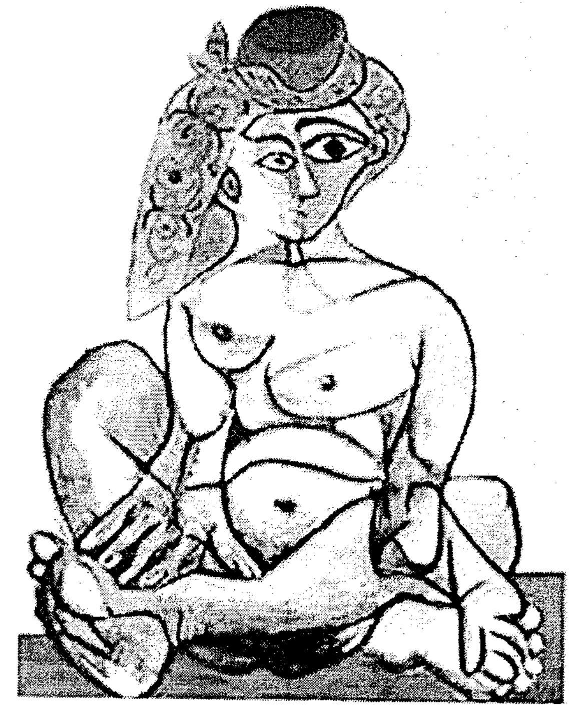
Non Symmetrical Data Analysis
New Methods and Applications

Carlo Lauro
Dipartimento di Matematica e Statistica
Università di Napoli "Federico II"

Monna-Lisa
by Leonardo

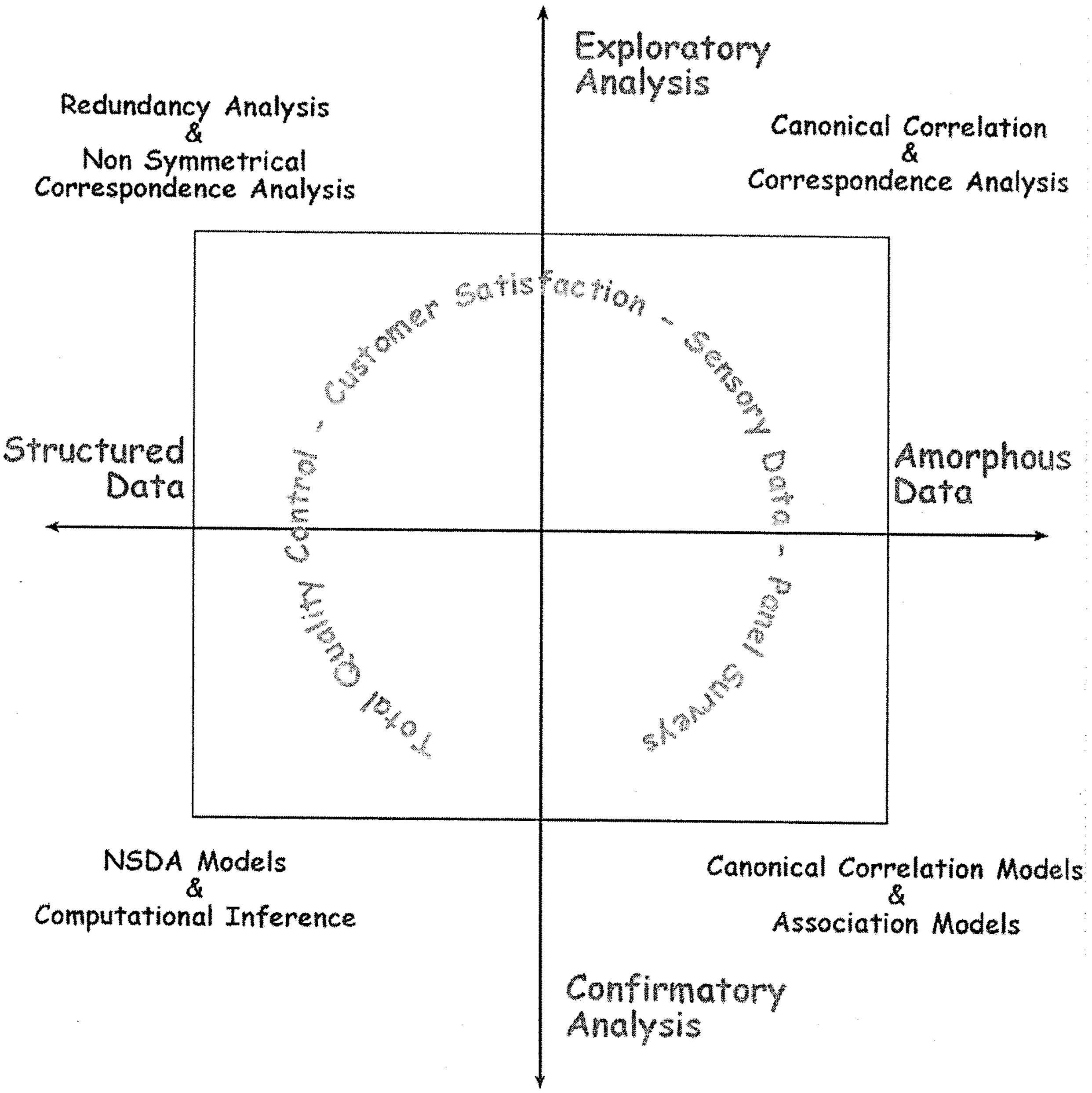
**NSDA ...
a different
point of view!!!**



designed by Enrico Cafaro

Nude woman with the Turkish cap
by P. Picasso

New Trends in Methods and Applications of Non Symmetrical Data Analysis



Exploratory Analysis

Redundancy Analysis & Non Symmetrical Correspondence Analysis

Canonical Correlation & Correspondence Analysis

Structured Data

Amorphous Data

Total Quality Control - Customer Satisfaction - Sensory Data - Panel Surveys

NSDA Models & Computational Inference

Canonical Correlation Models & Association Models

Confirmatory Analysis

EXPLORATORY ANALYSIS OF TWO OR MORE SETS OF VARIABLES

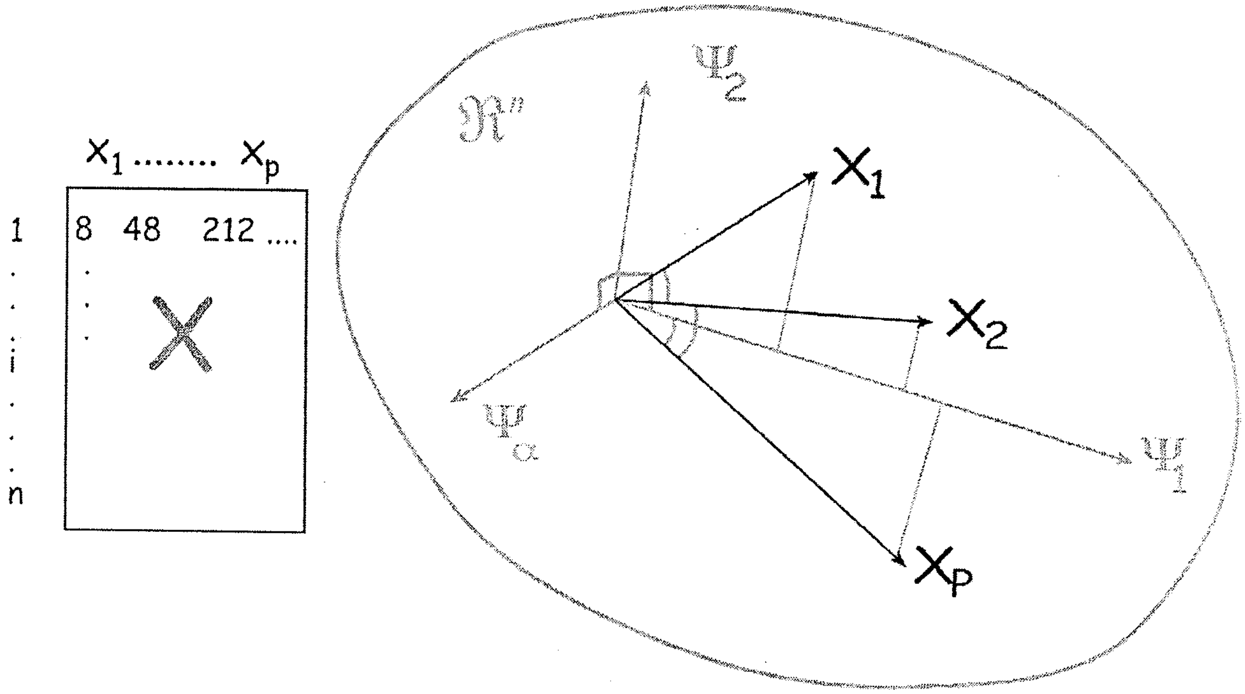
VARIABLES	SYMMETRICAL APPROACHES	NON SYMMETRICAL APPROACHES
Quantitative Two Sets	Canonical Correlation Analysis (Hotelling, 1936)	Principal Component Analysis of Instrumental Variables (Rao, 1964; Robert, Escoufier, 1976) Redundancy Analysis (Gleason, 1976; Van den Wollenberg, 1977) Principal Component Analysis onto a Reference Subspace (D'Ambra, Lauro, 1982a) Explanatory PCA (Obadia, 1982) Factorial Analysis of Structured Data (Sabatier, 1987) PLS2 (Tenenhaus, 1995)
More Sets	Generalised Canonical Correlation Analysis (Carrol, 1968; Kettenring, 1971) Foundations of MVA (Takeuchi, Yanai, Mukherjee, 1982)	Principal Component Analysis onto more than one Reference Subspace (D'Ambra, Lauro, 1982b)
Qualitative Two Sets	Method of Reciprocal averages (Horst, 1935) Optimal quantification theory (Guttman, 1941; Hayashi, 1950, 1952) Correspondence Analysis (Escofier, 1965; Benzécri, 1973)	Non symmetrical Correspondence Analysis (Lauro, D'Ambra, 1984) Redundancy analysis for Qualitative variables (Isräels, 1984) Canonical correspondence Analysis (Ter Braak, 1986)
More Sets	Multiple correspondence analysis (Benzécri, 1972; Lebart, 1975) Optimal quantification theory (Hayashi, 1952) Generalisations of CA in terms of projection operators (Yanai, 1986)	Non symmetrical multiple Correspondence analysis (Lauro D'Ambra, 1984) Redundancy analysis for Qualitative variables (Isräels, 1987)

Symmetrical Analyses

Principal Component Analysis (PCA, Pearson 1901)

Data Structure

1 set X of p **Numerical** Variables observed on n S.U.



