{ plik roadmap8.tex January 24, 2005}

8 Visualization of high-dimensional data

8.1 Visualization of individual data points using Kohonen's SOM

8.1.1 Typical Kohonen's map

The concept of visualization by Kohonen's maps is depicted in Figure 8.1, after Fig. 1. from Liao et. al. [3]. The SOM may be considered as the result of mapping performed by a network.

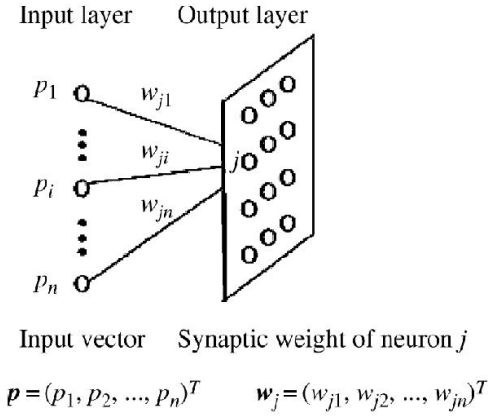

Figure 8.1: Network structure of a SOM. Graph from Liao et al. [3]. File papers/liao1.jpg

Some ideas of the representation – obtained from data located in multivariate space – are shown in Figures 8.2 and 8.3 (published in [5] as Figs 12 and 26). Data may be located in a sub-manifold of the data space. Then the SOM captures the location of the points in the sub-manifold.

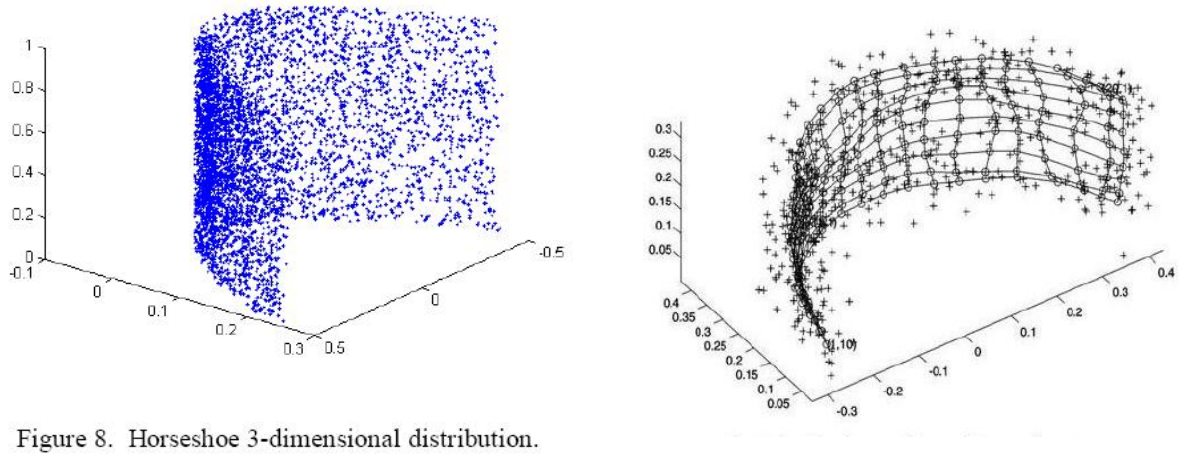

Figure 8.2: **Left**: The 3D horseshoe data as illustration of the fact that data in multivariate space may be *de facto* located in a quasi-linear manifold of lower dimension. **Right**: Data with Horseshoe distribution and resulting Kohonen's map. Graphs from [5]. File verl4.jpg, verl5.jpg

8.1.2 The new method proposed by Liao et.al. 2003

Let $\mathbf{p}$ denote a data vector $\in \mathcal{D}$.

The SOM is of size $m_1 \times m_2$. Let m denote the total number of codebook vectors (prototypes). Obviously $m = m_1 \cdot m_2$.

Let $\mathbf{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ denote the weights of the prototypes.

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ denote the map coordinates of subsequent prototypes, recorded column-wise.

