APPLICATION OF SELF ORGANIZED MAPS AND CURVILINEAR COMPONENT ANALYSIS TO THE DISCRIMINATION OF THE VESUVIUS SEISMIC SIGNALS

Stefano Masiello[1], Antonietta M. Esposito[1, 2], Silvia Scarpetta[1], Flora Giudicepietro[2], Anna Esposito[3], Maria Marinaro[1]
[1]Dipartimento di Fisica, Università di Salerno, INFN, and INFM Salerno. Italy
masiello@sa.infn.it, silvia@sa.infn.it, marinaro@sa.infn.it
[2]Osservatorio Vesuviano, INGV, Napoli. Italy
giudicepietro@ov.ingv.it, aesposito@ov.ingv.it
[3]Seconda Università di Napoli, IIASS, and INFM Salerno. Italy
iiass.annaesp@tin.it

Abstract – *This paper reports on the unsupervised analysis of seismic signals recorded by four stations situated on the Vesuvius area in Naples, Italy. The dataset under examination is composed of earthquakes and false events like thunders, quarry blasts and man-made undersea explosions. The goal is to use these specific data for comparing the performance of three projection methods that are well known to be able to exploit structures and organizes data, providing a framework for understanding and interpreting the relationships between data items, and suggesting simple descriptions of these relationships. The three unsupervised techniques under examination are: Principal Component Analysis (PCA), which is linear, Self-Organizing Map (SOM) and Curvilinear Component Analysis (CCA), which are nonlinear. The results show that, among the above techniques, SOM can better visualize the complex set of high-dimensional data allowing to discover their intrinsic clusters structure and eventually discriminate the earthquakes from the false events either natural (thunder) or artificial (quarry blast and undersea explosions).*

Key words – seismic signals, unsupervised clustering techniques

1 Introduction

This paper reports on the unsupervised discrimination of earthquakes and false events (like thunders, quarry blasts, and illegal undersea fishing explosions) recorded by four seismic stations on the Mt. Vesuvius area in Naples, Italy. The Vesuvius is a high risk volcano close to the city of Naples in the South of Italy. In this populated area (about 2 million people) volcano-tectonic earthquakes and transient signals due to external sources (human-made explosions, thunder, etc) are observed on a daily basis by human experts, through procedures based on the visual analysis of the spectral and temporal features of the detected signals. These procedures often rely on the time delays at which the signals produced by a single source arrive at the different recording stations. However, for small networks of recoding stations, these procedures can fail and produce false event detections when other sources, artificial and/or natural (such as thunders, and human made explosions in quarries and undersea) generate

signals similar to those produced by local earthquakes. Unfortunately, this is just the case in the area described above and consequently additional signal analysis must be performed in order to reduce the probability of false event detections. Therefore, an automatic high-performance strategy for discriminating earthquakes from the other transient signals could drastically reduce the workload of the community involved in the seismological monitoring of the area. Our approach in discriminating among these signals is based on unsupervised techniques that should allow the visualization of the intrinsic data structure and the clustering together of similar events. To this aims, we decided to compare on this specific data set, the performance of three different projection strategies characterized in terms of the different assumptions they make about the *representational structure* used to define clusters, and the *similarity measures* that describe the relationships between objects and clusters. The seismic events of our dataset were manually labelled by the experts on the basis of their experience in identifying seismic events and on the information they received by the port authorities and/or by private citizens signalling unauthorized undersea explosions. We have already faced this problem in a previous work [1] using a supervised learning algorithm that was able to discriminate, on the test set, more than 90 percent of the events described above. However, a supervised analysis always needs a dataset correctly labelled by the experts. This is not possible in many situations, due to the nature of the events which are of many different typologies and continuously changing and to the need of a heavy labelling work. In these cases, a good unsupervised strategy for the visualization (and discrimination) of the recorded signals may be more helpful than a supervised strategy.

1.1 Data Description

Around the Mt. Vesuvius area both earthquakes and false events, like artificial explosions and natural thunders, are recorded by a permanent seismic monitoring network, composed of ten analogical stations. Nine of them are deployed on the volcanic edifice whereas one seismic station is located in Nola, a town about 15 Km far from the crater axis. The seismic signals recorded by the remote stations are frequency modulated and transmitted via radio to the Vesuvius Observatory Monitoring Center [2]. The collected analogical signals are sampled at 100 Hz, stored on Personal Computers and made available to the experts for the analysis. Figure 1 shows the seismic monitoring network at Mt. Vesuvius.