Criterion

To adapt a line or a plane to a cloud of points in a hyperspace by finding linear combinations $\Psi_\alpha = Xu_\alpha$ α=1,...,p of the variables in X taking into account most of the variance of the variables themselves

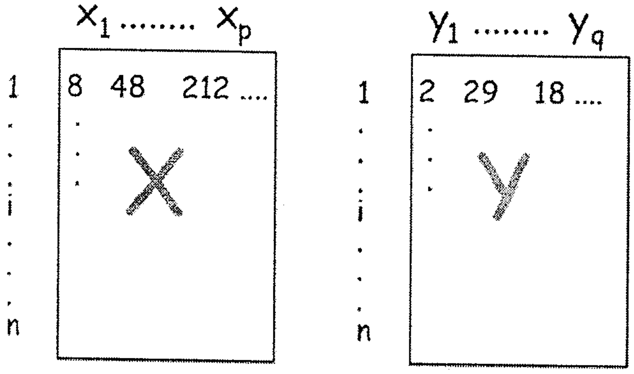
Solution

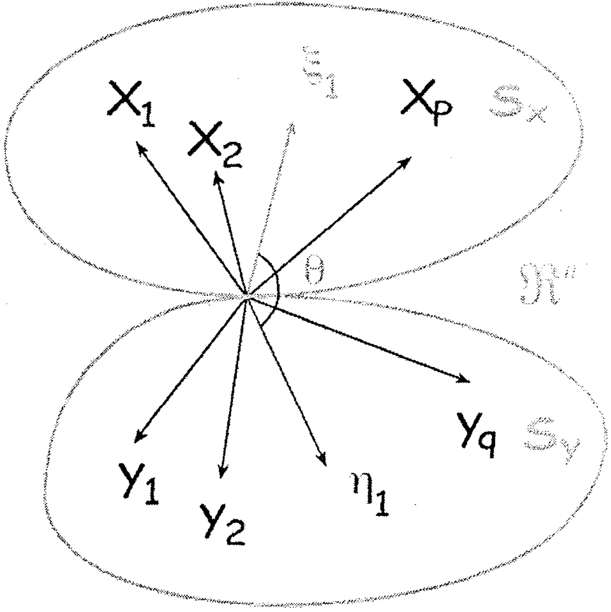
$$X'Xu_\alpha = \lambda_\alpha u_\alpha$$

Canonical Correlation Analysis (CCA, Hotelling 1936)

Data Structure

2 sets X and Y of, respectively, p and q **Numerical** Variables





Criterion

To find those linear combinations, $\xi_\alpha = Xa_\alpha$ and $\eta_\alpha = Yb_\alpha$, of the variables in X and Y showing the highest **correlation** coefficient among them

Solution

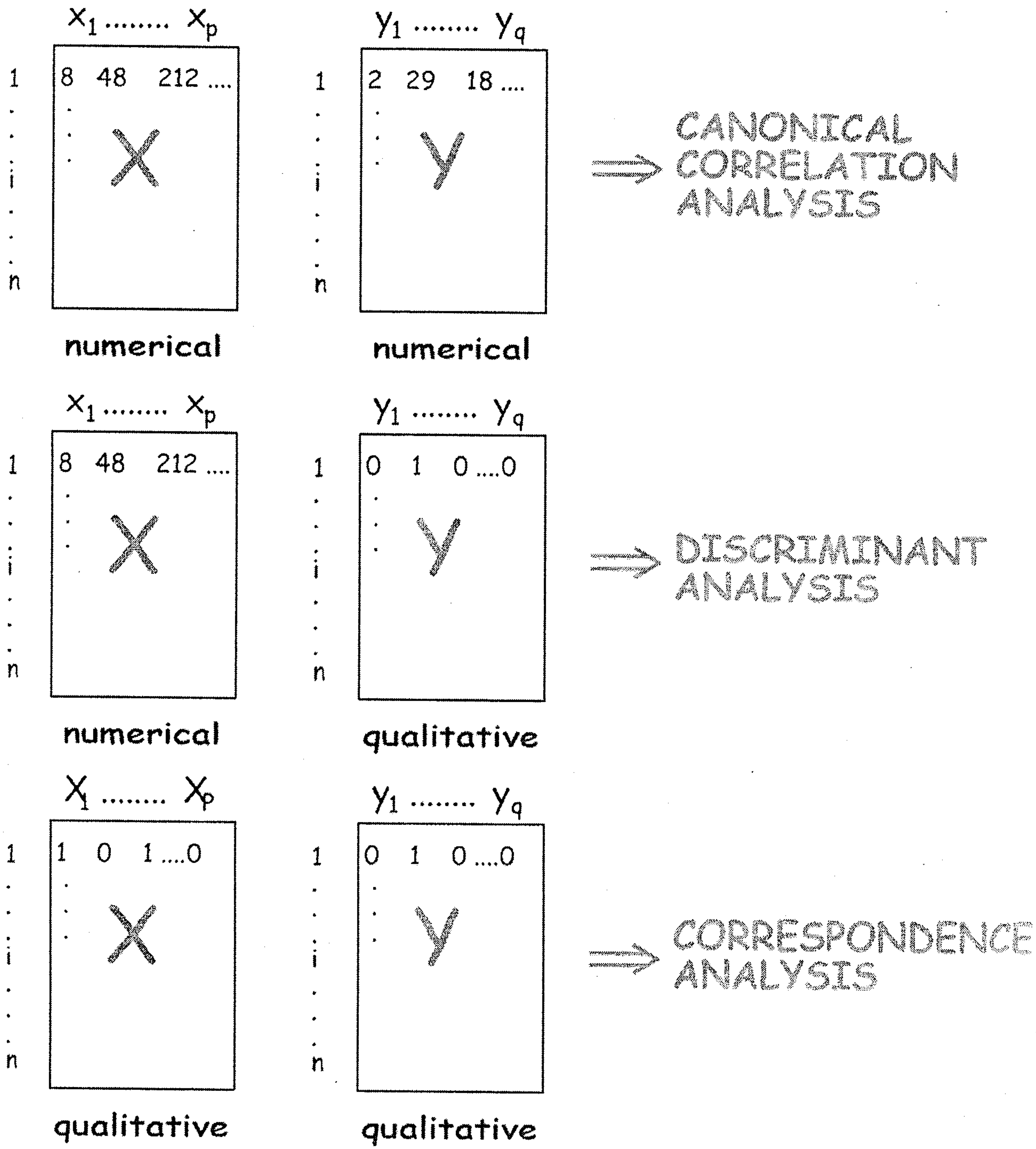
$$(X'X)^{-1}(X'Y)(Y'Y)^{-1}(Y'X)a_\alpha = \gamma_\alpha^2 a_\alpha$$

$$(Y'Y)^{-1}(Y'X)(X'X)^{-1}(X'Y)b_\alpha = \delta_\alpha^2 b_\alpha$$

with $\gamma = \delta = \cos\theta$

Note: Both the variables in X and Y are centred and divided by $n^{\frac{1}{2}}$ so that cross-product matrices define covariance matrices.

SPECIAL CASES OF CCA



⟹ CANONICAL CORRELATION ANALYSIS

⟹ DISCRIMINANT ANALYSIS

⟹ CORRESPONDENCE ANALYSIS

Non-Symmetrical Alternatives

PCA with Instrumental Variables (PCAIV, Rao 1964)

Data Structure

As in CCA but Y is considered to be **instrumental** for the explanation of X

Criterion

To substitute, moving from the joint-dispersion matrix, Y with a **lower-dimensional** matrix $T'Y$ that maximises its predictive efficiency for X

Solution

$$(Y'Y)^{-1}(Y'X)(X'Y)f_\alpha = \lambda_\alpha f_\alpha$$

so to minimise the residual-dispersion matrix

Problem

Clearly set in a Multivariate **Regression**-like framework so that no **geometrical** interpretation is available

Redundancy Analysis

(RA, van den Wollenberg 1977)

Criterion

Maximising the explained **variance** of the variables in X through the maximisation of the Redundancy Index (Stewart and Love, 1968):

$$\frac{tr(Y'X)(X'X)^{-1}(X'Y)}{tr(Y'Y)}$$

i.e. the average variance of the X variables accounted by the Y variables

Solution

RA comes to the same solution of PCAIV

Problem

As an alternative to CCA, RA suffers from the same interpretation drawbacks

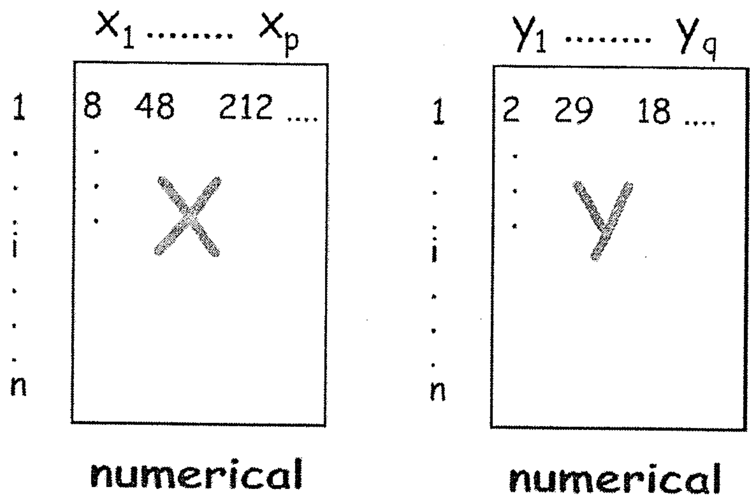
PCA onto a Reference Subspace

(PCAR, D'Ambra and Lauro 1982)

Data Structure

1 set X of p **Dependent** Variables

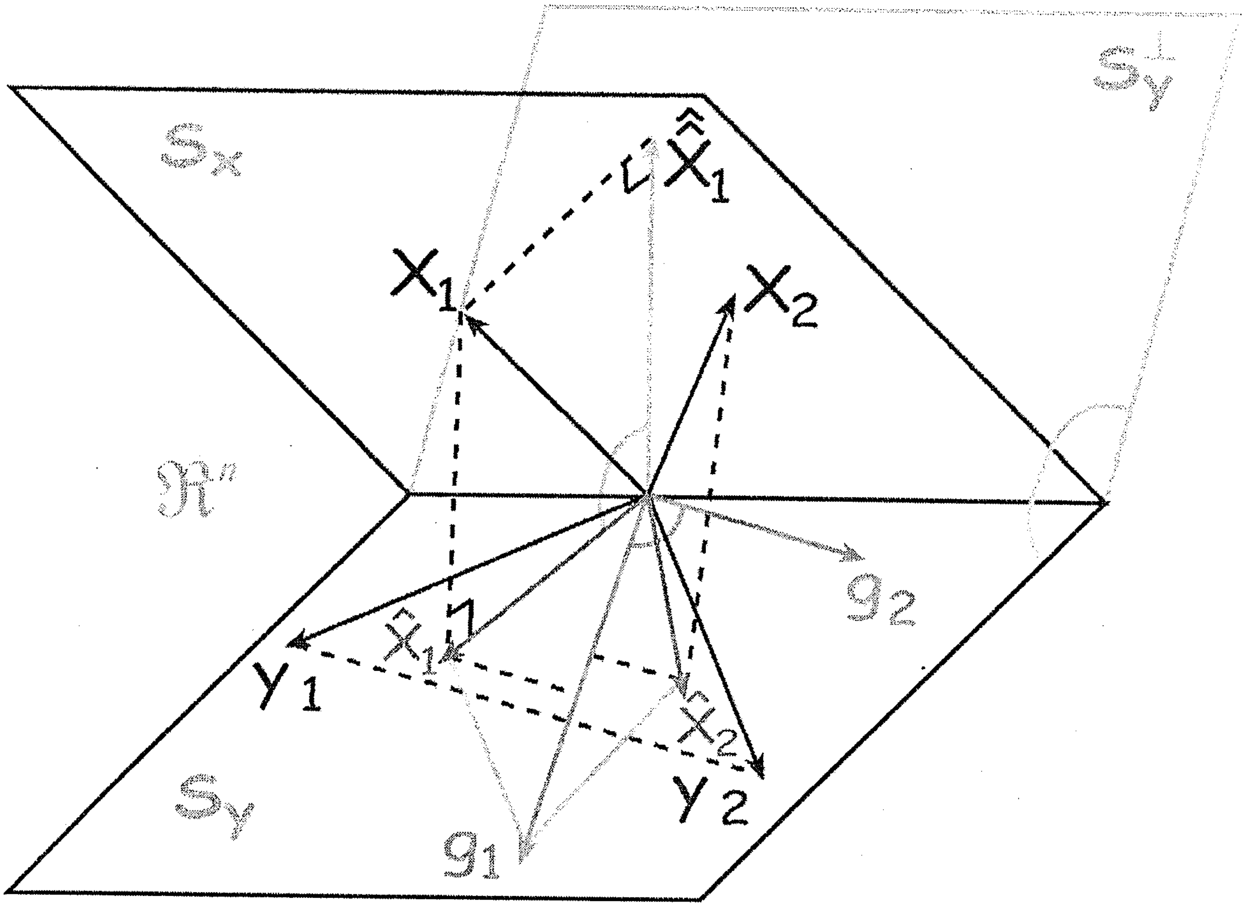
1 set Y of q **Explanatory** (or Reference) Variables



