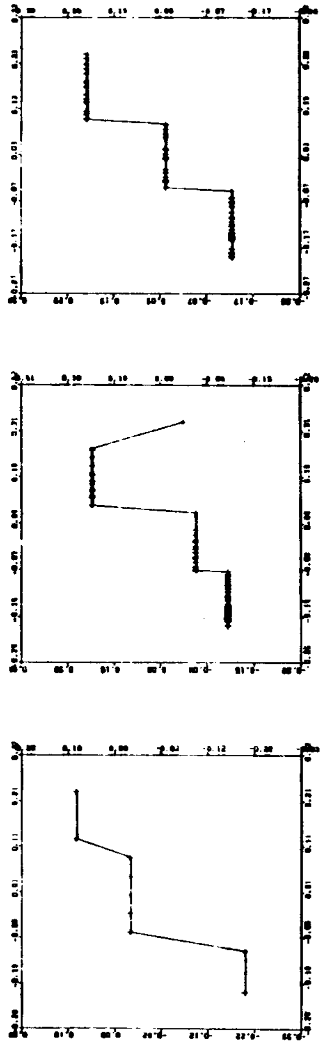


Figure [illegible] example [illegible]

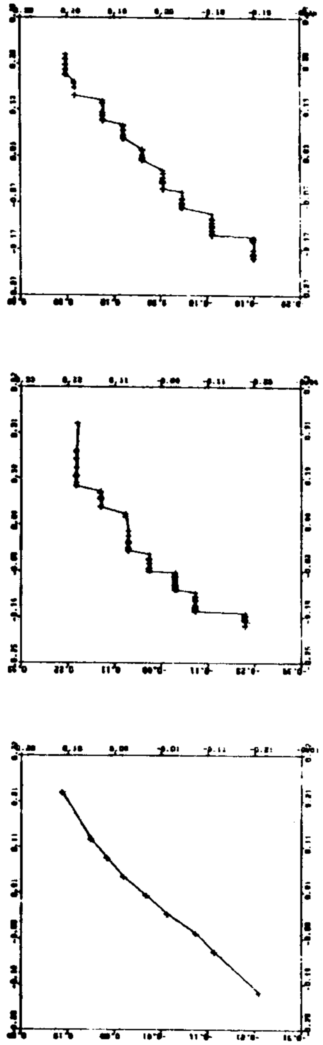


Figure 11.4 Gibbs-Wilson example fine step functions.

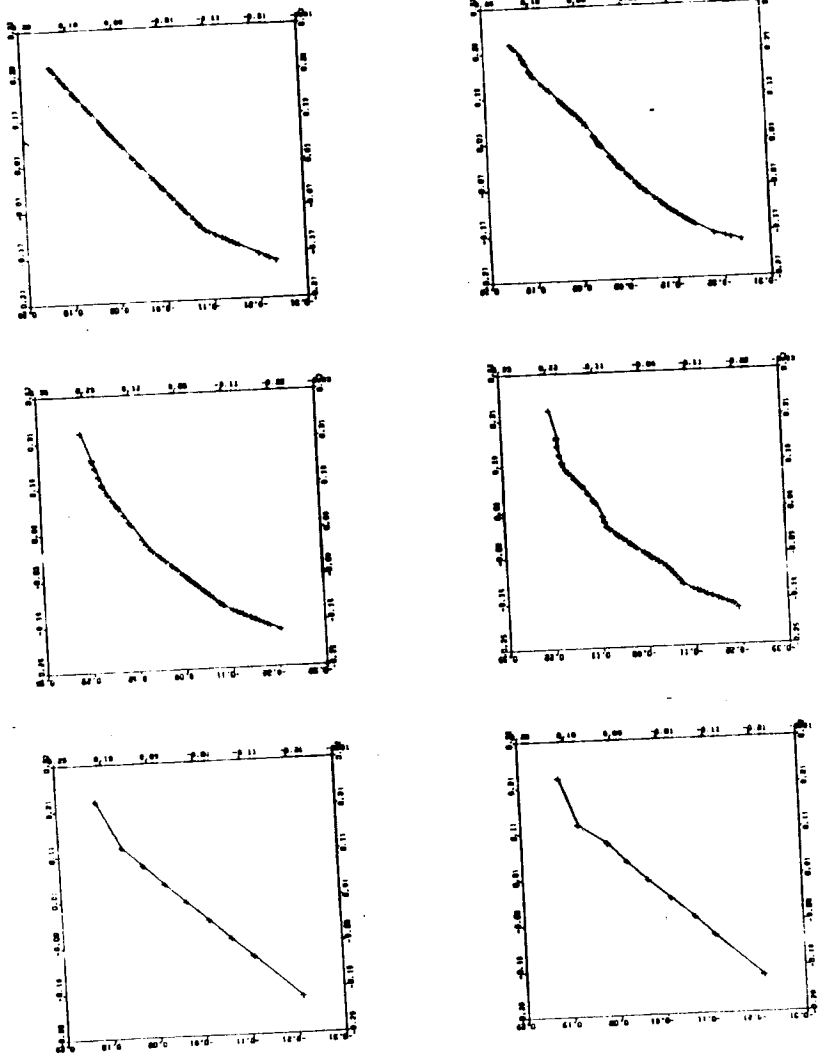
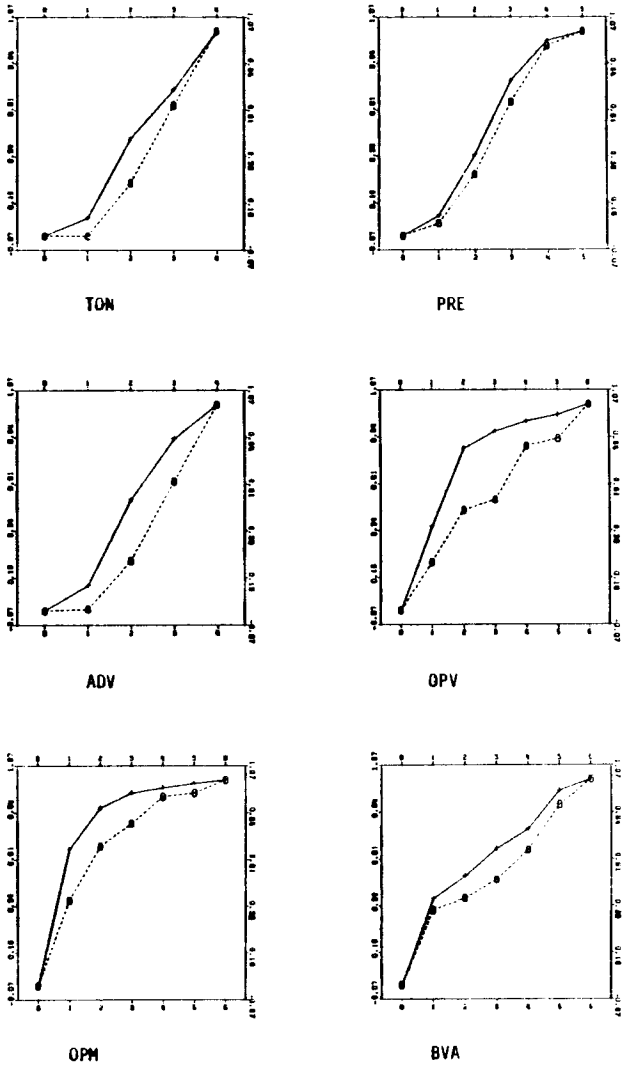
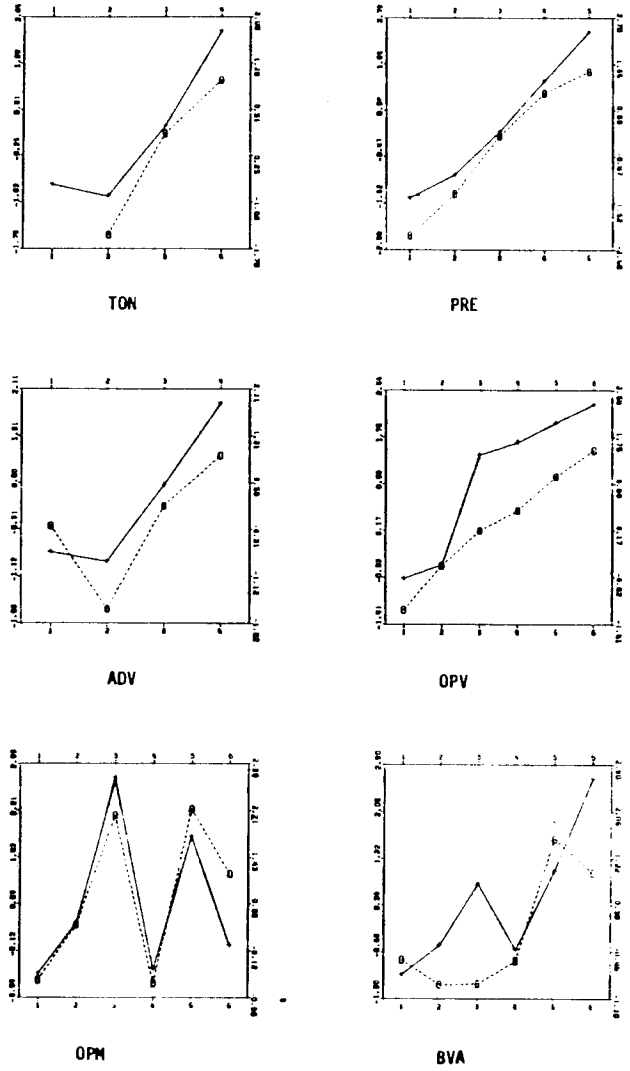


Figure II.5 Gibbs-Wilson example
cruder splines

Figure II.6 Gibbs-Wilson example
fine splines



figuur 1: kumulatieve verdelingen VJTJ(+) en SMVO(O).



figuur 2: CANALS-transformaties VJTJ(+) en SMVO(O).

	TON	PRE	ADV	OPV	OPM	BVA
TON	1.00					
PRE	.63	1.00				
ADV	.74	.68	1.00			
OPV	.37	.30	.34	1.00		
OPM	.24	.19	.24	.45	1.00	
BVA	.31	.28	.29	.60	.35	1.00

tabel 3a: korrelaties VJTJ uit De Jong e.a.

	TON	PRE	ADV	OPV	OPM	BVA
TON	1.00					
PRE	.63	1.00				
ADV	.80	.67	1.00			
OPV	.37	.27	.36	1.00		
OPM	.36	.21	.31	.48	1.00	
BVA	.33	.26	.33	.59	.40	1.00

tabel 3b: korrelaties SMVO uit De Jong e.a.

	VJTJ	SMVO
PRE	.22	.17
ADV	.55	.64
OPV	.09	.04
OPM	.01	.10
BVA	.03	.01
R^2	.59	.67

tabel 4: beta-gewichten en multipele korrelaties uit tabel 3.

142

	TON	PRE	ADV	OPV	OPM	BVA
TON	1.00					
PRE	.68	1.00				
ADV	.80	.69	1.00			
OPV	.44	.30	.36	1.00		
OPM	.27	.18	.22	.34	1.00	
BVA	.42	.31	.32	.56	.27	1.00

tabel 5a: korrelaties VJTJ uit CANALS.

	TON	PRE	ADV	OPV	OPM	BVA
TON	1.00					
PRE	.61	1.00				
ADV	.85	.59	1.00			
OPV	.38	.27	.35	1.00		
OPM	.23	.14	.22	.26	1.00	
BVA	.28	.20	.25	.47	.21	1.00

tabel 5b: korrelaties SMVO uit CANALS.

	VJTJ	SMVO
PRE	.21	.15
ADV	.57	.72
OPV	.10	.06
OPM	.05	.03
BVA	.10	.04
R^2	.70	.74

tabel 6: beta-gewichten en multipele korrelaties uit CANALS.

143

		VJTJ	BS1	BS2	BS3	BS4	BS5
TON:	VGLO:	-0.74	-0.79	-0.76	-0.75	-0.67	-0.88
	LBO :	-0.93	-0.89	-0.89	-0.83	-0.88	-0.98
	ULO :	0.24	0.20	0.23	0.13	0.12	0.32
	VHMO:	1.79	1.88	1.84	1.92	1.83	1.64
PRE:	1 :	-1.39	-1.32	-1.49	-1.32	-1.40	-1.40
	2 :	-0.89	-0.86	-0.89	-0.82	-0.90	-0.86
	3 :	0.09	-0.01	0.21	0.06	0.13	0.04
	4 :	1.27	1.42	1.31	1.30	1.14	1.24
	5 :	2.34	2.26	2.02	2.43	2.58	2.34
ADV:	VGLO:	-0.69	-0.72	-0.64	-0.70	-0.62	-0.95
	LBO :	-0.86	-0.80	-0.83	-0.76	-0.79	-0.89
	ULO :	0.45	0.39	0.41	0.25	0.25	0.60
	VHMO:	1.85	1.99	1.94	2.04	1.93	1.60
OPV:	LO :	-0.64	-0.83	-0.49	-0.49	-0.58	-0.71
	LBO :	-0.38	-0.09	-0.33	-0.55	-0.40	-0.19
	ULO :	1.23	1.51	1.12	1.61	1.56	0.23
	MBO :	1.72	0.31	0.65	1.15	2.20	2.60
	VHMO:	2.20	2.20	0.25	2.61	3.02	2.27
	HBO :	2.42	2.68	3.93	2.22	1.65	1.98
OPM:	LO :	-0.53	-0.41	-0.37	-0.52	-0.37	-0.47
	LBO :	0.36	0.01	-0.28	0.03	0.61	0.08
	ULO :	2.72	0.53	2.97	2.36	2.93	2.39
	MBO :	0.32	4.11	0.93	2.32	-2.35	-0.21
	VHMO:	1.83	3.34	2.11	2.85	-0.27	3.43
	HBO :	0.09	2.34	-0.09	1.95	-1.21	1.23
BVA:	1 :	-0.90	-0.88	-0.83	-0.88	-0.76	-0.93
	2 :	-0.35	-0.03	0.01	-0.37	-0.48	-0.27
	3 :	0.96	1.08	0.63	1.73	1.20	0.71
	4 :	-0.40	-0.60	-1.04	-0.12	-0.17	-0.40
	5 :	1.00	1.30	1.54	0.83	0.16	1.41
	6 :	2.43	1.54	1.48	1.34	2.94	1.79
β	PRE :	.24	.26	.21	.22	.20	.25
	ADV :	.69	.68	.70	.72	.74	.68
	OPV :	.13	.13	.14	.12	.12	.10
	OPM :	.06	.06	.08	.07	.07	.07
	BVA :	.11	.11	.13	.07	.12	.14
Rmult	:	.84	.84	.85	.86	.86	.85

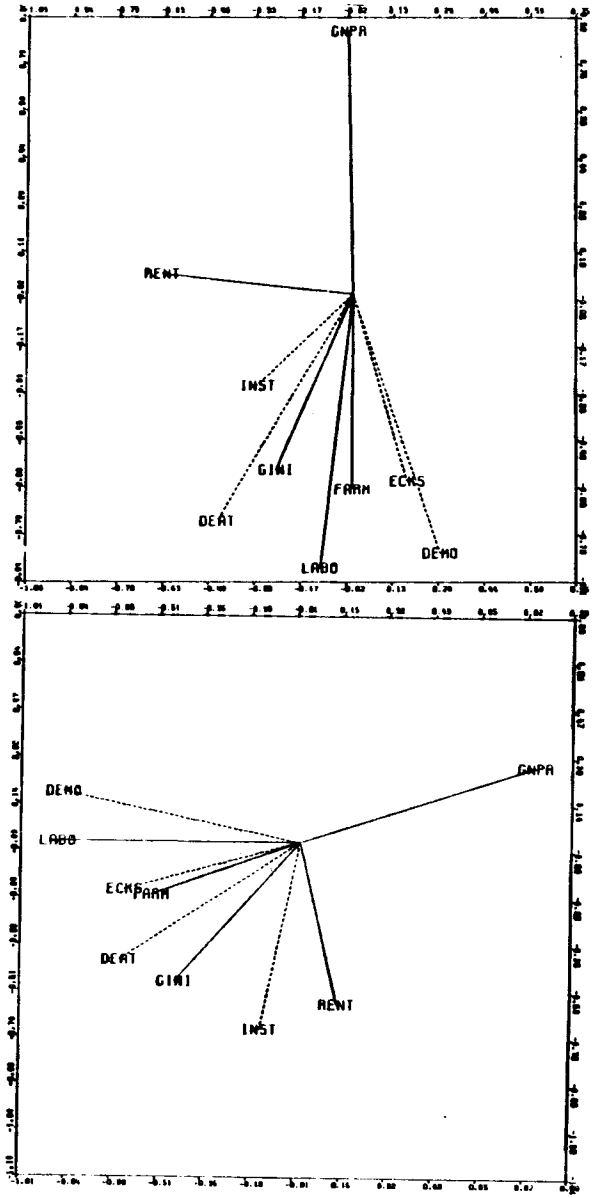
tabel 7: CANALS analyse op VJTJ en op vijf bootstrap-samples uit VJTJ.

