– PANTHÉON - SORBONNE –
UNIVERSITÉ PARIS 1

Centre de Recherche

SAMOS

Statistique Appliquée &
MOdélisation Stochastique

Treatment of qualitative variables by Kohonen algorithm. Simultaneous classification of modalities and individuals.

S. Ibbou

Prépublication du SAMOS N° 102

Septembre 1998

Treatment of qualitative variables by Kohonen algorithm. Simultaneous classification of modalities and individuals

Smaïl Ibbou *

Abstract

Qualitative variables occupy a significant place in data analysis and their processing is not completely obvious. Several questions arise when we try to deal with them: which coding of the variables to adopt, which distance to use for the variables and for the individuals? But one of the principal encountered difficulties relates to the classification of the individuals when the size of the data is very large. We propose here two methods, *KACM I* and *KACM II*, based on the Kohonen algorithm which provide classification and map similar to Multiple Correspondence Analysis projection. The second method *KACM II* is very fast from the point of view of the computing time.

1 Introduction

There are many aspects of categorical data problems. In this paper we consider the following situation with N "*individuals*" [1] that answer to Q questions. Each question has a finite number of "*modalities*" ("choices", "options") and is answered by only one modality. In the classical multiple correspondence analysis, the purpose is to see the relations between the modalities and to project them in a factorial subspace. We can also do the same projection for the individuals and even a simultaneous projection including the modalities of the variables[2]. The third operation that we can consider, in this context, is the segmentation of the individual space into homogeneous classes by means of a hierarchical classification algorithm or a k-means algorithm. In this study, we propose to realize this three operations (double projection and classification) with only one Kohonen based algorithm.

In the case of quantitative variables [1], the Kohonen algorithm makes a vectorial quantization of the input space described by the observations of d continuous variables. Using this self-organizing method, we obtain a classification and an organized map similar to a Principal Component Analysis projection.

*SAMOS Lab., Université Paris 1 Panthéon-Sorbonne, 90 rue de tolbiac 75634 Paris. ismail@univ-paris1.fr.

[1] The word "individual" is a generic term that represent any statistical object like household, companies, clients...

[2] Although this double projection is not always justified and accepted in the literature.

When we deal with qualitative variables, we seek to highlight the typology of the modalities and we try to emphasize the relations existing between the modalities of the variables. For example, if there is a considerable proportion of individuals that chose the modality "a" of the question "1" and the modality "b" of the question "2", then we will say that the modalities (1,a) and (2,b) are close, and that they attract each other. We would like to get them very close in the representation. Conversely if there is an important proportion of individuals who choose (1,a) and reject (2,b), then these modalities repulse each other, and we would like to observe distant representations. The goal is to represent this kind of observations in a global manner which takes into account all the modalities of all the variables.

It is clear that the Kohonen algorithm [5] which organizes the units by respecting the proximities in the input space can be appropriate for this type of treatment. The first method dealing with qualitative data by means of the Kohonen algorithm, goes back to 1993 and is due to [2]. Baptized "*Kouplet*", this method is convenient when the number of variables (or questions) Q, is equal to two. In 1995, we proposed [3], another Kohonen-like method, called "*KACM*", which makes it possible to classify the modalities of Q qualitative variables, where Q is equal or greater than two. We present here an improvement of this method that allows to represent on the same Kohonen map the individuals and the modalities which characterize them, when we have stored the answers of each individual. The result is the analogue of a simultaneous projection of the individuals and the variables provided by a classical Multiple Correspondence Analysis.

2 The classical Multiple Correspondence Analysis

Let us assume that we have the answers of N individuals to Q qualitative variables, with $Q \geq 2$. Each question q, $1 \leq q \leq Q$, has m_q modalities and we denote by $M = \sum_{q=1}^{Q} m_q$ the total number of modalities. The data table can be the *complete disjunctive* table or the *Burt* table and defined as follows.
Let us denote by $K_{(N \times M)}$ the matrix with N rows and M columns which corresponds to the *complete disjunctive table*:

$$K = (k_{ij}) \quad \text{with} \quad k_{ij} = \begin{cases} 1 & \text{if the individual } i \text{ chooses the modality } j \\ 0 & \text{otherwise} \end{cases}$$

We have $\forall i, 1 \leq i \leq N, \quad k_{i\cdot} = \sum_{j=1}^{j=M} k_{ij} = Q$. The marginal on the rows is constant equal to the number of questions. Let W be the diagonal matrix made up by the elements of the column margin.

