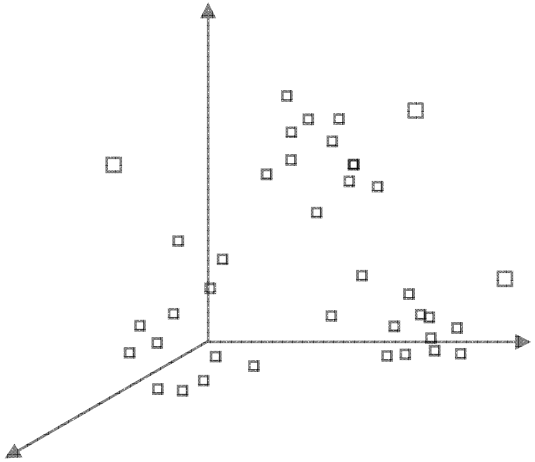
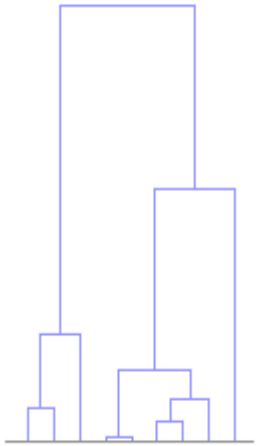


Correspondence & HIerarchical Cluster (CHIC) Analysis Software v1.0

Angelos Markos, George Menexes & Iannis Papadimitriou

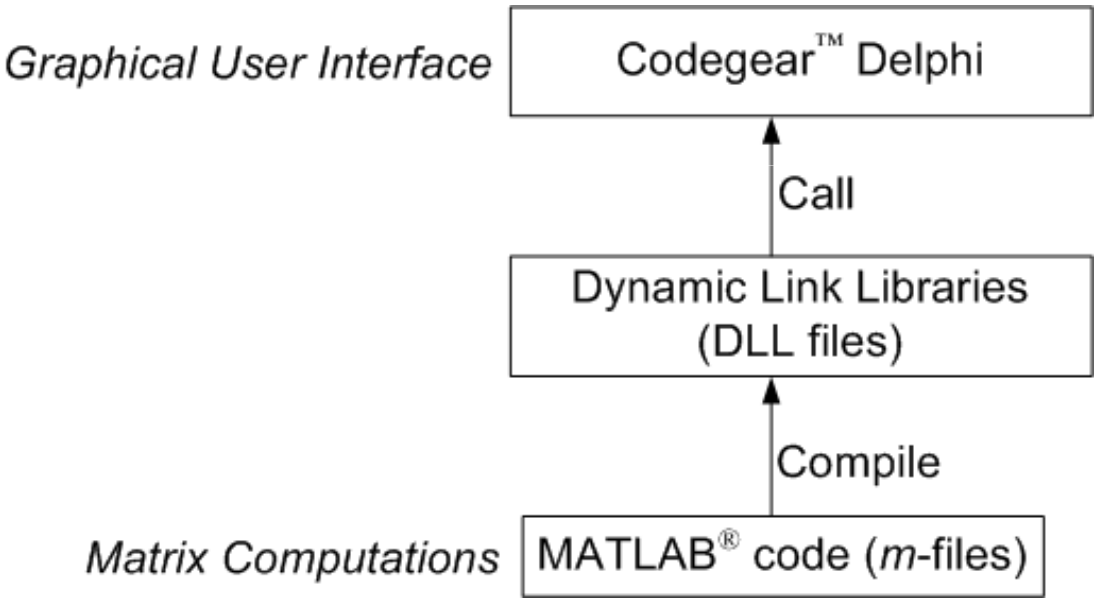
Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece

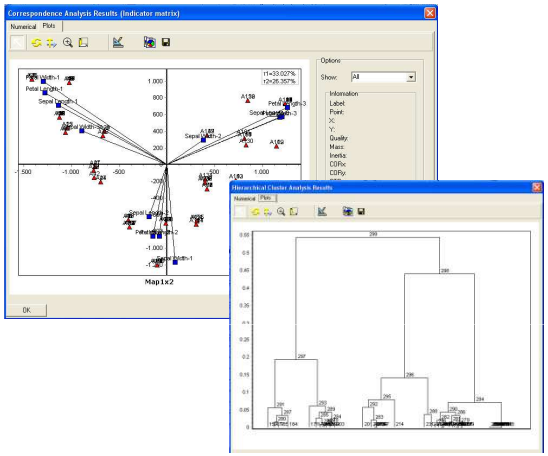


Outline

- What is CHIC Analysis?
- Motivation & Goals
- Data Entry & Data Management
- Interpretation Options
- The SVD Implementation of MCA
- Hierarchical Cluster Analysis as a complementary method
- Work in Progress

