Correspondence Analysis

Ying-Chao Hung
Department of Staistics
National Chengchi University
hungy@nccu.edu.tw

Introduction

- Correspondence Analysis (CA) is an exploratory multivariate technique that converts frequency-table data into graphical displays in which the rows and the columns of the table are displayed as points.
- Mathematically, CA decomposes the χ^2-measure of association of the table data into components in a manner similar to that of PCA for continuous data.

  ➔ Transform the χ^2-measure into a low-dimensional metric (or distance) measure.
- In CA, no model is introduced, no assumptions on the underlying stochastic mechanism that generated the data are made.

Pearson χ^2 Statistic

- Consider a 2-way contingency (frequency) table:

J

<table>
<tr><td></td><td colspan="4">Column</td></tr>
<tr><td rowspan="4">Row</td><td></td><td></td><td></td><td rowspan="3">$r(i)$</td></tr>
<tr><td></td><td>$F(i, j)$</td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td colspan="3">$c(j)$</td><td>N</td></tr>
</table>

I

where $F(i, j)$ is the frequency of row i with column j, $r(i)$ and $c(j)$ are the sums of row i and column j, respectively.

Pearson χ^2 Statistic

- If the row variable is independent of the column variable, the expected frequency of row i with column j is

$$E(i,j) = \frac{r(i)c(j)}{N}$$

(Note that under the "independence" assumption, $NP_{ij} = NP_i P_j$.

Thus, $N\hat{P}_{ij} = N\hat{P}_i\hat{P}_j = N\frac{r(i)}{N}\frac{c(j)}{N} = \frac{r(i)c(j)}{N}$.)

- The Pearson chi-squared statistic:

$$\chi^2 = \sum_{i,j}\frac{[E(i,j) - F(i,j)]^2}{E(i,j)}$$

Pearson χ^2 Statistic

- Note that if the quantity

$$\sum_{i,j} \frac{[E(i,j) - F(i,j)]^2}{E(i,j)}$$

is large, then the row variable tends to be not independent of the column variable.

Question: What is the relationship between row and column?

Two Types of CA

- **Simple CA**

  ➔ CA of contingency tables (i.e. 2-way tables)

- **Multiple CA (MCA)**

  ➔ Handle more than two categorical variables (i.e. 3-way, 4-way tables)

Simple Correspondence Analysis

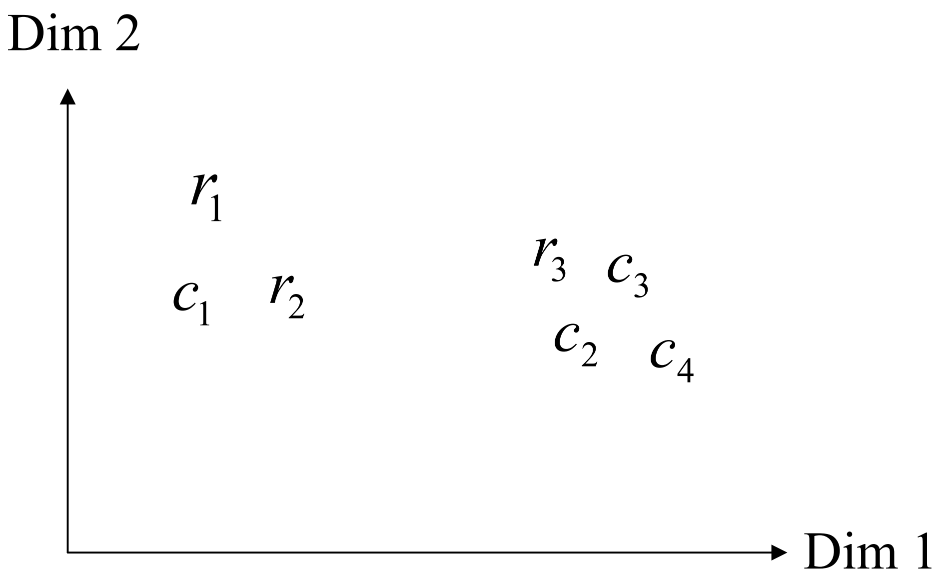
- Let F be a 2-way contingency (frequency) table:

<table>
<tr><td></td><td></td><td colspan="3">J</td><td></td></tr>
<tr><td></td><td></td><td colspan="4">Column</td></tr>
<tr><td rowspan="3">I</td><td rowspan="4">Row</td><td></td><td></td><td></td><td rowspan="3">$r(i)$</td></tr>
<tr><td></td><td>$F(i, j)$</td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td></td><td colspan="3">$c(j)$</td><td>N</td></tr>
</table>

where $F(i, j)$ gives the frequency of row i with column j.

Goal of CA

- CA finds a multi-dimensional displays of the dependences between the rows and the columns using distances.
- Example:



The χ^2 Distance

- Represent the dissimilarity between rows (or columns) using a χ^2 distance.

  For example, the χ^2 distance between row *i* and row *i'* is

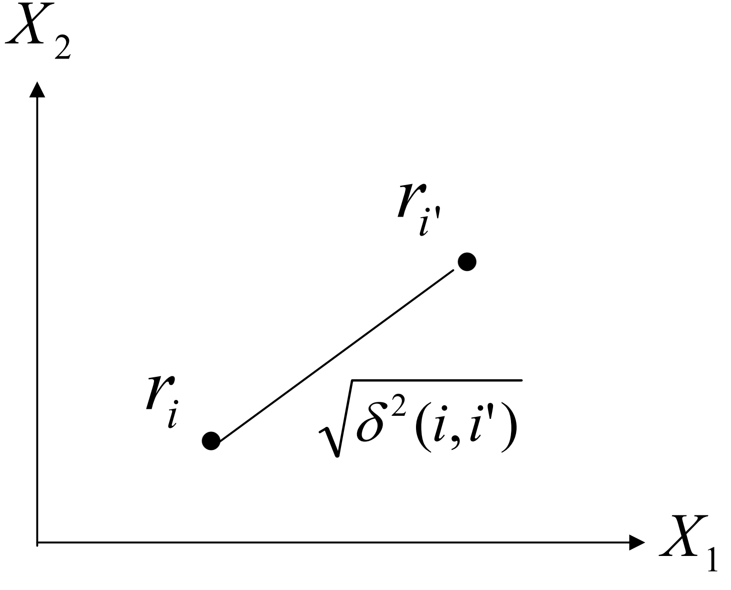
第i個row的分配

$$\delta^2(i,i') = \sum_{j=1}^{J} \frac{\left[F(i,j)/r(i) - F(i',j)/r(i')\right]^2}{c(j)/N}$$

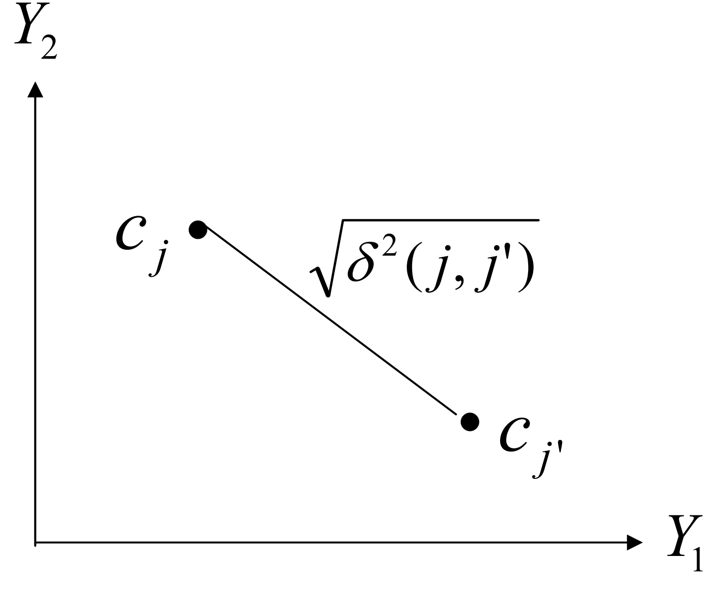
- Try to find a space X (for row scores) such that

$\delta^2(i,i')=$ the Euclidean distance between row i and row i' in X.