	GINI	FARM	RENT	GNPR	LA	INST	ECK	DEAT	DEM
ARGENTINA	86.3	98.2	32.9	374	25	13.6	57	217	2
AUSTRALIA	92.9	99.6	----	1215	14	11.3	0	0	1
BELGIUM	58.7	85.8	62.3	1015	10	15.5	8	1	1
BOLIVIA	93.8	97.7	20.0	66	72	15.3	53	663	3
BRAZIL	83.7	98.5	9.1	262	61	15.5	49	1	2
CANADA	49.7	82.9	7.2	1667	12	11.3	22	0	1
CHILE	93.8	99.7	13.4	180	30	14.2	21	2	2
COLOMBIA	84.9	98.1	12.1	330	55	14.6	47	316	2
COSTA RICA	88.1	99.1	5.4	307	55	14.6	19	24	2
CUBA	79.2	97.8	53.8	361	42	13.6	100	2900	3
DENMARK	45.8	79.3	3.5	913	23	14.6	0	0	1
DOMINICAN REP.	79.5	98.5	20.8	205	56	11.3	6	31	3
ECUADOR	86.4	99.3	14.6	204	53	15.1	41	18	3
EGYPT	74.0	98.1	11.6	133	64	15.8	45	2	3
EL SALVADOR	82.8	98.8	15.1	244	63	15.1	9	2	3
PHILIPPINES	56.4	88.2	37.3	201	59	14.0	15	292	3
FINLAND	59.9	86.3	2.4	941	46	15.6	4	0	2
FRANCE	58.3	86.1	26.0	1046	26	16.3	46	1	2
GUATEMALA	86.0	99.7	17.0	179	68	14.9	45	57	3
GREECE	74.7	99.4	17.7	239	48	15.8	9	2	2
UNITED KINGDOM	71.0	93.4	44.5	998	5	13.6	12	0	1
HONDURAS	75.7	97.4	16.7	137	66	13.6	45	111	3
IRELAND	59.8	85.9	2.5	509	40	14.2	9	0	1
INDIA	52.2	86.9	53.0	72	71	3.0	83	14	1
IRAQ	88.1	99.3	75.0	195	81	16.2	24	344	3
ITALY	80.3	98.0	23.8	442	29	15.5	51	1	2
JAPAN	47.0	81.5	2.9	240	40	15.7	22	1	2
YUGOSLAVIA	43.7	79.8	0.0	297	67	0.0	9	0	3
LUXEMBOURG	63.8	87.7	18.8	1194	23	12.8	8	0	1
LIBYA	70.0	93.0	8.5	90	75	14.8	8	0	3
NETHERLANDS	60.5	86.2	53.3	708	11	13.6	2	0	1
NICARAGUA	75.7	96.4	----	254	68	12.8	16	16	3
NEW ZEALAND	77.3	95.5	22.3	1259	16	12.8	0	0	1
NORWAY	66.9	87.5	7.5	969	26	12.6	1	0	1
AUSTRIA	74.0	97.4	10.7	532	32	12.8	4	0	2
PANAMA	73.7	95.0	12.3	350	54	15.6	29	25	3
PERU	87.5	96.9	----	140	60	14.6	23	26	3
POLAND	45.0	77.7	0.0	468	57	8.5	19	5	3
SPAIN	78.0	99.5	43.7	254	50	0.0	22	1	3
TAIWAN	65.2	94.1	40.0	102	50	0.0	3	0	3
URUGUAY	81.7	96.6	34.7	569	37	14.6	1	1	1
VENEZUELA	90.9	99.3	20.6	762	42	14.9	36	111	3
UNITED STATES	70.5	95.4	20.4	2343	10	12.8	22	0	1
WEST GERMANY	67.4	93.0	5.7	762	14	3.0	4	0	2
SOUTH VIETNAM	67.1	94.6	20.0	133	65	10.0	50	1000	3
SWEDEN	57.7	87.2	18.9	1165	13	8.5	0	0	1
SWITZERLAND	49.8	81.5	18.9	1229	10	8.5	0	0	1

Table 7.2.a Original data

ECON.VAR	POL.VAR
5 5 4 4 2	5 3 4 2
6 5 9 7 1	4 1 1 1
2 3 5 6 1	6 2 2 1
6 5 3 1 4	6 3 5 3
5 5 2 4 4	6 3 2 2
1 2 2 7 1	4 2 1 1
6 5 3 3 2	5 2 2 2
5 5 3 4 3	5 3 4 2
5 5 2 4 3	5 2 3 2
4 5 5 4 3	5 3 5 3
1 1 2 6 2	5 1 1 1
4 5 3 3 3	4 2 3 3
5 5 3 3 3	6 3 2 3
4 5 3 2 4	6 3 2 3
5 5 3 3 4	6 2 2 3
2 3 4 3 5	5 2 4 1
2 3 2 6 3	6 2 1 2
2 3 4 6 2	6 3 2 2
5 5 3 3 4	5 3 3 3
4 5 3 3 3	6 2 2 2
4 4 5 6 1	5 2 1 1
4 5 3 2 4	5 3 4 3
2 3 2 5 2	5 2 1 1
2 3 5 1 4	2 3 2 1
5 5 5 3 5	6 2 4 3
5 5 4 5 2	6 3 2 2
1 2 2 3 2	6 2 2 2
1 1 1 4 4	1 2 1 3
3 3 3 7 2	5 1 1 1
3 4 2 1 4	5 2 3 3
3 3 5 6 1	5 2 1 1
4 5 9 4 4	5 2 2 3
4 5 4 7 1	5 1 1 1
3 3 2 6 2	5 2 1 1
4 5 2 5 2	5 2 1 2
4 4 3 4 3	6 2 3 3
5 5 9 2 3	5 2 3 3
1 1 1 5 3	3 2 2 3
4 5 5 4 3	1 2 2 3
3 4 4 2 3	1 2 1 3
5 5 4 5 2	5 2 2 1
6 5 3 6 3	5 2 6 3
4 5 3 8 1	5 2 1 1
3 4 2 6 1	2 2 1 2
3 4 3 2 4	4 3 5 3
2 3 3 7 1	3 1 1 1
1 2 3 7 1	3 1 1 1

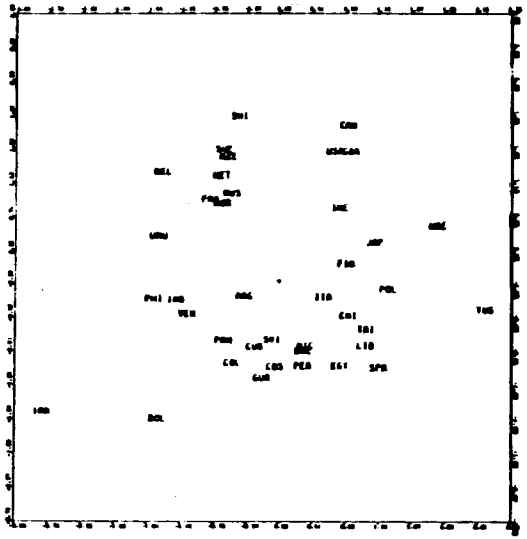
Table 7.2.b Discreticized data



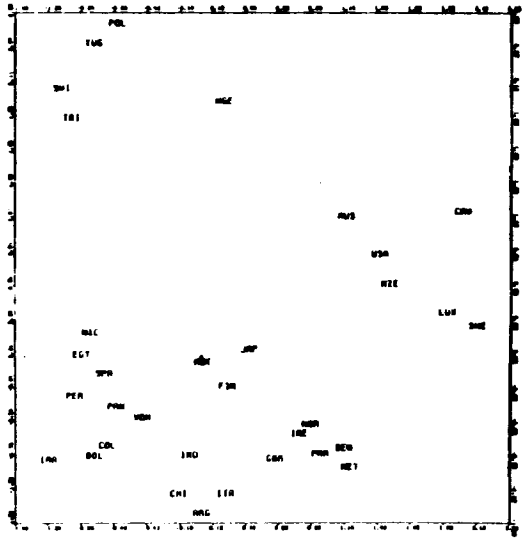
a Rankings

b Discrete

Figure 7.2 Correlations in the canonical space of economic variables

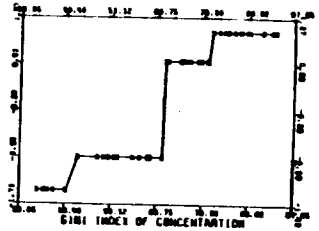


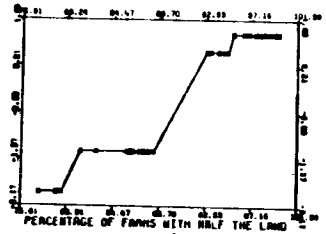
a Rankings

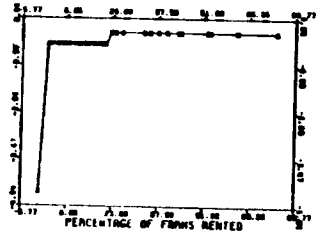


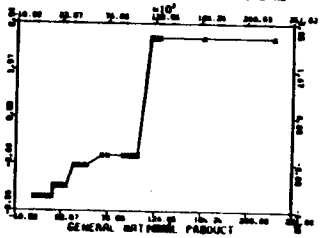
b Discrete

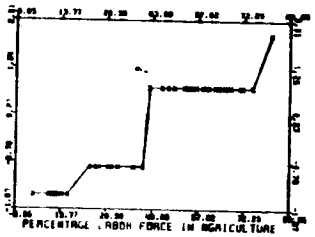
Figure 7.3 Canonical scores in the space of economic variables

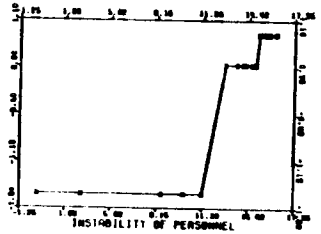


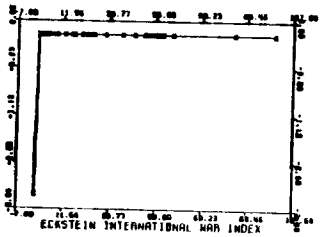


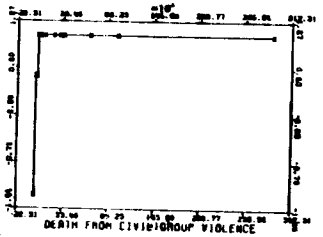


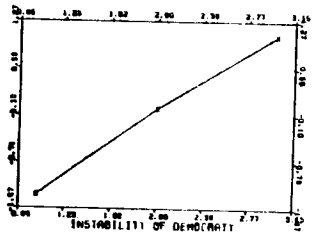












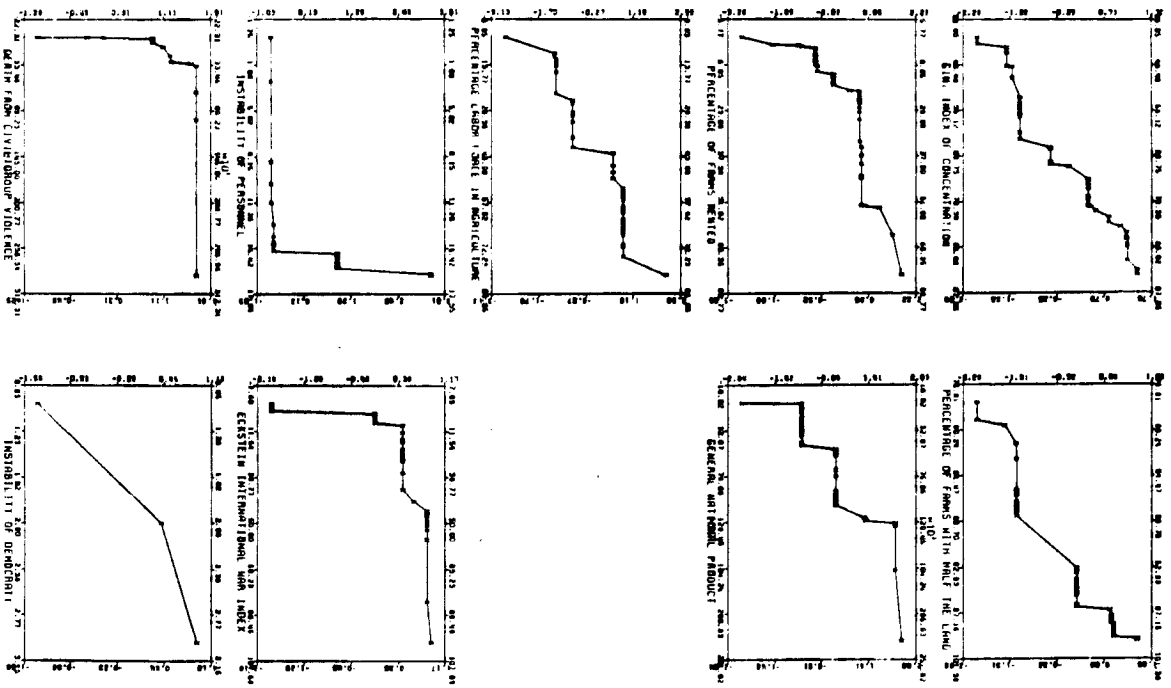


Figure 7.4.a Transformations of the rankings

ORDERING OF MEFV-CURVES RELATIVE TO RESPIRATORY SYMPTOMS

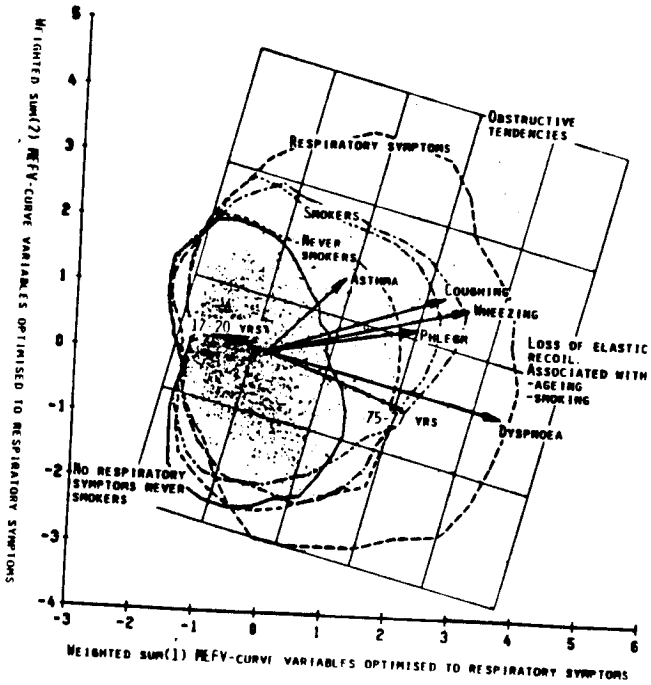


Figure 1.
SCATTERGRAM
- Most parsimonious ordering of MEFV-curves with respect to chronic respiratory symptoms. The ordering is expressed in terms of a specific weighted sums of the MEFV-curve variables (abscissa and ordinate). In this way each MEFV-curve gets its position in the scattergram.

