Multiple Correspondence analysis of a crosstabulations matrix using the Kohonen algorithm

Smail Ibbou , Marie Cottrell
SAMOS Université Paris 1
90, rue de Tolbiac F-75634 Paris Cedex 13
E-mail : ismail@univ-paris1.fr
E-mail : cottrell@univ-paris1.fr

1. Introduction

The multiple correspondence analysis is a statistical technique to handle qualitative variables and to try to show the correlations between several kinds of variables in a population sample. Classical methods like canonical analysis or factorial analysis are proven to be efficient to deal with this sort of problems. But they present some inconvenients : they are intrinsically linear and moreover they provide graphic representations wich have no true significance overall when there are more than two crossed variables. In an previous paper [3], M.Cottrell et al. had defined a new algorithm (KOUPLET) wich allows to deal with the case where Q equals two. This algorithm is inspired from the self organisation Kohonen algorithm. In this paper, we present another Kohonen-like algorithm to analyze the relations between Q qualitative variables with $Q \geq 2$.

2. The Problem and the Notations

Let us define the data and introduce the basic notations. Let us consider a $N - sample$ of individuals and Q variables or questions. Each question has m_q possible answers (or modalities). The individuals answer each question q $(1 \leq q \leq Q)$ by choosing only one modality among the m_q modalities. If we assume that $Q = 3$ and $m_1 = 3$, $m_2 = 2$ and $m_3 = 3$, then an answer of an individual could be $(0, 1, 0|0, 1|1, 0, 0)$, where 1 corresponds to the chosen modality for each question. Let us denote by M the total number of all the modalities : $M = \sum_{q=1}^{Q} m_q$. To simplify, we can enumerate all the modalities from 1 to M and denote by Z_i, $(1 \leq i \leq M)$ the column vector constructed by the N answers to the i-th modality. The k-th element of the vector Z_i is 1 or 0, according to the choice of the individual k (it is 1 if and only if the individual k has chosen the modality i). Then we can define a $(N \times M)$ matrix K as logical canonical matrix whose columns are the Z_i vectors. It is composed with Q

blocks where each $(N \times m_q)$ block contains the N answers to the question q. One has :

$$K = (Z_1, ..., Z_{m_1}, ..., Z_i, ...Z_M)$$

K	m_1		m_q		m_Q
1	0		0		... ...
	1		0		...
	0		1		...
	...				
	1		0		...
N	0		0		... ...

This matrix K gives the complete data and is called *complete disjunctive table.*It is essential if we want to remember who answered what. But if we only have to study the *relations between the Q variables (or questions),* we can sum up the data in a crosstabulations table, called *Burt matrix*, defined by $B = K'K$ where K' is the transposed matrix of K. The matrix B is a $(M \times M)$ symetric matrix and is composed of $Q \times Q$ blocks, such that the $(q \times r)$ block B_{qr} $(for\ \ 1 \leq q, r \leq Q)$ is the contingency table which crosses the question q and the question r. The block B_{qq} is a diagonal matrix, whose diagonal entries are the numbers of individuals who have respectively chosen the modalities $1, ..., m_q$, for the question q. The Burt table can be represented as below. It has to be seen as a generalized contingency table, when are more than 2 kinds of variables to simultaneously study.

B	Z_1	Z_j	Z_M
Z_1	n_1 0 0 0 n_i 0 0 0 n_{ml}		
Z_i		n_{ij}	
Z_M			. 0 0 0 . 0 0 0 n_M

From now, we denote by n_{ij} the entries of the matrix B, whatever are the questions which contain the modalities i or j. According to the data if i and j are two different modalities of same question, $n_{ij} = 0$ and if $i = j$, the entry n_{ii} is the number of individuals who chose the modality i. In that case, we use only one sub index and write n_i instead of n_{ii}. This number is nothing else than the sum of the elements of the vector Z_i. Each row of the matrix B characterizes *a modality of a question* (or variable). One can observe that for each row i (or column, B is symetric), $\sum_j n_{ij} = Qn_i$, since this number is repeated in each block of the matrix B and that
$\sum_i n_i = \sum_{q=1}^{Q} \sum_{l=1}^{m_q} n_l = NQ$. So the total sum of all the entries of B is
$b = \sum_{i,j} n_{ij} = Q \sum_{n_i} = Q^2 N$. One defines successively :

-the table F of the relative frequencies, with entry $f_{ij} = \frac{n_{ij}}{b}$,