Criterion

PCA of a suitable **image** of the variables in X obtained on the **reference** subspace by means of orthogonal **projection** operators

Geometrical Insight



Solution

$$X'P_yXg_\alpha = \lambda_\alpha g_\alpha$$

with k=1...min(rank(X), rank(Y))

where $P_y = y\left(y'y\right)^{-1}y'$

Variance Decomposition

$$\left(X'X\right) = \left(X'P_yX\right) + \left(X'X - X'P_yX\right)$$

Remarks

- The preliminary **geometrical** transformation of projection in PCAR may be meant as an optimal, or at least coherent, quantification or coding of the variables in according with the objective of the analysis

- The **predictive** efficiency in PCAR is measured by the percentage of explained variance $tr(X'P_YX)/tr(X'X)$ and this portion is just the one decomposed by PCAR.

- With respect to the PCAIV solution, PCAR has the same non-trivial eigenvalues, and simple relations exist among the eigenvectors **but** PCAR provides useful geometrical interpretation tools as well as the possibility to be generalised to more than 2 sets of variables.

PCAR Biplot Representations

As the matrix X'P_YX is **symmetric**, the following relations hold:

$$\mathbf{g}_\alpha'\mathbf{g}_\alpha = \lambda_\alpha \quad \text{and} \quad \mathbf{g}_\alpha'\mathbf{g}_{\alpha'} = 0 \quad \forall \alpha \neq \alpha'$$

- **Principal Components** calculation: $\mathbf{c}_\alpha = \mathbf{P}_Y\mathbf{X}\mathbf{g}_\alpha$

- For **interpretative** scopes, it is very helpful to set both statistical units and variables in the same geometry. Therefore, we ensure that the graphical display in the reduced space is a **biplot** by dealing with components normalised to 1:

$$\mathbf{c}_\alpha^* = \mathbf{P}_Y\mathbf{X}\mathbf{g}_\alpha / \sqrt{\lambda_\alpha}$$

- **Dependent** variables co-ordinates, as the correlation between the variables themselves and the principal components:

$$\mathbf{X}'\mathbf{c}_\alpha^* = \mathbf{X}'\mathbf{P}_Y\mathbf{X}\mathbf{g}_\alpha / \sqrt{\lambda_\alpha} = \mathbf{g}_{\underset{\alpha}{\underline{k}}}\sqrt{\lambda_\alpha}$$

- **Explanatory** variables co-ordinates:

$$\mathbf{Y}'\mathbf{c}_\alpha^* = \mathbf{Y}'\mathbf{X}\mathbf{g}_\alpha / \sqrt{\lambda_\alpha}$$

correlaciones entre las Y y las C.P.$_\alpha$ $= P_Y C_\alpha$

" " " " X $= P_{YX} g_\alpha$

Interpretation Property

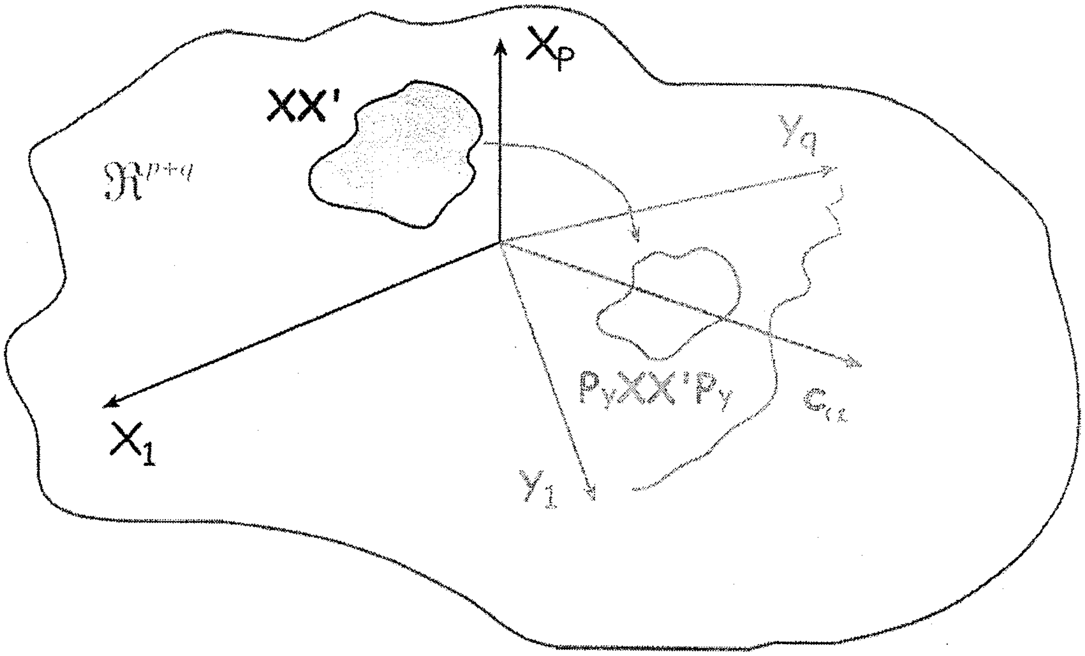
The image of the correlations between the two sets X and Y is reconstructed directly on the principal axes.

Differently from CCA, we are enabled to read on a unique factorial plane both the internal and external correlations

- The α-th **eigenvalue** of PCAR is the sum of squared correlations between the variables in X and the principal components of the reference subspace. Therefore, it represents the **explanatory power of the principal component**, associated with the k-th eigenvector, with respect to the dependent variables

$\lambda_\alpha = \sum_{j=1}^{p} r^2_{x_j c_\alpha}$ con $\boxed{c_\alpha = P_q^A g_\alpha}$

PCAR Dual Analysis



Characteristic Equation

$$P_y XX' P_y X g_\alpha = \lambda_\alpha P_y X g_\alpha$$

Remark 1

The **Principal Vectors**

$$P_y X g_\alpha$$

represent the **Principal Components** relative to the variables space

Remark 2

The **Principal Vectors** can be expressed as linear combinations of the variables in Y:

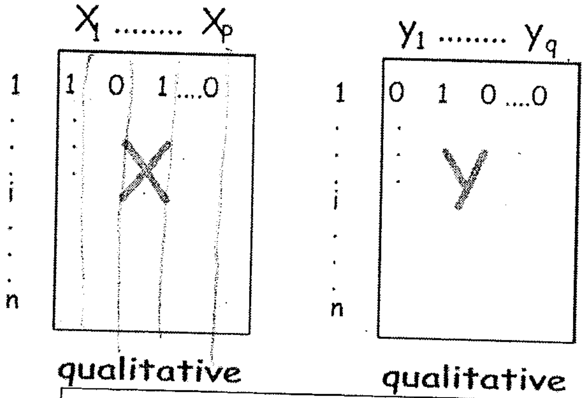
$$c_\alpha = P_y X g_\alpha = y\left[(y'y)^{-1} y' X g_\alpha\right] = y z_\alpha$$

PCAR for Different Data Structures

Dependent Variables	Explanatory Variables	Analyses
Numerical	Numerical	PCAR
Nominal	Nominal	Non Symmetrical Correspondence Analysis
Ordinal (Ranks)	Nominal (Experimental Design)	Factorial Conjoint Analysis
Numerical	Nominal (Experimental Design)	MANOVA Total Quality Control

Non Symmetrical Correspondence Analysis (NSCA) for Two Sets of Binary Dummy Variables (Lauro et al., 1984)

Data Structure



Characteristic Equation

$$\frac{1}{n}(X'P_yX - X'P_mX)u_\alpha = \mu_\alpha u_\alpha$$

with α=1...min(rank(X), rank(Y)) and P_m is the centring matrix

Interpretation

From the Huyghens theorem,
the total variability of X is decomposed as:

$$\frac{1}{n}(X'X - X'P_mX) = \frac{1}{n}(X'P_yX - X'P_mX) + \frac{1}{n}(X'X - X'P_yX)$$

Therefore, **NSCA aims at decomposing:**

$$\tau_{X.y} = \frac{tr(X'P_yX - X'P_mX)}{tr(X'X - X'P_mX)} = \frac{Explained Variability}{Total Variability}$$

Expression of Goodman-Kruskal's association index

Non Symmetrical Multiple Correspondence Analysis

(NS-MCA, Lauro et al., 1984, 1989, 1992)

Limitations of the classical MCA

- Symmetry (Survey Analysis)
- Interactions (Burt's Table diagonalizations)

3 Qualitative Variables X Y_1 Y_2

a) S_{Y1} and S_{Y2} are DISJOINTS

$$P_m^{\perp} X = P_m^{\perp} P_{Y1} X + P_m^{\perp} P_{Y2} X + P_m^{\perp} (\mathcal{I} - P_{Y1} - P_{Y2}) X$$

$$\boxed{\frac{1}{n} X' \left[\sum_{j=1}^{2} \left(P_{Y_j} - P_m \right) \right] X}$$

