ASSESSING SELF ORGANIZING MAPS VIA CONTIGUITY ANALYSIS

Ludovic Lebart
CNRS and GET-Télécom Paris.
lebart@enst.fr

Abstract - *Contiguity Analysis is a straightforward generalization of Linear Discriminant Analysis in which the partition of elements is replaced by a more general graph structure. Applied to the graph induced by a Self Organising Map (SOM), Contiguity Analysis provides a set of linear projectors leading to a representation as close as possible to the SOM map. As expected, such projectors may only concern local parts of the SOM maps. They allow us to visualize the shapes of the clusters and the pattern of the elements within each clusters. In some contexts, they provide confidence areas for elements via a standard partial bootstrap procedure.*

Key words – Contiguity analysis, Assessment of SOM, Bootstrap

1 Introduction

For many users, the Self Organizing Maps outperform both usual clustering techniques and principal axes techniques (principal components analysis, correspondence analysis, etc.). On the one hand, the displays of identifiers of units within rectangular or octagonal cells allow for clear and legible printings. On the other hand, the SOM grid, basically non-linear, can be viewed as a compromise between a high-dimensional set of clusters and the two-dimensional plane generated by any pairs of principal axes. Some attempts have been made to propose some assessment procedures. Let us mention, among other works, the algorithm of Kleiweg [5] that complements the map by both a progressive darkening of the edges (which indicates stronger differences between the concerned cells) and a drawing of the minimum spanning tree joining the centroids of the non-empty cells; Cottrel and Rousset [2] and Rousset and Guinot [11] propose in the same vein several noticeable improvements to visualize the distances between clusters. The present paper proposes, through Contiguity Analysis (briefly reminded in section 2), a set of linear projectors providing a representation as close as possible to a SOM (section 3 and 4). The sequence of processing is presented in section 5, with the help of an example of application. When a standard partial bootstrap procedure is applicable, we can then provide the clustered elements with confidence areas (ellipses) (section 6).

2 Principles of contiguity analysis

Let us consider a set of multivariate observations, (n observations described by p variables, leading to a (n,p) data matrix **X**), whose n rows have an *a priori* graph structure. Thus, the n observations are also the n vertices of a symmetric graph G, whose associated matrix is **M** ($m_{ii'}$

= 1 if vertices i and i' are joined by an edge, $m_{ii'} = 0$ otherwise). We denote by **N** the (n,n) diagonal matrix having the degree of each vertex i as diagonal element n_i (n_i stands here for n_{ii}). y is the vector whose i-th component is y_i. Note that: $n_i = \Sigma_{i'} m_{ii'}$. **U** designates the square matrix such that $u_{ij} = 1$ for all i and j.

2.1 Local variance $v^*(y)$ of a variable y

y being a random variable taking values on each vertex i of G, the local variance is:

$$v^*(y) = (1/n)\sum_{i=1}^{i=n}(y_i - m_i^*)^2, \quad \text{with:} \quad m_i^* = (1/n_i)\sum_{k=1}^{k=n_i} m_{ik} y_k$$

Note that if G is a complete graph (all pairs (i,i') are joined by an edge), *v*(y)* is nothing but *v(y)*, the classical empirical variance. When the observations are distributed randomly on the graph, both *v*(y)* and *v(y)* are estimates of the variance of y.

The contiguity ratio, analogue to the contiguity ratio of Geary [4], is written:

$$c^*(y) = v^*(y) \,/\, v(y), \text{ or : } \quad c^*(y) = \mathbf{y'(I - N^{-1}M)'\,(I - N^{-1}M)\,y \,/\, y'\,(I - (1/n)U)\,y}$$

2.2 Local covariance matrix

The contiguity ratio can be generalized :
- to different distances between vertices in the graph,
- to multivariate observations (both generalizations are dealt with in [7]).