$$W = Diag\{k_{\cdot 1}, \ldots, k_{\cdot i}, \ldots, k_{\cdot M}\}$$

where $k_{\cdot i} = \sum_{j=1}^{N} k_{ji}$. The table K is not always available. Sometimes the data can be summarized into a Burt table $B_{(M \times M)}$ defined by

$$B_{M\times M} = (b_{ij}) = K^t K$$

In this case, we loose a part of the information about the individuals answers but we keep the information regarding the relations between the modalities of the qualitative variables. It is a generalized contingency table made up with $Q \times Q$ blocks, and each block B_{qr} for $1 \leq q, r \leq Q$ represents the contingency table which crosses the question q and the question r. We briefly point out (see as an example [6]) that the Multiple Correspondence Analysis (MCA) is equivalent to a weighted Principal Component Analysis (PCA) performed on the row-profiles or column-profiles, obtained with a particular metric known as the Chi-square metric.

Indeed, for a MCA performed on the complete disjunctive table, we consider the matrix of the row-profiles (which is equal to the matrix $\frac{1}{Q}K$), on which we carry out a PCA, by using the metric defined by the matrix NQW^{-1} and the weight of the row-profiles defined by the matrix $\frac{1}{N}I_N$ (I_N is the identity matrix). In the case of the MCA on the Burt matrix, we use the matrix of the row-profiles equal to $\frac{1}{Q}W^{-1}B$, the metric defined by the matrix QNW^{-1} and the weight matrix defined by $\frac{1}{QN}W$. We have finally to do a factorization of the two inertia matrices :

$$\mathcal{I}_K = \frac{1}{Q}K^t K W^{-1} = (K^c)^t K^c, \quad \text{with} \quad K^c = (k^c_{ij}) \quad \text{and} \quad k^c_{ij} = \frac{k_{ij}}{\sqrt{k_{i\cdot}}\sqrt{k_{\cdot j}}}.$$

$$\mathcal{I}_B = \frac{1}{Q^2}BW^{-1}BW^{-1} = (B^c)^t B^c \quad \text{with} \quad B^c = (b^c_{ij}) \quad \text{and} \quad b^c_{ij} = \frac{b_{ij}}{\sqrt{b_{i\cdot}}\sqrt{b_{\cdot j}}}.$$

We introduce tow corrected input matrices (K_c and B_c) that includes the use of convenient metric and weight matrix. In our case, it is much easier to use uniform weighting and Euclidean distance, because in this way we can use the standard Kohonen algorithm. Moreover, although the theoretical demonstrations are done for an unspecified input distribution and any distance, the adaptive parameters, generally empirical, are tested and controlled in the case of the uniform distribution and the usual Euclidean distance. For this reason, and to be able to use a standard SOM (Self Organizing Map), that we come down to this case, by using a corrected Burt matrix or a corrected complete disjunctive table as explained above.

A second remark to be made at this stage concerns the relations existing between the rows of K^c and the rows of B^c. Indeed in the set of modalities, two modalities or more will be close if there is a large proportion of individuals that choose them simultaneously. In the same time, these individuals are grouped in the same region. We state (without a rigorous proof) that a subset of modalities is closer to a subset of individuals that have chosen these modalities than to a group of individuals that have not. In the figure 1, we have symbolized these two distances by d and D. Group M_1 of modalities is chosen by group of individuals I_1, group M_2 of modalities is chosen by group of individuals I_2. We maintain that $d < D$.

Furthermore the axis that maximize the two inertias are equal. Indeed, if a is eigenvector of $\mathcal{I}_K$ corresponding to the eigenvalue λ then a is an eigenvector of $\mathcal{I}_B$ corresponding to the eigenvalue λ^2. In fact we have $\mathcal{I}_B = \mathcal{I}_K \mathcal{I}_K$.