Given a (new) data vector $\mathbf{p}$ evaluate for j=1:p the *Responsibilities* $R_j(\mathbf{p})$ of the prototypes for that vector. The responsibilities are evaluated using the neighborhood function $N_c(j)$, with c being the winning prototype ('conqueror') for the presented data vector $\mathbf{p}$:

$$R_j(\mathbf{p}) = \frac{f(d_j)N_c(j)}{\sum_{j=i}^{m} f(d_i)N_c(i)} \tag{8.1}$$

where $d_j = \|\mathbf{p} - \mathbf{w}_j|\|, \quad f(d_j) = \exp(-d_j)$.

The neighborhood function was taken as the indicator function

$$N_c(j) \;=\; \begin{cases} 1, & \text{if } j \in N_c, \\ 0, & \text{if } j \notin N_c. \end{cases}$$

Then the map coordinates of $\mathbf{p}$ are

$$\mathbf{x_p} = \sum_{j=1}^{m} R_j(\mathbf{p})\mathbf{x}_j \tag{8.2}$$

The new method was applied by Liao et.a. for monitoring gearbox condition. The graph of gearbox transmission with two pairs of meshing gears is shown in Figure 8.4. The graph is taken from [3], p. 125 .

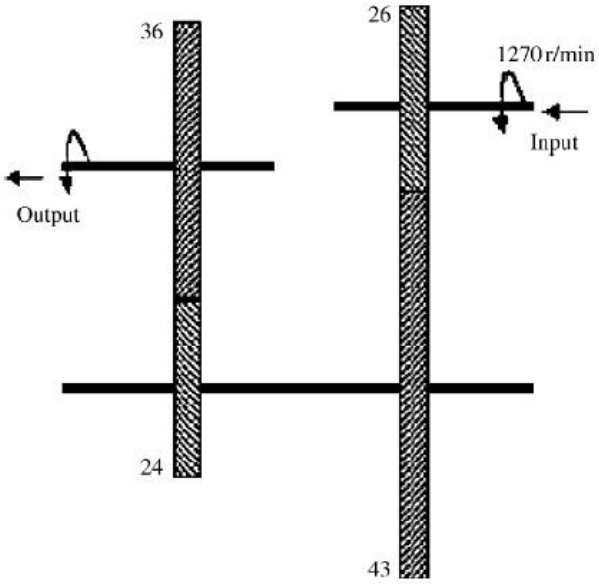

Figure 8.3: Graph displaying gearbox transmission with two pairs of meshing gears. The input rotating speed and teeth numbers are denoted in the plot. File papers/liao4.jpg

8.1.3 Example 1, Synthetic

Next two figures show an example from Liao et. al. [3]. The data contain objects belonging to 3 classes: 'circle', 'triangle' and 'diamond'. The traditional Kohonen SOM is shown in Figure 8.4 top (denoted in [3] as Fig. 4). The map was obtained using the U-mat technique. The visualization of the individual data points using the new technique is shown in the bottom plot.