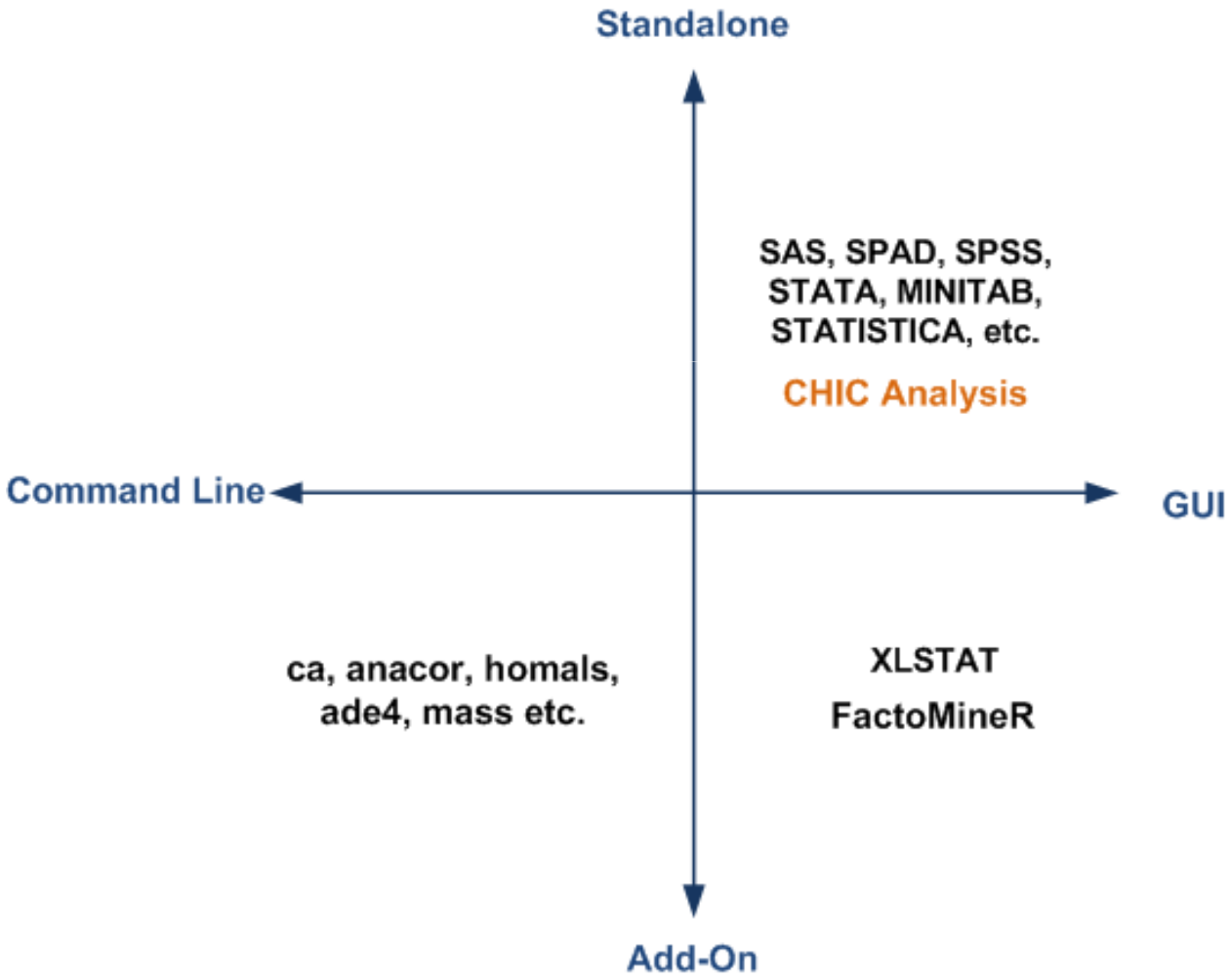
Development Tools



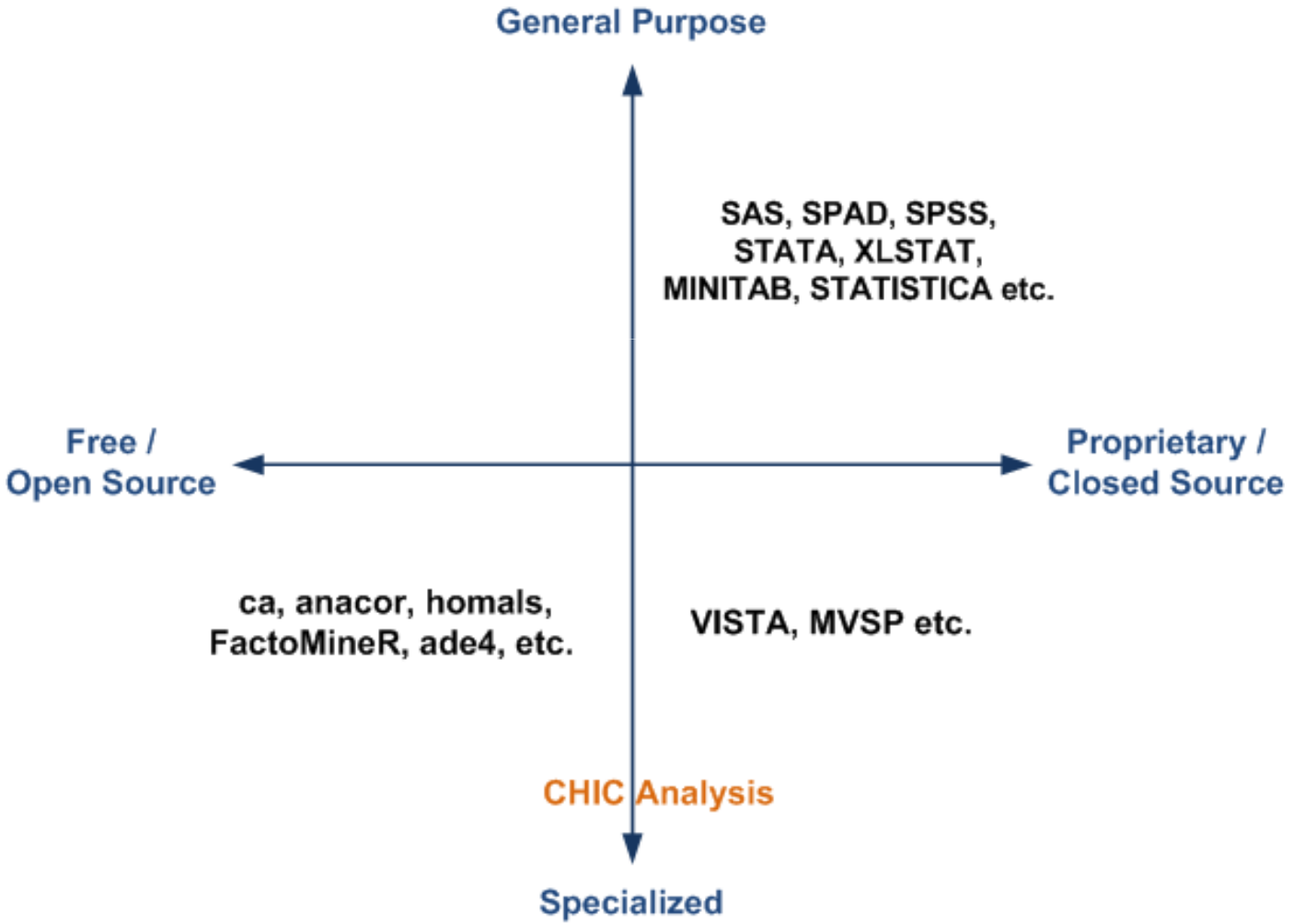


```
% Correspondence matrix
P = X./total;
% Standardized residuals matrix
%S = (P - ri'*cj)./(sqrt(ri'*cj));
S = stdresij/sqrt(total);
```

Corners of Statistical Software for CARME (1)



Corners of Statistical Software for CARME (2)