Remark: Extensions to More Sets of Variables

$$\frac{1}{n} X_i' \left[\sum_{j=1}^{q} \left(P_{Y_j} - P_m \right) \right] X_i \Rightarrow \sum_{j=1}^{q} P_{Y_j} \sum_i X_i X_i'$$

b) Analysis onto the Cartesian Product Space S_{Y12} by means of the Projection Operator:

$$P_{Y_{12}} = Y_{12} \left(Y_{12}' Y_{12} \right)^{-1} Y_{12}'$$

$$\boxed{\frac{1}{n} X' \left(P_{Y_{12}} - P_m \right) X}$$

Gray Williams' Multiple τ

c) Analysis onto the Cartesian Product Space S_{Y12} Orthogonal to S_{Y1}:

$$\boxed{\frac{1}{n} X' P_m^{\perp} \left(P_{Y_1}^{\perp} P_{Y_{12}}\right) X = \frac{1}{n} X' \left(P_{Y_{12}} - P_{Y_1}\right) X}$$

Gray Williams' Partial τ

d) Analysis onto the Interaction Sub-Space $S_{Y_1 \otimes Y_2}$:

$$\boxed{P_{Y_1 \otimes Y_2} = P_{Y_{12}} - P_{Y_1} - P_{Y_2} + P_m}$$

e) Analysis onto the Union Sub-Space $S_{Y_1 \cup Y_2}$:

$$\boxed{P_{Y_1 \cup Y_2} = P_{Y_1} + P_{Y_2 / Y_1} = P_{Y_2} + P_{Y_1 / Y_2}}$$

where $P_{Y_1 / Y_2} = P_{Y_2} Y_1 \left(Y_1' Y_1\right)^{-} Y_1' P_{Y_2}$

... Extensions $X_{i_1 \ldots i_p}$ with $i_1 \times i_2 \times \ldots \times i_p$ categories

NSCA for Contingency Tables

(Lauro et al., 1984)

Notations

$$F = \frac{1}{n} X'Y \qquad D_p = \frac{1}{n} X'X \qquad D_q = \frac{1}{n} Y'Y$$

$$f_p' = \left[f_{1.}, \dots, f_{p.}\right] \qquad f_q' = \left[f_{.1}, \dots, f_{.q}\right]$$

- NSCA studies the q **conditional distributions** fij/f.j with reference to the **independence hypothesis** fi.
- From a **geometrical view point** NSCA aims at studying the spread of the q column points around their centroid f_p in the space $\mathbb{R}^{p}$ spanned by the rows of $\mathbf{FD}_q^{-1}$

Interpretation

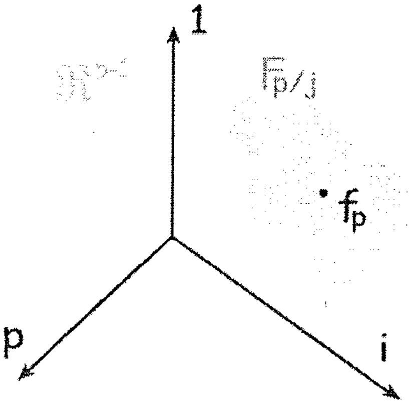
$$\tau_{X.Y} = \frac{tr\left(FD_q^{-1}F' - f_p f_p'\right)}{tr\left(D_p - f_p f_p'\right)}$$

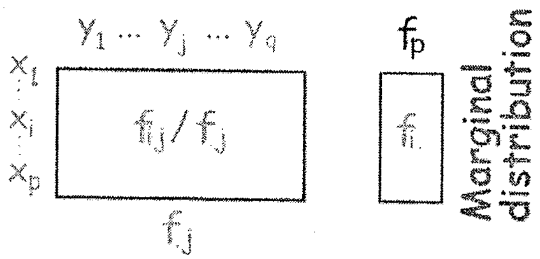
Solution

Eigen-analysis of the matrix:

$$FD_q^{-1}F'$$

Simple NSCA



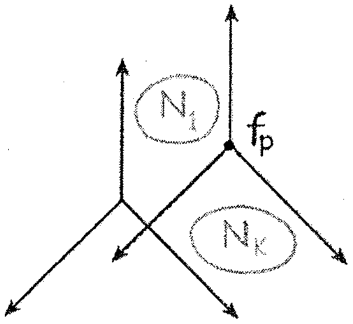


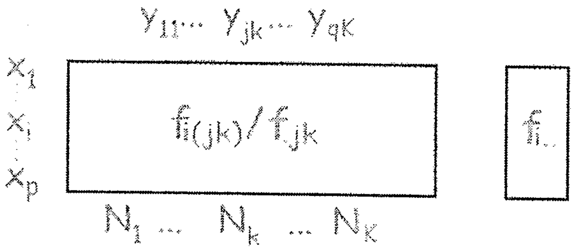
Profile matrix

	Metric	Criterion
NSCA ⟹	Euclidean	Goodman-Kruskal's τ
BCA ⟹	Chi-square	Pearson's φ^2

Multiple NSCA

Analysis with respect to the common centroid f_p



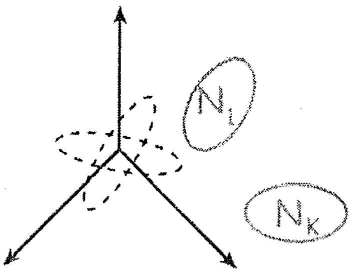

$y_{11} \cdots y_{jk} \cdots y_{qK}$
x_1
x_i
x_p
$f_{i(jk)}/f_{.jk}$
$f_{i..}$
$N_1 \ldots N_k \ldots N_K$

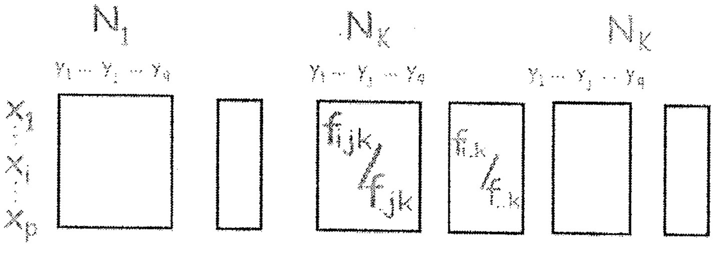
Weights $f_{.jk}$

Criterion: Gray-Williams Multiple τ

Partial NSCA

Analysis with respect to the strata centroids



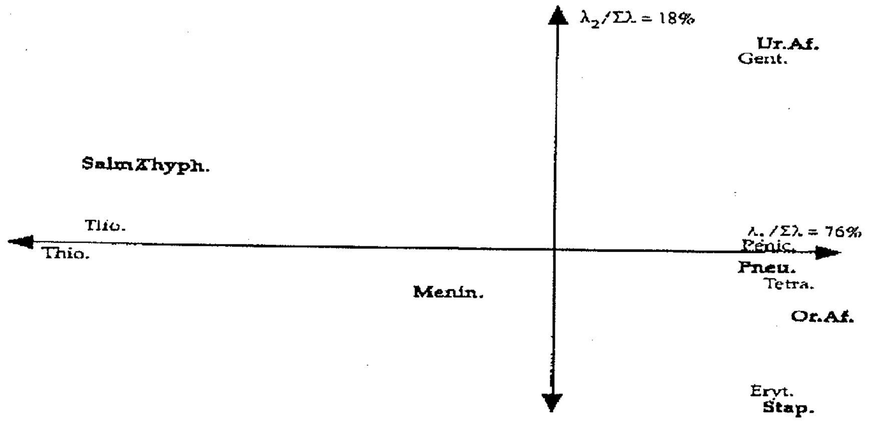

N_1
N_k
N_K
$y_1 \cdots y_j \cdots y_q$
$y_1 \cdots y_j \cdots y_q$
$y_1 \cdots y_j \cdots y_q$
x_1
x_i
x_p
$f_{ijk}/f_{.jk}$
$f_{i.k}/f_{..k}$

Criterion: Gray-Williams Partial τ

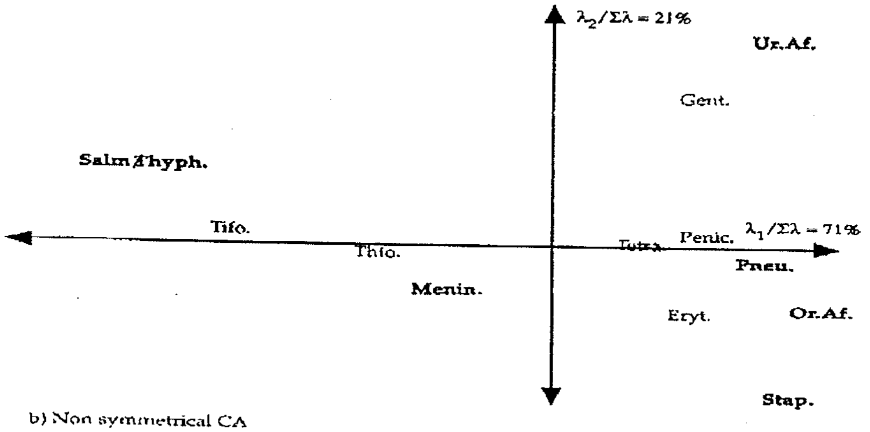
An Example on Medical Data

Simple NSCA

Disease ~ Medicine	Typh.	Salmon.	Oral Af.	Pneumo.	Mening.	Urin. Af.	Staphil.	Total
Penicillin	0	0	8	7	2	4	3	24
Typhom.	4	2	0	0	2	0	0	8
Tetracycl.	0	0	5	5	0	2	1	13
Erythroc.	0	0	3	2	0	0	3	8
Thioph.	2	1	0	0	0	0	0	3
Gentam.	0	0	3	3	1	6	0	13
Total	6	3	19	17	5	12	7	69



a) Symmetrical CA



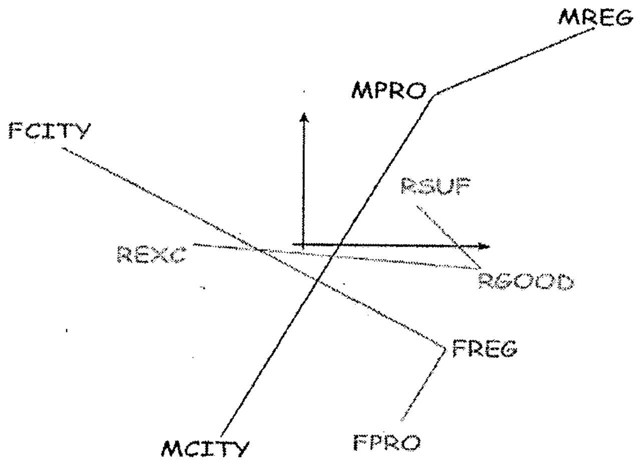
b) Non symmetrical CA

An Example on School Success

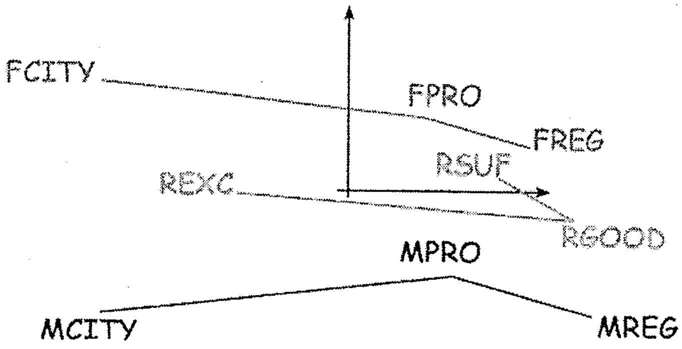
Multiple and Partial NSCA

	Male			Female			
Residence ~ Grade	CITY	PROV.	REGION	CITY	PROV.	REGION	Total
Sufficient	6	4	5	1	1	1	18
Good	17	8	9	3	3	4	44
Excellent	19	5	3	10	2	1	40
Total	42	17	17	14	6	6	102