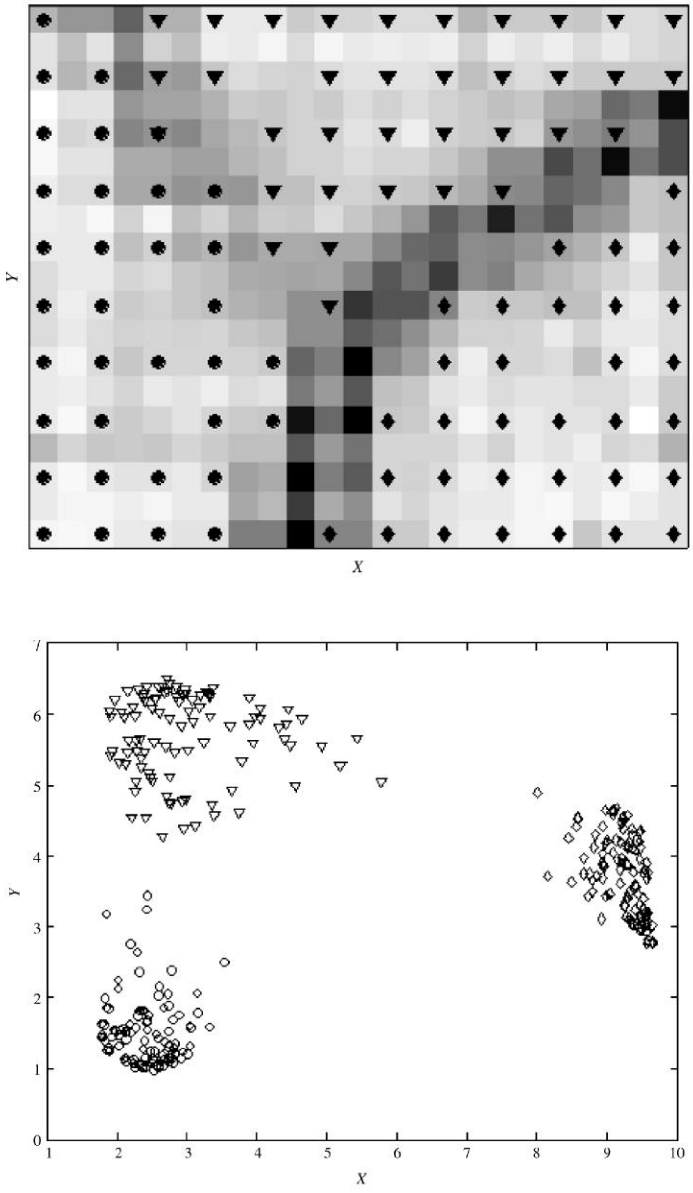

Figure 8.4: **Top**: Synthetic example of data containing 3 kinds of objects: circles, triangles and diamonds. **Top**: Kohonen SOM. **Bottom**: Visualization of individual data points by the new method Files liao2.jpg and liao3.jpg

8.1.4 Monitoring gearbox condition

Liao et al. have applied the method to monitoring the gearbox work. Some sample vibration signals are shown in Figure 8.5 Top plot. The signals were recorded during fixed time intervals, and the records were characterized by 7 derived features.

The proposed method of visualization clearly differentiates the three states of work of the gearbox.

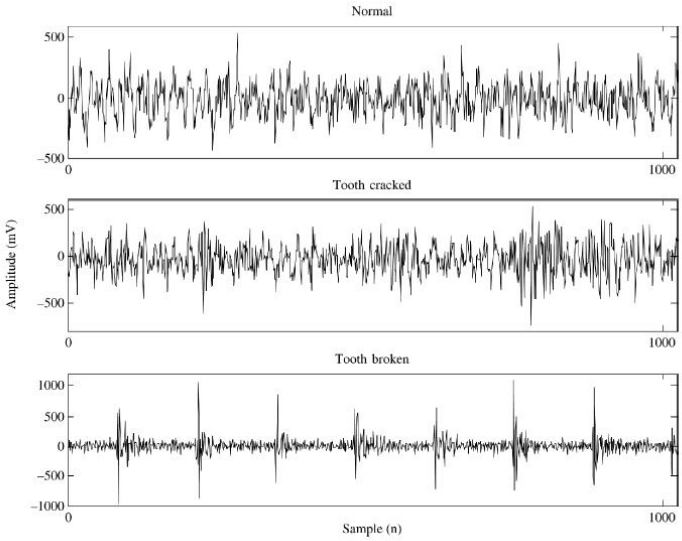

Fig. 7 Gearbox vibration signals measured under normal, tooth cracked and tooth broken conditions