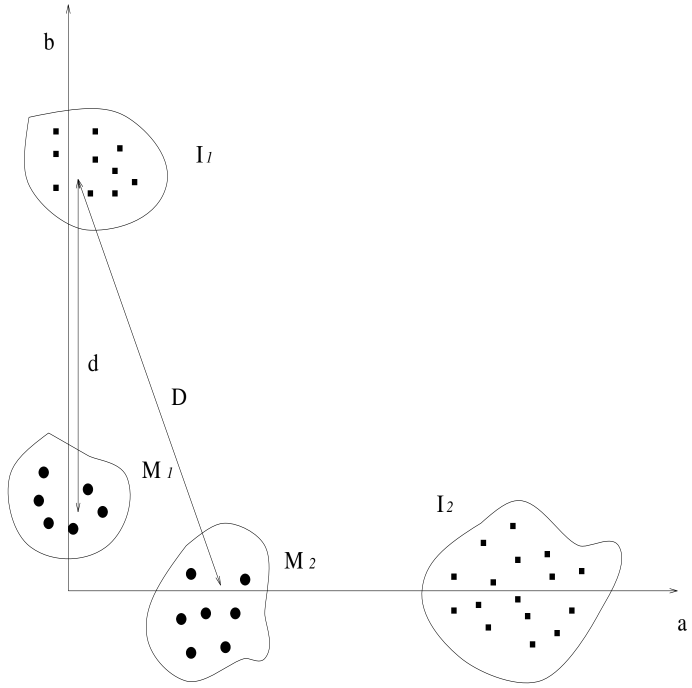


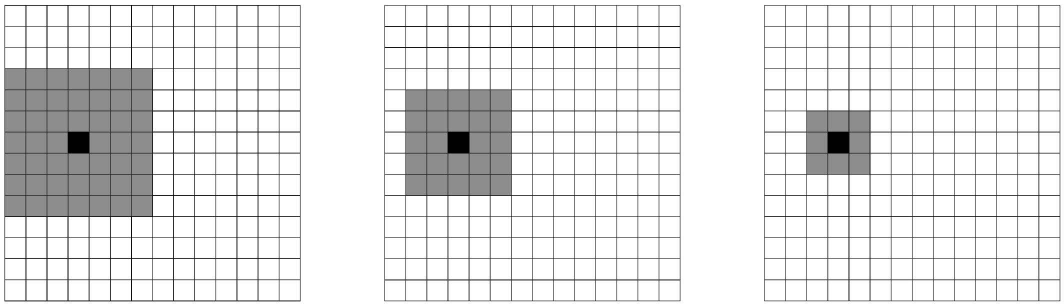
Figure 1: Schematization of two groups of modalities M_1 and M_2 and two corresponding groups of individual I_1, I_2. a and b indicate the two first principal axis.

Actually there remains a difficulty which is not completely solved in the literature, concerning the simultaneous projection of individuals and modalities in the classical multiple correspondence analysis. In fact, the Burt table is a symmetric matrix $M \times M$ where the rows and the columns represent the modalities. On the other hand the complete disjunctive table is a matrix $N \times M$ where the rows represent the individuals and the columns represent the modalities. In both cases, the column vectors represent the same modalities, except that in one case, they are vectors of $\mathbb{R}^M$ and in the other one they are vectors of $\mathbb{R}^N$. It is known that the classical MCA constructs "generic modalities" (principal components) which are linear combinations of the original modalities, i.e. of column vectors . But if the dimensions are distinct, the superposition of the new systems of axes (principal components) is not completely justified.

In our method *KACM*, this conflict of dimensions is not present, since only the distances of the rows of these two matrices, which all are vectors of $\mathbb{R}^M$, are taken into account.
We distinguish two variants *KACM I* and *KACM II* ; the first is using the corrected matrix K^c for the training of the network and the second the corrected matrix B^c.

3 The Self-Organizing Map algorithms

The Kohonen algorithm is an unsupervised neural algorithm that has the property to reproduce the topology of the input space on a network made up by U units arranged following a "*grid*" or a "*line*". Each unit u, $1 \leq u \leq U$, is characterized by a "*code-vector*" or a "*weight-vector*" y_t^u, with the same dimension as the input space. We define also a neighborhood function $V_{r(t)}$ depending of a radius $r(t)$ which is a time-decreasing function. For instance, the two figures below show the neighborhood function for three radii.



49 Neighbors | 25 Neighbors | 9 Neighbors

7 Neighbors | 5 Neighbors | 3 Neighbors

We present now the two original algorithms *KACM I* and *KACM II* to deal with the categorical data by providing a Kohonen map with individuals and variables.