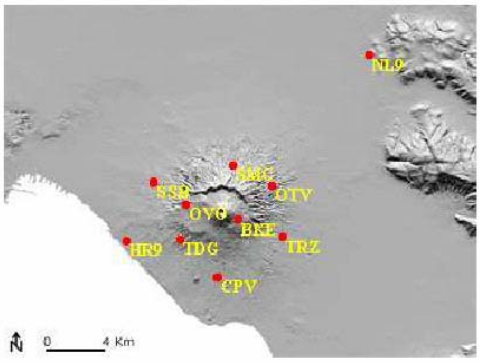

Fig. 1. Seismic monitoring network at Mt. Vesuvius.

Our dataset is composed of 961 events registered at four stations (CPV, NL9, TRZ, and BKE in Figure 1) described below:

- **The CPV station**, located on the coast of the Gulf of Naples, records earthquakes and man-made undersea explosions caused by illegal fisherman. The available dataset contains the recordings of 144 earthquakes and 247 explosions.
- **The NL9 station,** located in Nola, records seismic signals and man-made explosions in quarries. The available dataset contains the recordings of 109 earthquakes and 114 quarry explosions.
- **The TRZ station,** located on the basis of the Vesuvius, records earthquakes and man made explosions in quarries. The available dataset contains the recordings of 104 earthquakes and 103 explosions.
- **The BKE station,** located up on the Vesuvius crater, mainly records earthquakes and natural false events like thunders. The available dataset contains the recordings of 72 earthquakes and 68 thunders.

Each event is 22 seconds long and is described by a vector of 2200 components due to the 100 Hz sampling rate. The labelling, manually made by the experts, identified a total of 429 earthquakes, 247 undersea explosions, 114 quarry blasts recorded by the NL9 station, 103 quarry blasts recorded by the TRZ station, and 68 thunders, constituting the 5 classes we want to identify with the help of the unsupervised techniques described below.

1.2 Extraction of Seismic Features

Feature extraction is an important stage in any data analysis task. This step is performed in order to extract, from signals, significant information eliminating as much as possible redundancy, and obtaining a compact and significant data representation. To this aim, we use the Linear Predictive Coding (LPC) algorithm [3] to extract spectral features from the signals under examination, and a discrete waveform parametrization algorithm to extract amplitude versus time information.
The basic idea behind the LPC algorithm is to model each signal sample s_n as a linear combination of a certain number p of its past values as described in the equation below:

$$\overline{s_n} = \sum_{k=1}^{p} c_k s_{n-k} + G \qquad \textbf{(1)}$$

where c_k are the *prediction coefficients*, G is the *gain* and p represents the *model order*. The c_k estimation is obtained by an optimization procedure which tries to minimize the error between the real value of the signal sample at time t and its LPC estimate. The coefficients c_k efficiently encode the signal frequency features. Each recording was processed on a short time basis, dividing it in 15 overlapping analysis windows, each of 2.56 seconds long, and extracting a certain number p of LPC coefficients from each of them. The overlapping step was 1.28 seconds long. The choice of the model order p is problem dependent and is generally made estimating the LPC residual error over the dataset at the hand.

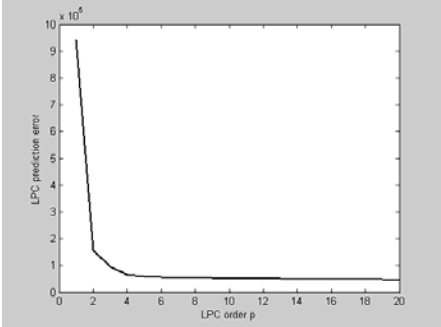

Fig. 2. LPC residual error as a function of the model order *p*. The curve has been averaged over all the events in the dataset.

Applying the LPC coding to each signal and computing the residual error averaged over the all dataset, it was possible to estimate the model order *p* that best fits our data in term of a balance between the compactness of the representation and the error made using such a representation. Figure 2 shows this error as a function of the order *p* for our data.
Based on the results showed in Figure 2 and on our previous experience [1] in using LPC processing for seismic signal analysis, the model order was set to p=10 as a good trade-off in maximizing the representation compression and minimizing the corresponding error. The time domain information, extracted from the discrete waveform, and computed as the properly normalized difference between the maximum and the minimum signal amplitude in a 1 second long analysis window was added to the data representation. The final encoding of each event in our dataset was a feature vector of 187 components constituted of 165 LPC coefficients and 22 time features. Moreover, the resulting feature vectors were *logarithmically* normalized since, as we will show below, normalization seems to improve the clustering both for the CCA and the SOM strategy.