This section is devoted to the second generalization: the analysis of sets of multivariate observations having an *a priori* graph structure. Such situation occurs frequently in geography, ecology, geology. The multivariate analogue of the local variance is now the local covariance matrix. With the previous notation, its entry cell $cov^*(y_j,y_{j'})$ is given by:

$$cov^*(y_j, y_{j'}) = (1/n)\sum_{i=1}^{i=n}(y_i - m_i^*)(y_{j'} - m_{j'}^*)^2$$

If **X** designates the (n,p) data matrix of the values of the p variables for each of the n vertices of the graph G described by its incidence matrix **M**, the local covariance matrix **V*** is :

$$\mathbf{V^* = (1/n)\,X'(\,I - N^{-1}M)'\,(\,I - N^{-1}M)\,X}$$

The diagonalization of the corresponding local correlation matrix (Local Principal Component Analysis) produces a description of the local correlations, which can be compared to the results of a classical PCA performed with the global correlation matrix [1]. If the graph is made of k disjoined complete subgraphs, **V*** coincide with the classical "within covariance matrix" used in linear discriminant analysis. If the graph is complete (associated matrix = **U**), then **V*** is the classical covariance matrix **V**.

2.3 Contiguity Analysis

Let **u** be a vector defining a linear combination $u(i)$ of the p variables for vertex i:

$$u(i) = \Sigma_j u_j y_{ij} = \mathbf{u'y}_i$$

The local variance of the artificial variable $u(i)$ is then, in matrix notations :

$$v^*(\mathbf{u}) = \mathbf{u}' \mathbf{V}^* \mathbf{u}$$

The contiguity coefficient of this linear combination can be written :

$$c^*(\mathbf{u}) = \mathbf{u}' \mathbf{V}^* \mathbf{u} \; / \; \mathbf{u}' \mathbf{V} \mathbf{u}$$

where **V** is the classical covariance matrix of vector **y**.

The search for **u** that minimizes c*(**u**) produces functions having the properties of "minimal contiguity": these functions are, in a sense, the linear combinations of variables the more continuously distributed on the graph.

Instead of assigning an observation to a specific class, (as it is done in classical linear discriminant analysis) these functions allows one to assign it in a specific area of the graph. Therefore, this technique (designated as Contiguity Analysis) can be use to discriminate between overlapping classes.

3 SOM and external associated graph

The Self Organizing Maps (SOM) proposed by Kohonen [6] aim at clustering a set of multivariate observations. The obtained clusters are often displayed as the vertices of a rectangular (chessboard like) or octagonal graph. The distances between vertices on the graph are supposed to reflect, as much as possible, the distances between clusters in the initial space. The algorithm is similar to the McQueen algorithm [10] in its on line version, and to the k-means algorithm in its batch version.

3.1 Graph associated with a SOM

Figure 1 represent a stylised symmetric matrix (70, 70) $\mathbf{M}_1$ associated to a SOM assigning n = 70 elements to k = 8 clusters. Rows and columns represent the same set of n elements (elements belonging to a same class of the partition form a subset of consecutive rows and columns). All the cells of the black sub-matrices contains the value 1. All the cells outside these black sub-matrices contains the value 0 . These 8 clusters have been obtained through a SOM algorithm from a square 3 x 3 grid (with an empty cluster).

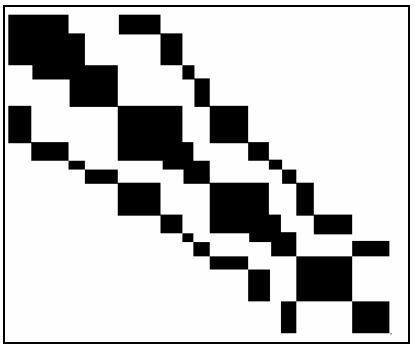

Figure 1 . Stylised incidence matrix $\mathbf{M}_1$ of the graph associated with a (3 x 3) SOM
(all the cells in the white [resp. black] areas contain the value 0 [resp. 1])

In figure 1, two elements i and j are linked (m_{ij} = 1) in the graph if they belong to a same cluster, or if they belong to contiguous clusters. Owing to the small size of the SOM grid, the diagonal adjacency is not taken into account. (e.g.: In the SOM of figure 3 below, elements belonging to cluster 7 are considered as contiguous to those of clusters 4 and 8, but not to the

elements of cluster 5). Similarly to matrices $\mathbf{M}_1$, a matrix $\mathbf{M}_2$ can be defined, that extends the definition of the edges of the graph to diagonal links, and, likewise, matrices $\mathbf{M}_3$ …. $\mathbf{M}_k$.