-the margins with entry $f_{i.} = \sum_j f_{ij}$ or $f_{.j} = \sum_i f_{ij}$,

-the table P of the profiles which sum to 1, with entry $P_{ij} = \frac{f_{ij}}{f_{i.}}$

The classical Multiple Correspondence Analysis (**MCA**), (see for example [1]), is a weighted Principal Component Analysis computed on the M row profiles of the matrix P, for the χ^2 metric between the rows, each row being weighted by $f_{i.}$. Let *r(i)* and *r(i')* be two row profiles of the matrix P. One has :

$$\chi^2(r(i), r(i')) = \sum_j \frac{1}{f_{.j}}(P_{ij} - P_{i'j})^2 = \sum_j \left(\frac{f_{ij}}{\sqrt{f_{.j}} f_{i.}} - \frac{f_{i'j}}{\sqrt{f_{.j}} f_{i'.}} \right)^2$$

So it is equivalent to compute profile matrix C whose entry is $c_{ij} = \frac{f_{ij}}{\sqrt{f_{.j} f_{i.}}}$, to consider the Euclidean distance between its rows and to inpute a weight $f_{i.}$ at each row. After realizing the Principal Component Analysis on this data matrix C, the classical Multiple Correspondence Analysis provides a simultaneous representation of the M vectors on a low dimensional space which gives some information about the relations between the Q variables. But as it is possible to use the Kohonen algorithm to get such a representation, (for which there is no more constraint of linearity of the projection), we propose to train a Kohonen network with these *corrected row profiles* as inputs, and a probability distribution $\pi = (f_{1.}, ..., f_{i.}, ..., f_{M.})$ and to study the resulting map to extract the relevant information about the relations between the Q variables.

3. The Kohonen algorithm and the KMCA

The Kohonen algorithm is an unsupervised algorithm that produces a feature map preserving the input topology of data. Let us briefly recall its definition. Consider a $k \times k$ network (bidimensional grid), where a topological neighborhood is defined in a homogeneous way around each unit.

Each unit u is represented by a weight vector $w(u)$ in R^M; the weights are initialized at random. The training at step t consists of
- presenting a stimulus $c(i)$ i.e a row of the *corrected row profile matrix C*, according to the probability distribution π.
- look for the winner unit, i.e that one which minimizes $||c(i) - w(u)||^2$ for all the units u.
- update the weights of the winner unit and its neighbors by
$w_{t+1}(u) - w_t(u) = \epsilon(t)\,(c(i) - w(u))$ for $u = u_0$ or neighbour of u.

The neighbourhood function and the adaptation parameter ϵ are decreasing-time functions. See [5] or [4] for the definitions and properties of this well-known and largely used algorithm. After training, each row profile $c(i)$ can be represented by its corresponding winner unit. Because of the topology preserving property of the Kohonen algorithm, the representation of the M inputs on the $K \times K$ grid emphasizes the *proximity* between the modalities of the Q questions (or variables).

We apply this new method called **Kohonen Multiple Correspondence Analysis (KMCA)** to several examples which are presented in the next sections.

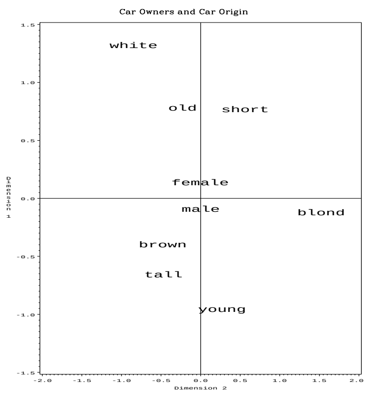
4. Examples

4.1. Physical characteristics

This example is extracted from the SAS examples (version 6.0). We compute a crosstabulations table of four variables : Age (Old, Young), Sex (Male, Female) Height (Tall, Short), Hair (White, Brown, Blond). So $Q = 4$ and $m_1 = 2$, $m_2 = 2$, $m_3 = 2$ and $m_4 = 3$. The *Burt table* is not writen down for simplicity. After training a 4×4 Kohonen gird, we get this map :

Blond		Young	Tall
Female			Brown
			Male
White	Old	Short	

We can compare it with the representation that we get by projection of the 9 vectors on the plane of the two first factorial directions :