ARROWS
- The position of a curve in the scattergram is related to the probability on chronic respiratory symptoms. These increase in the direction of the [illegible] arrows.
- Curves having a high probability for ASTHMA are [illegible] more upwards and somewhat to the right; [illegible] indicative for DYSPNOEA mainly to the right. [illegible] lungfunctions are found at the bottom left.

CONTOURS
- Subgroups (having identical age distributions) [illegible] indicated by way of percentile contours containing 95% of their values.
- SMOKERS (- - -, - . - . -) occupy an intermediate position between the subgroups NO RESPIRATORY SYMPTOMS NEVER SMOKERS (——) and RESPIRATORY SYMPTOMS (- - -). They contain more curves with a larger probability for one or more respiratory symptoms than NEVER SMOKERS (- . - . -).

KNOTTED LINE (——•——).
- The knotted line for age shows that with increasing age, curves are more probable associated with respiratory symptoms, especially DYSPNOEA. Independent of age, height is hardly related to curve-positions.

INTERPRETATION
- The association with smoking and especially ageing in direction of DYSPNOEA, suggests that loss of elastic recoil may be involved in the change in curve configuration towards the right.
- Relatively independent of the former trend [illegible] direction of the vector ASTHMA suggests that [illegible] more vertical direction obstructive tendencies increase from other causes than loss of elastic recoil alone.

ORDERING OF MEFV-CURVES RELATIVE TO AGE AND HEIGHT: PHASES

150

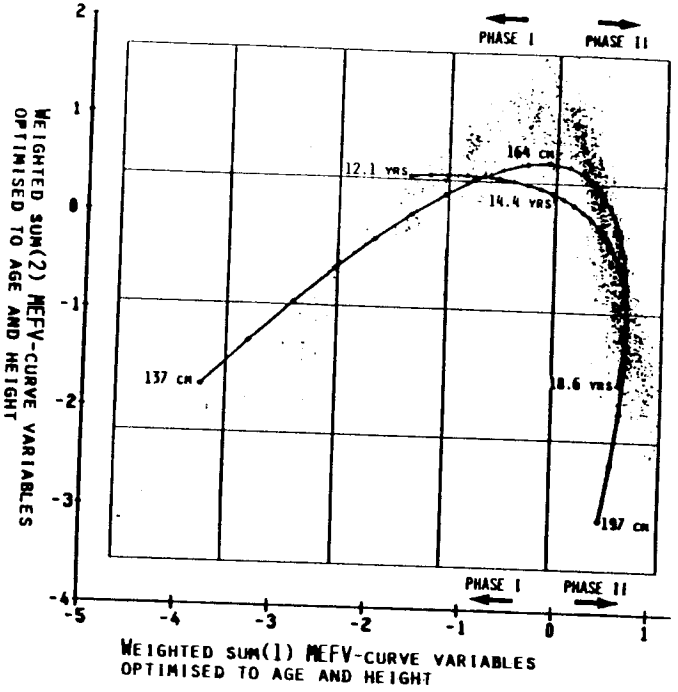


Figure 1
SCATTERGRAM
- Most parsimonious ordering of MEFV-curves with respect to age and height.
  The ordering is expressed in terms of 2 specific weighted sums of the MEFV-curve variables. In this way each MEFV-curve gets its position in the scattergram.

PHASES
- The scattergram has a boomerang-like shape suggesting two phases.
- Phase 1 covers mostly the age range 12-14.5 yrs. Curve variability is mainly related to differences in height, mostly 137-164 cm.
- Phase 2 concerns a larger age range mostly 14.5-18 yrs. Curve variability is also mainly related to differences in height, mostly 165-197cm.

151

of nomination and renomination on internal party processes could give central party organs a strong weapon with which to discipline deviant behaviour. The electorate gives a mandate to a party, not to individual members of the parliament: defiant members cannot bring their case to individual constituents but must satisfy themselves with such hearing as they can get within the party or leave.

5.1. *Description of the data*

In 1972 the members of the Dutch Parliament (MPs) were interviewed. Among other things, the MPs gave their opinions on a number of issues and their preference votes for the political parties. The issues concerned development aid, abortion, law and order, income differences, worker participation, taxation and defence. The opinions were measured on a nine-point scale of which the lowest and the highest category were described (Table 1). The party preferences were recorded in a table of rank

Table 1. The issues and the meanings of the lowest and the highest category

	1 DEVELOPMENT AID	
the government should spend more money on aid to developing countries	(1) ..... (9)	the government should spend less money on aid to developing countries
	2 ABORTION	
the government should prohibit abortion completely	(1) ..... (9)	a woman has the right to decide for herself about abortion
	3 LAW AND ORDER	
the government takes too strong action against public disturbances	(1) ..... (9)	the government should take stronger action against public disturbances
	4 INCOME DIFFERENCES	
income differences should remain as they are	(1) ..... (9)	income differences should become much less
	5 PARTICIPATION	
only management should decide important matters in industry	(1) ..... (9)	workers must also have participation in decisions important for industry
	6 TAXATION	
taxes should be increased for general welfare	(1) ..... (9)	taxes should be decreased so that people can decide for themselves how to spend their money
	7 DEFENCE	
the government should insist on shrinking the Western armies	(1) ..... (9)	the government should insist on maintaining strong Western armies

orders. The scores in this table tell us the rank order each member of the parliament gave to the different parties (2 = highest preference, 15 = lowest preference). The lowest score (2) was always used for the MP's own party. For our illustration we only consider the preferences for the four largest parties, which are:

PvdA – labour party (socialists) (39)
ARP – Anti Revolutionary Party (christian democrats) (13)
KVP – catholic party (christian democrats) (35)
VVD – 'liberal' party, referred to as conservatives (16)

The figure in parentheses is the number of MPs. The other parties in 1972 were

CHU – Christian Historical Union (christian democrats) (10)
D'66 – democrats 66 (liberals) (11)

	GINI	FARM	RENT	GNPR	LA	INST	ECK	DEAT	DEM
ARGENTINA	86.3	98.2	32.9	374	25	13.6	57	217	2
AUSTRALIA	92.9	99.6	-----	1215	14	11.3	0	0	1
BELGIUM	58.7	85.8	62.3	1015	10	15.5	0	1	1
BOLIVIA	93.8	97.7	20.0	66	72	15.3	53	663	3
BRAZIL	83.7	98.5	9.1	262	61	15.5	49	1	2
CANADA	49.7	82.9	7.2	1667	12	11.3	22	0	1
CHILE	93.8	99.7	13.4	180	30	14.2	21	2	2
COLOMBIA	84.9	98.1	12.1	330	55	14.6	47	316	2
COSTA RICA	88.1	99.1	5.4	307	55	14.6	19	24	2
CUBA	79.2	97.8	53.8	361	42	13.6	100	2900	3
DENMARK	45.8	79.3	3.5	913	23	14.6	0	0	1
DOMINICAN REP.	79.5	98.5	20.8	205	56	11.3	6	31	3
ECUADOR	86.4	99.3	14.6	204	53	15.1	41	18	3
EGYPT	74.0	98.1	11.6	133	64	15.8	45	2	3
EL SALVADOR	82.8	98.8	15.1	244	63	15.1	9	2	3
PHILIPPINES	56.4	88.2	37.3	201	59	14.0	15	292	3
FINLAND	59.9	86.3	2.4	941	46	15.6	4	0	2
FRANCE	58.3	86.1	26.0	1046	26	16.3	46	1	2
GUATEMALA	86.0	99.7	17.0	179	68	14.9	45	57	3
GREECE	74.7	99.4	17.7	239	48	15.8	9	2	2
UNITED KINGDOM	71.0	93.4	44.5	998	5	13.6	12	0	1
HONDURAS	75.7	97.4	16.7	137	66	13.6	45	111	3
IRELAND	59.8	85.9	2.5	509	48	14.2	9	0	1
INDIA	52.2	86.9	53.0	72	71	3.0	83	14	1
IRAQ	88.1	99.3	75.0	195	81	16.2	24	344	3
ITALY	80.3	98.0	23.8	442	29	15.5	51	1	2
JAPAN	47.0	81.5	2.9	240	40	15.7	22	1	2
YUGOSLAVIA	43.7	79.8	0.0	297	67	0.0	9	0	3
LUXEMBOURG	63.8	87.7	18.8	1194	23	12.8	8	0	1
LIBYA	70.0	93.0	8.5	90	75	14.8	8	0	3
NETHERLANDS	60.5	86.2	53.3	708	11	13.6	2	0	1
NICARAGUA	75.7	96.4	----	254	68	12.8	16	16	3
NEW ZEALAND	77.3	95.5	22.3	1259	16	12.8	0	0	1
NORWAY	66.9	87.5	7.5	969	26	12.6	1	0	1
AUSTRIA	74.0	97.4	10.7	532	32	12.8	4	0	2
PANAMA	73.7	95.0	12.3	350	54	15.6	29	25	3
PERU	87.5	96.9	----	140	60	14.6	23	26	3
POLAND	45.0	77.7	0.0	468	57	8.5	19	5	3
SPAIN	78.0	99.5	43.7	254	50	0.0	22	1	3
TAIWAN	65.2	94.1	40.0	102	50	0.0	3	0	3
URUGUAY	81.7	96.6	34.7	569	37	14.6	1	1	1
VENEZUELA	90.9	99.3	20.6	762	42	14.9	36	111	3
UNITED STATES	70.5	95.4	20.4	2343	10	12.8	22	0	1
WEST GERMANY	67.4	93.0	5.7	762	14	3.0	4	0	2
SOUTH VIETNAM	67.1	94.6	20.0	133	65	10.0	50	1000	3
SWEDEN	57.7	87.2	18.9	1165	13	8.5	0	0	1
SWITZERLAND	49.8	81.5	18.9	1229	10	8.5	0	0	1

Table 7.2.a Original data

	ECON.VAR					POL.VAR			
ARGENTINA	5	5	4	4	2	5	3	4	2
AUSTRALIA	6	5	9	7	1	4	1	1	1
BELGIUM	2	3	5	6	1	6	2	2	1
BOLIVIA	6	5	1	1	4	6	3	5	3
BRAZIL	5	5	2	4	4	6	3	2	2
CANADA	1	2	2	7	1	4	2	1	1
CHILE	6	5	3	3	2	5	2	2	2
COLOMBIA	5	5	3	4	3	5	3	4	2
COSTA RICA	5	5	2	4	3	5	2	3	2
CUBA	4	5	5	4	3	5	3	5	3
DENMARK	1	1	2	6	2	3	1	1	1
DOMINICAN REP.	4	5	3	3	3	4	2	3	3
ECUADOR	5	5	3	3	3	6	3	2	3
EGYPT	4	5	3	2	4	6	3	2	3
EL SALVADOR	5	5	3	3	4	6	2	2	3
PHILIPPINES	2	3	4	3	5	5	2	4	1
FINLAND	2	3	2	6	3	6	2	1	2
FRANCE	2	3	4	6	2	6	3	2	2
GUATEMALA	5	5	3	3	4	5	3	3	3
GREECE	4	5	3	3	3	6	2	2	2
UNITED KINGDOM	4	4	5	6	1	5	2	1	1
HONDURAS	4	5	3	2	4	5	3	4	3
IRELAND	2	3	2	5	2	5	2	1	1
INDIA	2	3	5	1	4	2	3	2	1
IRAQ	5	5	5	3	5	6	2	4	3
ITALY	5	5	4	5	2	6	3	2	2
JAPAN	1	2	2	3	2	6	2	2	2
YUGOSLAVIA	1	1	1	4	4	1	2	1	3
LUXEMBOURG	3	3	3	7	2	5	1	1	1
LIBYA	3	4	2	1	4	5	2	1	3
NETHERLANDS	3	3	5	6	1	5	2	1	1
NICARAGUA	4	5	9	4	4	5	2	2	3
NEW ZEALAND	4	5	4	7	1	5	1	1	1
NORWAY	3	3	2	6	2	5	2	1	1
AUSTRIA	4	5	2	5	2	5	2	1	2
PANAMA	4	4	3	4	3	6	2	3	3
PERU	5	5	9	2	3	5	2	3	3
POLAND	1	1	1	5	3	3	2	2	3
SPAIN	4	5	5	4	3	1	2	2	3
TAIWAN	3	4	4	2	3	1	2	1	3
URUGUAY	5	5	4	5	2	5	2	2	1
VENEZUELA	6	5	3	6	3	5	2	4	3
UNITED STATES	4	5	3	8	1	5	2	1	1
WEST GERMANY	3	4	2	6	1	2	2	1	2
SOUTH VIETNAM	3	4	3	2	4	4	3	5	3
SWEDEN	2	3	3	7	1	3	1	1	1
SWITZERLAND	1	2	3	7	1	3	1	1	1

Table 7.2.b Discreticized data

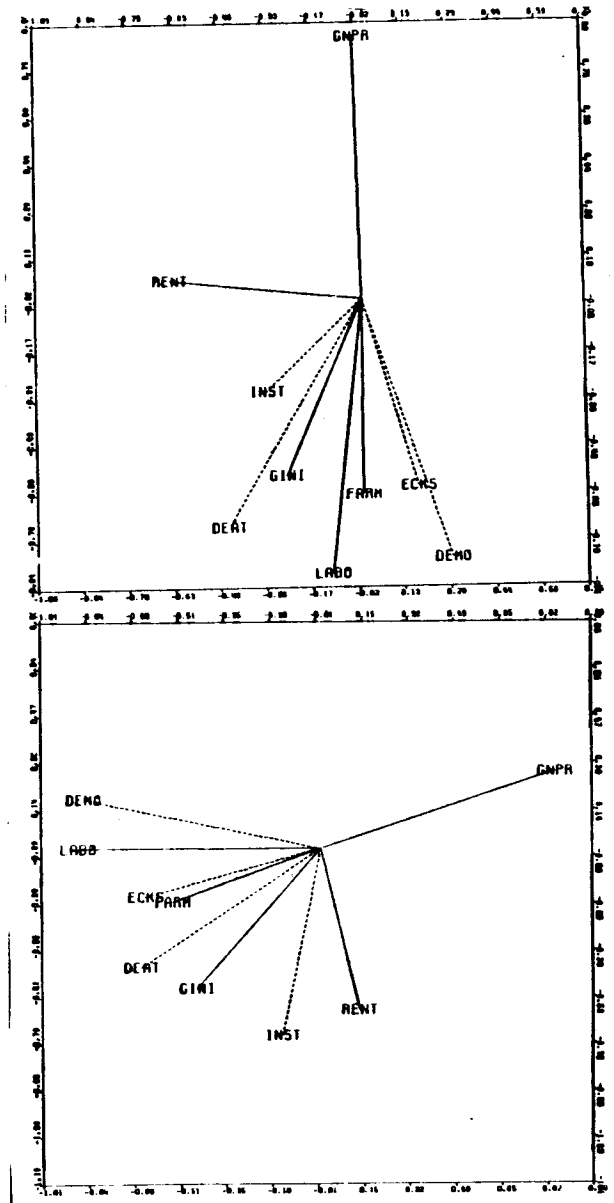
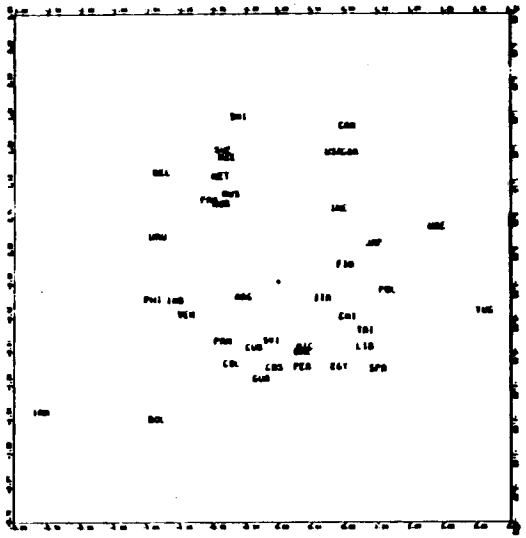
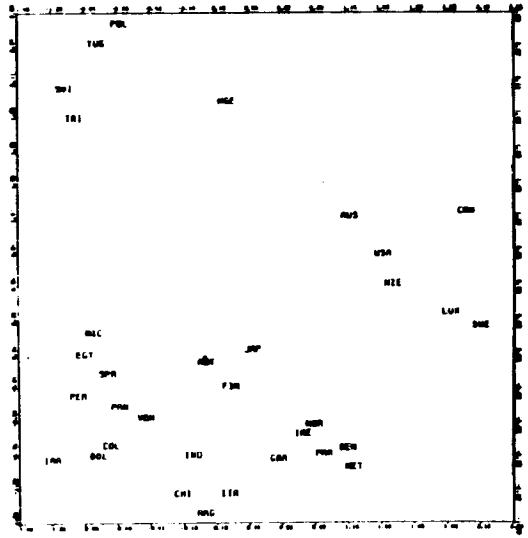