Motivation & Goals

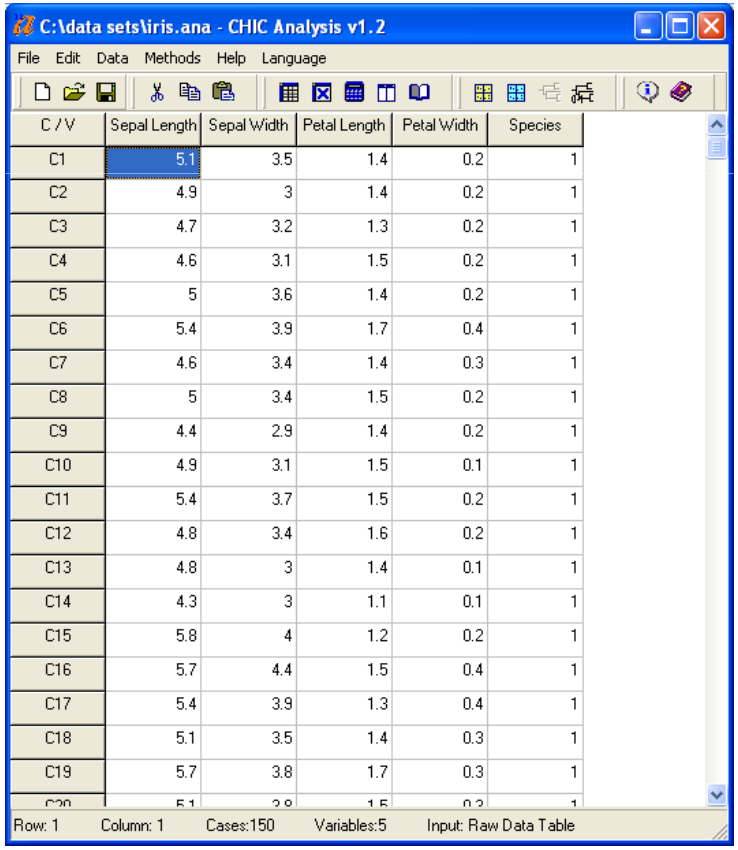
- Educational purposes

- A CA software towards interpretation:
  - Options for both empirical interpretation and statistical inference
  - A synthesis of contributions (French School & the GIFI System)
  - Direct construction and analysis of Burt subtables
  - Options for Symmetric Plots and Biplots

Data Entry & Data Management (1)

- The Data Spreadsheet

(a) Raw data table



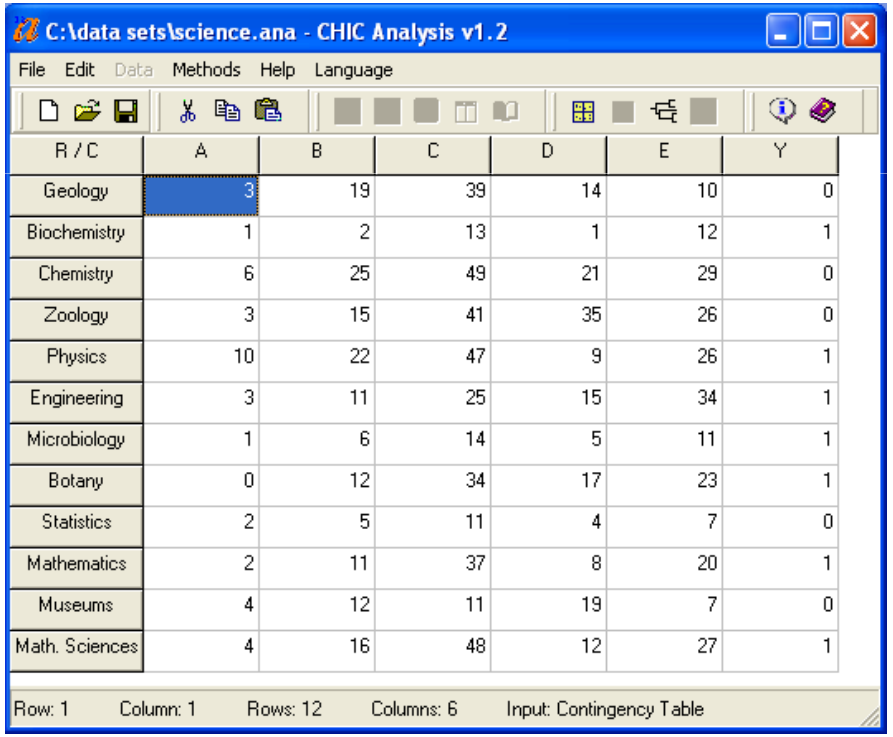
C:\data sets\iris.ana - CHIC Analysis v1.2

File Edit Data Methods Help Language

C / V	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
C1	5.1	3.5	1.4	0.2	1
C2	4.9	3	1.4	0.2	1
C3	4.7	3.2	1.3	0.2	1
C4	4.6	3.1	1.5	0.2	1
C5	5	3.6	1.4	0.2	1
C6	5.4	3.9	1.7	0.4	1
C7	4.6	3.4	1.4	0.3	1
C8	5	3.4	1.5	0.2	1
C9	4.4	2.9	1.4	0.2	1
C10	4.9	3.1	1.5	0.1	1
C11	5.4	3.7	1.5	0.2	1
C12	4.8	3.4	1.6	0.2	1
C13	4.8	3	1.4	0.1	1
C14	4.3	3	1.1	0.1	1
C15	5.8	4	1.2	0.2	1
C16	5.7	4.4	1.5	0.4	1
C17	5.4	3.9	1.3	0.4	1
C18	5.1	3.5	1.4	0.3	1
C19	5.7	3.8	1.7	0.3	1

