Visualization of textual data: unfolding the Kohonen maps.

Ludovic Lebart

CNRS - GET - ENST
46 rue Barrault,
75013, Paris, France
(e-mail: ludovic.lebart@enst.fr)

Abstract. The Kohonen self organizing maps (SOM) can be viewed as a visualisation tool that performs a sort of compromise between a high-dimensional set of clusters and the 2-dimensional plane generated by some principal axes techniques. The paper proposes, through Contiguity Analysis, a set of linear projectors providing a representation as close as possible to a SOM map. In so doing, we can assess the locations of points representing the elements via a partial bootstrap procedure.
Keywords: Contiguity analysis, Kohonen maps, SOM, Bootstrap.

1 Introduction

For many users of visualisation tools, the Kohonen self organising maps (SOM) outperform both usual clustering techniques and principal axes techniques (principal components analysis, correspondence analysis, etc.). Indeed, the displays of identifiers of words (or text units) within rectangular or octagonal cells allow for clear and legible printings. The SOM grid, basically non-linear, can then be viewed as a compromise between a high-dimensional set of clusters and the planes generated by any pairs of principal axes. One can regret however the absence of assessment procedures and of valid statistical inference as well. The 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 map (section 3 and 4). An example of application is given in section 5. Via a partial bootstrap procedure, we can now provide these representations with the projection of confidence areas (e.g. ellipses) around the location of words (section 6).

2 Brief reminder about contiguity analysis

Let us consider a set of multivariate observations (n observations described by p variables, leading to a (n, p) matrix $\mathbf{X}$), having an *a priori* graph structure. The n observations are also the n vertices of a symmetric graph $\mathcal{G}$, whose associated (n, n) matrix is $\mathbf{M}$ ($m_{ii'} = 1$ if vertices i and i' are joined by an edge, $m_{ii'} = 0$ otherwise). We denote by $\mathbf{N}$ the (n, n) diagonal matrix

having the degree of each vertex i as diagonal element n_i (n_i stands here for n_{ii}). $\mathbf{y}$ is the vector whose i^{th} component is y_i. Note that: $n_i = \sum_{i'} m_{ii'}$. $\mathbf{U}$ designates the square matrix such that $u_{ij} = 1$ for all i and j. y being a random variable taking values on each vertex i of a symmetric graph $\mathcal{G}$ the *local variance* of y, $v^*(y)$, is defined as:

$$v^*(y) = (1/n) \sum (y_i - m_i^*)^2$$

where: $m_i^* = (1/n_i) \sum_{i'} m_{ii'} y_{i'}$. It is the average of the adjacent values of vertex i. Note that if $\mathcal{G}$ 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 Geary contiguity ratio [Geary, 1954]), is written: $c^*(y) = v^*(y)/v(y)$. It can be generalized : a) to different distances between vertices in the graph, b) to multivariate observations (both generalizations are dealt with in: [Lebart, 1969]). This section is devoted to the second generalization: multivariate observations having an *a priori* graph structure. The multivariate analogue of the local variance is now the local covariance matrix $\mathbf{V}^*$, given by (using the previously defined notation):

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

The diagonalization of the corresponding local correlation matrix (Local Principal Component Analysis) [Aluja Banet and Lebart, 1984] produces a description of the local correlations that can be compared to the results of a PCA . Comparisons between correlation matrices (local and global) can be done through Procustean Analysis (see: [Gower and Dijksterhuis, 2004]). If the graph is made of k disjoined complete subgraphs, $\mathbf{V}^*$ coincide with the classical *within covariance matrix* used in linear discriminant analysis. If the graph is complete (associated matrix = $\mathbf{U}$ defined above), then $\mathbf{V}^*$ is the classical global covariance matrix $\mathbf{V}$.

Let $\mathbf{u}$ be a vector defining a linear combination $u(i)$ of the p variables for vertex i:

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

The local variance of $u(i)$ is: $v^*(u) = \mathbf{u}'\mathbf{V}^*\mathbf{u}$. The contiguity coefficient of $u(i)$ can be written: $c(u) = \mathbf{u}'\mathbf{V}^*\mathbf{u}/\mathbf{u}'\mathbf{V}\mathbf{u}$. Contiguity Analysis is the search for $\mathbf{u}$ that minimizes $c(u)$. It produces linear functions having the properties of "minimal contiguity". Instead of assigning an observation to a specific class, (case of discriminant analysis) these functions allows one to assign it in a specific part of the graph. Therefore, Contiguity Analysis can be used to discriminate between overlapping classes.