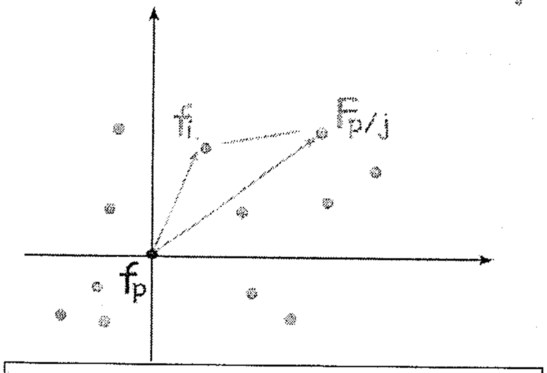
Multiple NSCA



Partial NSCA
sex stratum



Interpretation of NSCA Displays



Distance from the origin

- A row-modality (marginal) far from the origin indicates a **dependence** of that modality **from** the column-character
- A column-profile far from the origin indicates a great **influence** of the related column-modality on the behaviour of the dependent (row) variable

Distance between rows

Analogies with respect to column character **dependency**

Distance between columns

Similar profiles indicate similar influence on the dependent variable

Angles *(no distances)* between row & column profiles

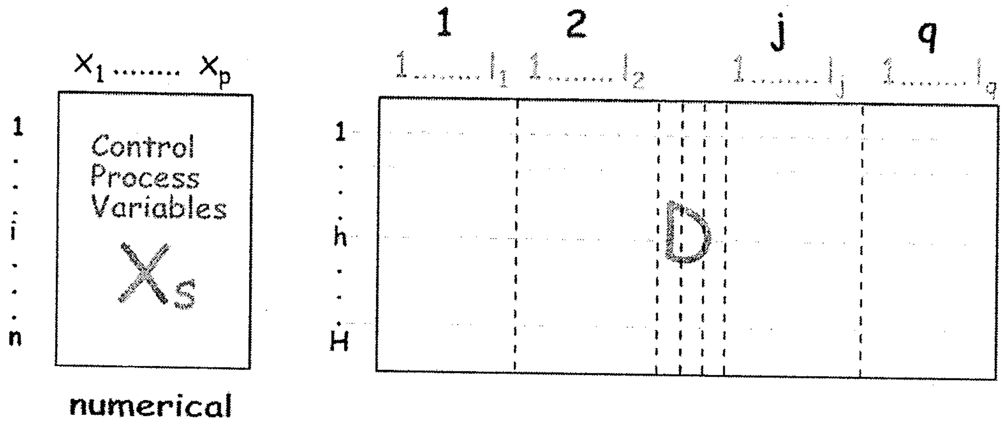
A **large cosine** indicates strong influence of the column-modality on the row-modality

Comparison of descriptive factorial methods for two-way contingency tables analysis

	Correspondence Analysis (CA)	*Non-symmetrical CA*
Analysis in $\mathbb{R}^{p-1}$		
Coordinates	$f_{ij}/f_{.j}$	$f_{ij}/f_{.j}$
Weight	$f_{.j}$	$f_{.j}$
Distance from centre of mass	$\sum_i \frac{1}{f_{i.}}\left(f_{ij}/f_{.j} - f_{i.}\right)^2$	$\sum_i \left(f_{ij}/f_{.j} - f_{i.}\right)^2$
Metric	χ^2	Euclidean
Inertia	$\sum_j f_{.j}\sum_i \frac{1}{f_{i.}}\left(f_{ij}/f_{.j} - f_{i.}\right)^2$	$\sum_j f_{.j}\sum_i \left(f_{ij}/f_{.j} - f_{i.}\right)^2$
Index	$0 \leq \phi^2 \leq \min(p,q)-1$	$0 \leq \tau \leq f_{i.} \leq 1$
Characteristic Equations	$\sum_j f_{.j}\sum_{i'} \frac{1}{f_{i'.}}(\frac{f_{ij}}{f_{.j}} - f_{i.})(\frac{f_{ij}}{f_{.j}} - f_{i'.})u_{i'\alpha}$ $= \lambda_\alpha u_{i'\alpha}$	$\sum_j f_{.j}\sum_{i'} (\frac{f_{ij}}{f_{.j}} - f_{i.})(\frac{f_{ij}}{f_{.j}} - f_{i'.})u_{i'\alpha}$ $= \lambda_\alpha u_{i'\alpha}$
Constraints	$\mathbf{u}'_\alpha \mathbf{D}_p \mathbf{u}_\alpha = 1, \mathbf{u}'_\alpha \mathbf{D}_p \mathbf{u}_{\alpha'} = 0$	$\mathbf{u}'_\alpha \mathbf{u}_\alpha = 1, \ \mathbf{u}'_\alpha \mathbf{u}_{\alpha'} = 0$
Factorial Coordinates	$\psi_\alpha = \sqrt{\lambda_\alpha}\mathbf{D}_p^{-1}\mathbf{u}_\alpha$	$\psi_\alpha = \sqrt{\lambda_\alpha}\mathbf{u}_\alpha$
Transition formulae to $\mathbb{R}^{q-1}$	$\mathbf{v}_\alpha = 1/\sqrt{\lambda_\alpha}\,\mathbf{F}'\mathbf{D}_p^{-1}\mathbf{u}_\alpha$	$\mathbf{v}_\alpha = 1/\sqrt{\lambda_\alpha}\,\mathbf{F}'\mathbf{u}_\alpha$
Constraints $\mathbb{R}^{q-1}$	$\mathbf{v}'_\alpha \mathbf{D}_q \mathbf{v}_\alpha = 1, \mathbf{v}'_\alpha \mathbf{D}_q \mathbf{v}_{\alpha'} = 0$	$\mathbf{v}'_\alpha \mathbf{D}_q \mathbf{v}_\alpha = 1, \mathbf{v}'_\alpha \mathbf{D}_q \mathbf{v}_{\alpha'} = 0$
Coordinates $\mathbb{R}^{q-1}$	$\varphi_\alpha = \sqrt{\lambda_\alpha}\mathbf{D}_q^{-1}\mathbf{v}_\alpha$	$\varphi_\alpha = \sqrt{\lambda_\alpha}\mathbf{D}_q^{-1}\mathbf{v}_\alpha$
Reconstruction Formulae	$f_{ij} = f_{.j}f_{i.}\left(1 + \sum_{\alpha}^{M^*} \lambda_\alpha \mathbf{u}_{\alpha i}\mathbf{v}_{\alpha j}\right)$	$f_{ij} = f_{.j}\left(f_{i.} + \sum_{\alpha}^{M^*} \lambda_\alpha \mathbf{u}_{\alpha i}\mathbf{v}_{\alpha j}\right)$

Principal Matrices Analysis onto an Experimental Design *(PMAD, Lauro et al., 1997)*

Data Structure


1
2
j
q
1........l1 1........l2 1........lj 1........lq
x1 .... xp
1 . . i . . n
Control Process Variables
Xs
1 . . h . . H
D
numerical

s= 1...S control samples or bootstrap replications

h= 1...H Experimental **conditions**

j= 1...q Experimental **factors** $\sum_{j=1}^{q} l_j = L$ levels of the factors

Aim

To build **non parametric control charts** taking into account of a possible different behaviour of the control variables through representing the statistical units **rearrenged** according to their own experimental pattern

$$X_s^* = T_s^{-1} M_s' X_s$$

where $M_{ihs} = \begin{cases} 1 \text{ if the i - th unit} \in \text{h - th condtion} \\ \quad 0 \text{ otherwise} \end{cases}$

and $T_s = diag(M_s' 1)$

PMAD Solution

Step 1

Each X_s^* is **projected** onto the subspace spanned by the columns of the **experimental matrix D**:

$$A_s = P_s X_s^*$$

where $P_s = D(D'T_sD)^{-1}D'T_s$

Step 2

The **three-way structure** determined by the A_s's is analysed by a Principal Matrices Analysis.
In particular, **Co-Chart** are built for the experimental conditions based on:

$$Co = \sum_{s=1}^{S} u_{r_b} X_s^*$$

where u_{r_b} is the b-th element of the eigenvector corresponding to the r-th eigenvalue of the so-called **matrix IS** having as general element:

$$tr\left(X_s^{*'} X_s^*\right)$$

A PMAD Application: Buffon's Beams

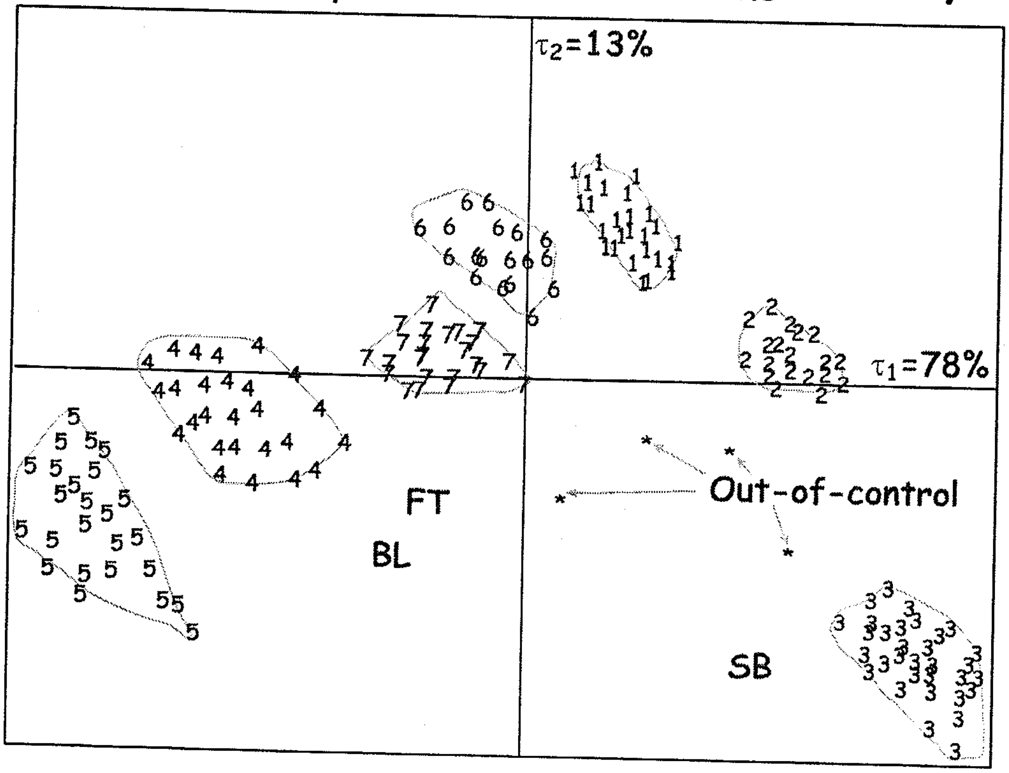
76 Items

6 variables:

Factors	Width	2 levels
	Length	2 levels
	Weight	3 levels
Responses	Breaking Load	
	Failure Time	
	Sag of the Beam at First Crack	

Experimental Conditions: **12**

Observed Experimental Conditions: **7**



Control Chart based on experimental design
95% limits,200 bootstrap replications

A Multidimensional Approach to Conjoint Analysis

(Lauro et al., 1997)

- **Conjoint Analysis** deals with preference judgements, expressed by **individuals** (judges) about a set of **stimuli** (products or services). Stimuli are described by several attributes at different levels

- Conjoint Analysis aims at evaluating the **relative importance** of the levels of each attribute in establishing the known global preference associated to the different stimuli, by means of a decompositive approach

- We mainly cope with the so called **Metric Conjoint Analysis** approach in which the multiple linear regression model is used in order to estimate the **part-worth coefficient** of each level

In order to improve the **interpretation** of the Conjoint Analysis results we propose an alternative approach to CA in the context of

Multidimensional Data Analysis

- Optimal Synthesis of the Conjoint Analysis results
- Geometrical approach
- Visualization and interpretation of subsets of models

The individual regression model estimated by the

Metric approach to Conjoint Analysis is:

$$\boxed{X_p = b_{p1}D_1 + \ldots + b_{pl}D_l + \ldots + b_{pL}D_L + e_p}$$

where:

- X_p is the centred preference vector of the *g*-th judge;
- D_l is the *l*-th level of the generic attribute;
- b_{pl} is the individual *utility coefficient* for the level D_l;
- e_p is the error term;

Data Structure

Design Matrix

n Stimuli; L Levels;   q Attributes

$$L = \sum_{j=1,q} l_j$$

$$\mathbf{D} = \begin{array}{c|ccc:ccc:c:ccc}
 & D_1^1 & \cdots & D_{l_1}^1 & D_1^2 & \cdots & D_{l_2}^2 & \cdots & D_1^q & \cdots & D_{l_q}^q \\
\hline
S_1 & 1 & \cdots & 0 & 0 & \cdots & 1 & \cdots & 1 & \cdots & 0 \\
S_2 & 0 & \cdots & 1 & 1 & \cdots & 0 & \cdots & 0 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
S_n & 0 & \cdots & 1 & 1 & \cdots & 0 & \cdots & 0 & \cdots & 1
\end{array}$$

Preference Judgements Matrix

n Stimuli; p Judgements

$$\mathbf{X} = \begin{array}{c|cccccc}
 & X_1 & X_2 & \cdots & X_i & \cdots & X_p \\
\hline
S_1 & n & 2 & \cdots & 2 & \cdots & 1 \\
S_2 & 2 & 1 & \cdots & n & \cdots & n \\
\vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
S_n & 1 & n & \cdots & 1 & \cdots & 2
\end{array}$$

COMPUTATIONAL ASPECTS

Consider the multivariate **regression** model in matrix notation:

$$\mathbf{X} = \mathbf{DB} + \mathbf{E}$$

Due to the peculiar structure of D, it can be easily seen that its rank is equal to *(L-q)*, and consequently one cannot compute the **inverse** of the matrix **D'D**. . .

Assuming that D is an orthogonal design, a particular solution estimates **B** by means of the **inverse diagonal terms of D'D**, so that:

$$\hat{\mathbf{B}} = \Delta_D^{-1}\mathbf{D}'\mathbf{X} \qquad \text{where} \qquad \Delta_D = Diag(D'D)$$

As a **synthesis** of utility coefficients

we propose an approach based on the decomposition of the explained preference-judgements variability

$$trace\left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)=trace\left(\mathbf{X}'D\Delta_D^{-1}D'\mathbf{X}\right)$$

➩ it follows from the **characteristic equation**:

$$\mathbf{X}'\mathbf{D}\left(\Delta_D^{-1}\right)\mathbf{D}'\mathbf{X}\mathbf{v}_\alpha=\lambda_\alpha\mathbf{v}_\alpha$$

under the usual orthonormality constraints

$$\mathbf{v}'_\alpha\mathbf{v}_\alpha=1 \qquad \mathbf{v}'_\alpha\mathbf{v}_{\alpha'}=0$$

the direction cosines in Vα lead to the principal judgments

$$\hat{\mathbf{X}}\mathbf{v}_\alpha=\mathbf{D}\Delta_D^{-1}\mathbf{D}'\mathbf{X}\mathbf{v}_\alpha$$

An application on the preference for a *Cup of Coffee*

A case-study from the Philip Morris Award 1996

60 judges expressed their preferences on 9 different kinds of coffee described by 3 attributes

Taste	moderate bitter	very bitter	
Flavour	Weak	intense	
Strength	weak	moderate	strong

Design Matrix

Attributes Stimuli	Taste	Flavour	Strength
S1	Moderate bitter	intense	moderate
S2	Moderate bitter	weak	weak
S3	Moderate bitter	weak	strong
S4	Moderate bitter	intense	strong
S5	Moderate bitter	intense	weak
S6	very bitter	weak	weak
S7	very bitter	weak	moderate
S8	Moderate bitter	weak	moderate
S9	very bitter	weak	strong

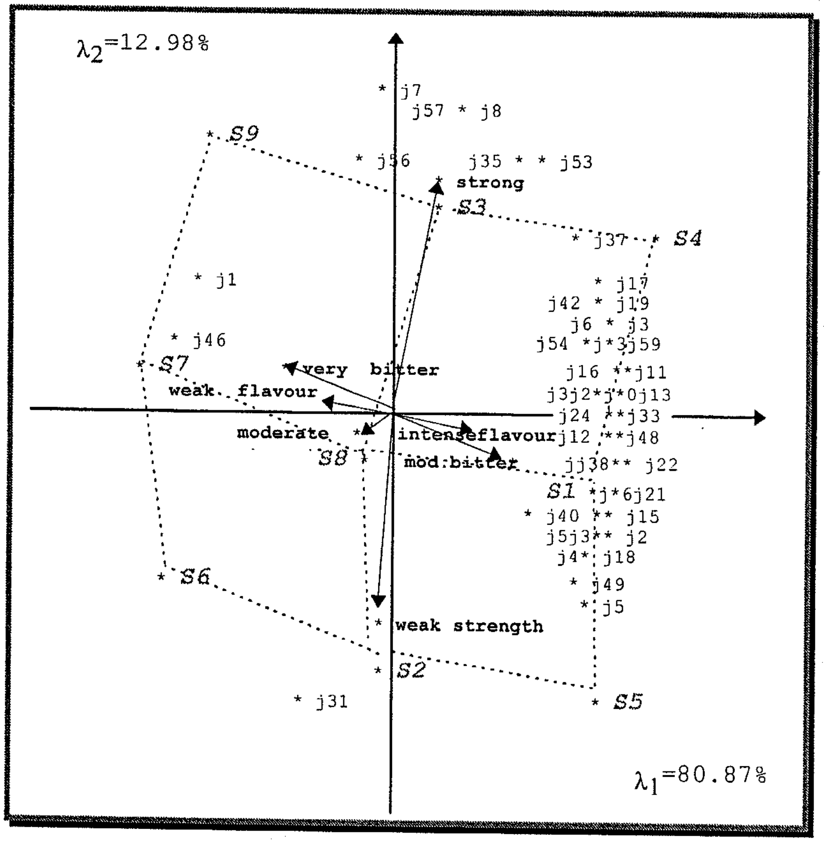
PCAR and Conjoint Analysis results

Levels	Axis 1	Axis 2	Averaged utility
(1) moderate bitter	1.77	-0.54	0.98
(2) very bitter taste	-2.51	0.71	-0.98
(3) weak flavour	-1.77	0.11	-1.01
(4) strong flavour	2.51	-0.21	1.01
(5) weak strength	-0.23	-3.39	-0.23
(6) moderate strength	-0.69	0.14	-0.24
(7) strong strength	0.92	3.46	0.47

Eigenvalue	%
15.64	80.87
2.51	12.98
0.71	3.66
0.48	2.49

Attribute Importance	Axis 1	Axis 2	Averaged utility
Taste	38.98	5.83	34.05
Flavour	37.71	13.59	31.08
Strength	23.31	80.58	34.87

Representation on the first factorial plan



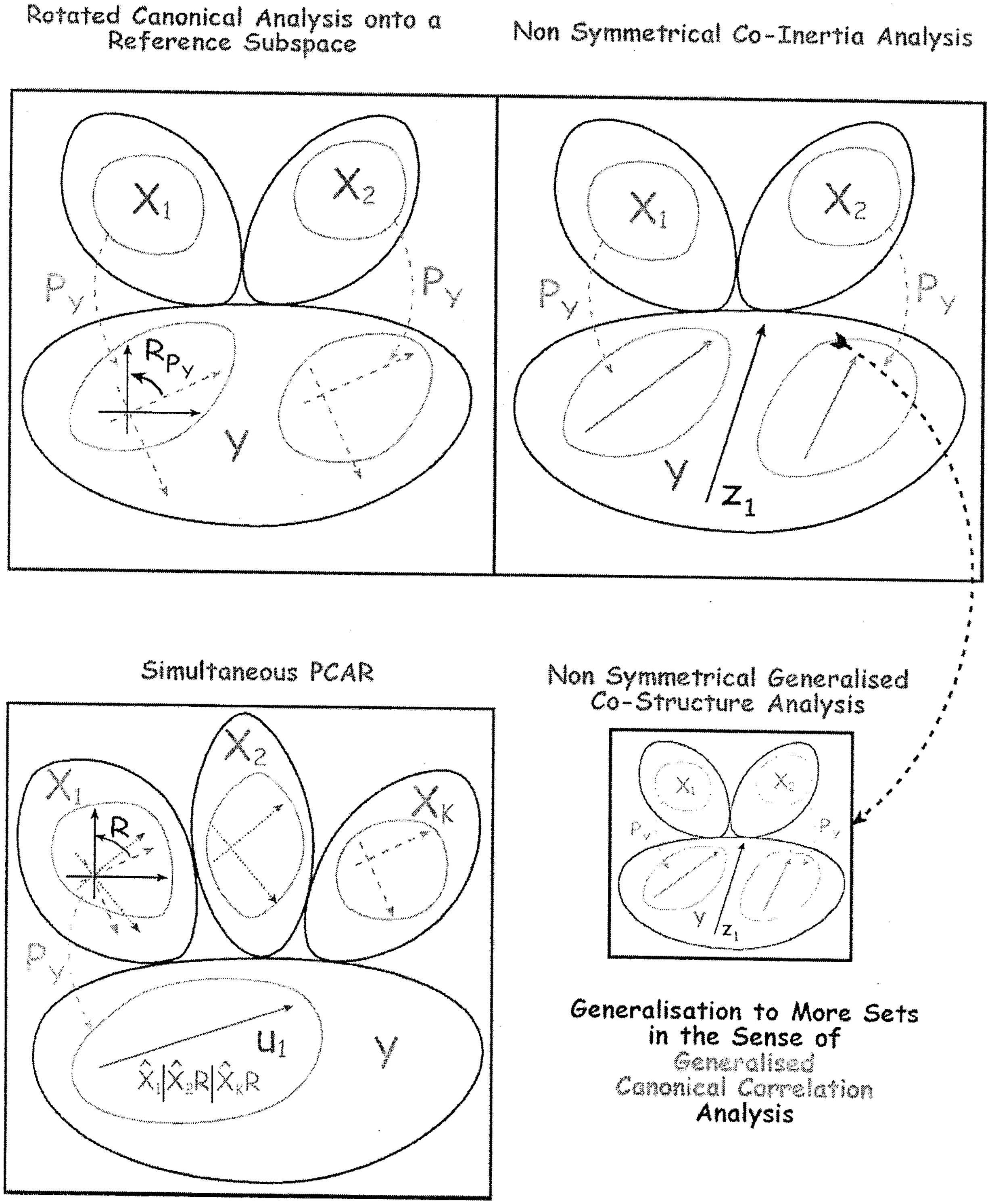
On the **1st Factorial Plan** the stimulus S4