Geometric Illustration



(row scores)



(column scores)

Solving *X* and *Y*

- Denote the expected frequency matrix by E which has the elements

$$E(i,j) = \frac{r(i)c(j)}{N}$$

- Consider the singular value decomposition (SVD) of the following matrix

類似chi sq統計量

$$D_r^{-1/2}(F-E)D_c^{-1/2} = K\Lambda L'$$

where

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \ddots \end{pmatrix},\ D_r = \begin{pmatrix} r(1) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & r(I) \end{pmatrix},\ D_c = \begin{pmatrix} c(1) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & c(J) \end{pmatrix},$$

and $K'K = L'L = I$.

Solving X and Y

- Note that matrix K contains row scores corresponding to the row category while matrix L contains column scores corresponding to the column category.
- The solution of (X, Y) is given by

$$\begin{cases} \widetilde{X} = N^{1/2} D_r^{-1/2} K\Lambda & \text{(space for row scores)} \\ \widetilde{Y} = N^{1/2} D_c^{-1/2} L\Lambda & \text{(space for column scores)} \end{cases}$$

- The relationship between $\widetilde{X}$ and $\widetilde{Y}$ are given (after some algebra) by

$$D_r^{-1}(F-E)\tilde{Y}\Lambda^{-1} = \tilde{X} \quad \text{or} \quad D_c^{-1}(F-E)'\tilde{X}\Lambda^{-1} = \tilde{Y}.$$

Remarks

- The dimension of the solution is $\min(I-1, J-1)$.
- Can show

$$\tilde{X}' D_r \tilde{X} = \tilde{Y}' D_c \tilde{Y} = N \cdot \Lambda^2$$

The Pearson χ^2 statistic is

$$\text{trace}(\tilde{X}' D_r \tilde{X}) = \text{trace}(\tilde{Y}' D_c \tilde{Y}) = N \cdot \boxed{\text{trace}(\Lambda^2)}$$

(this is also known as total inertia in the French literature)

- The scores of rows in $\tilde{X}_1$ have the maximum correlation with the scores of columns in $\tilde{Y}_1$.

➔ $\text{Corr}(\tilde{X}_1, \tilde{Y}_1) = \lambda_1$, the 1st canonical correlation.

Remarks

- To interpret the result, one can plot $(\tilde{X}, \tilde{Y})$ (or rescale them)
- The proportion of the total inertia accounted by the first dimension is

$$\lambda_1^2 \Big/ \left(\sum \lambda_i^2\right)$$

類似PCA, 比例=chi sq 統計量可以被解釋的百分比
決定dim的個數

➔ The interpretation of the relationship revealed in the 1st dimension is the most important.

- Instead of plotting $(\tilde{X}, \tilde{Y})$, the following pairs are plotted in R:

$$\begin{cases} X = D_r^{-1/2} K \Lambda & \text{(row scores)} \\ Y = D_c^{-1/2} L \Lambda & \text{(column scores)} \end{cases}$$

Example (Fisher Data)