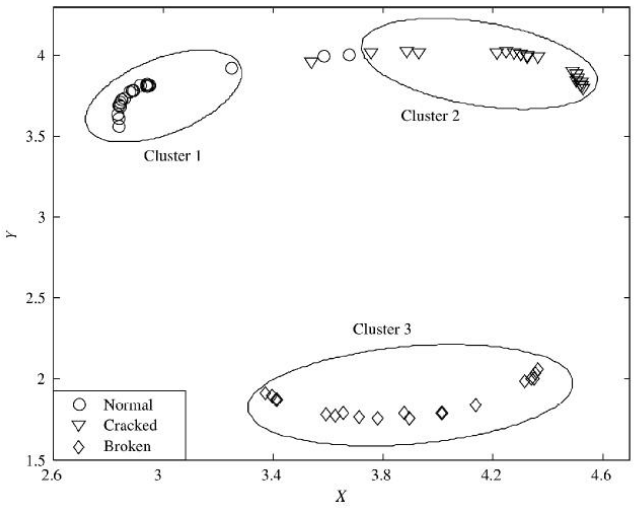

Figure 8.5: **Top**: Gearbox vibration signals measured under 'normal', 'tooth cracked' and 'tooth broken' conditions. **Bottom**: Visualization of individual data points by the new method. Files liao5.jpg and liao6.jpg

8.2 The concepts of *volume*, *distance* and *angle* in multivariate space

Some weird facts about the concepts of 'volume', 'distance' and 'angle' in multivariate space:

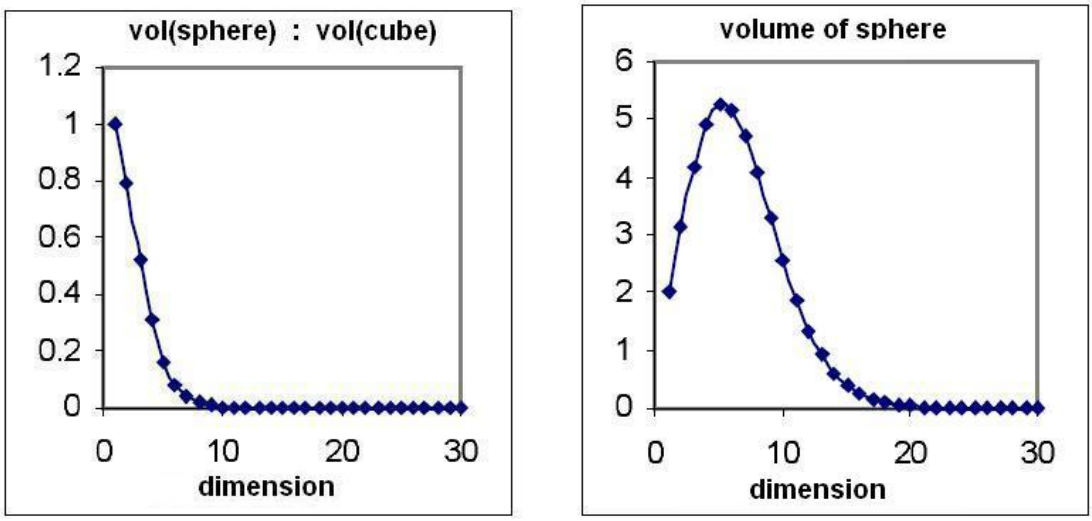

Figure 8.6: Weird facts about high-dimensional data I. [5], p.26. **Left**: Ratio between the volume of a sphere and the volume of a cube (length of an edge equal to the diameter of the sphere) versus the dimension of the space. **Right**: Volume of the sphere (radius=1) versus the dimension of the space. Files verl2.jpg, verl3.jpg

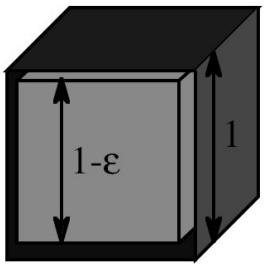

Figure 2. The external shell of a hypercube contains almost all the available volume (see text).

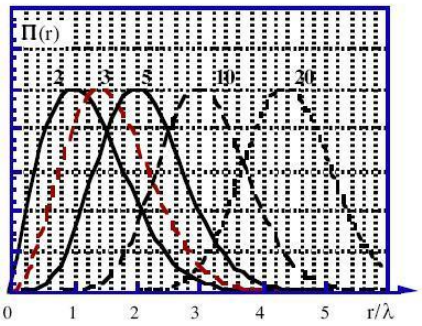

Figure 3. The probability density of a point at distance r from the center is, maximal at $\sqrt{d-1}$

Figure 8.7: Weird facts about high-dimensional data. **Left**: External shell of a hypercube. **Right**:Probability density of a point at distance r from the center. Files herault02.jpg, herault02a.jpg

Fact 1. The volume of a hyper-sphere of unit radius goes to zero as dimension growth. The volume of a hyper-sphere of radius r in d dimensions is given by

$$V(d) = \frac{\pi^{d/2}}{\Gamma(d/2+1)}\, r^d\,.$$

Making plot $V(r,d)$ versus d while keeping r constant (e.g. $r=1$), one may notice, that with increasing d the volume rapidly decreases. So, in higher dimensions, a unit sphere is nearly empty.