3.1 The KACM I algorithm : Training by Complete and Disjunctive Table

Let us suppose that the network is chosen, i.e. the number of units U and its architecture (line or grid) are chosen. The code vectors of the units are vectors of $\mathbb{R}^M$ and initialized at random. With the rows of this matrix K^c, we proceed to the training of the network:

- At each iteration we choose, uniformly among all the rows of the matrix K^c, one row k_l^c:

$$k_l^c = \left(\frac{k_{l1}}{\sqrt{k_{l\cdot}}\sqrt{k_{\cdot 1}}}, \ldots\ldots, \frac{k_{lM}}{\sqrt{k_{l\cdot}}\sqrt{k_{\cdot M}}}\right).$$

- We look for the winner unit u_0 among all the units in the lattice, which is the unit realizing the minimum of the usual distance.

$$u_0 = \arg\min_u \|y_t^u - k_l^c\|$$

- We update the code vectors of the unit u_0 and its neighbors by the standard formula.

$$y_{t+1}^u = y_t^u + \epsilon(t)(k_l^c - y_t^u)$$

These steps are repeated about 4 or 5 times over the total number of the inputs. We begin with a large radius $r_(t)$ and we decrease it to zero. The adaptive parameter $\epsilon(t)$ verifies the Robbins Monroe conditions: $\sum_t \epsilon(t) = \infty$ and $\sum_t \epsilon(t)^2 < \infty$.

After training, each individual is classified in the network by assignation to his winning unit. We then obtain a Kohonen map where only the individuals are classified. To represent the modalities on the same Kohonen lattice, we proceed as follows: with the rows of the matrix B^c, we classify the modalities in the map by assigning each one of them to the closest unit. For example the modality p, $1 \leq p \leq M$, corresponding to the row vector b_p^c will be allocated to the unit u_p

$$u_p = \arg\min_u \|y_T^u - b_p^c\|$$

where y_T^u is the final value of the weight-vector y^u after the training step.

Although the training was not made with the rows corresponding to the modalities, the network obtained can be used to classify the modalities, although it is not a good vectorial quantifier for the modalities. In this method is that groups of individuals which resemble each other for having chosen the same modalities will attract these same modalities.

3.2 The KACM II algorithm : Training by the table of Burt

The second method consists in using only the rows of the matrix B^c to train the lattice. This one being involved, we classify then the modalities represented by the rows of the matrix B^c. To classify the individuals on the same map we use the rows of the matrix K^c.

The KACM II algorithm

- For each iteration

  We choose uniformly among all the rows of the matrix B^c, one row b_l^c:

$$b_l^c = \left(\frac{b_{l1}}{\sqrt{b_{l\cdot}}\sqrt{b_{\cdot 1}}}, \ldots\ldots, \frac{b_{lM}}{\sqrt{b_{l\cdot}}\sqrt{b_{\cdot M}}}\right).$$

- We look for the winner unit u_0 among the whole of the units of the network which is the unit that carries out the minimum of the usual distance :

$$u_0 = \arg\min_{u \in R} ||y_t^u - b_l^c||$$

- We update the code vectors of the unit u_0 and its neighbors by

$$y_{t+1}^u = y_t^u + \epsilon(t)(b_l^c - y_t^u)$$

After training, we classify the modalities (represented by the rows of B^c) in the network by assigning each of them to the nearest unit u of the network. To represent the individuals on the same Kohonen map we proceed as *KACM I* but using the rows of the matrix K^c. For example the individual j, $1 \leq j \leq N$, corresponding to the row vector k_j^c will be affected to the unit u_p

$$u_p = \arg\min_{u} ||y_T^u - k_j^c||$$

where y_T^u is the final value of the weight-vector y^u after the training step.
Generally the number of modalities is not very large; the training of the network is consequently very fast. This method is very interesting and very computing time saving. In fact, if we consider the case of very large data files (as in insurance companies or marketing data), it happens that in this type of data base, the disjunctive complete table has hundred variables but hundreds of thousands (or more) individuals. It can take several hours to classify the individuals into groups by a hierarchical classification. Using this method (*KACM II*), it is sufficient to compute the Burt matrix and to train a Kohonen network with its rows. After this, it is easy to classify the individuals in the lattice by seeking the winner units related to the rows of the matrix K^c. Now let us illustrate this method with an example.

4 Example

This example is taken from [6]. The data consist in a characterization of 27 races of dogs by the 7 following variables: **Velocity**(Small, Average, big), **Size** (Small, Average, big), **Weight** (Small, Average, big), **Affection** (Affectionate, Non-affectionate), **Intelligence** (Small, Average, big), **Aggressiveness** (Aggressive, Non-aggressive), **Function** (Assistance, Hunting, Company).