Row: 1 Column: 1 Cases:150 Variables:5 Input: Raw Data Table

(b) Contingency table



C:\data sets\science.ana - CHIC Analysis v1.2

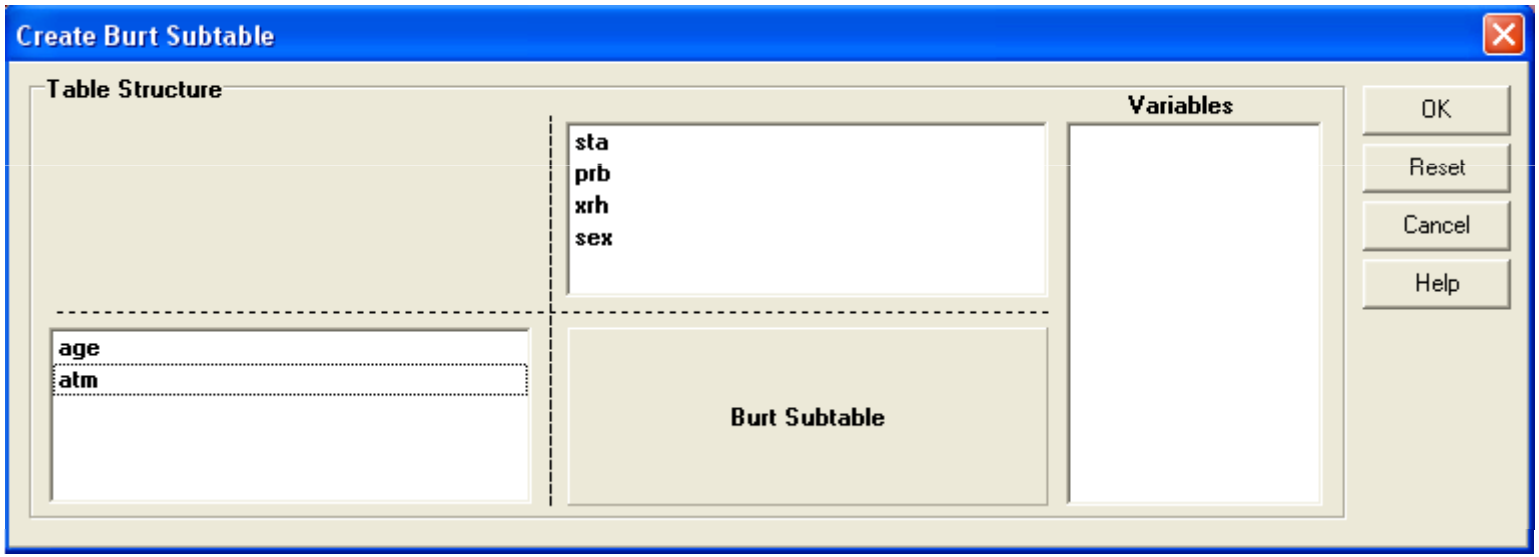
File Edit Data Methods Help Language

R / C	A	B	C	D	E	Y
Geology	3	19	39	14	10	0
Biochemistry	1	2	13	1	12	1
Chemistry	6	25	49	21	29	0
Zoology	3	15	41	35	26	0
Physics	10	22	47	9	26	1
Engineering	3	11	25	15	34	1
Microbiology	1	6	14	5	11	1
Botany	0	12	34	17	23	1
Statistics	2	5	11	4	7	0
Mathematics	2	11	37	8	20	1
Museums	4	12	11	19	7	0
Math. Sciences	4	16	48	12	27	1

Row: 1 Column: 1 Rows: 12 Columns: 6 Input: Contingency Table

Data Entry & Data Management (2)

- Burt Subtable construction



Simple Correspondence Analysis Options (1)

- Significance criteria of individual points

Active points

- Mass, Inertia, COR,CTR, QLT, SQCOR, Best

Supplementary points

- Inertia, COR, QLT

Simple Correspondence Analysis Options (2)

- Selection of significant axes
  - Scree Plot
  - Statistical significance tests by
    - Nishisato
    - van de Geer
    - Greenacre

- Additional Options
  - Expected frequencies
  - Residuals, Standardized residuals, Adjusted standardized residuals
  - Chi-square contributions
  - Reconstruction matrix (chi-square goodness of fit)

Multiple Correspondence Analysis Options (1)

- Significance criteria of individual points

Active

- Mass, Inertia, COR, SQCOR, CTR, SumCTR, QLT, Best, Discrimination Measures (absolute**,** relative)

Supplementary

- Inertia, COR, QLT, Test values

Multiple Correspondence Analysis Options (2)

- Selection of significant axes
  - Scree Plot
  - Principal inertias > $1/p$
  - Cronbach's alpha
  - Statistical significance test by
    - Nishisato
- Inertia Adjustment
  - Greenacre
  - "Interesting" Inertia

Symmetric Plots & Biplots (1)

- Normalization Options
  - Principal (PN) — Symmetric Plot
  - Row Principal (RPN) — Biplot
  - Column Principal (CPN) — Biplot
  - Symmetrical (SN) - SPSS — Biplot

Symmetric Plots & Biplots (2)

- Biplot Axes (Normalization: RPN, CPN, SN)