4. Linear projectors onto the "best SOM plane"

The binary incidence matrices $\mathbf{M}_1$, and $\mathbf{M}_2$ can easily be obtained as a by-product of the SOM algorithm. If a Contiguity Analysis using the graphs G1 or G2 (associated matrices $\mathbf{M}_1$, or $\mathbf{M}_2$) is then performed, the principal planes strive to reconstitute the positions of the clusters in the SOM map. In the initial p-dimensional space, the SOM map can be represented by the graph whose vertices are the centroids of the clusters. Those vertices are joined by an edge if the corresponding clusters are contiguous in the grid used in the algorithm. This graph in a high dimensional space will be partially or totally unfolded by the contiguity analysis. The following example will show the different phases of the procedure.

5. The sequence of operations: an example

The selected data set deals with survey data. The questionnaire [12] consists of 210 words that respondents must rate (on a 7-items scale) according to the pleasure or displeasure they experience at the mention of each of these words.

TO_CONSOLE SOUL SACRED PERFECTION PATIENCE MODERATION **7**	REASON HONOUR ATTACHMENT **8**	TO_FERTILISE TO_ADMIRE SOFTNESS PURITY PEACE COUNTRYSIDE CONFIDENCE CHILDHOOD **9**
TRADITION TO_THINK TO_FORBID NOBLE JUSTICE GOD CERTAINTY ABSOLUTE **4**	TO_WRITE TO_ORDER TO_MASTER SCIENCE PRACTICAL INVENTOR CREATOR BOOK **5**	TO_PRODUCE TO_INHERIT PRECIOUS GOLD FREE ELEGANCE BLUE **6**
WALL UNKNOWN TO_CLIMB TO_BREAK SPACE RIGID METALLIC GUN FIRE DANGER **1** ARMOUR	SPEED RIVER PEAK MOUNTAIN MOON MATERIAL ISLAND **2** ADVENTURER	WATER TO_SWIM TO_DREAM SUBLIME SENSUAL PRUDENCE NUDITY ESCAPE DESIRE CARNAL **3** ANIMAL

Figure 2. A (3 x 3) SOM showing the 70 words, rated in the 3360 responses

The pattern obtained in the space spanned by the six first principal axes of a Principal Component Analysis of the (3360 x 70) data table appears to be stable over time, and similar in several European countries. We run the example on a subset of 70 words and 10 principal coordinates derived from a preliminary PCA performed on a subset of 300 respondents.

Figure 3 simultaneously represents the projections of the 70 variables (words) together with the nine centroids of the nine clusters produced by a classical SOM algorithm (square 3 x 3 grid) onto the plane spanned by the two first principal components.

Figure 4 represents the plane spanned by the two first axes of the contiguity analysis using the matrix $\mathbf{M}_1$. We can check that the graph describing the SOM map (the vertices of which C_1, C_2, …C_9 are the centroids of the elements of the corresponding cells of figure 3), is, in this particular case, a satisfactory representation of the initial map. The pattern of the nine centroids is similar to the original grid exemplified by figure 2.