1.2 PCA, SOM and CCA Algorithms

Clustering techniques are widely used for the analysis and the visualization of complex sets of data. These techniques may be distinguished in two classes. In the former are linear methods like Principal Component Analysis (PCA) [4] or the classical Multidimensional Scaling (MDS) [4]. In the latter class are nonlinear methods like Self-Organizing Map (SOM) [5] or nonlinear variants of MDS, like the recently proposed Curvilinear Component Analysis (CCA) and Curvilinear Distance Analysis (CDA) [6, 7].
PCA finds the axes of maximum variance of the input data and represent them by a linear projection onto the subspace spanned by the principal axes [4]. CCA [6, 8] instead performs a nonlinear dimensionality reduction and representation in two steps: (1) a vector quantization (VQ) of the input data into k quantized n-dimensional prototypes and (2) a nonlinear projection of these quantized vectors onto a p-dimensional output space. After learning the quantized prototypes, the prototype pairs (x_i,y_i) are used to interpolate the continuous mapping between the n-dimensional input space X and the p-dimensional output space Y. The nonlinear mapping is done minimizing the cost function

$$E = 1/2 \sum_i \sum_j (X_{ij} - Y_{ij})^2 F(Y_{ij}, \lambda) \qquad \textbf{(2)}$$

where $X_{ij}=d(x_i,x_j)$ and $Y_{ij}=d(y_i i,y_j)$ are the Euclidean distances between the quantized vectors and the output vectors, respectively, and $F(Y_{ij},\lambda)=exp(-Y_{ij}/\lambda)$ is a weighting function that

favour the preservation of the topology on different scales depending on the λ value. A drawback of this procedure is that the values of the neighbourhood parameter λ and the learning rate η have to be properly chosen by the user. While the identification of η requires no particular attention, on the contrary, the neighbourhood parameter λ is critical. As noted also in [2] the CCA performance critically depends on the choice of the λ value and its decreasing time-speed: if λ decreases too slowly, the nonlinear dependencies are not well unfolded, whereas, a fast decrease compromises the CCA convergence.
A help in the choice of the CCA parameters comes from the use of the *dydx* plot. The *dydx* plot shows the joint distribution of the distances $dx=X_{ij}$ and $dy= Y_{ij}$ in the input and output space respectively. In this representation, a perfect match ($X_{ij} \approx Y_{ij}$) clusters the points around the identity function. A locally good mapping is shown by a distribution close to the identity function near the origin (local projection), while unfolding is revealed by bent and spread data where $dy>dx$ in average. Even though the *dydx* plot can be of some help to the user, the CCA dependency on the critical parameters raises several difficulties, as we will see in the paragraph below.
As the CCA, the bi-dimensional Kohonen Self-Organizing Map (SOM) performs a non-linear mapping of an n-dimensional input space onto a two-dimensional regular grid of processing units known as *neurons*. A prototype vector is associated with each node. The fitting of the prototype of each node is carried out by a sequential regression process that minimizes the differences between each input vector and the corresponding winning node's prototype. Namely, at each time step t = 1,2,... a sample ***x(t)*** is extracted and the winner index c (best match) is identified by the constraint described in the equation below:

$$\forall i, \left\| \mathbf{x}(t) - \mathbf{m}_c(t) \right\| \leq \left\| \mathbf{x}(t) - \mathbf{m}_i(t) \right\| \qquad \textbf{(3)}$$

where ***x(t)*** is the feature vector of the signal extracted at the time step t, and $m_i(t)$ is the prototype of the node i. Once the best match is identified, all prototypes are updated according to:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{c(x),i}(\mathbf{x}(t) - \mathbf{m}_i(t)) \qquad \textbf{(4)}$$