Figure 7.2 Correlations in the canonical space of economic variables

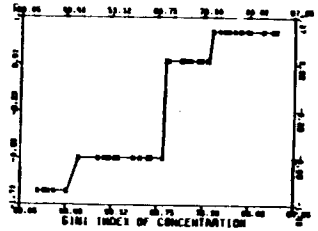


a Rankings

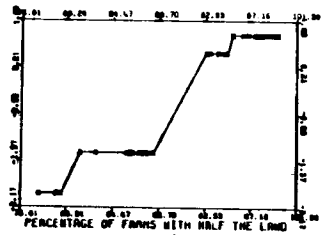


b Discrete

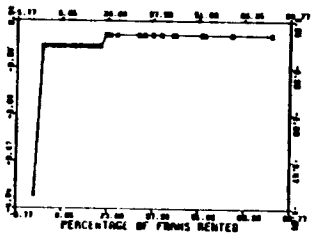
Figure 7.3 Canonical scores in the space of economic variables



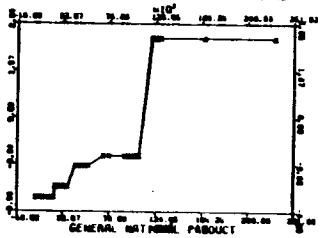
GINI INDEX OF CONCENTRATION



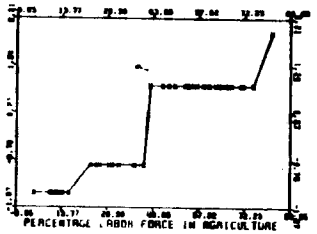
PERCENTAGE OF FARMS WITH HALF THE LAND



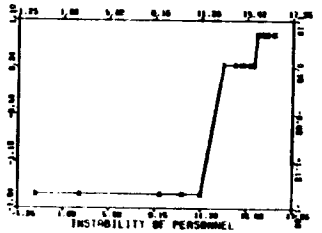
PERCENTAGE OF FARMS RENTED



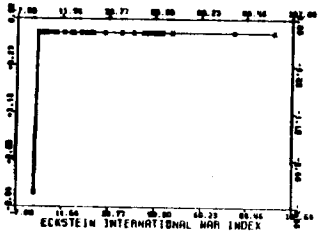
GENERAL NATIONAL PRODUCT



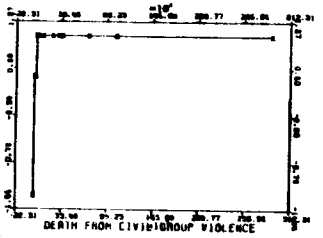
PERCENTAGE LABOR FORCE IN AGRICULTURE



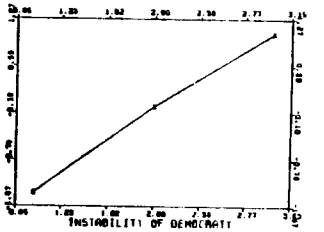
INSTABILITY OF PERSONNEL



EDERSTEIN INTERNATIONAL WAR INDEX



DEATH FROM CIVIL GROUP VIOLENCE



INSTABILITY OF DEMOCRACY

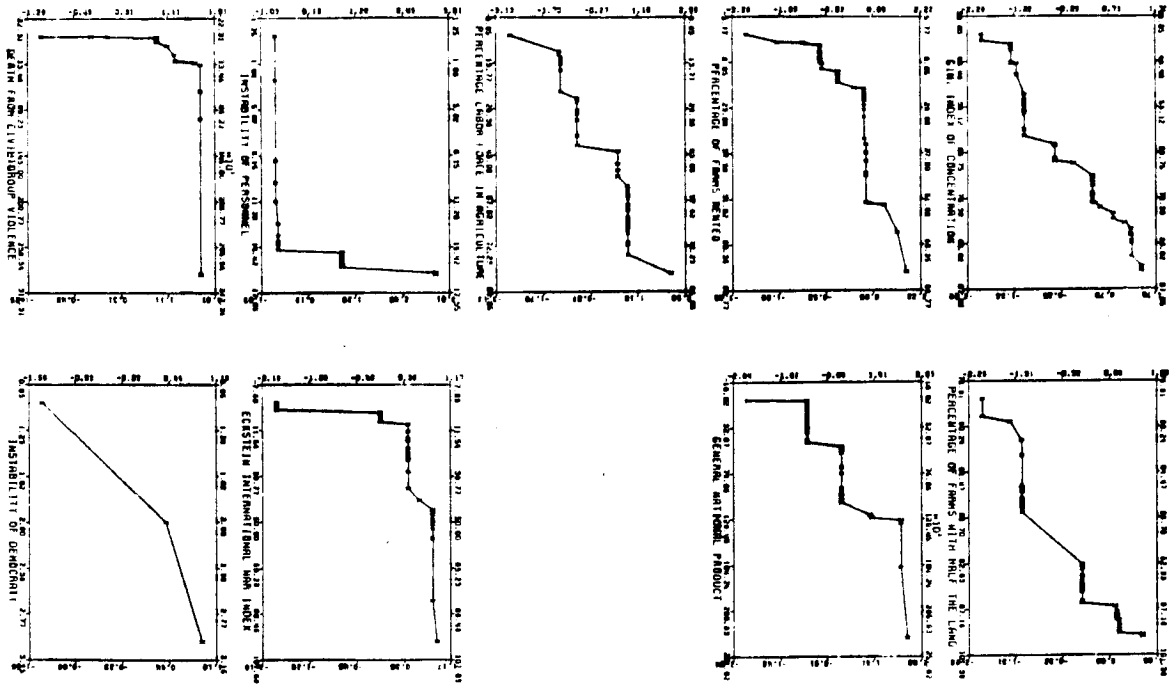


Figure 7.4.a Transformations of the rankings

ORDERING OF MEFV-CURVES RELATIVE TO RESPIRATORY SYMPTOMS

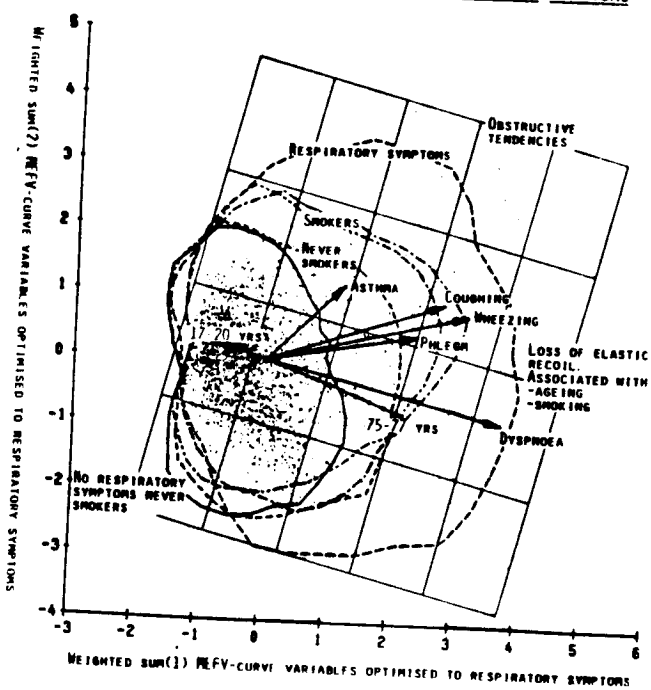


Figure 1.
SCATTERGRAM
- Most parsimonious ordering of MEFV-curves with respect to chronic respiratory symptoms.
The ordering is expressed in terms of a specific weighted sums of the MEFV-curve variables (abscissa and ordinate). In this way, each MEFV-curve gets its position in the scattergram.

ARROWS
- The position of a curve in the scattergram is related to the probability on chronic respiratory symptoms. These increase in the direction of the labelled arrows.
- Curves having a high probability for ASTHMA are ... more upwards and somewhat to the right; those indicative for DYSPNOEA mainly to the right. ... lungfunctions are found at the bottom left.

CONTOURS
- Subgroups (having identical age distributions) are indicated by way of percentile contours containing 95% of their values.
- SMOKERS (- - -, - - - -) occupy an intermediate position between the subgroups NO RESPIRATORY SYMPTOMS/NEVER SMOKERS (——) and RESPIRATORY SYMPTOMS/SMOKERS (- - - -). They contain more curves with a larger probability for one or more respiratory symptoms than NEVER SMOKERS (- - - -).

KNOTTED LINE (——•——).
- The knotted line for age shows that with increasing age, curves are more probable associated with respiratory symptoms, especially DYSPNOEA. Independent of age, height is hardly related to curve positions.

INTERPRETATION
- The association with smoking and especially ageing in the direction of DYSPNOEA, suggests that loss of elastic recoil may be involved in the change in curve configuration towards the right.
- Relatively independent of the former trend the direction of the vector ASTHMA suggests that in more vertical direction obstructive tendencies increase (see other figures) than loss of elastic recoil alone.

ORDERING OF MEFV-CURVES RELATIVE TO AGE AND HEIGHT; PHASES

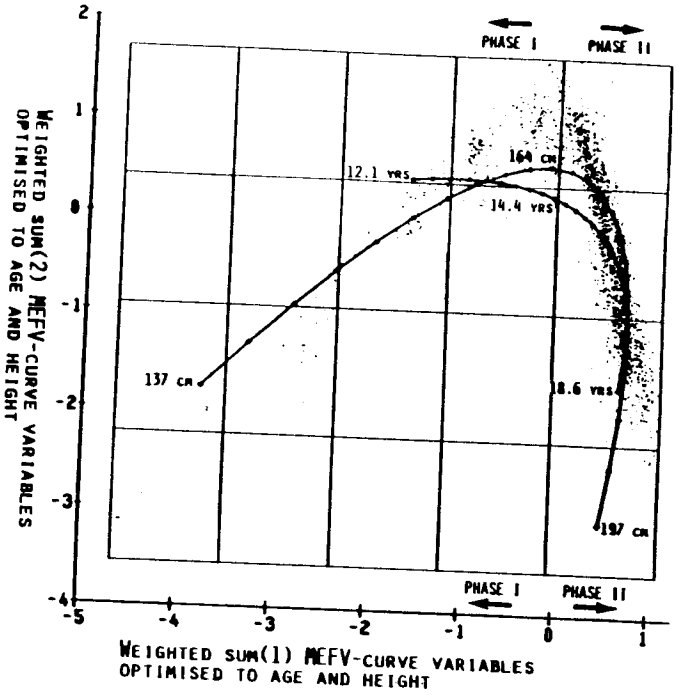


Figure 1
SCATTERGRAM
- Most parsimonious ordering of MEFV-curves with respect to age and height.
  The ordering is expressed in terms of 2 specific weighted sums of the MEFV-curve variables. In this way each MEFV-curve gets its position in the scattergram.

PHASES
- The scattergram has a boomerang-like shape suggesting two phases.
- Phase 1 covers mostly the age range 12-14.5 yrs. Curve variability is mainly related to differences in height, mostly 137-164 cm.
- Phase 2 concerns a larger age range mostly 14.5-18 yrs. Curve variability is also mainly related to differences in height, mostly 165-197cm.

of nomination and renomination on internal party processes could give central party organs a strong weapon with which to discipline deviant behaviour. The electorate gives a mandate to a party, not to individual members of the parliament: defiant members cannot bring their case to individual constituents but must satisfy themselves with such hearing as they can get within the party or leave.