3 SOM maps and associated graphs

The self organizing maps (SOM maps) [Kohonen, 1989] aim at clustering a set of multivariate observations. The obtained clusters are 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. Let us summarize the principles of the algorithm:

The size of the graph, and consequently, the number of clusters are chosen *a priori* (for example: a square grid with 5 rows and 5 columns, leading to 25 clusters). The algorithm is similar to the MacQueen algorithm [MacQueen, 1967] in its on-line version, and to the k-means algorithm [Forgy, 1984] in its batch version. Let us consider n points in a p-dimensional space (rows of the (n, p) matrix $\mathbf{X}$). At the outset, to each cluster k is assigned a provisional centre C_k with p components (e.g.: chosen at random). For each step t, the element $i(t)$ is assigned to its nearest provisional centre $C_{k(t)}$. Such centre, together with its neighbours on the grid, is then modified according to the formula: $C_{k(t+1)} = C_{k(t)} + \varepsilon(t)(i(t) - C_{k(t)})$. In this formula, $\varepsilon(t)$ is an adaptation parameter ($0 \leq \varepsilon \leq 1$) which is a (slowly) decreasing function of t, as those usually involved in stochastic approximation algorithms. The process is reiterated, and eventually stabilizes, but the partition obtained may depend on the initial choice of the centres. In the batch version of the algorithm, the centres are updated only after a complete pass of the data.
Figure 1 represent a stylised symmetric matrix (70, 70) $\mathbf{M}_0$ associated to a partition of n=70 elements in k=8 classes (or 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). The graph consists of 8 cliques. All the cells of the black blocks contains the value 1. All the cells outside these diagonal blocks contains the value 0 . The 8 classes of the previous partition have been obtained through a SOM algorithm from a square 3 x 3 grid (with an empty class).

The left hand side matrix of figure 1 does not take into account the topology of the grid: links between elements do exist only within clusters. In the right hand side of figure 1, two elements i and j are linked ($m_{ij} = 1$) in the graph if they belong either to a same cluster, or to contiguous clusters. Owing to the small size of the SOM grid (figure 2), the diagonal adjacency is not taken into account. (e.g.: 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 $\mathbf{M}_0$ and $\mathbf{M}_1$, a matrix $\mathbf{M}_2$ can be defined, that extends the definition of the edges of the graph to diagonal links. In the simple example of figure 3, the elements of cluster 7, for example, are considered as contiguous to the elements of clusters 4, 8, and 5.

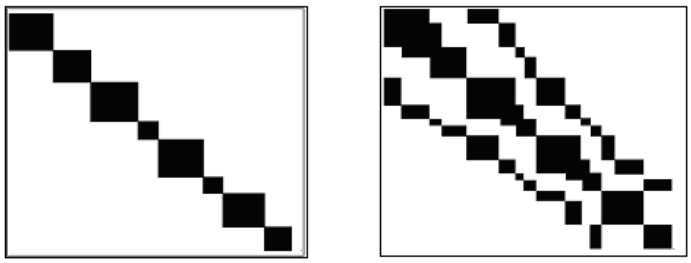

Fig. 1. Stylised incidence matrices $\mathbf{M}_0$ of the graph associated with a simple partition (left), and $\mathbf{M}_1$, relating to a SOM map (right) (all the cells in the white areas contain the value 0 whereas those in the black areas contain the value 1)

7	8	9
4	5	6
1	2	3

Fig. 2. The *a priori* SOM grid

4 Linear projectors onto the best SOM plane

The matrices $\mathbf{M}_0$, $\mathbf{M}_1$, and $\mathbf{M}_2$ can be easily obtained as a by-product of the SOM algorithm. In the case of contiguity analysis involving the graph G_0 the associated matrix of which is $\mathbf{M}_0$, the local variance coincide with the "within variance", and the result is a classical linear discriminant analysis of Fisher (LDA). In the plane spanned by the two first principal axes, the clusters are optimally located in the sense of the LDA criterion. In the cases of contiguity analysis using the graphs G_1 or G_2 (associated matrices $\mathbf{M}_1$, or $\mathbf{M}_2$), 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 schematised 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 An example of application

An open-ended question has been included in a multinational survey conducted in seven countries (Japan, France, Germany, Italy, Nederland, United Kingdom, USA) in the late nineteen eighties [Hayashi *et al.*, 1992]. The respondents were asked : "What is the single most important thing in life for you?" . The illustrative example is limited to the British sample. The

counts for the first phase of numeric coding are as follows: Out of 1043 responses, there are 13 669 occurrences (tokens), with 1 413 distinct words (types). When the words appearing at least 25 times are selected, there remain 9815 occurrences of these words, with 88 distinct words. In this ex-