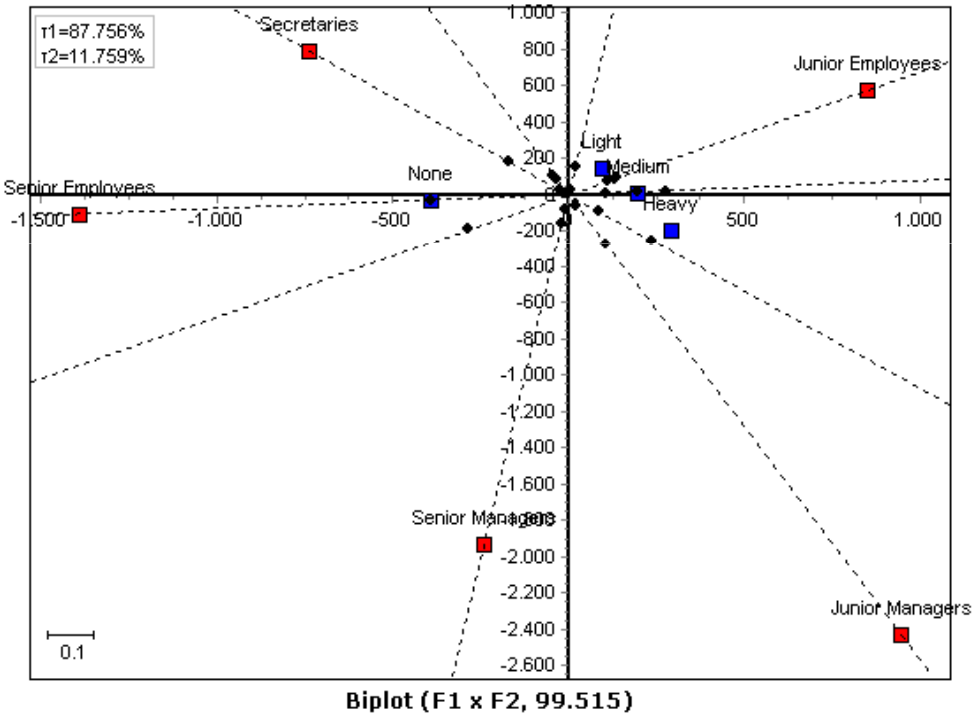
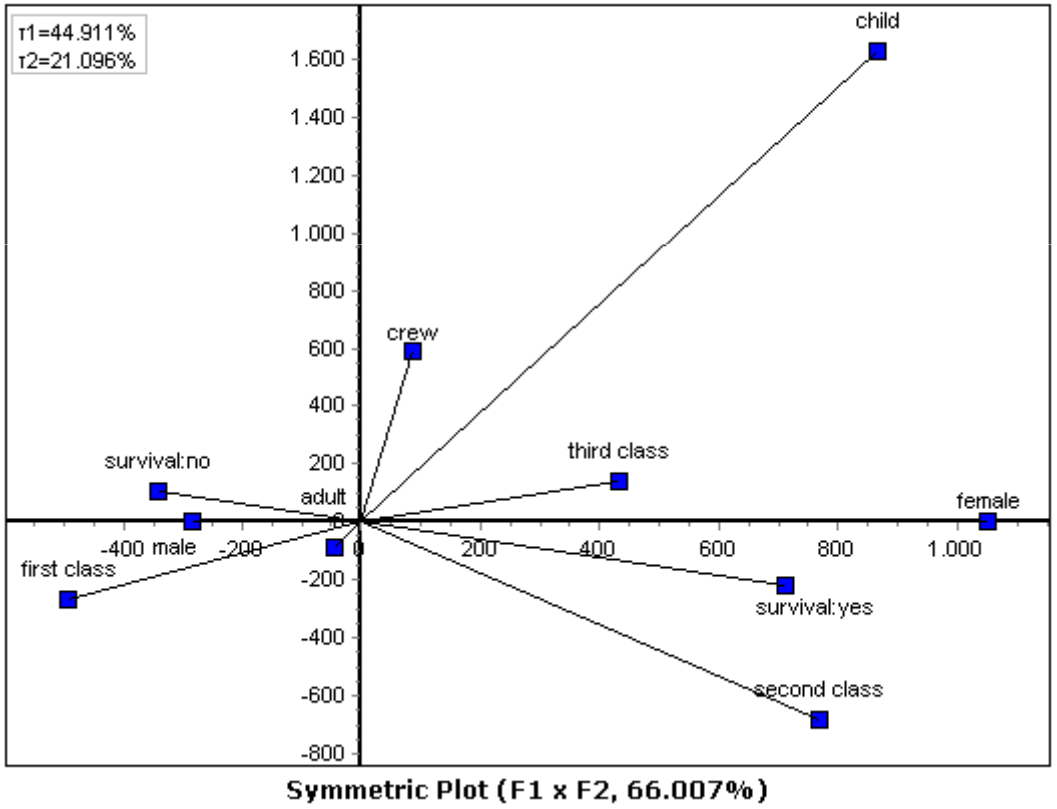