Beauceron Alsatian			Fox-Hound			Mastiff	Saint-Bernard Newfoundland
	Doberman				German-Dog		Bull-Mastiff
			Greyhound				
Basset			Pointer		Setter		Gascogne
						Spaniel-F	
Chihuahua Pekinese		Poodle		Collie		Labrador	Spaniel-B
Bull-Dog Dachshund		Fox-Terrier		Cocker		Boxer	Dalmatien

Table 1: Kohonen map on individuals; the learning is done with the rows of the corrected matrix K^c

Company Cocker		*Affect*		*Ave Intell*			*Ave Size* Dalmatien Labrador Boxer
Fox-Terrier				*Non-Agress* Collie		*Ave Veloc*	Spaniel-B
Sma Weight Bull-Dog dachshund	Poodle				Spaniel-F		
	Sma Size	Chihuahua Pekinese			*Hunting* Setter	*Ave Weight*	
			Basset		Fox-Hound Gascogne Greyhound	Pointer	*Big Intell* Beauceron Alsatian Doberman
				Sma Intell			*Big Veloc*
	Sma Veloc						*Big Size*
			Aggressive	*Big Weight* Bull-Mastiff German-Dog Mastiff Saint-Bernard Newfoundland	*Assistance*		*Non-Affect*

Table 2: Kohonen map on individuals and modalities; the learning is done with the rows of the corrected matrix K^c

Sma Size *Sma Weight* Chihuahua Pekinese dachshund	Bull-Dog		Basset	*Sma Intell*			*Big Weight* Mastiff Saint-Bernard Newfoundland
Poodle Fox-Terri			*Sma Veloc*			German-Dog	Bull-Mast
Company							*Assistance*
Affect	Cocker	*Ave Intell*			*Aggressive*		
Ave Size Boxer Dalmatien Labrador	*Ave Velocity* Spaniel-B			*Big Intell* Beauceron Alsatian			*Non-Affect*
				Pointer	Doberman		
	Non-Agress Collie		*Hunting* Spaniel-F GrBlGasco Greyhound Setter	*Ave Weight*	*Big-Velocity* Fox-Hound		*Big-Size*

Table 3: Learning done with the rows of the matrix B^c

In these three maps, we have got the good clusters. Indeed we have for example small size, company and Poodle, Dachshund in the same region of the map, hunting with Setter and Greyhound etc.

5 Conclusion

The results of this algorithm are very satisfactory and promising: on the examples for which we have applied the algorithms *KACM I* and *KACM II*, we quickly obtained a very good representation of the relations between the variables. These algorithms have the advantage of comparing the profiles by using only the distance between these profiles.

It is a method of natural classification which does not use a linear approximation which can make the interpretation of the results sometimes difficult. The maps which we obtain are less precise than classical projections of the MCA, but they summarize very well the various relations (attractions, repulsions) between modalities.

Even if the simultaneous representation of the various modalities of the variables and the individuals does not have a rigorous justification, it gives nevertheless good results. The methods *KACM I* and *KACM II* have the advantage of producing on the one hand a Kohonen map comparable with a traditional MCA projection and on the other hand a classification of the individuals compared to the modalities. Moreover the method *KACM II* is particularly cheap in computing time, which is a considerable advantage compared to other classical classification methods such as a hierarchical classification.

References

[1] F. Blayo and P. Demartines. Data analysis : How to compare Kohonen neural networks to other technics ? In Prieto, editor, *Proc of IWANN 91*, Lectures Notes in Computer Science, pages 469–476. Springer-Verlag, 1991.

[2] M. Cottrell, P. Letremy, and E. Roy. Analyzing a contingency table with Kohonen maps : a Factorial Correspondence Analysis. In J. Cabestany, J. Mira, and A. Prieto, editors, *Proc of IWANN 93*, pages 305–311. Springer-Verlag, 1993.

[3] S. Ibbou and M. Cottrell. Multiple Correspondence Analysis of a crosstabulations matrix using the Kohonen algorithm. In M. Verleysen, editor, *Proc of ESANN'95*, pages 27–32. D facto Bruxelles, 1995.

[4] Smaïl Ibbou. *Classification, analyse des correspondances et méthodes neuronales.* PhD thesis, Université Paris 1 Panthéon-Sorbonne, SAMOS, Paris, 1998. http://www.univ-paris1.fr/SAMOS/.

[5] T. Kohonen. *Self-organization and associative memory.* Springer, New York Berlin Heidelberg, 1984. 3^{rd} edition 1989.

[6] G. Saporta. *Probabilités Analyses des données et statistique.* Editions Technip, Paris, 1990.