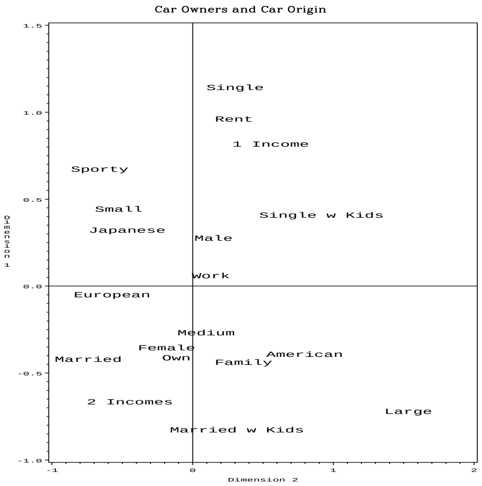
We observe that the main conclusions are the same : association between Tall, Brown and Young, between White, Old and Short, etc.

4.2. Cars and their owners

(Example coming from SAS 6.0). We cross five variables: origin (American, Japanese, European), size of car (Small, Medium, Large), type of car (Family, sporty, Work vehicle), home ownership (Owns, Rents), marital/family status (Single, Married, Single and living with children, Married and living with children), income (1-income, 2-income), sex (Male, Female). One has $M = 19$. The Kohonen map is below:

1-income Single	Rent			Single-w-kids
	Sporty		European	Work
Japanese Small Male				Large
		Female		American
2-income Married		Own	Medium	Family Married-w-kids

The classical Multiple Correspondence Analysis gives:



We can deduce the same main conclusions: 2-income with Married, Married with kids near Family car, Sporty car near Single, Male, Small car or Japenese, etc.

4.3. Historical monuments

We study this example with $Q = 2$ to compare this method with the KOUPLET method defined in [3], since there are here 2 variables. We cross the variable Historical monuments with ten categories (Prehistoric, Historic, Caste, Military, Cathedral, Church, Chapel, Monastery, Public equipement, Private equipement, Diverse) and the variable Owners with six categories (Town, Territorial and administrative division of France, State, Public, Private, no determined).

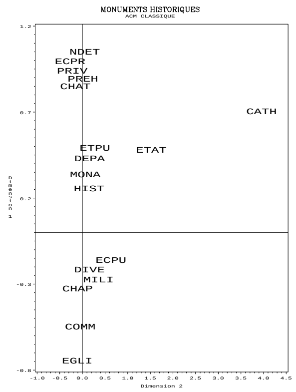
The resulting Kohonen map is:

PREH NDET		MILI	CHAP	EGLI COMM
		HIST		
ECPR			ECPU	DIVE ETPU
PRIV		ETAT		
CHAT		CATH		MONA DEPA

The KOUPLET map is:

PREH NDET				EGLI
			CHAP	COMM
PRIV CHAT ECPR		HIST	MILI	DIVE
		MONA		ECPU
ETAT CHAT		DEPA		ETPU

The classical map is below and the conclusions are mainly the same:



5. Discussion

The first results that we get are very promissing: on the simple examples that we study, the KMCA method very quikly provides good representations of the relations between several qualitative variables. The main advantage is that there is no arbitrary choice of the representation. The M vectors in R^M which correspond to the modalities are correctly classified by the network and the map is realized in a very natural way. However, it is well-known that the classical representations uses a strong approximation which can make an interpretation of the relations very difficult. The disadvantage of the KMCA method is that there remains an open problem in defining some quality criteria, to know for example which is well represented or not on the map. One of our objectives will be to go on in this way to provide such quantitative performance indices.

References

[1] Benzecri J.P., *L'analyse des données, T.2, L'analyse des correspondances*, Dunod, Paris, 1973.

[2] Blayo F., Demartines P. : Data Analysis : How to compare Kohonen neural networks to other technics ? In *Proc of IWANN 91*, Prieto Ed., L.N.C.S., Springer, 469-476, 1991.

[3] Cottrell M., Letremy P., Roy E., : Analysing a contingency table with Kohonen maps : a Factorial Correspondence Analysis In *Proc of IWANN 93*, J.Cabestany,J.Mary,A.Prieto Eds., Springer, 305-311, 1993.

[4] Cottrell M., Fort J.C, Pagès G., : Two or three things that we know about the Kohonen algorithm, In *Proc of ESANN 94*, M.Verleysen ED., D Facto, Bruxelles.

[5] Kohonen T., *Self-organisation and Associative Memory*, Springer, 1989.