5.1. *Description of the data*

In 1972 the members of the Dutch Parliament (MPs) were interviewed. Among other things, the MPs gave their opinions on a number of issues and their preference votes for the political parties. The issues concerned development aid, abortion, law and order, income differences, worker participation, taxation and defence. The opinions were measured on a nine-point scale of which the lowest and the highest category were described (Table 1). The party preferences were recorded in a table of rank

Table 1. The issues and the meanings of the lowest and the highest category

	1 DEVELOPMENT AID	
the government should spend more money on aid to developing countries	(1) ..... (9)	the government should spend less money on aid to developing countries
	2 ABORTION	
the government should prohibit abortion completely	(1) ..... (9)	a woman has the right to decide for herself about abortion
	3 LAW AND ORDER	
the government takes too strong action against public disturbances	(1) ..... (9)	the government should take stronger action against public disturbances
	4 INCOME DIFFERENCES	
income differences should remain as they are	(1) ..... (9)	income differences should become much less
	5 PARTICIPATION	
only management should decide important matters in industry	(1) ..... (9)	workers must also have participation in decisions important for industry
	6 TAXATION	
taxes should be increased for general welfare	(1) ..... (9)	taxes should be decreased so that people can decide for themselves how to spend their money
	7 DEFENCE	
the government should insist on shrinking the Western armies	(1) ..... (9)	the government should insist on maintaining strong Western armies

orders. The scores in this table tell us the rank order each member of the parliament gave to the different parties (2 = highest preference, 15 = lowest preference). The lowest score (2) was always used for the MP's own party. For our illustration we only consider the preferences for the four largest parties, which are:

PvdA – labour party (socialists) (39)
ARP – Anti Revolutionary Party (christian democrats) (13)
KVP – catholic party (christian democrats) (35)
VVD – 'liberal' party, referred to as conservatives (16)

The figure in parentheses is the number of MPs. The other parties in 1972 were:

CHU – Christian Historical Union (christian democrats) (10)
D'66 – democrats 66 (liberals) (11)

correlations between the optimally scaled variables and the canonical variates, the *canonical loadings*, are standard output of the CANALS program. As the optimally scaled variables and the canonical variates are standardized, the canonical loadings correspond to the *projections* of the optimally scaled variables on the canonical variates. We do not give the figures for the canonical spaces here. Because the canonical correlations of both the ordinal and numerical solution are rather high the canonical loadings of the first set are very similar to the canonical loadings of the second set. Therefore we averaged the canonical variates over the sets so as to get one figure for the two sets together and computed the correlations between the optimally scaled variables and the mean canonical variates. These correlations, also called *component loadings* analogous to principal components analysis, can be computed from the canonical loadings when multiplied by $\sqrt{[(1+\lambda)/2]}$, the canonical loadings of the optimally scaled variables and canonical variates of the same set give the component loadings of this set. For the first dimension λ is λ_1 and for the second dimension λ is λ_2 (see Table 2). As we also standardized the mean canonical variates the component loadings correspond to the projections of the optimally scaled variables in the mean canonical space (Fig. 3). The horizontal axis is the first mean canonical variate and the vertical axis is the second mean canonical variate. We see in Fig. 3 that 'income differences' and 'participation' are the most important issues after which 'taxation', 'defence' and 'law and order' follow. 'Abortion' and

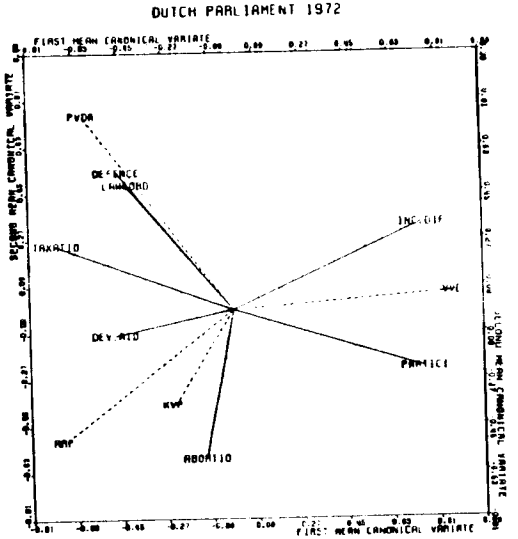


Figure 3. Component loadings: the projections of the optimally scaled variables in the mean canonical space (ordinal solution)

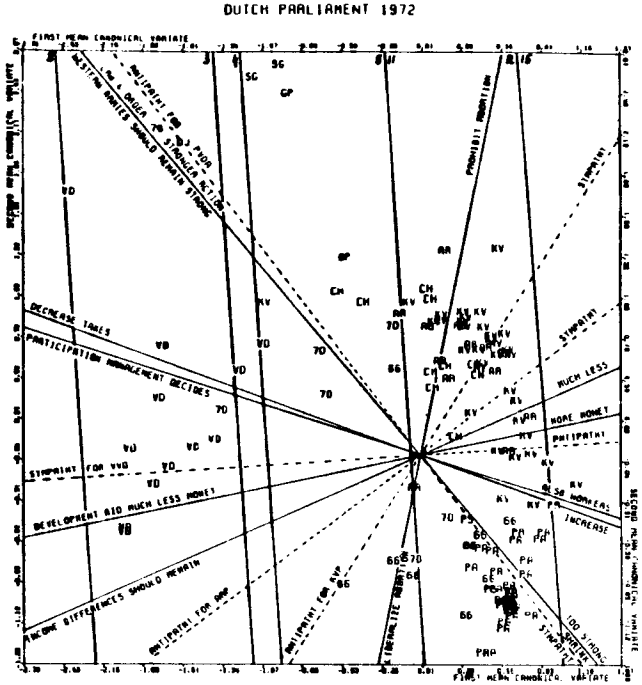


Figure 5. The MPs projected in the mean canonical space (ordinal solution). The individuals are labelled by their party membership (PA = PvdA, AR = ARP, KV = KVP, VD = VVD, CH = CHU, 66 = D'66, 70 = DS'70, PS = PSP, PR = PPR, GP = GPV, SG = SGP, DJ). The lines through the origin are the directions of the optimally scaled variables. The parallel lines demonstrate the position of the category numbers of the preference for the VVD.

MPs are more regularly spread over Fig. 6. In an ordinal analysis the scores get a chance to grow close or extreme, which is not possible in a numerical analysis. So in an ordinal solution clustering and 'extremism' can be stronger than in a numerical solution of the same data. But the plot of the component loadings of the numerical solution was clearer than that of the ordinal solution. This is because in the numerical analysis the direction is the most important way to differentiate between the different MPs. In essence we find the same configuration in Fig. 5 as in Fig. 6, but especially on the item 'abortion' and the preference for the KVP and ARP the conservatives are not so divided. The christian democratic MPs still think differently

5.4. *The category quantifications*

To get an impression of the transformations of the variables, we give a plot of the original category numbers against the category quantifications (Fig. 7). The points are connected to show the monotone transformation of each variable (see Section 4.2). For the issue 'development aid' we see for instance that the MPs who agree with

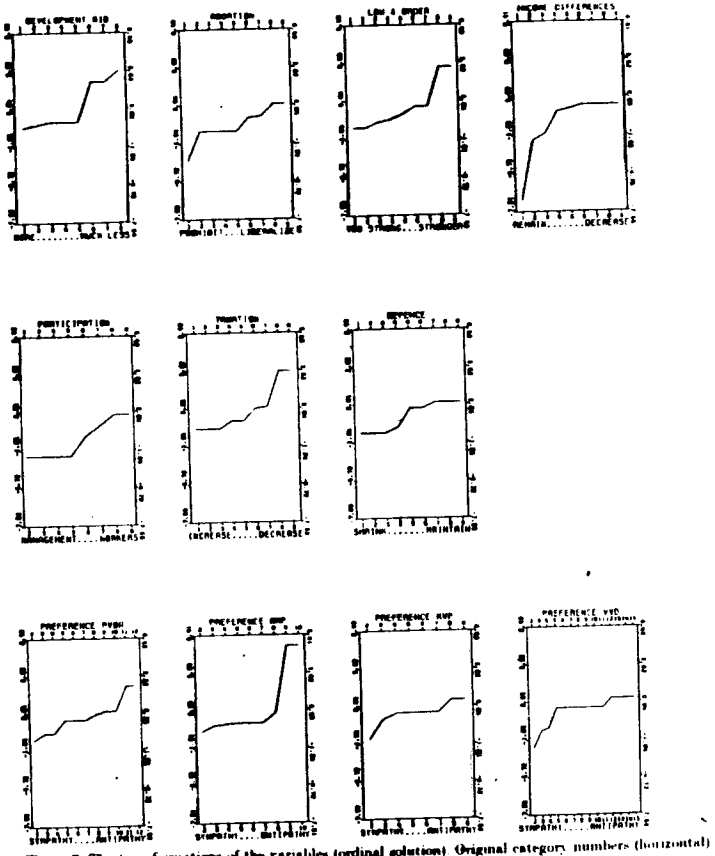


Figure 7. The transformations of the variables (ordinal solution). Original category numbers (horizontal) against category quantifications (vertical).

Summary of Thursday

Yesterday we discussed the use of **B-splines** as a form of fuzzy coding, in connection with NLPCA. We also discussed various alternative criteria of the form $\mu = \Phi(R)$ that can be used to define other types of NLPCA. Theory was developed which shows how the various forms were related.

Finally we discussed the use of additivity restrictions on the quantifications to introduce sets of variables, and thus other forms of MVA such as regression, (M)ANOVA, canonical correlation analysis, and so on.

We ended the talk yesterday by giving various examples of the use of additivity restrictions. We shall continue by giving some more.

Example 28 [ISRAELS, PM, 49, 1984, 331-346] 156

and Activity. Also the explanatory variables are considered to be qualitative variables, corresponding with the way these variables are measured in the survey. In the first column of Table 2 the meaning of the categories is given, and in the last column the number of people falling in each category. The total sample consisted of 4 108 people, all being 15 years or older.

In the second column the scale values $c^{(0)}$ for Satisfaction are given as well as the regression coefficients for the dummy-variables, i.e. the vector d. The scale values for Satisfaction are determined in such a way that Satisfaction becomes a quantitative variable with mean zero and variance one. For the explanatory variables neither holds. By

TABLE 2

Qualitative Regression Analysis of Satisfaction on Marital status, Schooling, Income and Activity

Variable	Coefficients (c and d)	Coefficients after centration	Standardized scale values	Number of people
Satisfaction				
Not too satisfied	-3.43			264
Rather satisfied	-0.81			539
Satisfied	0.25	idem	idem	1 857
Very satisfied	0.63			1 088
Extremely satisfied	0.53			360
Marital status				
Married	-0.21	0.04	0.26	2 940
Widowed	-0.56	-0.32	-2.34	268
Divorced	-1.17	-0.92	-6.72	53
Single	-0.21	0.03	0.24	847
Schooling				
Low	0.13	0.03	0.68	2 492
Medium	0.09	-0.02	-0.42	1 027
High	0.03	-0.07	-1.88	401
Unknown	—	-0.10	-2.69	188
Income (× Dfl. 1,000)				
< 21	-0.11	-0.12	-1.24	1 356
21 - < 40	0.07	0.06	0.55	1 689
> 40	0.19	0.18	1.81	454
Unknown	—	-0.01	-0.11	609
Activity				
Employed	0.17	0.04	0.17	1 987
Unemployed	-1.00	-1.13	-4.96	65
Not able to work	-0.86	-0.99	-4.35	118
Retired	0.26	0.13	0.59	314
Student	0.17	0.04	0.19	336
Housewife	0.18	0.06	0.25	1 203
Unknown	—	-0.13	-0.56	85

R^2 .085 [ind .0624, ridit .0588]

N = 4108

TABLE 3

Qualitative Regression Analysis of Happiness on Marital status, Schooling, Income and Activity

Variable	Coefficients (c and d)	Coefficients after centration	Standardized scale values	Number of people
Happiness				
Unhappy	-5.41			11
Not too happy	-4.17			92
Not happy, not unhappy	-1.64	idem	idem	561
Happy	0.28			2 555
Very happy	0.73			889
Marital status				
Married	-0.14	0.09	0.46	2 940
Widowed	-0.84	-0.61	-2.94	268
Divorced	-1.16	-0.93	-4.49	53
Single	-0.31	-0.08	-0.37	847
Schooling				
Low	0.03	0.00	0.28	2 492
Medium	0.04	0.01	0.71	1 027
High	-0.00	-0.03	-2.48	401
Unknown	—	-0.03	-2.24	188
Income (× Dfl. 1,000)				
< 21	-0.08	-0.12	-1.22	1 356
21 - < 40	0.10	0.07	0.68	1 689
> 40	0.20	0.16	1.63	454
Unknown	—	-0.04	-0.39	609
Activity				
Employed	0.15	-0.01	-0.08	1 987
Unemployed	-0.12	-0.28	-2.17	65
Not able to work	-0.50	-0.66	-5.10	118
Retired	0.24	0.07	0.57	314
Student	0.27	0.10	0.80	336
Housewife	0.23	0.06	0.47	1 203
Unknown	—	-0.16	-1.27	85

R^2 = .0754

Because $Ey_i y_i'$ is a diagonal matrix we can construct a $(\Sigma r_i) \times (\Sigma r_i)$ diagonal matrix Δ_y with block-diagonal submatrices $Ey_i y_i'$. The diagonal of Δ_y contains the probabilities of falling in the various categories. Defining $c = (c_1', \ldots, c_m')'$, (4.2) turns into

$$c'\Delta_y c = 1. \tag{4.3}$$

For the time being, we do not put separate normalizations on each of the c_i-vectors. This kind of normalizations can be done afterwards. Notice that the parameters $c_i^{(0)}$ can be considered as a product of a scaling parameter $c_i^{*(0)}$, with norm 1 for each variable, and a

158

TABLE 4

Qualitative Redundancy Analysis of Satisfaction and Happiness on Marital status, Schooling, Income and Activity

Variable	Coefficients (c and d)	Coefficients after centration	Standardized scale values
Satisfaction			
Not too Satisfied	-2.46		-3.34
Rather satisfied	-0.66		-0.90
Satisfied	0.16	idem	0.22
Very satisfied	0.51		0.69
Extremely satisfied	0.41		0.55
Happiness			
Unhappy	-3.71		-5.49
Not too happy	-2.79		-4.14
Not happy, not unhappy	-1.11	idem	-1.65
Happy	0.19		0.28
Very happy	0.49		0.73
Marital status			
Married	-0.23	0.09	0.39
Widowed	-0.98	-0.65	-2.76
Divorced	-1.63	-1.30	-5.52
Single	-0.36	-0.03	-0.12
Schooling			
Low	0.12	0.02	0.61
Medium	0.10	-0.01	-0.16
High	0.02	-0.08	-2.10
Unknown	—	-0.10	-2.71
Income (* Dfl. 1,000)			
< 21	-0.14	-0.18	-1.24
21 - < 40	0.12	0.09	0.61
≥ 40	0.28	0.24	1.73
Unknown	—	-0.03	-0.23
Activity			
Employed	0.21	0.02	0.08
Unemployed	-0.81	-1.00	-4.07
Not able to work	-0.98	-1.17	-4.75
Retired	0.35	0.15	0.62
Student	0.30	0.10	0.42
Housewife	0.27	0.08	0.33
Unknown	—	-0.19	-0.78

159

Example

Satisfaction $\approx$ .137 Marital Status + .039 Schooling + .100 Income + .227 Activity

Happiness $\approx$ .266 Marital Status + .014 Schooling + .098 Income + .130 Activity

.737 Satisfaction + .676 Happiness $\approx$ .236 Marital Status + .037 Schooling + .141 Income + .247 Activity

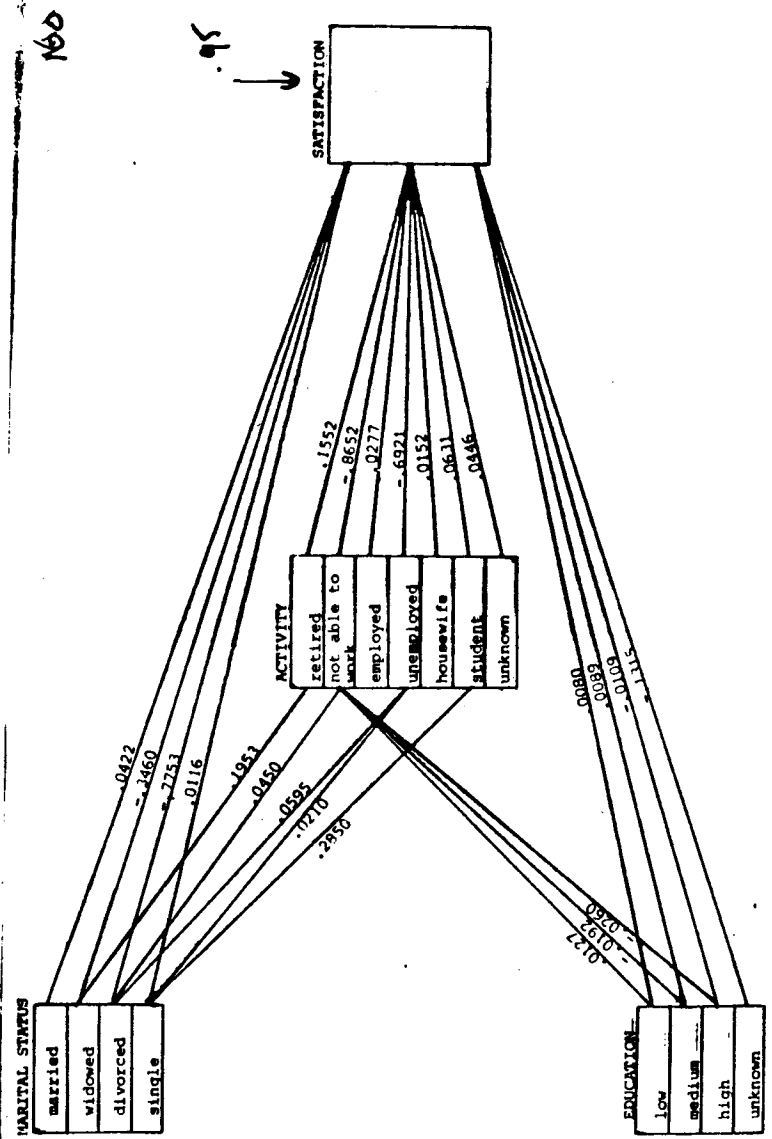


Figure 2. Path diagram of Satisfaction, Activity, Marital status and Education

TABLE 1

Variables from the study of chronic lung disease

set 1	VLA:	Residence, (1) Vlagtwedde, (2) Vlaardingen.
set 2	SMO:	Smoking, (1) never smoker, (2) ex-smoker, (3) current smoker.
	RATE:	Rate of smoking (amount of tabacco), (1) never smoker, (2) low rate, ..., (9) high rate.
	PER:	Time period smoked, (1) never smoker, (2) short period, ..., (13) long period.
	TIME:	Time since last cigaret, (1) never smoker, (2) long ago, ..., (5) recently, (6) current smoker.
set 3	AGE:	Age discretisized into periods of 3.5 years, (1) age 19-22.5, ..., (10) age 52.5-56.
	SEX:	Sex, (1) male, (2) female.
set 4	COU:	Coughing, (1) no, (2) persistent.
	PHLE:	Phlegm, (1) no, (2) persistent.
	DYS:	Dyspnoea or shortage of breath, (1) no, (2) slight/moderate, (3) severe.
	WHE:	Wheezing, (1) never, (2) ever, (3) severe.
	AST:	Asthma, (1) ever, (2) never.