Table 1. Biplot Interpretation (CPN)

Senior Managers	Heavy	None	Medium	Light
Distance	1.790	1.873	1.982	2.103
Cos^2	0.140	0.039	0.998	0.307
Junior Managers	Heavy	None	Medium	Light
Distance	2.469	2.540	2.636	2.744
Cos^2	0.390	0.780	0.615	0.109
Senior Employees	Heavy	None	Medium	Light
Distance	1.373	1.383	1.403	1.434
Cos^2	0.043	0.025	0.754	0.675
Junior Employees	Light	Medium	None	Heavy
Distance	0.936	1.009	1.082	1.141
Cos^2	1.000	0.874	0.204	0.430
Secretaries	Light	Medium	None	Heavy
Distance	0.987	1.058	1.129	1.187
Cos^2	0.404	0.718	0.085	0.949

Symmetric Plots & Biplots (3)

- Position Vectors (Normalization: PN)



Symmetric Plot (F1 x F2, 66.007%)

A note on the SVD Implementation of MCA

Step 1 - Calculate the matrix $\mathbf{S_B}$ of standardized residuals

$$\mathbf{S_B} = \mathbf{diag}\left(\mathbf{r_B}\right)^{-1/2}\left(\mathbf{P_B} - \mathbf{r}\mathbf{c}^{\mathbf{T}}\right)\mathbf{diag}\left(\mathbf{c_B}\right)^{-1/2}$$

Step 2 - Calculate the SVD of $\mathbf{S_B}$

$$\mathbf{S_B} = \mathbf{U_B}\mathbf{D_B}\mathbf{V_B^T}$$

Steps 3 & 4 - Standard coordinates $\mathbf{\Phi_Z}$ of rows and $\mathbf{\Gamma_Z}$ of columns

$$\mathbf{\Phi_Z} = \frac{\mathbf{1}}{\boldsymbol{p}}\mathbf{Z}\mathbf{\Gamma_Z}\mathbf{D_B^{-1/2}}$$

$$\mathbf{\Gamma_Z} = \mathbf{diag}\left(\mathbf{r_Z}\right)^{-1/2}\mathbf{U_B} = \mathbf{\Gamma_B}$$

Steps 4 & 5 - Principal coordinates $\mathbf{F_Z}$ of rows and $\mathbf{G_Z}$ of columns

$$\mathbf{F_Z} = \frac{\mathbf{1}}{\boldsymbol{p}}\mathbf{Z}\mathbf{\Gamma_Z} = \mathbf{\Phi_Z}\mathbf{D_B^{1/2}}$$