Fact 2. The ratio between the volumes of a sphere and a cube of the same radius tends towards zero with increasing dimension.

Fact 3. The ratio of a sphere of radius $1 - \epsilon$ and 1 tends toward zero. Thus the outer shell of sphere will contain more and more of volume of the entire sphere. The same is true for hyper-spheres and hyper-ellipsoids as well.

8.3 CCA, Curvilinear Component Analysis

8.3.1 The principle of CCA

Figure 8.8 shows the principle of constructing curvilinear components (graph taken from [1], p. 175, Fig. 7).

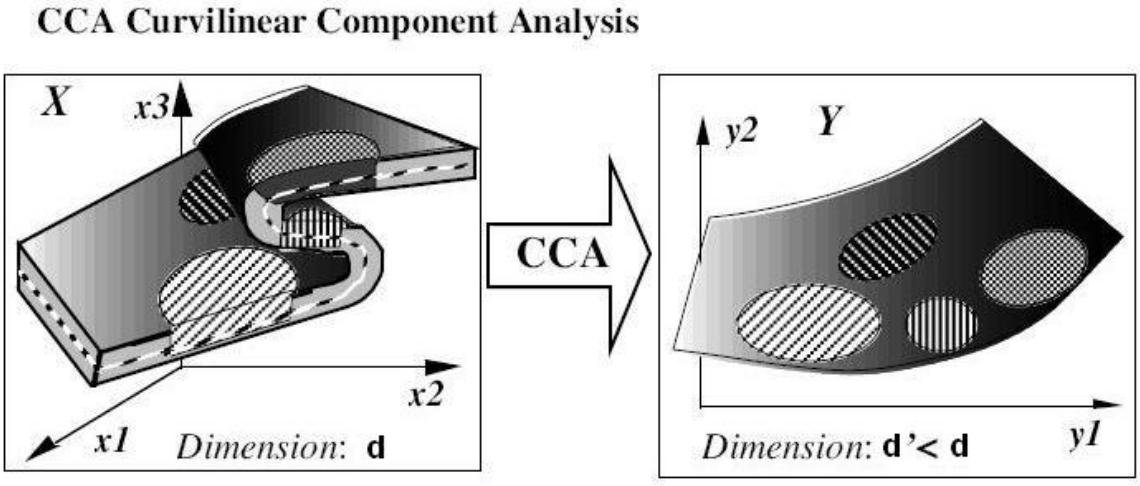

Figure 8.8: Principle of the CCA algorithm. After a possible vector quantization of the input space ($\mathcal{X}$) in d dimensions, the local topology of the input average manifold is mapped into an output representation space ($\mathcal{Y}$) of dimension $d' < d$. File papers/herault02b.jpg

8.3.2 Scene classification

An image is analyzed by a bank of spatial filters, according to four orientation and 5 frequency bands, ranging from very low spatial frequencies to medium ones. The global energies of the 20 filters' outputs constitute a 20–dimensional feature space. Using CCA, Herault et.al. obtained the visualization shown in Figure 8.9. The obtained organization of the data is in surprising accordance with some semantic meaning: Natural/Artificial for each side of the grey line, and and Open/Closed along this line.

Figure 8.10 shows the perceptual organization of the same grey level images set according to perceptual similarities captures from a psychological experiment. In the experiment, each subject was asked to give similarity marks (1 to 10) between images.