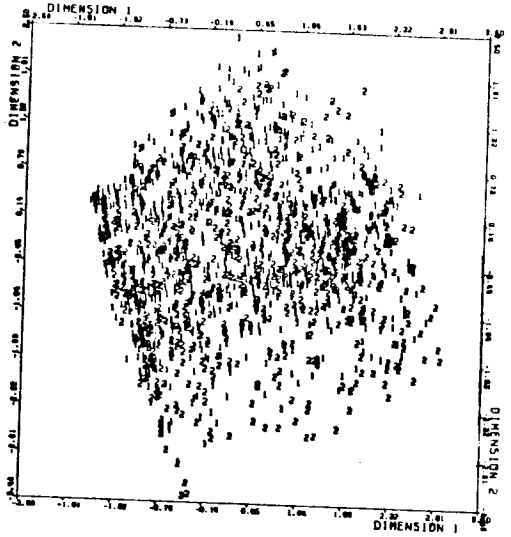


Figure 2 Object scores 1 Vlagtwedde
2 Vlaardingen

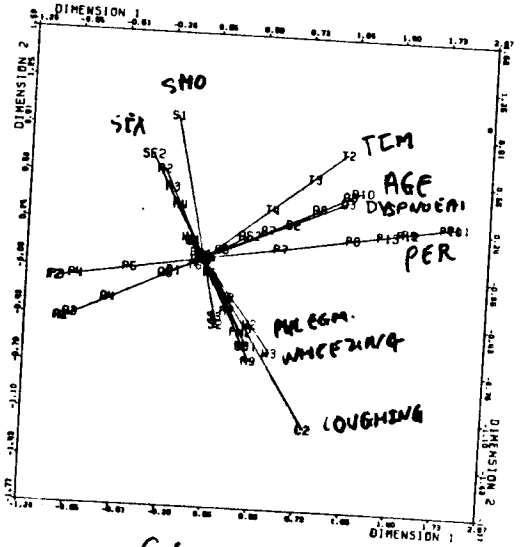


Figure 3 Category quantifications

The OVERALS algorithm (by implication this algorithm can also be used for the various special cases).

(a) Start with X_0, Z_0

(b) Loop over variables

$$\min \; tr\,[X - \sum_{j\in J} G_j Y_j]'[X - \sum_{j\in J} G_j Y_j] =$$

$$\min \; tr\,\Big\{[X - \sum_{\substack{j\in J \\ j\neq l}} G_j Y_j] - G_l Y_l\Big\}'\{\ldots\}$$

(c) first multiple quantification

$$Y_l \leftarrow D_l^{-1}\Big[X - \sum_{\substack{j\in J \\ j\neq l}} G_j Y_j\Big].$$

Observe that the centroid principle is not satisfied in the same way as before!

(d) Then

adjust for cone & rank one restrictions if necessary

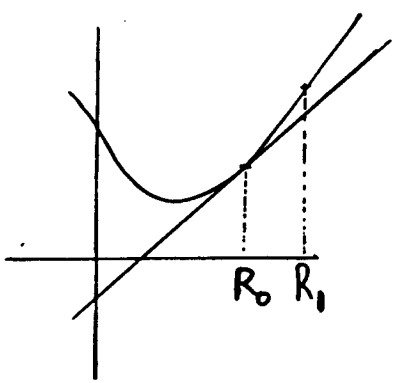
(e) Find new X at end of variable loop.

This is, of course, very similar to the PRINCIPALS algorithm, with just one additional complication.

We shall also briefly discuss the future alternative algorithm based on the Burt table. We want to maximize

$$\mu = \Phi\,[R(y)]$$

over quantifications. All variables are assumed to be single [rather-one], multiple variables and sets must be introduced through copies and additivity restrictions.



Thus $\Phi(R) \geq \Phi(R_0) + \mathrm{tr}\, G_0 (R - R_0)$

Now suppose we maximize $\mathrm{tr}\, G_0 R$ over R (i.e. over $y_1, \ldots, y_m$), and this gives R_1. Then

$$\Phi(R_1) \geq \Phi(R_0) + \mathrm{tr}\, G_0 (R_1 - R_0)$$
$$\geq \Phi(R_0)$$

Thus by maximizing $\mathrm{tr}\, G_0 R$ we increase Φ. This gives a convergent algorithm.

How does $\operatorname{tr} G_0 R$ look. We can write $r_{je} = y_j' G_e y_e$; s $y_j' D_j y_j = 1$. Thus $\operatorname{tr} G_0 R = \sum_j \sum_e y_j' C_{je} y_e \cdot g^0_{je}$. Thus if we define $\tilde{C}_0$ to be the Burt-table with each subtable multiplied by g^0_{je}, then $\operatorname{tr} G_0 R = y' \tilde{C}_0 y$, which must be maximized over all y satisfying $y_0' D_j y_0 = 1$ (and cone + additivity restrictions, i.e. cone restrictions).

If G_0^{**} is positive-semi definite additional simplification is possible. Then $\hat{C}_0$ is psd, and $y' \hat{C}_0 y$ is convex. We now apply the same trick (majorization)

$$y' \hat{C}^0 y \geq y_0' \hat{C}^0 y_0 + 2 y_0' \hat{C}^0 [y - y_0]$$

As a consequence we need to maximize the linear function $y' \hat{C}^0 y_0$ over y satisfying the restrictions. This can be done for each j separately by minimizing, letting $\hat{y}_0 = \hat{C}^0 y_0$,

$$[y_j - \hat{y}^0_j]' D_j [y_j - \hat{y}^0_j]$$

over $y_j \in K_j$ (the cone), and then setting the new y_j equal to the normalized projection.

Some examples of G_0.

If $\Phi(R) = \lambda_1 + \cdots + \lambda_p$ then $g^0_{je} = \sum_{s=1}^{p} \alpha_{js} \alpha_{es}$

If $\Phi(R) = \sum\sum r_{je}$ then $g^0_{je} \equiv 1$

If $\Phi(R) = \frac{1}{2}\sum\sum r^2_{je}$ then $g^0_{je} = r_{je}$

If $\Phi(R) = -|R|$ then $g^0_{je} = r^{je}$

Analysis of statistical stability

There are quite a few results on analytic and algebraic stability in Gifi.

Techniques used
- delta method
- Bootstrap
- Jackknife.

We shall first explain these techniques briefly. They are essentially nonparametric, and apply to all functions of counts (frequencies).

Suppose we have n repeated independent replications of a discrete random variable, taking m possible values. This defines a vector

p_n of relative frequencies (proportions). We know that

WLLN (a) $\quad p_n \xrightarrow{P} \pi$

CLT (b) $\quad n^{1/2}[p_n - \pi] \xrightarrow{L} \mathcal{N}(0, \Pi - \pi\pi')$.

Now suppose ψ is a function of p [in the unit simplex of $\mathbb{R}^m$] with values in $\mathbb{R}^t$. If ψ is continuous (at π). Then

$$\psi(p_n) \xrightarrow{P} \psi(\pi).$$

If ψ is differentiable (at π), with derivative G_π, then

$$n^{1/2}[\psi(p_n) - \psi(\pi)] \xrightarrow{L} \mathcal{N}(0, G_\pi V G_\pi').$$

Another result which has been known at least since Laplace.

Thus: any diff function of proportions is asymptotically normal

If ψ is continuously differentiable in a neighborhood of $\bar{\pi}$, then $G_{\bar{\pi}}' V G_{\bar{\pi}}$ can be estimated.

$$G_{\hat{p}}' [\hat{p} - p\hat{p}'] G_{\hat{p}}' \xrightarrow{P} G_{\bar{\pi}}' [\bar{\pi} - \pi\pi'] G_{\bar{\pi}} .$$

This defines the Delta Method.

Everything computed in Gifi is a function of the profile frequencies [even if splines are used] and consequently the delta method can be used. We merely need

- expressions for the derivatives
- programs for computing them.

This means that there is a routine way in which standard errors and confidence ellipsoids can be computed for our techniques [at least if you are willing to assume that the data are some random sample]. If the data are nonrandom the confidence regions can still be interpreted as measures of stability.

Most of the things computed in Gifi are functions of the Burt table C. We can write

$$C = \sum_{\mu=1}^{m} p_\mu g_\mu g_\mu'$$

with $g_{j\mu}$ the profile-indicators. We can write $\psi(p)$ in the form $\psi(C(p))$ and obtain (by the chain rule)

$$\frac{\partial \psi_s}{\partial p_\mu} = \sum_{je} \frac{\partial \psi_s}{\partial c_{je}} g_{\mu j} g_{\mu e}$$

This often simplifies matters. Many other things are functions of the standardized correlations $r_{je} = y_j' C_{je} y_e$. Then

$$\frac{\partial \Phi}{\partial p_\mu} = \sum_j \sum_e \frac{\partial \Phi}{\partial r_{je}} \frac{\partial r_{je}}{\partial p_\mu} =$$

$$= \sum_j \sum_e \frac{\partial \Phi}{\partial r_{je}} \left[y_j' C_{je} \frac{\partial y_e}{\partial p_\mu} + y_j' \frac{\partial C_{je}}{\partial p_\mu} y_e + \frac{\partial y_j'}{\partial p_\mu} C_{je} y_e \right].$$

If (model!) we have (in the population) linear bivariate regression then $C_{je} y_e = r_{je} D_j y_j$ and $C_{ej} y_j = r_{je} D_e y_e$. Thus

$$\frac{\partial \Phi}{\partial p_\mu} = \sum_j \sum_e \frac{\partial \Phi}{\partial r_{je}} \left[r_{je}\, y_j' D_j \frac{\partial y_j}{\partial p_\mu} + r_{je}\, y_e' D_e \frac{\partial y_e}{\partial p_\mu} + \right.$$

$$\left. + \left[y_j' g_{j\mu} \right] \left(y_e' g_{e\mu} \right) \right].$$

But we used the restrictions $y_j' D_j y_j = y_e' D_e y_e = 1$.

Thus

$$y_j' D_j \frac{\partial y_j}{\partial p_\mu} + \left(\frac{\partial y_j}{\partial p_\mu} \right)' D_j y_j + y_j' \frac{\partial D_j}{\partial p_\mu} y_j = 0$$

or

$$y_j' D_j \frac{\partial y_j}{\partial p_\mu} = -\frac{1}{2} y_j' \frac{\partial D_j}{\partial p_\mu} y_j$$

It follows that

$$\frac{\partial \Phi}{\partial p_\mu} = \sum_j \sum_e \frac{\partial \Phi}{\partial r_{je}} \left[y_j' \frac{\partial C_{je}}{\partial p_\mu} y_e - \frac{1}{2} y_j' y_j' \frac{\partial D_j}{\partial p_\mu} y_j - \frac{1}{2} y_e y_e' \frac{\partial D_e}{\partial p_\mu} y_e \right]$$

But this is the same derivative as the one we find if the y_j are known scores (not computed).

Implications for – PRIMALS [first step] or any other first-step method.

Implications in combination with LISREL. (for instance).

ANALOR. example 5. 174

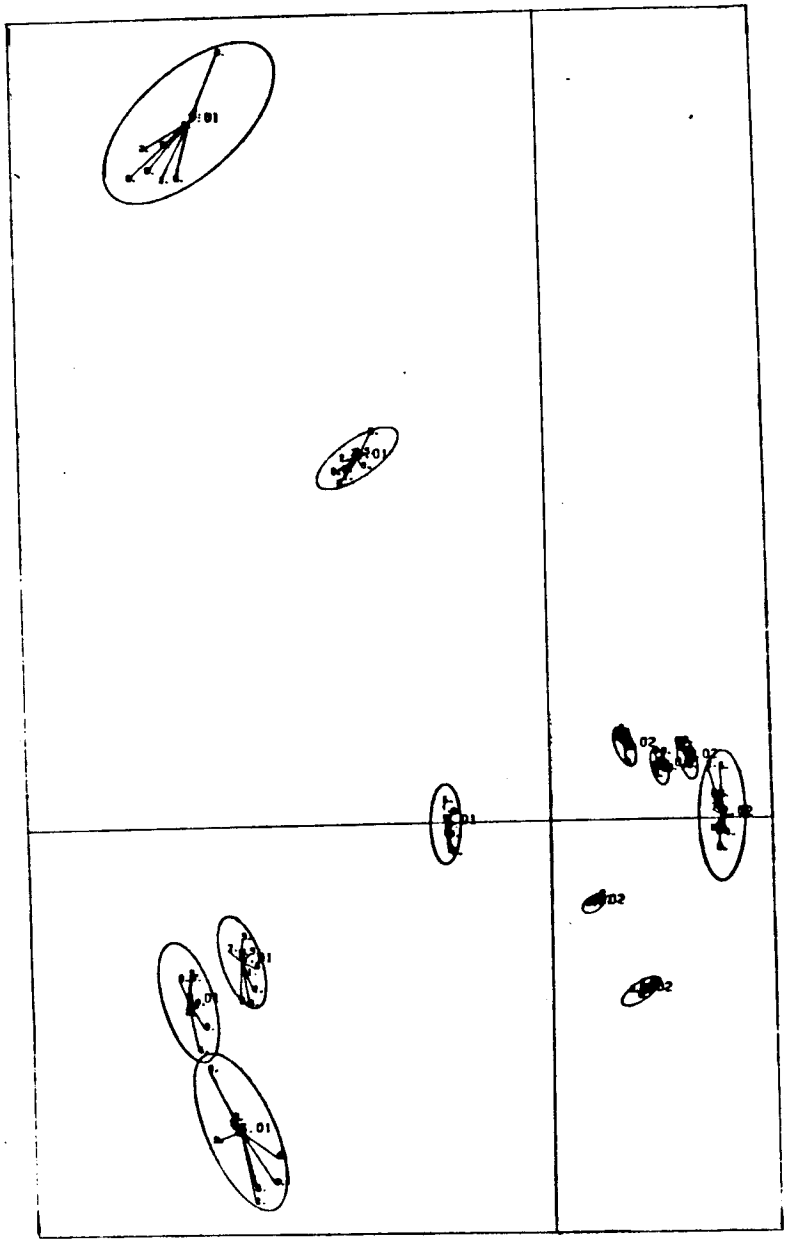


Figure 12.4 Stability of the ANAPROF solution of the Sugiyama data: 95%-ellipses and 10 bootstraps

Future = Derive the necessary derivatives for an application of the delta method to all techniques

[done for HOMALS
ANALOR ← - optimize
PRIMALS
—— ISRAELS
ANAPROF ← optimize]

= apply two-step model to LISREL etc [and improve two-step estimates in order to construct efficient ones]. Construct test for linearity of all regressions.