$$\mathbf{G_Z} = \mathbf{\Gamma_Z}\mathbf{D_B^{1/2}}$$

Hierarchical Cluster Analysis Algorithm

Input

Contingency Table, Indicator Matrix

Implementation

Ward's linkage criterion, Benzécri's chi-square distance

Output

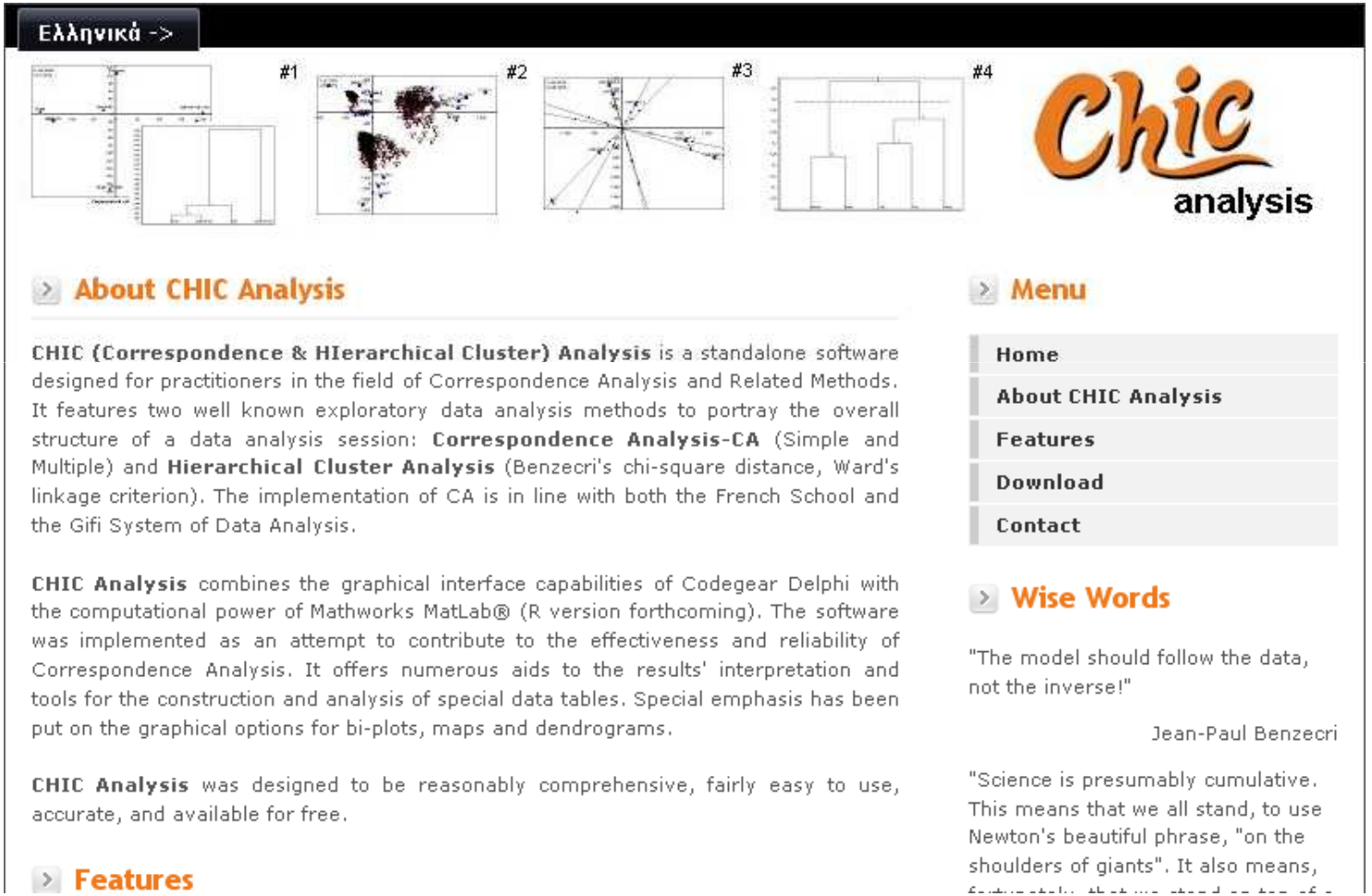
- Dendrogram
- Agglomeration schedule
- Cluster Membership
- Variable contributions to the characterization of clusters
- Variable contributions to the division of clusters

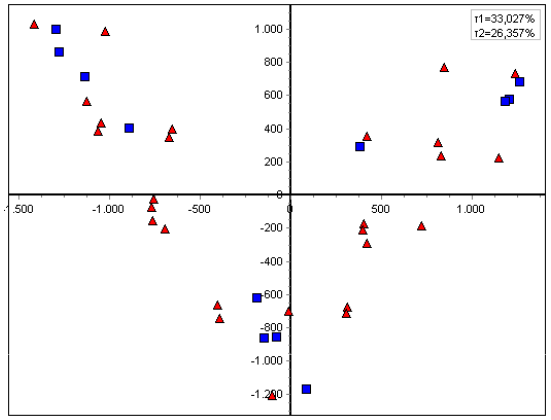
VACOR

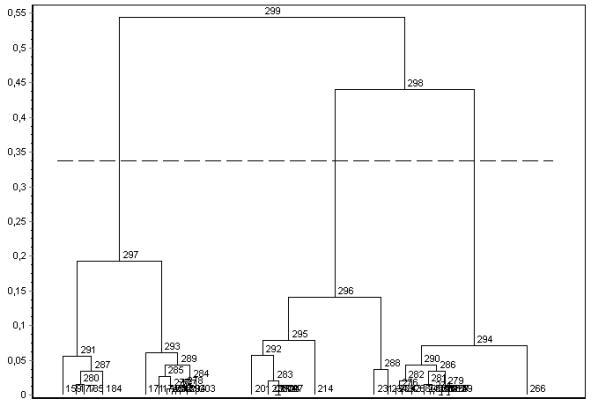
Work in Progress

- Take advantage of the common mathematical foundation of many multivariate data analysis methods, as a basis for writing unified software
- Correspondence Analysis Variations (Canonical, Joint, Subset, Non-Symmetrical, Regularized)
- Aspects of Stability
- A forthcoming **R** version

http://amarkos.gr/en/research/chic




r1=33,027%
r2=26,357%
1.000
800
600
400
200
0
-200
-400
-600
-800
-1.000
-1.200
-1.500
-1.000
-500
0
500
1.000


0,55
0,5
0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
299
298
297
296
295
294
291
287
280
293
289
285
284
292
283
288
290
286
282
281
279
214
266

#fin