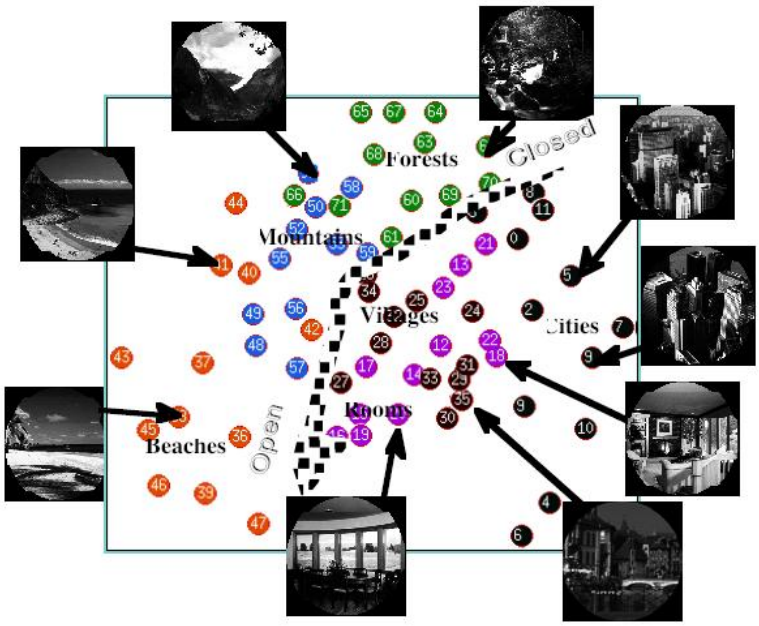

Figure 8.9: Scene analysis, natural vs. man-made. Natural: beaches, mountains, forests. Man-made: cities, villages, rooms. Input: 20–dimensional data vectors obtained by the mean energies in four orientations and five spatial bands. Open/closed File papers/herault02c.jpg

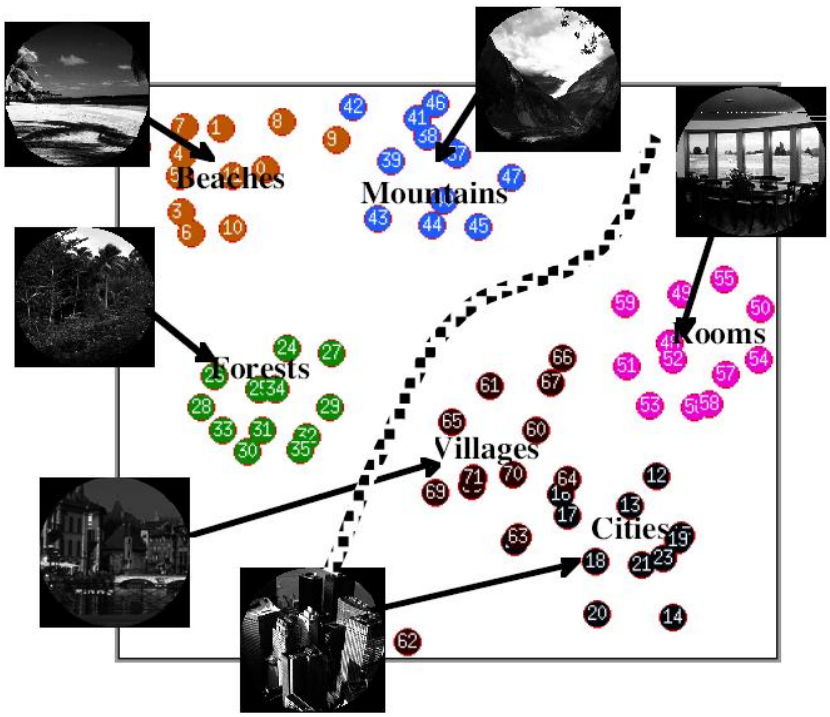

Figure 8.10: Scene analysis. Organization obtained with an input psychological distance matrix collected on the same data set with 20 human subjects. File papers/herault02d.jpg

References

[1] J. Herault, A. Guerin–Dugue, P. Villemain, and unsolved questions. ESANN'2002 proceedings, Bruges (Belgium), d-side publi., 173–184.

[2] J. Herault, Aude Oliva, Anne Guerin-Dugue, Scene Categorisation by Curvilinear Component Analysis of Low Frequency Spectra. ESANN'1997 proceedings, Bruges (Belgium), d-side publi., 96–96.