Hair Color

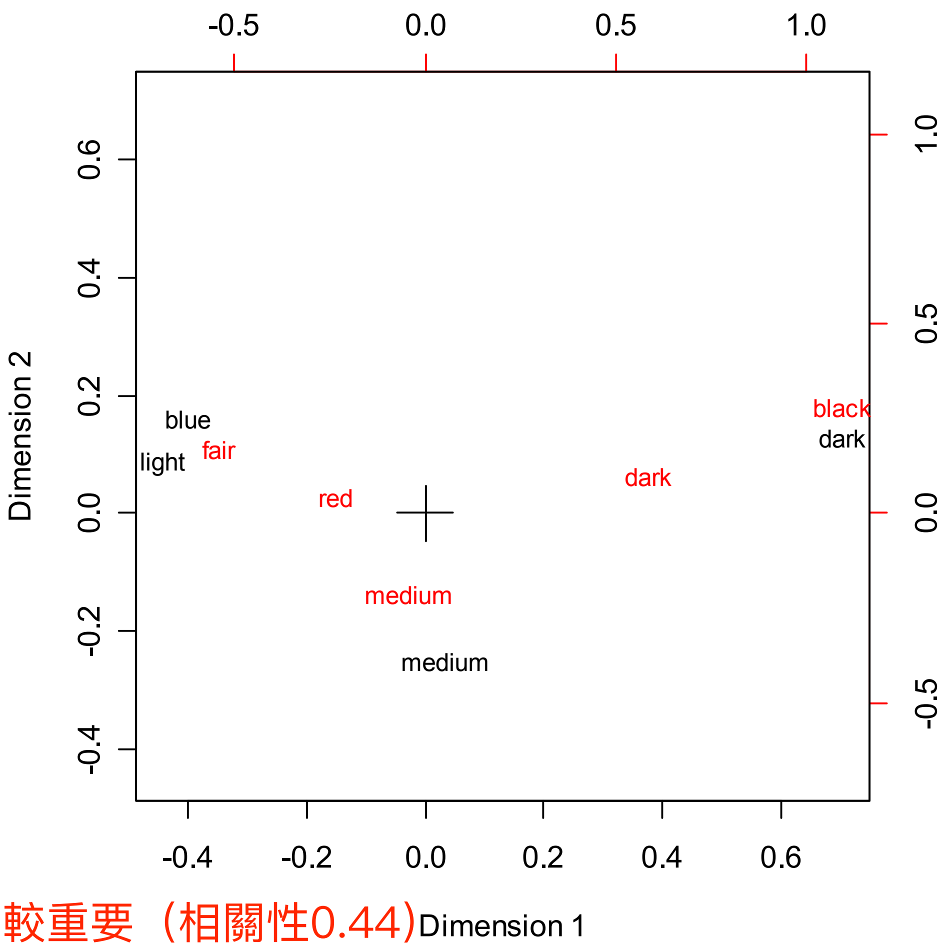
Eye Color

	fair	red	medium	dark	black
blue	326	38	241	110	3
light	688	116	584	188	4
medium	343	84	909	412	26
dark	98	48	403	681	85

Numerical Results

- $\mathrm{Corr}(\widetilde{X}_1, \widetilde{Y}_1) = \lambda_1 = 0.446,\ \mathrm{Corr}(\widetilde{X}_2, \widetilde{Y}_2) = \lambda_2 = 0.173,$
  $\mathrm{Corr}(\widetilde{X}_3, \widetilde{Y}_3) = \lambda_3 = 0.029.$
- The total inertia explained by the 1st dimension is 87%, while the 2nd dimension explains almost the rest of 13%.
- The first 2 dimensions are adequate to explain the relationship (whatever it is) between the rows and columns.

The 2-D Solution(from R)



橫向相對位置越近比較重要（相關性0.44）

Multiple Correspondence Analysis

- MCA handles more than two categorical variables.
- Example:

	X_1	X_2	X_3	$\cdots$	X_J
1	A	a	I		
2	A	b	I		
3	A	b	II		
	B	a	II		
	B	b	III		
	.	.	.		
	.	.	.		
N	.	.	.		

Data Matrix

- How do we present the data matrix ?
- Suppose we have *N* objects and *J* categorical variables, the variable *j* has k_j categories.
- <u>Idea</u>: using an indicator matrix G_j to represent the *j*-th column vector (i.e. the *j*-th variable) X_j
- Let G_j be a matrix with elements $G_j(i,t)=1$ if object *i* belongs to category *t* ; otherwise $G_j(i,t)=0$.

$$G_j = \begin{array}{c|cccc} & 1 & 2 & & k_j \\ \hline 1 & 1 & 0 & \cdots & 0 \\ 2 & 0 & 0 & \cdots & 1 \\ \vdots & & & & \\ N & & & & \end{array} \Rightarrow G = \left[G_1 \mid G_2 \mid \cdots \mid G_J\right]_{N\times(\sum k_j)}$$