where $h_{c(x),i}$, is the *neighbourhood function*, a decreasing function of the distance between the *i-th* and the *c-th* node on the map. The neighbourhood kernel can be written in terms of the Gaussian function $h_{c,i}=\eta_t * exp(-d_{ci}^2/2\sigma_t^2)$, where η_t is the scalar valued learning rate, σ_t the neighbourhood radius at step t, d_{ci} the distance between the c and i neurons on the map grid. Both η_t and σ_t are time monotonically decreasing functions and their exact forms are not critical [9]. The SOM algorithm carries out two important operations: (a) a clustering of the input data into nodes; and (b) a local spatial ordering of the map in the sense that the prototypes are ordered on the grid such that similar inputs fall in topographically close nodes. Such an ordering of the data facilitates the understanding of data structures. The clustering performed by the SOM becomes more visible by displaying on the map the Euclidean distances between prototype vectors of neighbouring nodes through grey levels. In such a way, the SOM gives a good representation of the cluster structure, by graphically depicting on the map both the density of the data and the Euclidean distances among prototypes.

2 Results

The three clustering methods described above were applied on the dataset under examination, using a bi-dimensional representation. Results are reported for each technique using either logarithmically normalized or non normalized data.

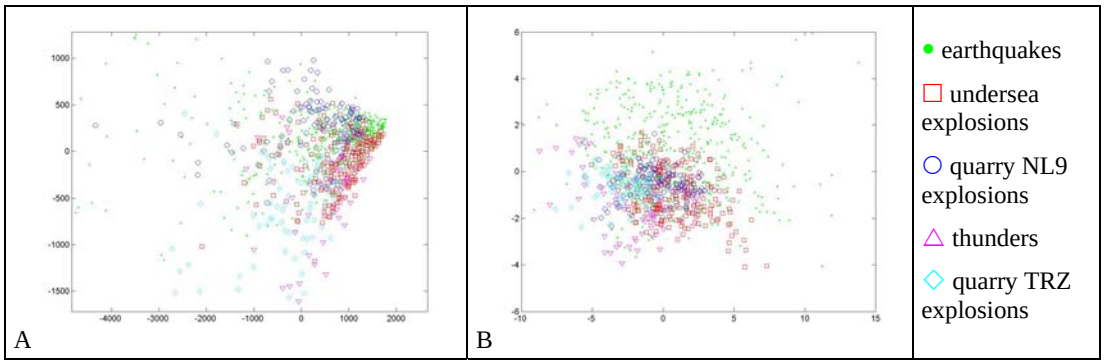

Fig.3. Bi-dimensional PCA projection of the (A) non normalized and (B) logarithmically normalized data. Filled circles indicate earthquakes, empty circles and empty diamonds quarry blasts recorded by the NL9 and the TRZ stations respectively, empty triangles are thunders, and empty squares undersea explosions. (Color figures available on-line)

Figure 3 displays the clustering obtained using PCA on the non normalized (Figure 3A) and logarithmically normalized (Figure 3B) data. The legend is made exploiting the labelling performed by the experts: the filled circles indicate volcanic earthquakes recorded by all the stations; the empty circles and empty diamonds are quarry explosions recorded by the NL9 and the TRZ stations respectively; the empty triangles are thunders, and the empty squares undersea explosions. The results in Figure 3 show that the projection performed by the PCA mixes the different signals all together and doesn't permit to discriminate among them. This can be due to the difficulty of this unsupervised algorithm to capture the peculiar characteristics of our data, since these characteristics may not be related to the maximum variance directions.
Figure 4 shows the bi-dimensional representation obtained using the CCA algorithm on the non normalized data.

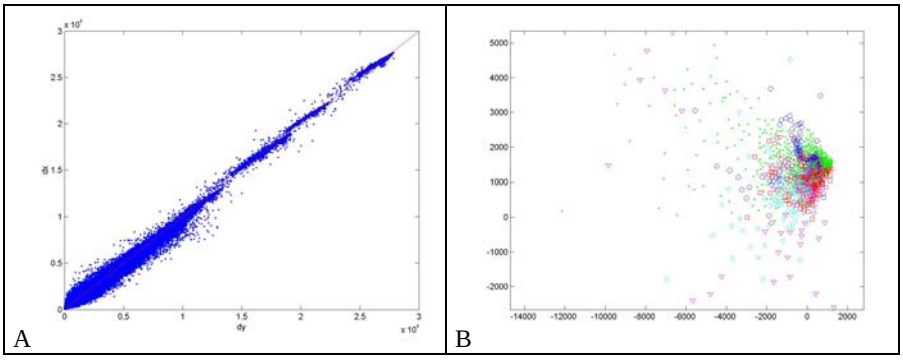

Fig. 4. (A) The CCA *dydx* plot on the non normalized data. (B) The two-dimensional projection of the not-normalized data obtained using the CCA algorithm. The Figure can be read using the legend in Figure 3. (Color figures available on-line).