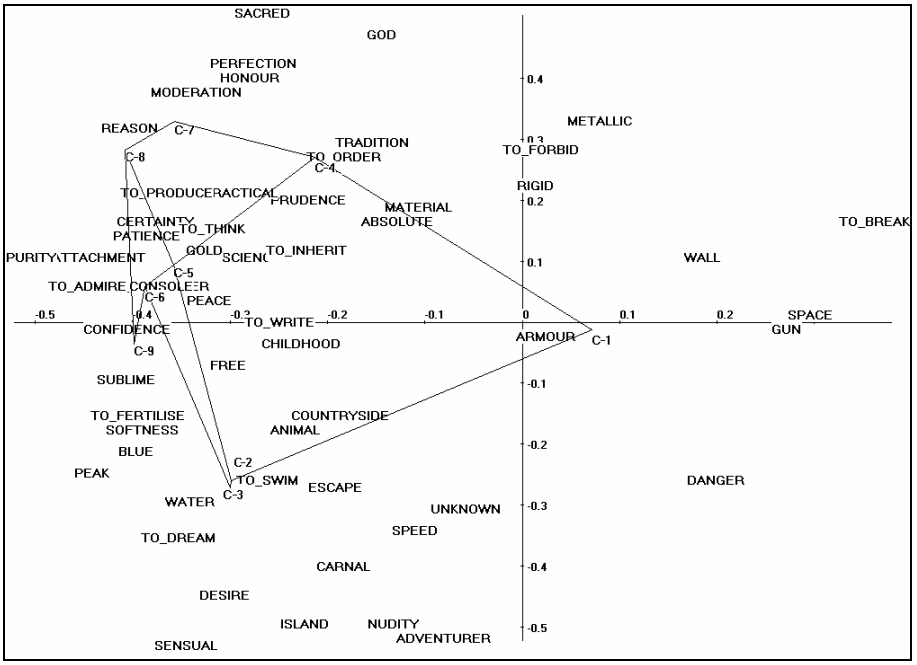

Figure 3. Principal plane of a PCA. The points C1, C2, …C9 represent the centroids of the 9 clusters derived from the SOM map.

The background of figure 5 is identical to that of figure 4. It contains in addition the convex hulls of the nine clusters C1, C2, …, C9.. Each of those convex hulls correspond exactly (if we except some double or hidden points) to a cell of Figure 2. We note that these convex hulls are relatively well separated.

In fact, figure 5 contains much more information than Figure 2, since we have now an idea of the shapes and sizes of the clusters, of the degree to which they overlap. We are now aware of their relative distances, and, another piece of information missing in Figure 2, we can observe the configurations of elements within each cluster.

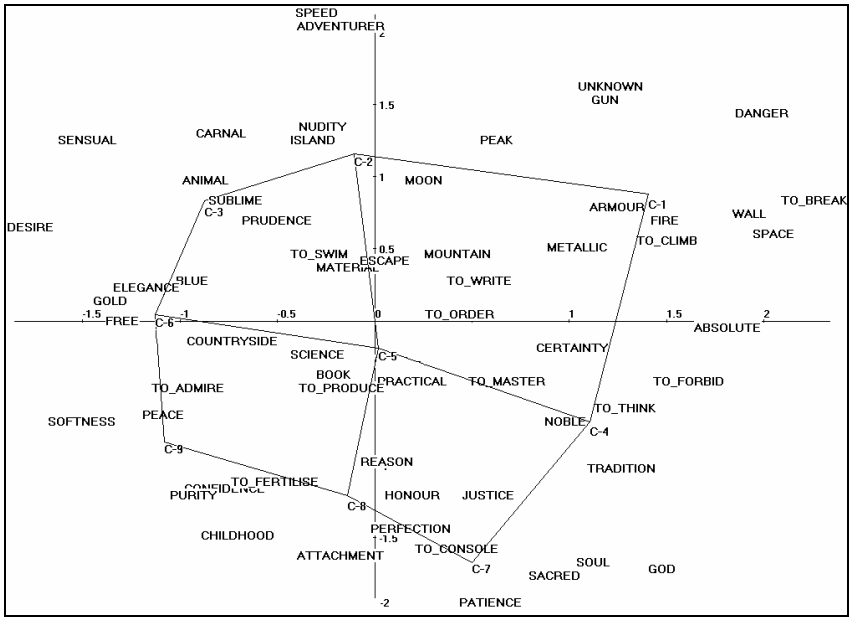

Figure 4. Principal plane of the contiguity analysis using matrix M1. The points C1, C2, …C9 represent the centroids of the 9 clusters derived from the SOM map.