The Burt Table

- Calculate the Burt table: $C = G'G$
- Example:

	X_1	X_2	X_3
1	A	C	E
2	B	C	F
3	B	C	E
4	A	C	F
5	B	D	G

$$N = 5,$$

$$k_1 = 2, k_2 = 2, k_3 = 3.$$

$$\Rightarrow \; G = \begin{array}{c} \begin{array}{ccccccc} X_1 & & X_2 & & X_3 & & \end{array} \\ \left[\begin{array}{cc:cc:ccc} 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \end{array}\right] \end{array}, \qquad G' = \left[\begin{array}{ccccc} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ \hdashline 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ \hdashline 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right].$$

The Burt Table

$$C = G'G = \begin{array}{c} X_{11} \\ X_{12} \\ X_{21} \\ X_{22} \\ X_{31} \\ X_{32} \\ X_{33} \end{array} \begin{bmatrix} 2 & 0 & & & & & \\ 0 & 3 & & & & & \\ & & 4 & 0 & & & \\ & & 0 & 1 & & & \\ & & & & 2 & 0 & 0 \\ & & & & 0 & 2 & 0 \\ & & & & 0 & 0 & 1 \end{bmatrix} \underset{= 7\times 7}{\left(\sum k_j\right)\times\left(\sum k_j\right)}$$

➔ MCA corresponds to perform CA on the Burt table *C*.

MCA Solution-1

- Since C is a squared matrix, consider

Eigen decomposition

$$J^{-1}D^{-1/2}\left(C - Duu'DN^{-1}\right)D^{-1/2} = B\Lambda B'$$

where $D = \mathrm{diag}(C)$, $J = \#$ of variables, $N = \#$ of objects,

$$u = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \ \Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \ddots \end{pmatrix}, \quad B = \text{eigenvectors}.$$

- Since C is symmetric, only category points are given:

$$X = Y = N^{1/2}D^{-1/2}B\Lambda.$$

MCA Solution-2

- To plot objects (X) and categories (Y) simultaneously, recall that in simple CA we consider

$$J^{-1/2} L G D^{-1/2} = U \Lambda V'$$

where L is a centering operator that leaves G in deviations from its column means (i.e. LG = each element in G minus its column mean).

- Similar to the "Biplot" in PCA, we can set: $\begin{cases} X = U\Lambda \\ Y = V' \end{cases}$
- Some algebra shows that the elements of Y are the centroids of all objects belonging to that particular category.
- The maximum number of MCA dimensions = $\left(\sum_{j=1}^{J} k_j\right) - J$.

Example: Mammals Dentition

- The data is taken from Hartigan's book (Clustering Algorithms, 1975), where dental characteristics are used to classify mammals.
- Variables:

TI: Top incisors; 1: 0 incisors, 2: 1 incisor, 3: 2 incisors, 4: 3 or mor e incisors
BI: Bottom incisors; 1 : 0 incisors, 2: 1 incisor, 3: 2 incisors, 4: 3 in cisors, 5: 4 incisors
TC: Top canine; 1: 0 canines, 2: 1 canine
BC: Bottom canine; 1: 0 canines, 2: 1 canine
TP: Top premolar; 1: 0 premolars, 2: 1 premolar, 3: 2 premolars, 3: 2 prem olars, 4: 3 premolars, 5: 4 premolars
BP: Bottom premolar; 1: 0 premolars, 2: 1 premolar, 3: 2 premolars, 3: 2 premolars, 4: 3 premolars, 5: 4 premolars
TM: Top molar; 1: 0-2 molars, 2: more than 2 molars
BM: Bottom molar; 1: 0-2 molars, 2: more than 2 molars