(moderate bitter, intense flavour, strong strength),

seems to individuate the Ideal Cup of Coffee

PCAR in Comparative Studies

Geometrical Insights



Rotated Canonical Analysis onto a Reference Subspace

Non Symmetrical Co-Inertia Analysis

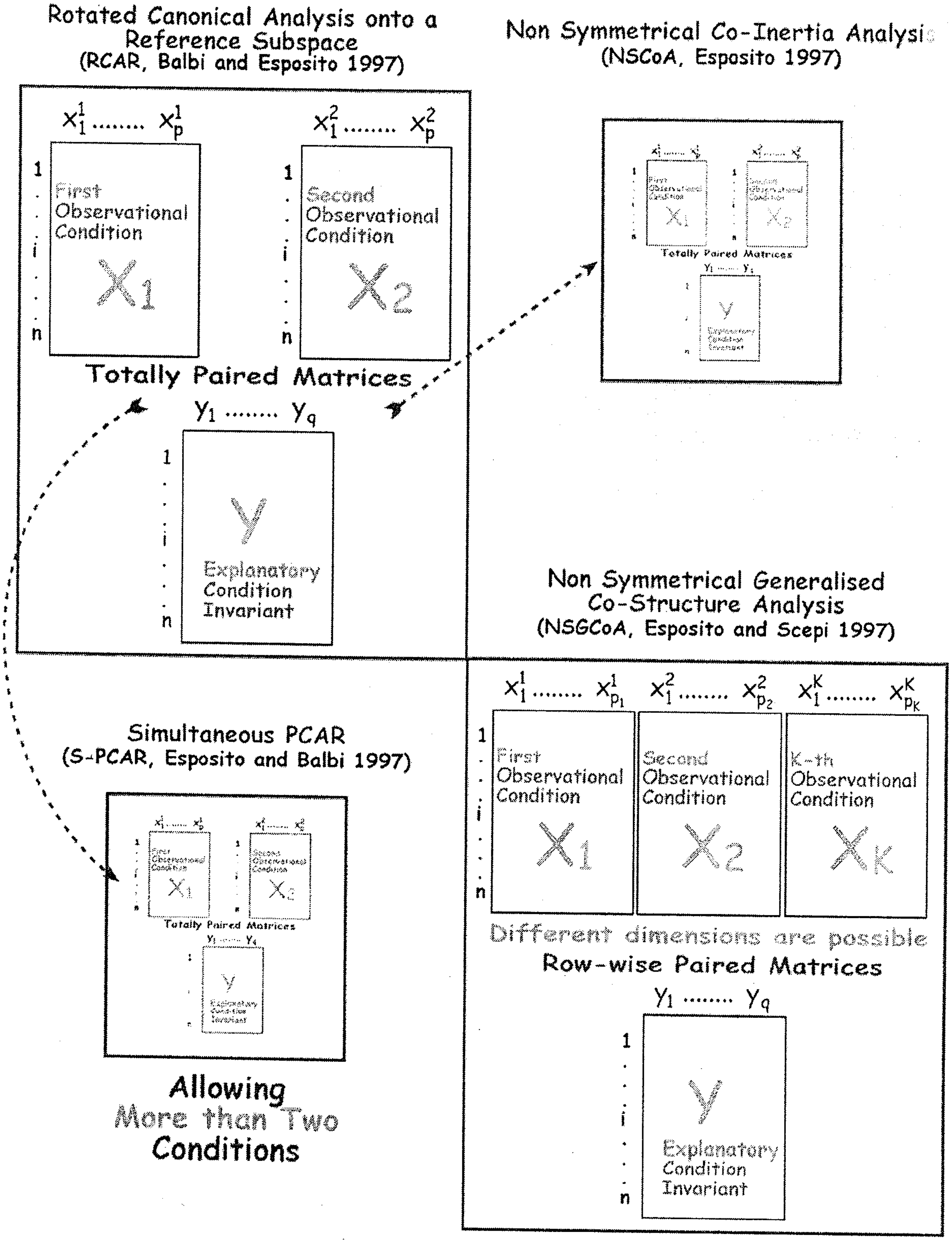
Simultaneous PCAR

Non Symmetrical Generalised Co-Structure Analysis

Generalisation to More Sets in the Sense of Generalised Canonical Correlation Analysis

PCAR in Comparative Studies

Data Structures



Rotated Canonical Analysis onto a Reference Subspace (RCAR, Balbi and Esposito 1997)

$x_1^1 \ldots\ldots x_p^1$
$x_1^2 \ldots\ldots x_p^2$
1 . . i . . . n
First Observational Condition
X_1
Second Observational Condition
X_2
Totally Paired Matrices
$y_1 \ldots\ldots y_q$
Y
Explanatory Condition Invariant

Non Symmetrical Co-Inertia Analysis (NSCoA, Esposito 1997)

First Observational Condition
X_1
Second Observational Condition
X_2
Totally Paired Matrices
Y
Explanatory Condition Invariant

Simultaneous PCAR (S-PCAR, Esposito and Balbi 1997)

First Observational Condition
X_1
Second Observational Condition
X_2
Totally Paired Matrices
Y
Explanatory Condition Invariant

Non Symmetrical Generalised Co-Structure Analysis (NSGCoA, Esposito and Scepi 1997)

$x_1^1 \ldots\ldots x_{p_1}^1$
$x_1^2 \ldots\ldots x_{p_2}^2$
$x_1^K \ldots\ldots x_{p_K}^K$
First Observational Condition
X_1
Second Observational Condition
X_2
K-th Observational Condition
X_K
Different dimensions are possible
Row-wise Paired Matrices
$y_1 \ldots\ldots y_q$
Y
Explanatory Condition Invariant

Allowing More than Two Conditions

PCAR Solutions in Comparative Studies

RCAR

1) Projection Step:

$$P_y X_1 \qquad P_y X_2$$

2) Rotation of $P_y X_2$ towards $P_y X_1$:

$$R_{P_y} = X_2' P_y X_1 \left(X_1' P_y X_2 X_2' P_y X_1 \right)^{-\frac{1}{2}}$$

so to have $P_y X_2 R_{P_y}$

3) Core of the Analysis:

CCA between $P_y X_1$ and $P_y X_2 R_{P_y}$

Aim: To represent the principal structure of similarity once the variability of each dependent set is decomposed

NSCoA

1) Projection Step:

$$P_y X_1 \qquad P_y X_2$$

2) Criterion to Maximise:

$$\mathrm{cov}(P_y X_1 z_1, P_y X_2 z_1)$$

3) Singulare Value Decomposition of:

$$\left(X_2' P_y X_1 + X_1' P_y X_2 \right) / 2$$

Aim: To identify a common structure to all conditions w.r.t. which the statistical units configurations are compared

Note: Compromise between two separate PCAR's and a global CCA on projected data

S-PCAR

1) Rotation of all X_k's towards X_1:

$$R = X_2' X_1 \left(X_1' X_2 X_2' X_1 \right)^{-\frac{1}{2}}$$

so to have $X_1, X_2 R, \dots, X_K R$

2) Projection Step:

$$P_y [X_1 \mid X_2 R \mid \dots \mid X_K R]$$

3) Core of the Analysis:

PCA on $P_y [X_1 \mid X_2 R \mid \dots \mid X_K R]$

Aim: To detect the differences in the overall structure of dependent variables and then to explain these differences in terms of the explanatory variables

NSGCoA

1) Projection Step:

$$P_y X_1 \qquad P_y X_2 \dots\dots P_y X_K$$

2) Criterion to Maximise:

$$\sum_k \pi_k (P_y X_k w_k \mid z_1)^2$$

3) Core of the Analysis:

PCA on $P_y [X_1 \mid X_2 \mid \dots \mid X_K]$

Aim: The same as NSCoA but extended to multiple just row-wise paired matrices

In RCAR and S-PCAR both inter- and intra- conditions variabilities are represented

Applications in Comparative Studies

RCAR **Sensory Data Analysis** Comparing, on the basis of a common structure, the judgements expressed by different groups of tasters w.r.t. the organoleptic features of a product	NSCoA **Customer Satisfaction** Measuring the gap between perceived and expected quality by the customers of a product/service w.r.t. a pre-defined set of scenarios
S-PCAR **Multivariate Quality Control** Comparing the really observed quality characteristics with the in-control situation and explain the eventual differences with respect to the process variables	NSGCoA **Non Parametric MVQC Charts** The whole set of quality characteristics may be split into differently sized groups according to a specified expert's criterion **Panel Data** A questionnaire is submitted to different samples in different occasions

An RCAR Example

4 Dependent Variables:

Judgements on: Sight
Taste
Smell
Aftertaste

of the *"Tocai friulano"* Italian wine produced by 22 wineries

Condition 1: **Experts** judgements (rotation reference)