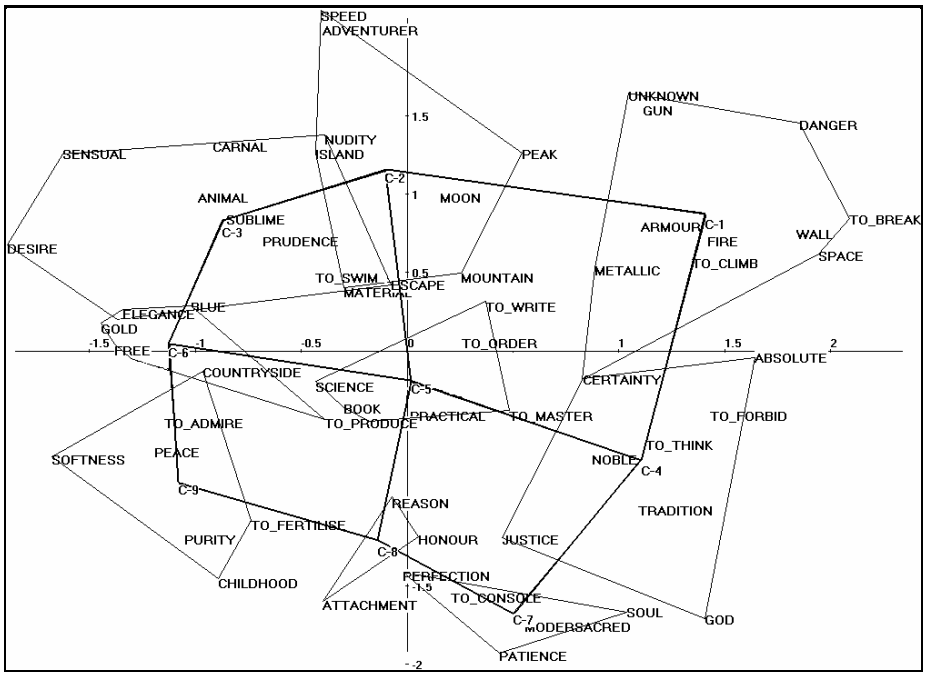

Figure 5. Principal plane of the contiguity analysis using matrix $\mathbf{M}_1$, with both the centroids of the 9 clusters and their convex hulls.

6. The assessment of SOM maps through partial bootstrap

We are provided at this stage with a tool allowing us to explore a continuous space. We can take advantage of having a projection onto a plane to project the bootstrap replicates of the original data set. We could do it onto a higher dimensional space, although the outputs are much more complicated in that case. This projection of replicates can be done in the framework of a partial bootstrap procedure. In the context of principal axes techniques (such as singular values decomposition, principal component analysis, correspondence analysis, and also contiguity analysis), *Bootstrap* resampling techniques [3] are used to produce confidence areas on two-dimensional displays. The bootstrap replication scheme allows one to draw confidence ellipses for both active elements (i.e.: elements participating in building principal axes) and supplementary elements (projected *a posteriori*).

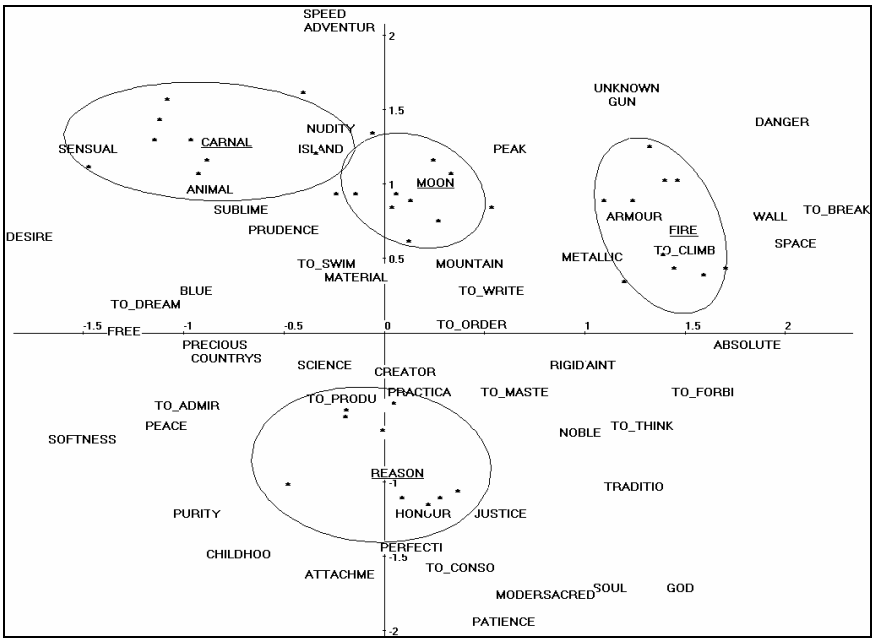

Figure 6. Bootstrap ellipses of confidence of the 5 words: CARNAL, MOON, FIRE, REASON in the same principal contiguity plane as in figure 4 and 5.