Figure 4A shows the *dxdy* plot obtained by using appropriate values for the η_t and λ_t parameters on the non normalized data. Figure 4B shows the projection obtained by using the CCA algorithm.

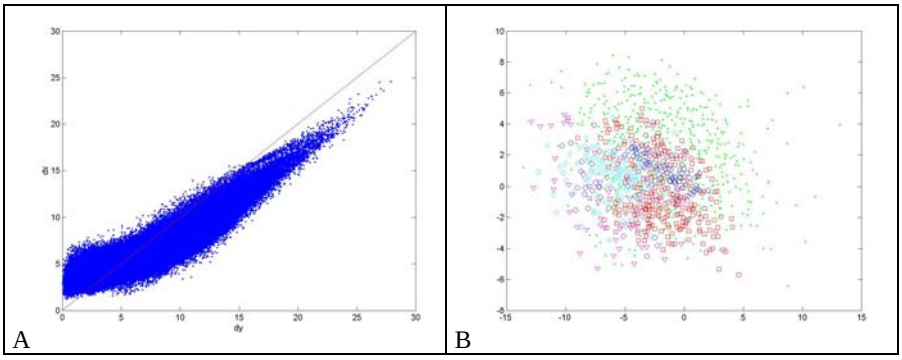

Fig. 5. (A) The CCA *dydx* plot on the logarithmically normalized data. (B) The two-dimensional projection of the normalized data obtained using the CCA algorithm. The Figure can be read using the legend in Figure 3. (Color figures available on-line).

As it can be seen, the CCA projection does not allow to discriminate among the classes of signals under examination since the principal curvilinear components of the non normalized data are not discriminative of our classes (see Figure 4B). Logarithmically normalizing the data and applying the CCA algorithm, the CCA is able to better discriminate the different classes under study, as it can be seen in Figure 5.

Figure 5A shows the *dxdy* plot obtained by using appropriate values for the η_t and λ_t parameters on the normalized dataset. Figure 5B shows the bi-dimensional projection obtained by using the CCA algorithm on these data. As it can be seen, even though the obtained clustering looks better than the previous one, a considerable amount of overlaps among the different classes of signals still remains.

Contrarily to the CCA, the clustering performed by the SOM is not critically dependent on parameters. In our experiments, the SOM learning parameters have been settled in agreement with the prescriptions reported in [9] and the input data were logarithmically normalized. The resulting map is shown in Figure 6 and is composed of 26x12=312 neurons. Each node is a prototype vector and its size represents the number of feature vectors associated with that prototype. The distances among the prototypes are visualized on the map using a grey level colouring. According to this colouring, large distances between two prototypes correspond to dark grey colour levels on the grid and indicate that the two prototypes and the associated feature vectors are very different. The shapes of the prototype on the map are used to indicate the different classes of events. Therefore, stars indicate earthquakes, circles and diamonds quarry blasts recorded by the NL9 and the TRZ stations respectively, triangles indicate thunders, and squares undersea explosions. Shape overlaps indicate that different types of signals belong to the same node. As it can be seen from Figure 6, each class of signals is clustered on a particular zone of the map and the overlaps between classes are less in comparison to those obtained either with the PCA or the CCA algorithm. We can conclude that

the clustering performed by the SOM algorithm can better separate the five classes of signals identified by the experts.