Data Summary

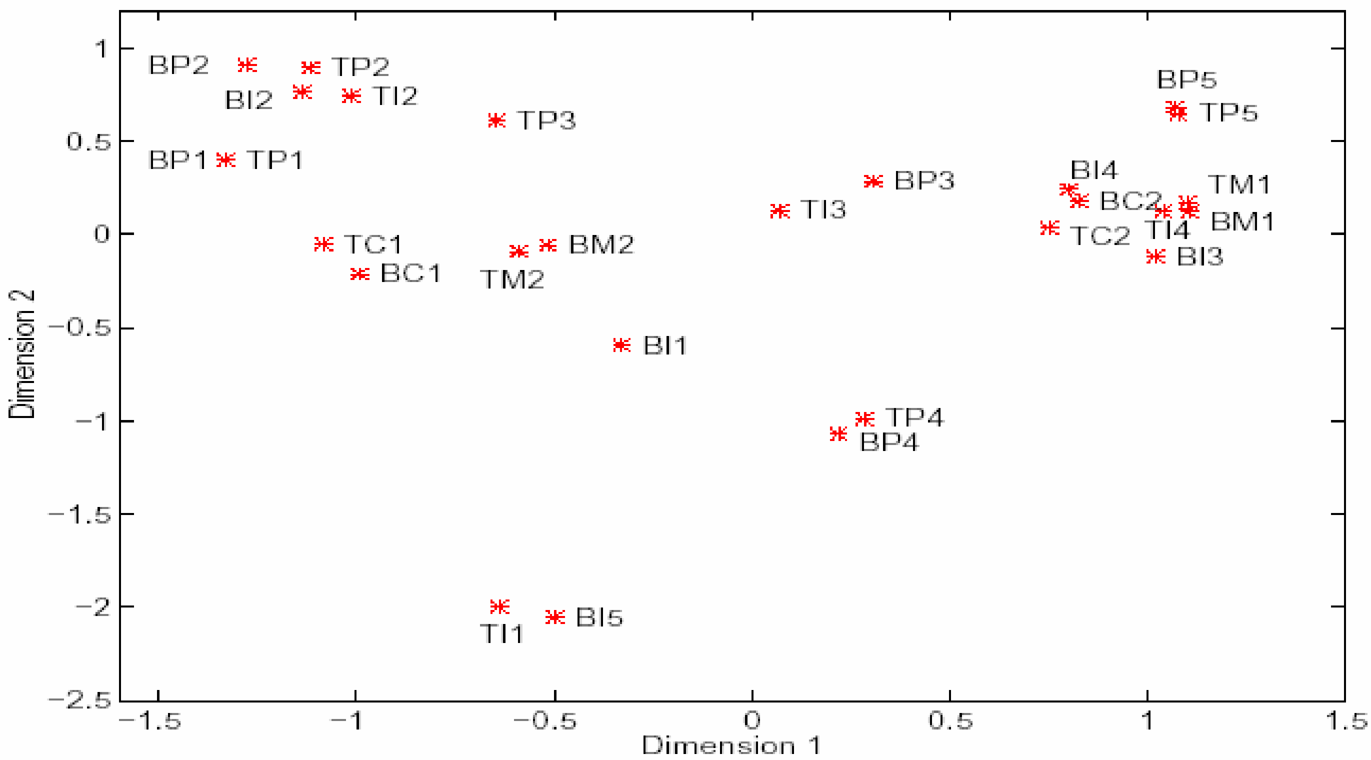
	Categories				
Variable	1	2	3	4	5
TI	15.2	31.8	13.6	39.4	
BI	3.0	30.3	7.6	43.9	15.2
TC	40.9	59.1			
BC	45.5	54.5			
TP	9.1	10.6	18.2	39.4	22.7
BP	9.1	18.2	15.2	36.4	21.2
TM	34.8	65.2			
BM	31.8	68.2			

TABLE 1. Mammals teeth profiles (in %, N=66)