[3] G. Liao, S. Liu, T. Shi and G. Zhang, Gearbox condition monitoring using self-organizing feature maps. Proc. Instn. Mech. Engrs Vol. 218 Part C, 119–129. J. Mechanical Engineering Science, ©IMechE 2004,

[4] M. Verleysen, D. François, G. Simon, W. Wertz, On the effects of dimensionality on data analysis with neural networks. J. Mirra (Ed.): IWANN 2003, LNCS 2687, 105–112.

[5] M. Verleysen, Machine Learning of High-Dimensional Data, PHD Thesis, Universite Catholique de Louvain, Laboratoire de Microelectronique, Louvain-la-Neuve, 2000.

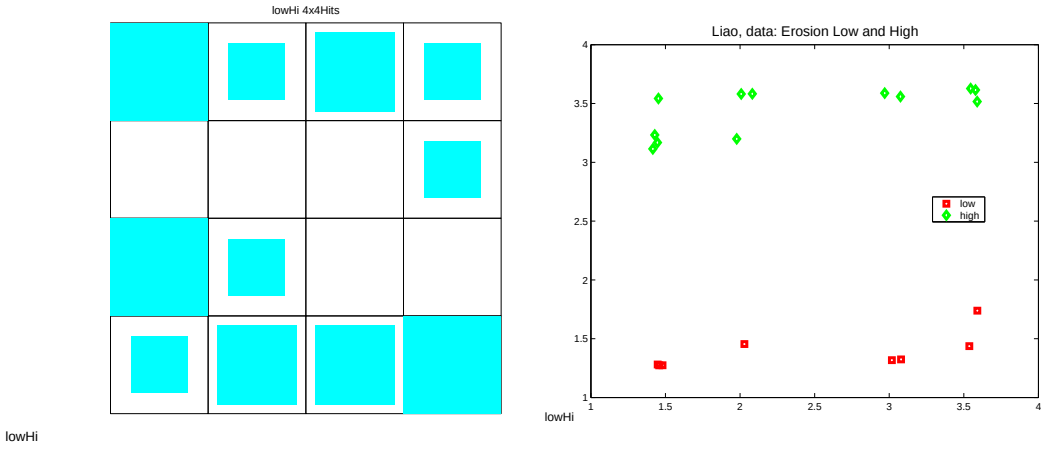

Figure 8.11: Erosion data LowHigh. Left: Kohonen map. Right: Liao' visualization. .... Files maplow.eps, liaoLH.eps

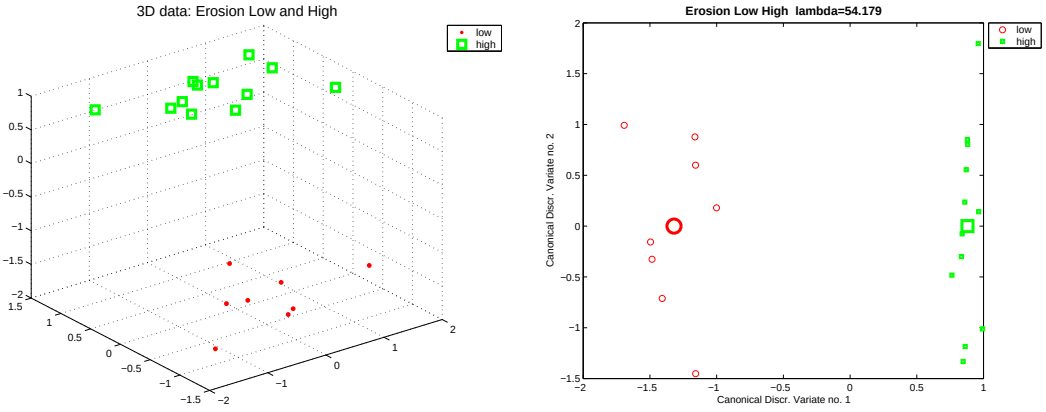

Figure 8.12: Erosion data LowHigh. Left: 3D plot. Right: Canonical discrimination. .... Files 3DLH.eps, canLH.eps