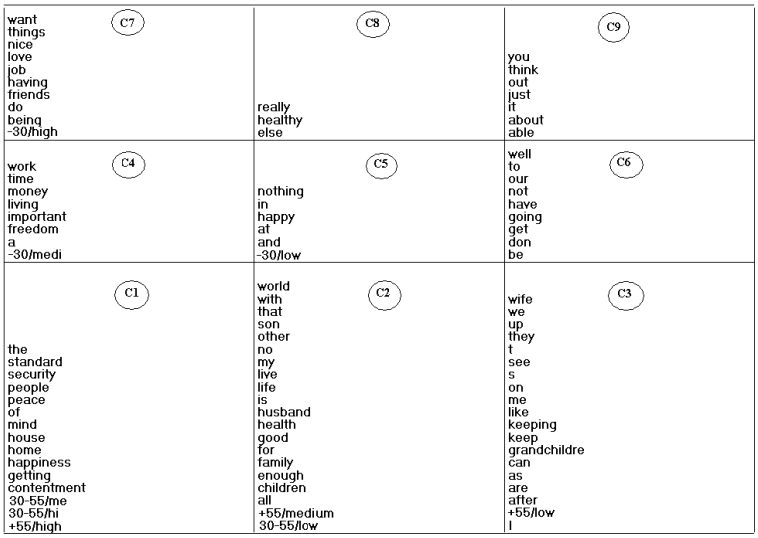

Fig. 3. A (3 x 3) Kohonen map applied to the words used in the 1043 responses

ample we focus on a partitioning of the sample into 9 categories, obtained by cross-tabulating age (3 categories) with educational level (3 categories). The nine identifiers combine age categories (-30, 30-55, +55) with educational levels (low, medium, high). Note that the SOM map (figure 3) provides a simultaneous representation of words and of categories of respondents. This is due to the fact that the input data are the coordinates provided by a correspondence analysis of the lexical contingency table cross-tabulating the words and the categories. 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 C1, C2, ...C9 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 3. 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 3. We note that these convex hulls are relatively well separated. In fact, figure 5 contains much more information than figure 3, since we have now an idea of the shapes and sizes of the clusters, of the

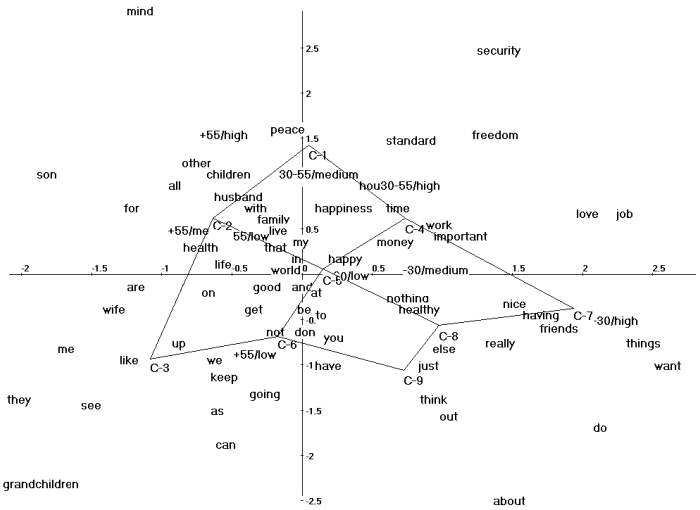

Fig. 4. Principal plane of the contiguity analysis using matrix $\mathbf{M}_1$. The points C1, C2, ...C9 represent the centroids of the 9 clusters derived from the SOM map.

degree to which they overlap. We are now aware of their relative distances, and, another piece of information missing in figure 3, we can observe the configurations of elements within each cluster.

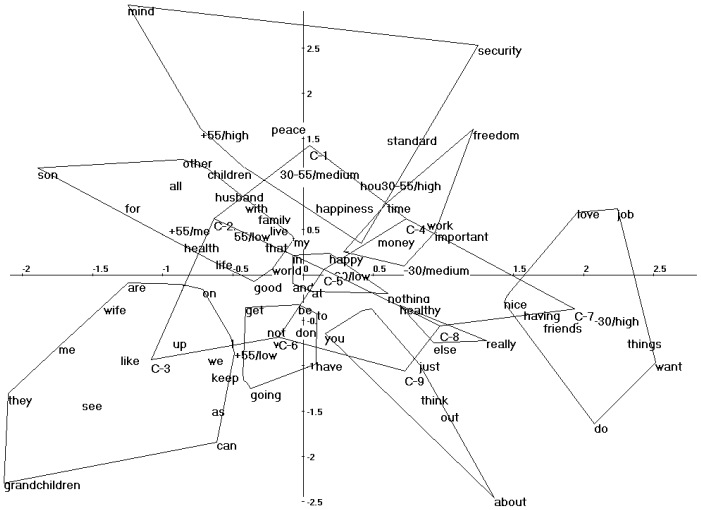

Fig. 5. Principal plane of the contiguity analysis using matrix $\mathbf{M}_1$, with both the centroids of the 9 clusters and their convex hulls