In the example of the previous section, the clustered variables (words) are the rows and columns of a correlation matrix. The perturbation of such matrix under a bootstrap re-sampling procedure leads to new coordinates for the *replicated* rows. Without re-computing the whole contiguity analysis for each replicated sample (conservative procedure of total bootstrap), one can project the replicated rows as supplementary elements on a common reference space, exemplified above by figures 4 and 5. Always on that same space, figure 6 shows a sample of the replicates of five points (small stars visible around the words CARNAL, MOON, FIRE, REASON) and the confidence ellipses supposed to contain approximately 90% of these replicated points. Such procedure of partial bootstrap gives satisfactory estimates of the relative uncertainty about the location of points. Although the background of figures 5 and 6 are the same, it is preferable, to keep the results legible, to draw the confidence ellipses on a distinct scattering diagram. It can be seen for instance that the location of *carnal* is rather fuzzy. That word could belong to other neighbouring clusters as well.

7. Conclusions

We have intended to immerse the Self Organizing Map into an analytical framework (the linear algebra of contiguity analysis) and into an inferential setting as well (re-sampling techniques of bootstrap). That does not put into question the undeniable qualities of clarity of the SOM maps. But it may perhaps help to assess the obtained representations: like most exploratory tools, they may help to uncover rapidly and at low cost some features and patterns. However, they should be complemented by other statistical procedures if deeper interpretation is needed.

[*The computations and figures have been carried out by using the software* DTM *that can be freely downloaded, together with the data set serving as an example, from* www.lebart.org.]

References

[1] T. Aluja Banet, L. Lebart (1984), Local and Partial Principal Component Analysis and Correspondence Analysis, *COMPSTAT Proceedings*, p. 113-118, Vienna, Physica Verlag.

[2] M. Cottrell, P. Rousset (1997), The Kohonen Algorithm: a powerful tool for analysing and representing multidimensional qualitative and quantitative data, *in: Biological and Artificial Computation : From Neuroscience to Technology*, J. Mira, R. Moreno-Diaz, J. Cabestany, (eds), Springer, p. 861-871.

[3] B. Efron, R. J. Tibshirani (1993), *An Introduction to the Bootstrap*, New York, Chapman and Hall.

[4] R. C. Geary (1954), The Contiguity Ratio and Statistical Mapping, *The Incorporated Statistician*, **vol. 5**, p.115-145.

[5] P. Kleiweg (1996), *Een inleidende cursus met practica voor de studie*, Alfa-Informatica. Master's thesis, Rijksuniversiteit Groningen.

[6] T. Kohonen (1989), *Self-Organization and Associative Memory*. Berlin, Springer Verlag.

[7] L. Lebart (1969), Analyse Statistique de la Contiguité, *Publications de l'ISUP*, XVIII, p. 81-113.

[8] L. Lebart (2000), Contiguity Analysis and Classification, In: W. Gaul, O. Opitz and M. Schader (eds), *Data Analysis*. Berlin, Springer, p. 233--244.

[9] L. Lebart L. (2004), Validation techniques in Text Mining. In: *Text Mining and its Application*, S. Sirmakensis (ed.), Berlin- Heidelberg, Springer Verlag, p. 169-178.

[10] J. B. MacQueen (1967), Some methods for classification and analysis of multivariate observations. *Proc. Symp. Math. Statist. and Probability (5th),* Berkeley, **1**, Berkeley, Univ. of California Press, p. 281-297.

[11] P. Rousset, C. Guinot (2002), Visualisation des distances entre les classes de la carte de Kohonen pour le développement d'un outil d'analyse et de représentation des données. *Revue de Statistique Appliquée*, p. 35-47.

[12] J.-F. Steiner and O. Auliard (1992), La sémiometrie: un outil de validation des réponses. In: *La Qualité de l'Information dans les Enquêtes / Quality of Information in Sample Surveys*, ASU (eds), Paris, Dunod, p. 241-274.