Next Bootstrap. This is very easily explained. We know that

$$p_n | \pi \xrightarrow{L} \eta(0, V(\pi))$$

Now suppose we draw a random sample, with replacement, from the multinomial with parameters (p_n, n). This gives $\hat{p}_n$. Now

$$\hat{p}_n | p_n \xrightarrow{L} \eta(0, V(p_n))$$

and because $V(p_n) \xrightarrow{P} V(\pi)$ we can use $\hat{p}_n | p_n$ to approximate $p_n | \pi$. This amounts to drawing random samples from your own data and making "inferences" to the "population" on the basis of this. In many recent publications (E from, Beran, Freedman, Diaconis) it has been shown

that Bootstrap methods are robust and efficient in many circumstances.

Examples
- Back to Japan
- Smoothing
- Bias correction $2\Phi(p_n) - \Phi(\hat{p}_n)$.

Future research (in progress)
- how many samples are needed.
- how best implemented.

Jackknife We can be brief.

$$\Phi_\mu(p_n) \triangleq \Phi\left[\frac{n p_n - e_\mu}{n-1}\right] = \Phi\left[p_n + \frac{1}{n-1}[p_n - e_\mu]\right]$$

$$\approx \Phi(p_n) + \frac{1}{n-1}\frac{\partial \Phi}{\partial p_n}(p_n - e_\mu)$$

$$n\,\Phi(p_n) - (n-1)\,\Phi_\mu(p_n) \approx \Phi(p_n) - \frac{\partial \Phi}{\partial p_n}(p_n - e_\mu)$$

pseudo-values $\tilde{\phi}_\mu(p_n)$

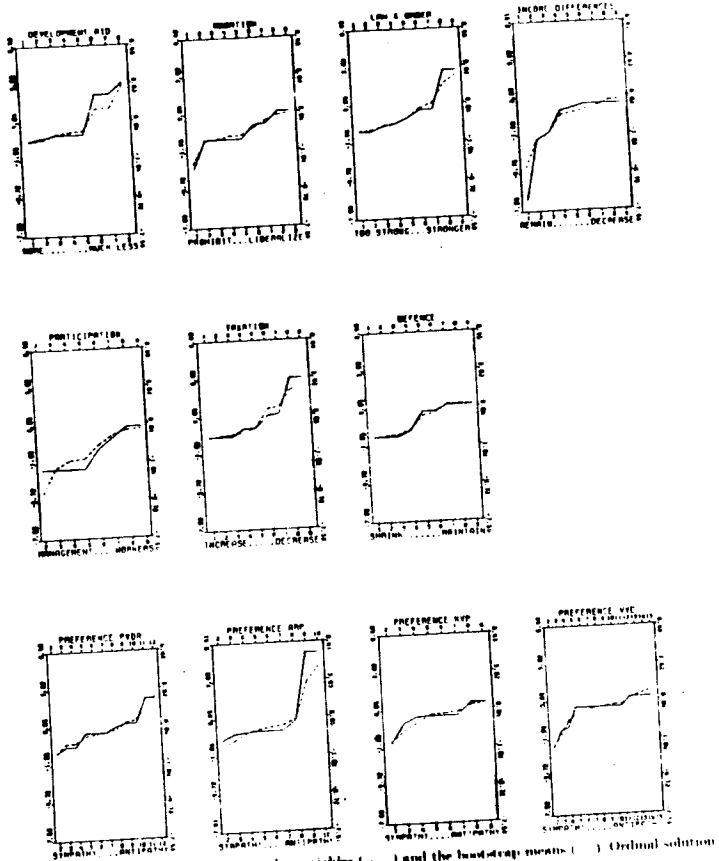


Figure 9. The transformations of the variables ( ) and the bootstrap means ( ). Ordinal solution

advantage of this technique is the fact that we are free to choose the measurement level of each variable separately, so that we do not have to impose strong restrictions on the analysis unnecessarily. The CANALS technique is a real extension and improvement of the technique proposed by Young *et al.* (1976). The alternating least squares algorithm of the CANALS program is a very nice way of avoiding eigenvalue computation and computing results rather quickly and precisely. Another advantage

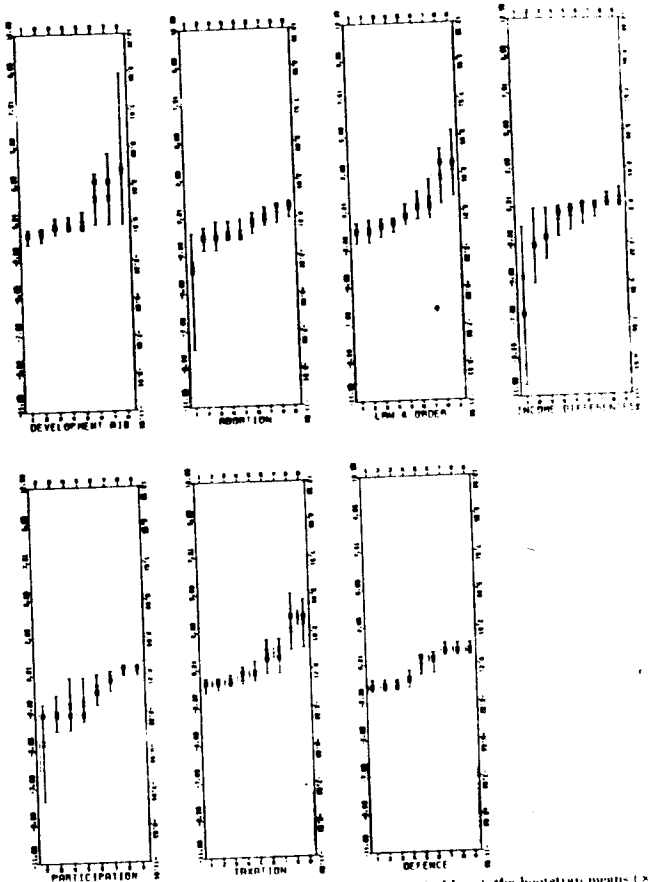


Figure 8(a). The range of the category quantifications (vertical lines), the bootstrap means (×) and the original values (o) for the issues

$$\sum_\mu f_\mu \tilde{\phi}_\mu (R_n) \approx \Phi(R_n)$$

$$\sum_\mu f_\mu \left[\tilde{\phi}_\mu (R_n) - \bar{\Phi}(R_n) \right]^2 \approx$$

$$g'(R_n)\, V(R_n)\, g(R_n)$$

which is the delta method variance estimate. Thus delete each profile, compute the pseudo-values and their variance.

Future: Comparison Bootstrap & Jackknife.

BS & JK are both useful as data analytic techniques, even if there is no random sampling.

Significance testing of generalized canonical correlations

De Leeuw & Vd Burg, Data analysis & Informatics IV

We use permutation methods. If variables are partitioned into sets we leave set 1 intact. Then permute all observations in set 2 randomly, same in set 3, etc. This creates the null-distribution. There are various ways to assess it

- enumeration (Fisher exact test)
- use asymptotic normality of the Burt table (as in Lebart, O'Neill for CA)
- Monte Carlo — via C
  - directly

182

TABLE 3: Generalized canonical correlations: empirical values (ev) and generated values at 5, 25, 50, 75 and 95 percent. Random permutation method. From Year to Year data, multiple nominal, single nominal and numerical.

	mult nominal		single nom		numerical	
	dim1	dim2	dim1	dim2	dim1	dim2
ev	.744	.570	.735	.530	.714	.496
5	.549	.532	.533	.511	.501	.482
25	.554	.542	.543	.521	.511	.494
50	.559	.545	.549	.527	.518	.499
75	.564	.550	.554	.532	.525	.504
95	.570	.557	.565	.541	.532	.514

TABLE 4: Generalized canonical correlations: empirical values (ev) and generated values at 5, 25, 50, 75 and 95 percent. C-matrix method. From Year to Year data, multiple nominal and numerical.

	multiple nom		numerical	
	dim1	dim2	dim1	dim2
ev	.744	.570	.714	.496
5	.553	.545	.509	.500
25	.560	.550	.514	.503
50	.565	.554	.521	.506
75	.569	.557	.525	.504
95	.574	.562	.535	.514

canonical correlation problems with p=2 were computed separately. The eigenvalues are in the first row of table 5, while the rest of the table shows percentiles of the PD (all estimated by the random permutation method). In the S & M analysis probability plots [given in figure 2] show a large deviation from normality for the first eigenvalue, all eigenvalues are very clearly significant, however. All estimated PD's have small variance, we see that $\lambda_{(95)}-\lambda_{(5)}<.03$, except for dimension 1 of S & M (which has a very light tail on the right, $\lambda_{(95)}-\lambda_{(75)}=.035$).

183

TABLE 6: Generalized canonical correlations: empirical values (ev) and generated values at 5, 25, 50, 75 and 95 percent. Random permutation method and C-matric method. Russett data, multiple nominal and numerical

	Random permutation method				C-matri method			
	mult nom		numerical		mult nom		numerical	
	dim1	dim2	dim1	dim2	dim1	dim2	dim1	dim2
ev	.815	.737	.687	.462	.815	.737	.687	.462
5	.724	.695	.462	.397	.748	.710	.457	.408
25	.754	.714	.499	.426	.775	.732	.480	.426
50	.773	.731	.525	.445	.795	.745	.500	.443
75	.794	.752	.545	.467	.809	.757	.519	.463
95	.830	.778	.576	.488	.841	.782	.562	.488

CONCLUSIONS

It appears from our examples that both the random permutation method and the C-matrix method give a fairly good approximation to significance probabilities and permutation distributions. In order to be sure, we shall have to apply enumeration somewhat more extensively. We also need to generalize the exact computation of G_s to the case of more than two sets. But from our experience so far, we can say that the Monte Carlo methods give the correct indication of the order of significance, and they show that with multiple options and not too many individuals the eigenvalues have to be very high to be significant. It is not surprising, for instance, that Gifi (1981) had great difficulty in interpreting the second dimension in the solution for the Russett data. Our approximations to the permutation distribution guard us against chance capitalization, and against trying to interpret effects which are not really there.

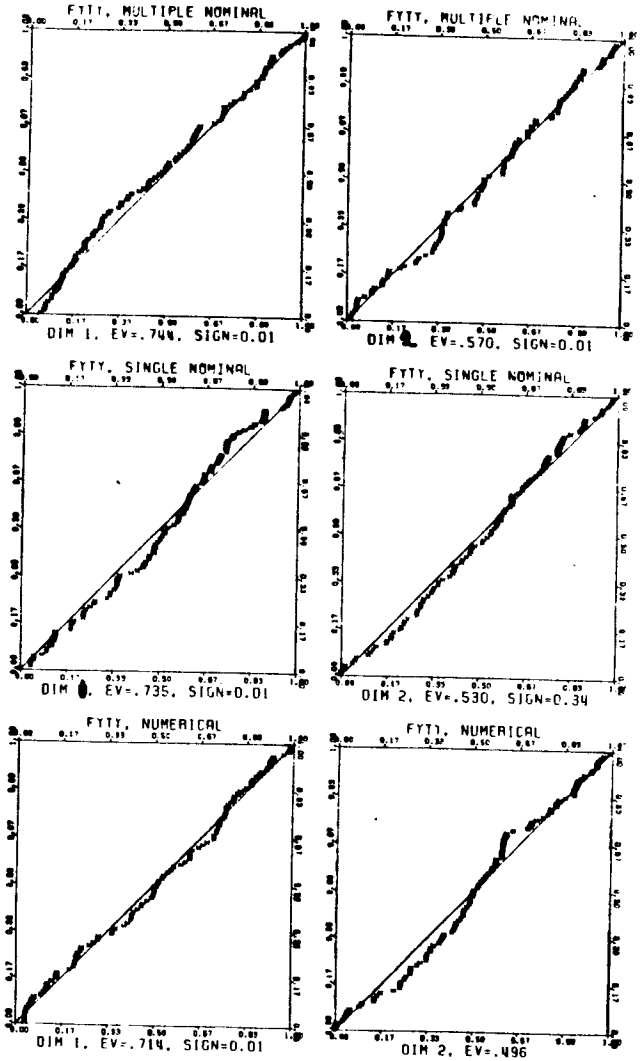


FIGURE 1: From Year to Year random permutations. Probability plot of eigenvalues.

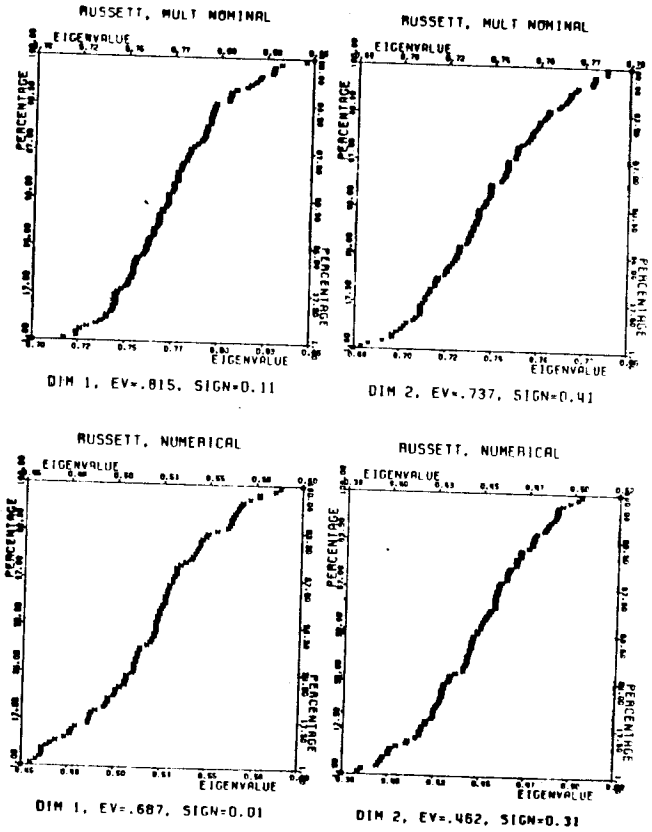


FIGURE 3: Russett random permutations. Eigenvalues (ordered) against percentages.

Possibilities for partial canonical correlation analysis

It can be incorporated in OVERALS by using copies of a variable in different sets [in all sets]. Using p copies again means multiple elimination, using one copy means single elimination. We must take care, however, that the variable has the same quantification in all sets in which it occurs.

Possibilities for "causal" analysis

We depart from the Gifi-system, with its heavy emphasis on component analysis. The more general loss function we now study is

$$\sigma(Y) = \sum_{j=1}^{m} \operatorname{tr} [G_j Y_j - \sum_{\ell \in I_j} G_\ell Y_\ell]' [\ldots\ldots].$$

Here $j = 1, \ldots, m$ are endogeneous variables and I_j is the set of "causes" of variable j. For normalization purposes it is convenient to require $Y_j' D_j Y_j = I$ for all j, and to write

$$\sigma(Y, B) = \sum_{j=1}^{m} \operatorname{tr} [G_j Y_j - \sum_{\ell \in I_j} G_\ell Y_\ell B_{\ell j}]' [\ldots\ldots]$$

Let us explain the idea behind this loss function.