6 Assessing 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 (and possibly onto a higher dimensional space, although the outputs are much more complicated in that case) to project the bootstrap replicates of the original data set. This can be done in the framework of a partial bootstrap procedure. In the context of principal axes techniques (such as SVD, PCA, and also contiguity analysis), Bootstrap resampling techniques [Efron and Tibshirani, 1993] 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).

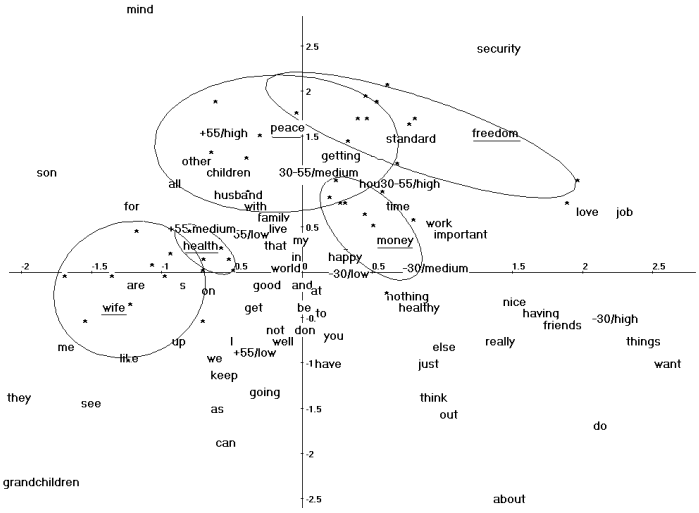

Fig. 6. Bootstrap ellipses of confidence of the 5 words: *freedom, health, money, peace, wife* in the same principal contiguity plane as in figure 4 and 5

In the example of the previous section, the words are the rows of a contingency table. The perturbation of such table under a bootstrap re-sampling procedure leads to new coordinates for the replicated rows. Without recomputing 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 *freedom, health, money, peace, wife*) and the confidence ellipses that contain approximately 90 % of these replicated points. Such procedures of partial bootstrap [Lebart,

2004] give 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 figure. It can be seen for instance that the words $freedom$ and $money$, both belonging to cluster C4, have different behaviours with respect to the re-sampling variability. The location of $freedom$ is much more fuzzy. That word could belong to some neighbouring clusters as well.

7 Conclusion

We have intended to immerse the SOM maps, obtained through an algorithm often viewed as a black box, 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 question the undeniable qualities of clarity and readability of the SOM maps. But it may perhaps help to assess their scientific status: like most exploratory tools, they help to rapidly uncover some patterns. However, they should be complemented with statistical procedures whenever deeper interpretation is needed.

References

[Aluja Banet and Lebart, 1984]T. Aluja Banet and L. Lebart. Local and partial principal component analysis and correspondence analysis. In M. Novak T. Havranek, Z. Sidak, editor, *COMPSTAT Proceedings*, pages 113–118, 1984.

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

[Forgy, 1984]R. Forgy. Cluster analysis of multivariate data : efficiency versus interpretability of classifications. In *Biometric Society Meetings,*, page 768, 1984.

[Geary, 1954]R. C. Geary. The contiguity ratio and statistical mapping. *The Incorporated Statistician*, pages 115–145, 1954.

[Gower and Dijksterhuis, 2004]J. C. Gower and G. B. Dijksterhuis. *Procustes Problem*. Oxford Statistical Science Series, Oxford, 2004.

[Hayashi *et al.*, 1992]C. Hayashi, T. Suzuki, and M. Sasaki. *Data Analysis for Social Comparative Research: International Perspective*. North-Holland, Amsterdam, 1992.

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

[Lebart, 1969]L. Lebart. Analyse statistique de la contiguité. *Publications de l'ISUP*, pages 81–112, 1969.

[Lebart, 2004]L. Lebart. Validation techniques in text mining. In S. Sirmakensis, editor, *Text Mining and its Application*, pages 169–178, 2004.

[MacQueen, 1967]J. B. MacQueen. Some methods for classification and analysis of multivariate observations. *Proc. Symp. Math. Statist. and Probability (5th)*, pages 281–297, 1967.