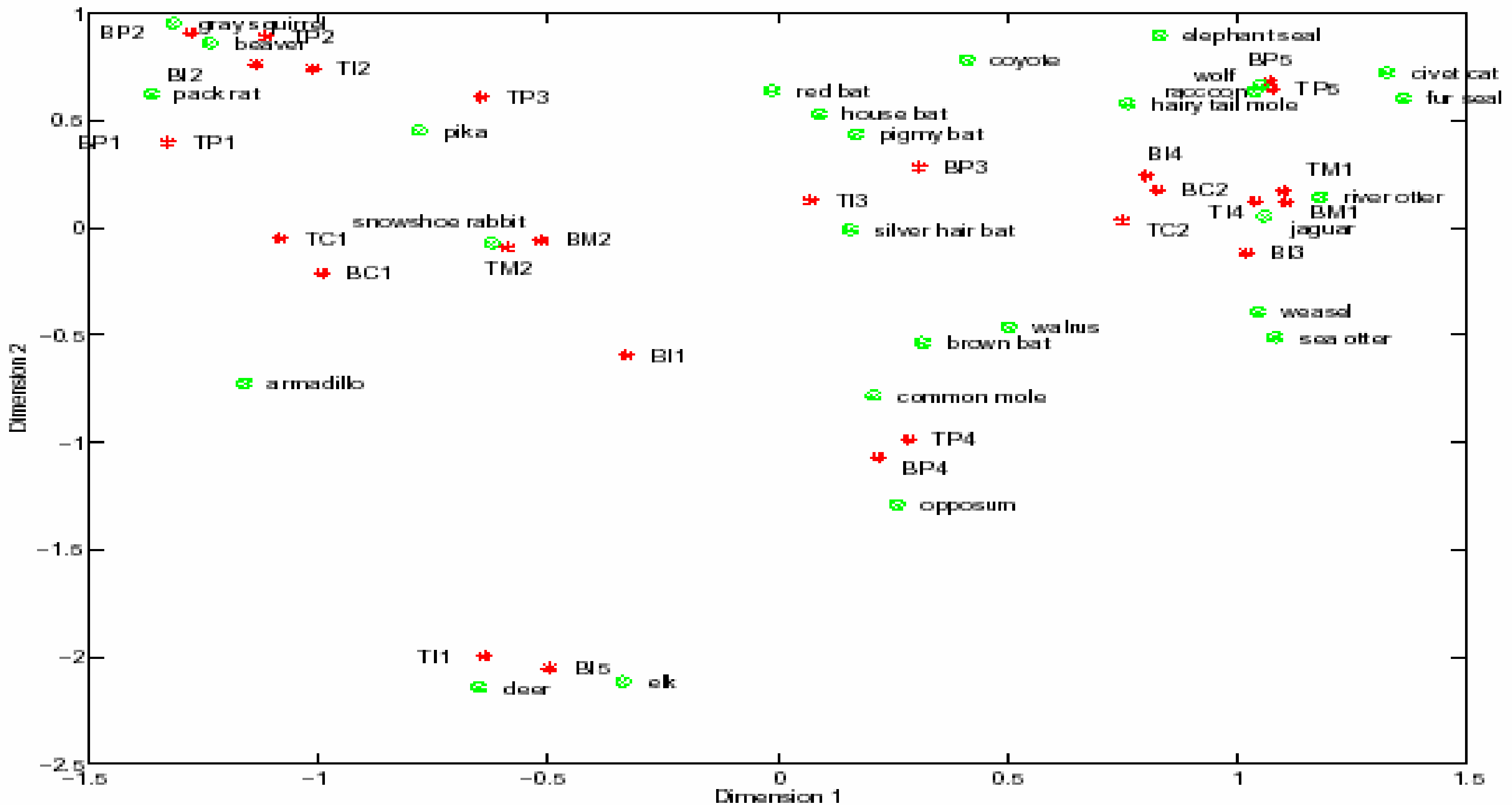
Armadillo	11111122	Skunk	44224411
Pika	32113322	River Otter	44225411
Snowshoe Rabit	32114322	Sea Otter	43224411
Beaver	22113222	Jaguar	44224311
Marmot	22113222	Ocelot	44224311
Groundhog	22113222	Cougar	44224311
Prairie Dog	22113222	Lynx	44224311
Ground Squirrel	22113222	Fur Seal	43225511
Chipmunk	22113222	Sea Lion	43225511
Gray Squirrel	22112222	Walrus	21224411
Fox Squirrel	22112222	Grey Seal	43224411
Pocket Gopher	22112222	Elephant Seal	32225511

A 2-D Solution for Categories

- Check out the singular values, it reveals that a 2-D solution is adequate.