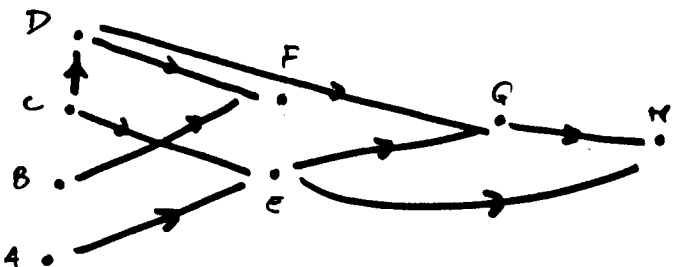
a) A causal model is a digraph.

b) Exogeneous variables are sources, or transmitters.

c) The relationship pictured in the graph is "is a direct cause of".

d) The chains in the graph define "is a cause of".

e) Causal graphs are transitive



	DIRECT CAUSES	CAUSES	LEVEL	PREDECESSORS
A	ϕ	ϕ	0	ϕ
B	ϕ	ϕ	0	ϕ
C	ϕ	ϕ	0	ϕ
D	C	C	1	A,B,C
E	A,C	A,C	1	A,B,C
F	B,D	C,B,D	2	A,B,C,D,E
G	D,E	A,C,D,E	2	A,B,C,D,E
H	G,E	A,C,D,E,G	3	A,B,C,D,E,F,G

What does a path diagram "mean" in semi-quantitative terms.

(a) A variable is **independent** of its predecessors, given its direct causes. Thus

$$E \perp_{A,C} B \qquad D \perp_{C} A, B$$

$$F \perp_{B,D} A, C, E \qquad G \perp_{D,E} A, B, C.$$

$$H \perp_{G,E} A, B, C, D, F$$

(b) Variables of the same level are **independent** given their predecessors

$$D \perp_{A,B,C} E \qquad F \perp_{A,B,C,D,E} G$$

This is still not entirely quantitative. We have to specify how we represent "variables" and how we mean "independence".

- var = subspace, ind = orth of subspaces.
  - var = vector, ind = orthogonality of vectors
    - var ∈ cone, ind = orthogonality of vectors
      - classical, log-linear analysis — identical for multivariate normal
      - single (rank-one) quantifications.

The prediction-oriented approach checks (by using, in our case, a least squares loss function) if, and in how far, a variable is in the space spanned by its direct causes.

Thus we count residuals to be zero, not orthogonal.

Key result: in recursive path models the predictive and model-fitting approaches for numerical variables are equivalent [Wold, 1966]. If there is missing information this is no longer true. Now missing information can be missing for various reasons

(a) proper
(b) non-numerical (ordinal, nominal)
(c) latent

The basic idea behind our approach to latent variables is that they are variables with a very primitive measurement level (they are continuous single nominal! no restriction at all).

This means that we can simply incorporate "latent" into our hierarchy of measurement levels.

Future: - Compare the various ways of making qualitative path models quantitative.

- Construct PATHALS algorithms, compare with LISREL and PLS. *

* Already been compared in detail, but not very well. The main difference is how the missing information is estimated. In PLS / GIFI by LS and OS ⇒ inconsistency

IN LISREL /ML by conditional expectation (EM - algorithm) ⇒ consistency

New version of Gifi [published by DSWO-press, estimated early 1987]
In three volumes.

Nonlinear Multivariate Data Analysis

Annotated References

Breiman, L., & Friedman, J.H. Estimating Optimal Transformations for Multiple Regression and Correlation. J. Amer. Statistical Association, 80, 1985, 580-619.

Advanced mathematical treatment of regression with optimal (smoothed) scaling by alternating least squares (called alternating conditional estimation).

Gifi, A. Non-Linear Multivariate Analysis. Leiden, The Netherlands: Dept. of Data Theory, University of Leiden. 1981.

Preliminary version of a book covering most nonlinear multivariate techniques. Uses ALS throughout.

Gittins, R. Canonical Analysis. Berlin: Springer-Verlag. 1985.

Reviews statistical and algorithmic theory for Canonical Analysis. Contains many ecological examples worked out in detail.

Greenacre, M.J. Theory and Applications of Correspondence Analysis. New York: Acedemic Press. 1984.

Discusses Corrrespondence Analysis from the French point of view. Comprehensive, including algorithmic and statistical aspects, with examples.

Heiser, W. Unfolding Analysis of Proximity Data. Leiden, The Netherlands: Dept. of Data Theory, University of Leiden. 1981.

Discusses Correspondence Analysis as a way of performing Multidimensional Unfolding (i.e., from the Multidimensional Scaling viewpoint).

Lebart, L., Morineau, A., & Warwick, K.M. Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices. New York: Wiley & Sons. 1984.

Translation of a French book on Correspondence Analysis (and cluster analysis). Elementary.

de Leeuw, J. Canonical Analysis of Categorical Data. Leiden, The Netherlands: DSWO-Press, 1985 (reprint of 1973 dissertation)

The progenitor of the current work on Nonlinear Multivariate Data Analysis. Algebraic.

de Leeuw, J. Nonlinear Principal Components Analysis. In: Caussinus, H. (ed.), COMPSTAT 1982, Vienna, Physica-Verlag, 1982.

Concise review of various approaches to nonlinear principal components analysis, including correspondence analysis.

de Leeuw, J. The Gifi system for nonlinear multivariate analysis. In: Diday, E. (ed.) Data Analysis and Informatics, Vol 3, Amsterdam: North Holland Publising Company. 1983.

Very concise review paper covering most of the material in this course. Discursive overview. Difficult but not mathematical.

Mardia, K.V., Kent, J.T., & Bibby, J.M. Multivariate Analysis. New York: Acedemic Press. 1979.

Covers nearly all the models discussed in this course, but without nonlinear data transformations. Advanced. Mathematical.

Marriott, F.H.C. The Interpretation of Multiple Observations. New York: Academic Press. 1974.

Beautifully written concise (117pp) overview of linear multivariate analysis. Requires elementary matrix algebra only.

Meulman, J. Homogeneity analysis of incomplete data. Leiden, The Netherlands: DSWO-press, 1982.

Clear introduction to homogeneity analysis (multiple correspondence analysis). Discusses missing data.

Perreault, W.D., & Young, F.W. Alternating Least Squares Optimal Scaling: Analysis of Nonmetric Data in Marketing Research. J. Marketing Research, 17, 1980, 1-13.

Overview of ALS methods as they apply to marketing research.

Stone, C.J. Additive Regression and other nonparametric models. Annals of Statistics, 13, 1985, 689-705.

Discusses nonlinear regression and conjoint analysis methods in a nonparametric statistical framework. Difficult and advanced.

Tenenhaus, M., & Young, F.W. An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and other methods for quantifying categorical multivariate data. Psychometrika, 50, 1985, 91-120.

Compares the many different proposals for analysis methods which are all equivalent to Correspondence Analysis.

Young, F.W., & Sarle, W.S. Exploratory Multivariate Data Analysis. Cary, NC: SAS Institute, Inc., 1983.

Notes for a course presented to those who wish to apply linear exploratory multivariate data analysis. Detailed examples. Little mathematics. Some nonlinear treatment.

Young, F.W. Quantitative Analysis of Qualitative Data. Psychometrika, 46, 1981, 357-388.

Overview of the first five years of developments in nonlinear multivariate data analysis. Algorithmic. Intermediate mathematics.

APPENDIX A:

ILLUSTRATIVE DATA SETS USED IN NLMVA COURSE

(1) 1980 Car data, taken from consumer reports, analyzed by Winsberg and Ramsay, our analysis by Van Rijckevorsel.

33 objects (cars)

5 variables: price in $
engine size
gas consumption city
gas consumption highway
weight

(2) baby data, taken from Shirley, The first two years, 1931. Analyzed by De Leeuw.

20 objects (babies)

71 variables (weeks) each with the categories stepping - standing - walking with help - walking alone - not yet stepping.

(3) skiing resort data, taken from tourist information, analyzed by Heiser.

115 objects (skiing resorts)

6 binary variables: suitable for beginners
good for walking
lots of amusement
for advanced skiers
round trips on skis
cross country skiing

(4) banks and industries, analyzed earlier by Levine, our analysis by Heiser.

cross table of 70 indutries and 14 banks, counting the number of board members they have in common.

(5) japanese religion data, taken from Sugiyama. Our analysis by Gifi, Heiser.

4243 objects

6 variables (religious practices): attending services
visiting graves
read religious books
visit shrines and temples
keep talisman
draw fortune

(6) American states data, from social indicator statistics, previously analyzed by Wainer, our analysis by De Leeuw and Meulman.

50 objects (states)

7 variables: population
income
illiteracy rate
life expectancy
homicide rate
percent high school graduates
number of days below freezing point

(7) Vlaardingen-Vlagtwedde data. Analysis by Van der Burg, Van Pelt.

4000 objects

variables: set 1: Vlaardingen -Vlagtwedde
set 2: smoking habits
- ever smoked
- rate of smoking
- period of smoking
- how long ago
set 3: background
- sex
- age
set 4: symptoms
- coughing
- phlegm
- dyspnoea
- wheezing
- asthma
set 5: maximum expiratory flow volume curve measurements
- PEF
- FVC
- etc

(8) uterine cervix data

126 objects (slides)

7 variables (pathologists) with categories - negative
- atypical squamous hyperplasia
- carcinoma in situ
- squamous carcinoma with early

stromal invasion
- invasive carcinoma

(9) Cities, villages, etc. Analyzed by De Leeuw.
792 objects (towns, cities, villages in Netherlands)
8 variables: social backwardness,
percentage lower technical education
percentage unemployed
public library present
number of types of schools
percentage of unemployed without proper schooling
percentage of young enemployed
number of journal/magazine subscriptions per family

(10) Event history data. Analyzed by De Leeuw, Van der Heijden, Kreft.
940 objects (individuals)
1440 variables (minutes) with categories - work, including school
- being at home
- shopping
- travelling
- other, including visits, sports, culture.

(11) Whale data. Analyzed by Van der Burg.
30 objects (whale , porpoise , dolphin species)
15 variables (neck, snaut, head, beak, dorsal fin, flippers, teeth, feeding, spouthole, colour, cervical vertebrae, lachrymal & jugal bones, habitat, throat, head bones).

(12) Suicide data. Taken from German social statistics, analyzed by Van der Heijden and De Leeuw.
about 50000 objects (sucessfull suicides)
three variables (sex, age, cause of death)

(13) Hartigan's Hardware. From the book by Hartigan, analyzed by Gifi.
24 objects (nails, tacks, screws, bolts).
6 variables (thread, head, head indentation, bottom, length, brass).

(14) Multiple choice. Data from Introductory Psychology Course, analysis by Gifi.
190 objects (students)
30 variables (items)

(15) Crime and fear. Data from Cozijn and Van Dijk, analyzed by Gifi, and by Van de Geer and Meulman.
1216 objects (respondents)
10 variables, first six from a questionaire on crime prevention methods, last four background
1. Re-ducation of offenders
2. Locking-up offenders
3. More severe punishments
4. Social work and rehabilitation
5. Labor camps
6. Better employment for potential offenders

First six variables five-point rating scaling from 5 = 'very effective' to 1 = 'very ineffective'.

7. Religion (5 cats)
8. Voting (10 cats)
9. Occupational status (high to low, 7 cats)
10. Age (low to high, 6 cats)

(16) Votes in Leiden. Analyzed by Van der Heijden.
Contingency table of 58 districts x 9 political parties, number of votes.

(17) Books. Constructed and analyzed by Gifi.
Contingency table of 20 books x 7 topics, number of pages.

(18) Intelligence vs schooling. Dutch Army Data, analyzed by Meester and De Leeuw.
Contingency table of 8 levels of schooling x 6 levels of intelligence, cells number of conscripts.

(19) Shoplifting data. Dutch Central Bureau of Statistics Data, analyzed by Israels and Sikkel.
Compound contingency table of (2 x 9) sex-age combinations by 13 types of goods stolen. Cells number of suspected persons.

(20) Spot patterns. Psychophysical Data from Guilford. Analyzed by Gifi. Contingency table with 23 cards with varying number of spots by nine 'equal-appearing' response categories.

(21) Munsingen-Rain data. Archeological data, due to Hodson. Analyzed by Meulman.
Presence-absence data. 70 varieties of objects are/are not in each of 59 graves.

(22) School-career data. From the SMVO-survey of the Central Bureau of Statistics.
5464 objects (pupils in secondary education)
5 variables describing school career, all with 25 categories
6 passive background variables
- province
- sex
- advice teacher
- occupational status father
- score arithmetic test
- score language test

(23) Journal preference data. Collected by Roskam. Analyzed by Gifi.
Ranking data. 39 psychologists rank 10 psychology journals.

(24) cylinder data. Using an idea of Thurstone. Analyzed by Gifi, and by Van Rijckevorsel.
20 objects (cylinders)
10 variables, all functions of height and radius

(25) macroeconomical variables. Analyzed by Van Kooten and Van Rijckevorsel.
31 objects (years),
6 variables (indices) - factor shares,
- unemployment
- profits
- institutional constraints
- prices
- public sector

(26) GLO 64-65 data. Collected by Dutch Central Bureau of Statistics. Analyzed by Meester and De Leeuw.
10455 objects (pupils)
6 variables - test score 6th grade
- teachers advice 6th grade
- choice of secondary education
- final form of secondary education
- occupational status father
- sex

(27) Gibbs-Wilson data. From the work of Willard Gibbs, analyzed by Gifi, analyzed earlier by E.B. Wilson.
65 objects (experiments)
3 variables - temperature
- pressure
- density

(28) Well-being research. Data collected by Dutch Central Bureau of Statistics. Analysis by Israels.
??? Objects
6 variables - satisfaction (5 cats)
- happiness (5 cats)
- marital status (4 cats)
- schooling (4 cats)
- income (4 cats)
- activity (7 cats)

(29) Comparative school careers. Two subsamples from the GLO 64-65 and the SMVO 1977 cohorts, data by Netherlands Central Bureau of Statistics. Analysis by De Leeuw, Van der Burg, and Bettonvil.
1788 + 1519 objects (pupils)
6 variables: choice of secondary education (5 cats)
test score 6th grade (6 cats)
advice teacher 6th grade (5 cats)
school level father (7 cats)
school level mother (7 cats)
fathers profession (7 cats)

(30) Russett data. Analyzed by Gifi.
47 objects (countries)
9 variables - Gini index of income inequality
- percentage of farmers owning half of the land, starting from below
- percentage of farmers that rent all their land
- gross national product per capita
- percentage of labour force in agriculture
- average number of years in office of first executive
- total number of violent internal war accidents in 1946-1961.
- total number of people killed due to internal conflinct
- stable democracy, unstable democracy, dictatorship

(31) POLP data, interview of members of Dutch parliament. Analyzed by Van der Burg and De Leeuw.
138 objects (MP's)
7 + 4 variables. First seven nine-point rating scales, measuring attitude towards development aid, abortion, law and order, income differences, participation of workers in industrial decision making, tax increases, money for defence. Next four: rank number of PvdA (socialists), ARP (christian democrats, protestant), KVP (christian democrats, catholic), VVD (liberals, conservative) in a rank order of all 15 parties.

(32) Population and emancipation. Data of the Netherlands Social Cultural Planning Bureau, analyzed by Kreft en De Leeuw.
4693 objects (random sample)
4 + 2 + 2 variables. First four attitude questions concerning birth control, and the role of government in giving information about birth control. Then the background variables sex x political choice (2 x 6 categories) and age (7 categories), and two attitude questions concerning emancipation.