Condition 2: **Ordinary Consumers** judgements

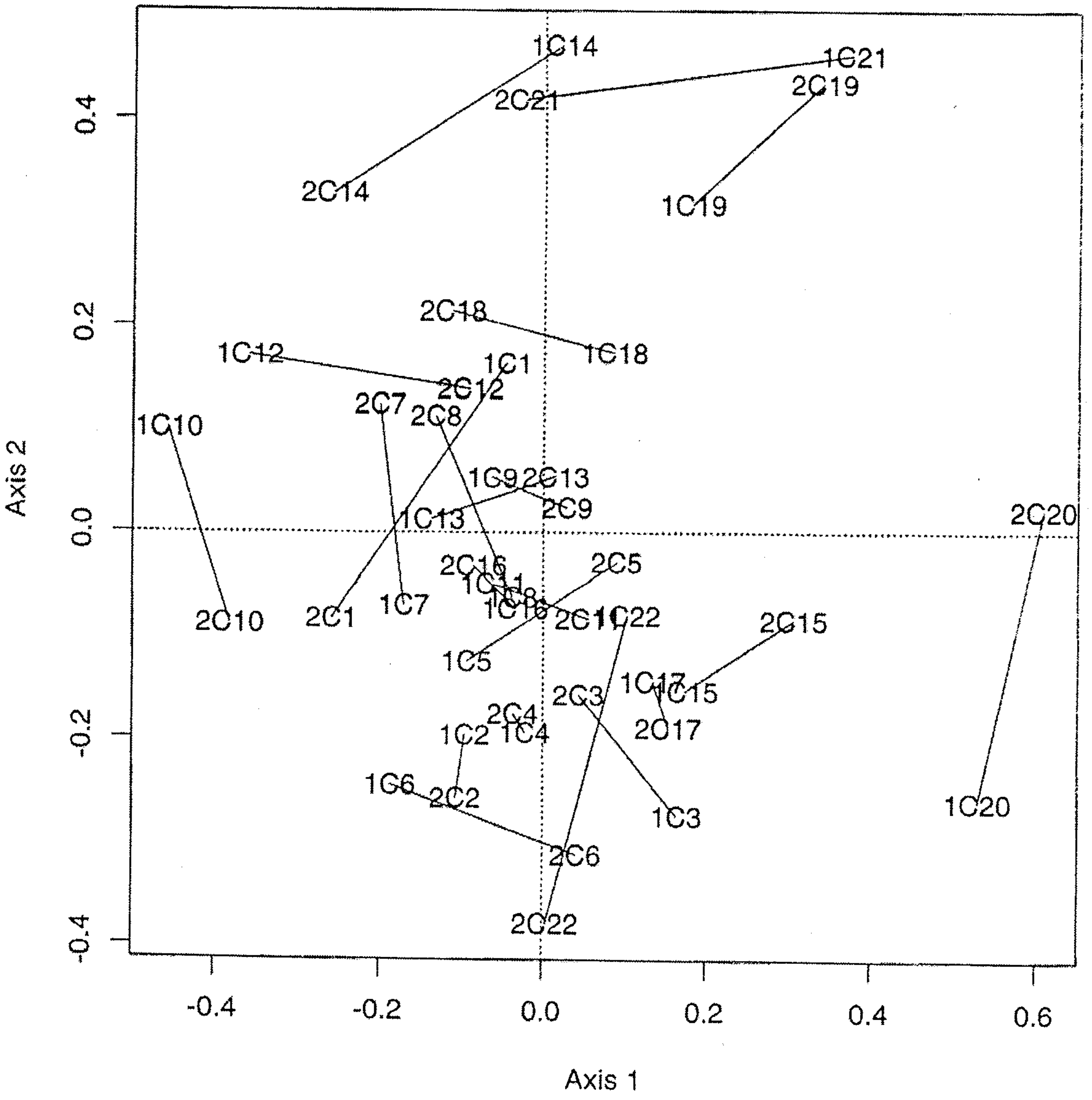
7 Explanatory Variables:

Physical-chemical features of the 22 wines:

Alcohol
pH
Sugar
Methanol
Free Sulphur Dioxide
Optical Absorbency
Ethyl Acetate

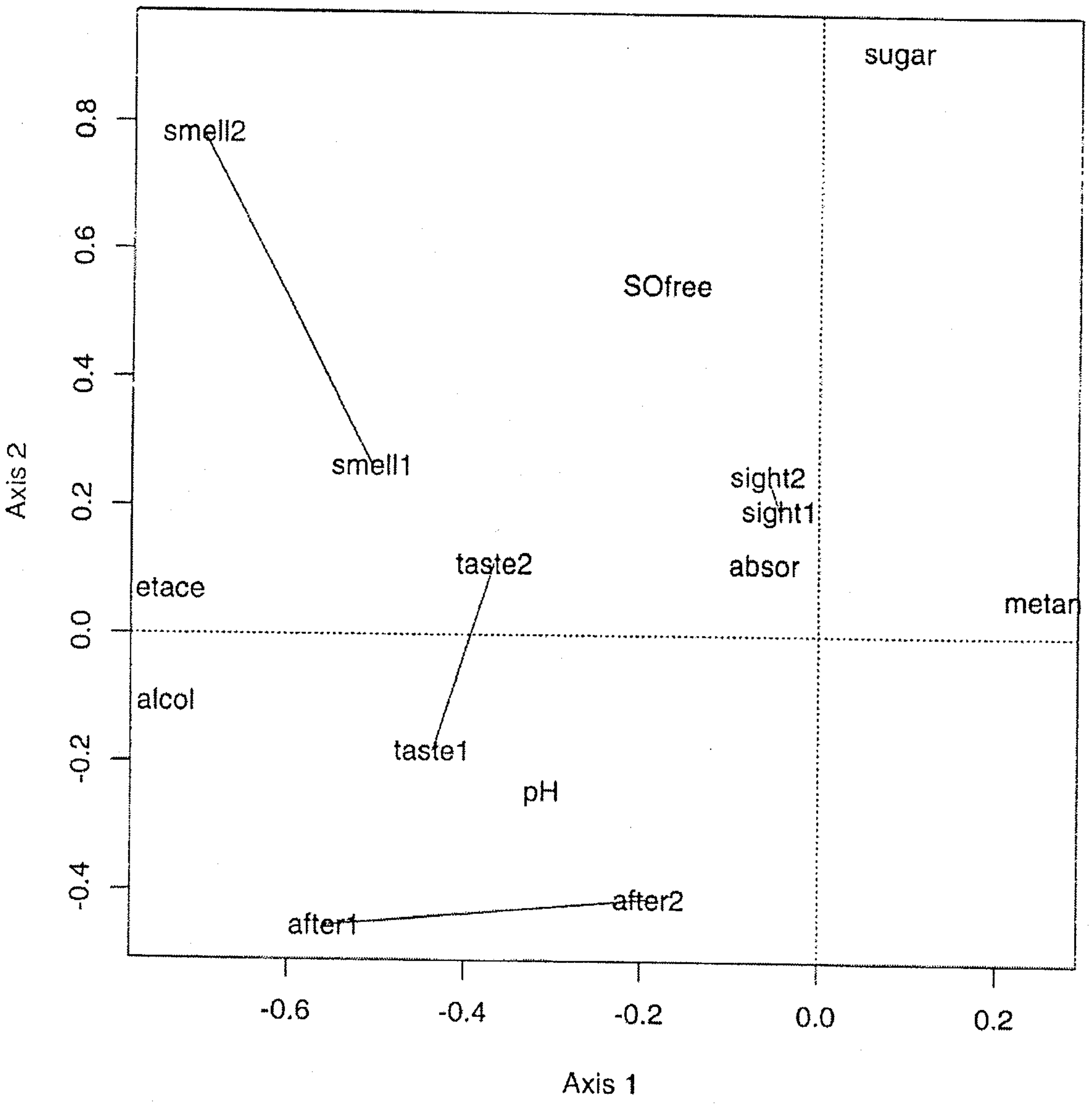
Representation of Paired Wineries

Individual-Points



Representation of Judgements

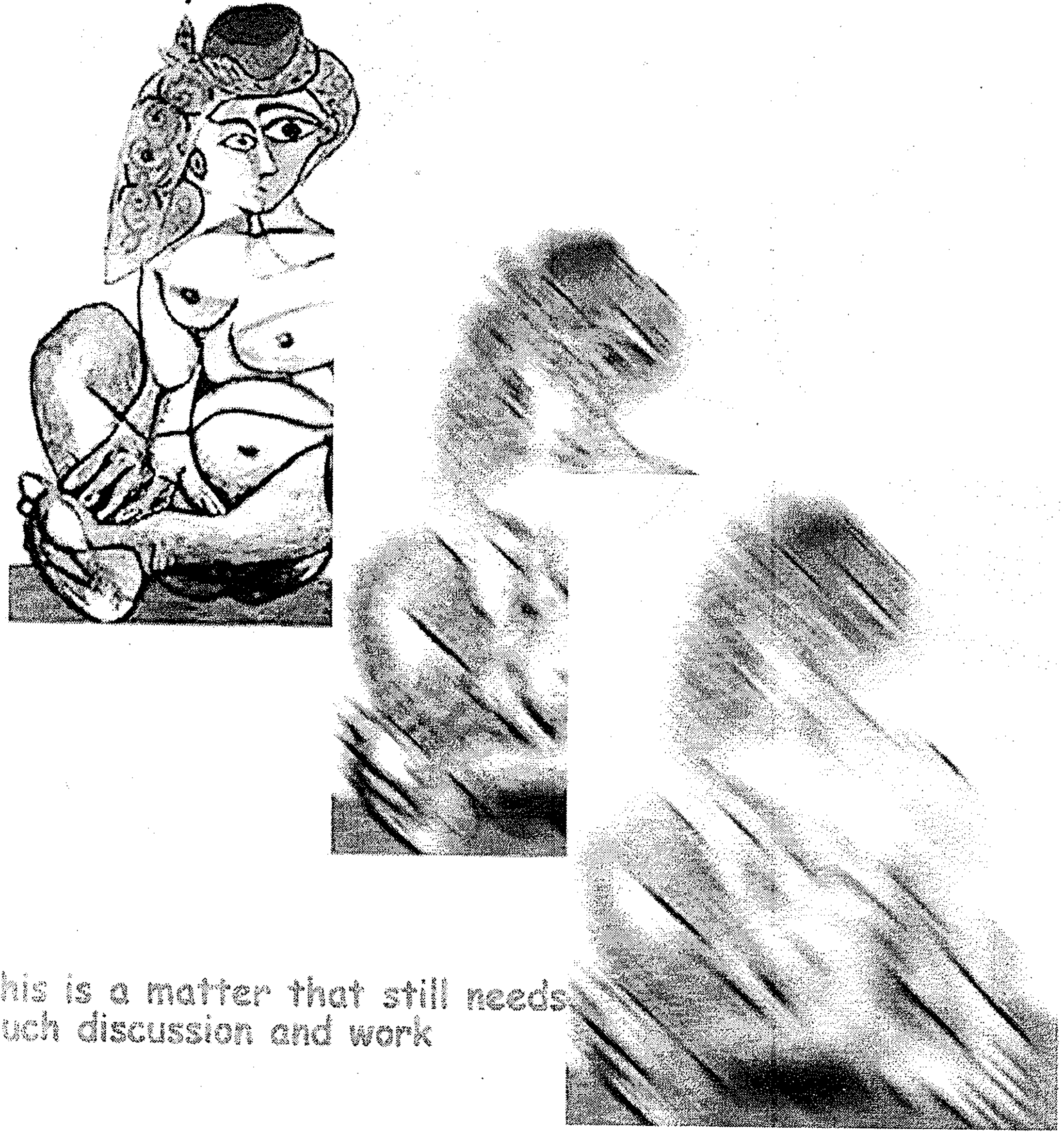
Variable-Points



CONCLUDING REMARKS AND PERSPECTIVES

The time dimension has always represented a challenge for data analysts. In fact, though easily to consider from a technical point of view, it often lacks of a proper interpretation.

Geometrically based techniques usually consider this dimension only implicitly, thus interpreting the ordinal feature of time a posteriori on the graphical displays regardless of its being a real variable that should be taken into account in the core of the analysis.



This is a matter that still needs much discussion and work