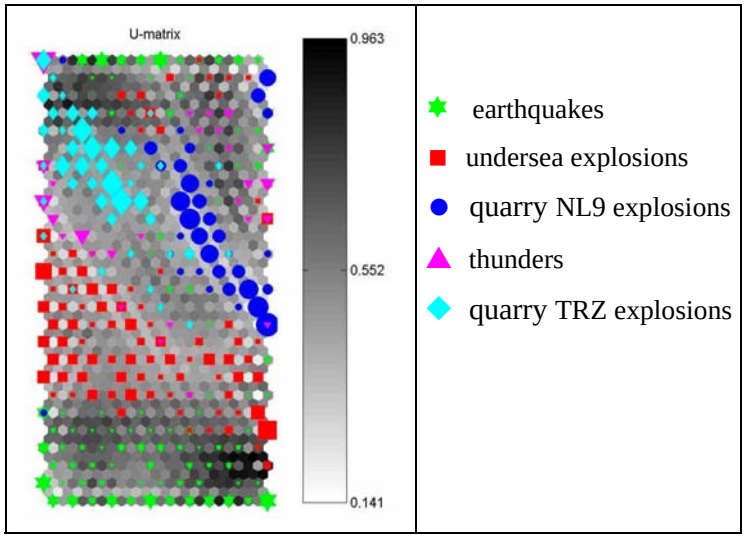

Fig.6. The SOM Map obtained on the normalized data. The grid is composed by 26x12=312 nodes. Stars indicate earthquakes, circles and diamonds quarry blasts recorded by the NL9 and the TRZ stations respectively, triangles indicate thunders, and squares undersea explosions. (Color figures available on-line).

3 Conclusions and Remarks

In the previous section, three unsupervised projection techniques have been applied to a data set composed of five classes of seismic events represented through a 187-feature vector encoding both spectral and time domain information. Our aim was to try to identify among them, the one that can better visualize on a bi-dimensional plane the hidden structure of our data, such that the resulting clustering can be helpful to the experts for the automatic labelling of the events under study. The unsupervised techniques considered were the PCA, the CCA and the SOM.

It has been shown that, among the above techniques, the SOM algorithm, exploiting information on the local topology of the vector prototypes, gives the best performance being able to group the 5 classes of events in separated clusters with minor overlaps than those obtained either with the PCA and/or the CCA algorithm. The poor performance of the PCA algorithm can be due to the difficulty of this linear unsupervised algorithm to capture the peculiar characteristics of our dataset which may not be related to the maximum variance directions. Moreover, the poorer performance of the CCA algorithm, seems to be due to its critical dependence on the choice of the parameter λ and on its decreasing time-speed. As noted in [2, 7], if λ decreases too slowly, the nonlinear dependencies are not well unfolded, whereas, a fast decrease compromises the CCA convergence. This could be overwhelmed introducing the CCA with geodetic (curvilinear) distance, also called Curvilinear Distance Analysis (CDA) [7] that has been proved in many cases to perform better than the CCA and to be not critically dependent from the choice of the λ value. Work is in progress to check the above hypothesis applying the CDA algorithm to our specific dataset.

References

[1] S. Scarpetta, F. Giudicepietro, E.C. Ezin, S. Petrosino, E. Del Pezzo, M. Martini, M. Marinaro (2005), Automatic Classification of Seismic Signals at Mt. Vesuvius Volcano Italy Using Neural Networks, *BSSA Bullettin of Seismol. Soc. of America*, **vol. 95**, p.185-196.

[2] J.A. Lee, A Lendasse, N. Donckers, M Verleysen (2000), A Robust Nonlinear Projection Method, *Proceedings of European Symposium on Artificial Neural Networks (ESANN'00)*, Bruge, ISBN 2-930307-00-5, p.13-20.

[3] J. Makhoul (1975), Linear Prediction: a Tutorial Review, *Proceeding of IEEE*, p.561-580.

[4] I.T. Jollife (1986), *Principal Component Analysis*, New York, Springer Verlag.

[5] T. Kohonen (1997), *Self-Organizing Maps*, Series in Information Sciences, **vol. 30**, Springer, Second edition.

[6] P. Demartines, J. Herault (1997), Curvilinear Component Analysis: A Self-Organizing Neural Network for Nonlinear Mapping of Data Sets, *IEEE Transactions on Neural Networks*, **vol. 8(1)**, p.148-154.

[7] J.A. Lee, A Lendasse, M Verleysen (2004), Nonlinear Projection with Curvilinear Distances: Isomap versus Curvilinear Distance Analysis, *Neurocomputing*, **vol. 57**, p. 49-76.

[8] J. Hérault, A. Guérin-Dugué, P. Villemain (2002), Searching for the Embedded Manifolds in High-Dimensional Data, Problems and Unsolved Questions *Proceedings of European Symposium on Artificial Neural Networks (ESANN02)*, Bruges, ISBN 2-930307-02-1, p.173-184.

[9] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen (1996). SOM_PAK: The Self-Organizing Map Program Package. *Report A31. Helsinki University*, Finland. Also available at http://www.cis.hut.fi/research/som_lvq_pak.shtml.