Objects vs Categories



Frequency on Object Locations

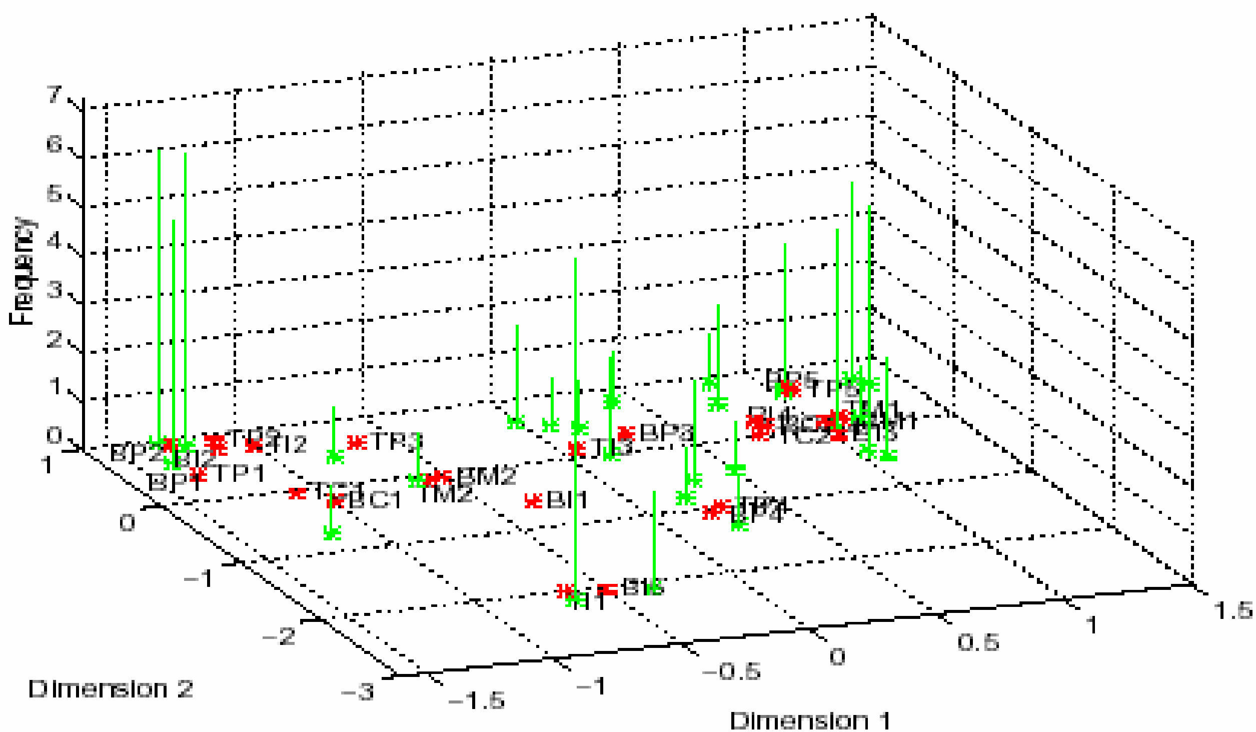


FIGURE 3. Category quantifications (red) and object scores (green) (height of the object scores shows how many mammals share the particular teeth profile)

Summary-1

- MCA can be thought as the joint analysis of all the two-way tables composing the Burt table.
- The problem of MCA is that the total inertia is usually high while the percentages of inertia along the principal axes are invariably low.

  Possible alternatives are:
  - Joint Correspondence Analysis (Greenacre, M. 1988) – consider off-diagonal blocks of C.
  - Homogeneity Analysis (Gifi, A. 1990)
  - Analysis of Profile Frequencies (ANAPROF) – using a different matrix F instead of C.

Summary-2

- XLSTAT is a commercial statistical package which can implement CA and MCA in a Microsoft Excel environment..