

# Linear mappings of local data structures

Mayer Aladjem and Its'hak Dinstein

*Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel*

Received 23 September 1991

## *Abstract*

Aladjem, M. and I. Dinstein, Linear mappings of local data structures, Pattern Recognition Letters 13 (1992) 153-159.

Two methods for linear mapping of multidimensional data in the case of unsupervised learning are proposed. The first method maximizes the mean square density gradient of the projected samples with the intention of compressing the clusters. The second method is based on the  $k$ -NN technique and obtains a map of the scatter of the neighbor clusters. An experiment with the classical Iris data shows the mapping accuracy of the latter method.

*Keywords.* Interactive pattern recognition, mapping of multidimensional data, cluster analysis, local data structure.

## 1. Introduction

Mapping methods (Siedlecki et al. (1988)) are especially important in the analysis of multivariate data. In general, the mapping consists of finding a transformation

$$h: x_j \rightarrow y_j, \quad j=1, 2, \dots, N, \quad (1)$$

where  $x_j$  is the given sample set of real vectors in the  $n$ -dimensional space  $\mathbb{R}^n$  ( $n > 2$ ) and  $y_j$  is the set of projections into  $\mathbb{R}^2$ . In unsupervised learning, the goal is to find linear or nonlinear transformations having minimal distortion of data structure. These mappings allow an approximative view of data structure and are to be used in conjunction with cluster analysis methods that can detect natural groupings or clusters of the patterns.

In unsupervised learning, the most popular mapping techniques are based on principal component analysis and multidimensional scaling. It is well

known that the principal component analysis is appropriate for simple data structures. For complicated structures, the multidimensional scaling is proved to be more effective. Its principal disadvantage is the lack of an analytical expression that ties the coordinates of the real vector  $x_j$  in  $n$ -dimensional space with the coordinates of its planar representative. Because of this, it is impossible for new samples to be projected onto previously obtained data mapping. This warns us that its use in pattern recognition is limited.

It should be pointed out that the technique based on the  $k$ -NN approach for scatter estimation was applied for supervised pattern classification successfully by Fukunaga and Mantock (1983), Huan Zhen-hua (1984) and Aladjem (1991). The results obtained in the present paper show that it is suitable for unsupervised pattern classification as well. The main advantage of this technique is its flexibility for adaptation to various data structures. This is achieved by the control of the degree of data localization through variation of the number of the nearest neighbor members.

This paper presents mappings that, as opposed to classical principal component mapping, express

This work was supported in part by the Israeli Ministry of Science and Technology under contract number 3528, and in part by the Paul Ivanier Center for Robotics and Production Management, Ben Gurion University of the Negev, Israel.

not the global but the local structure of the data. They provide an analytical expression of the mapping transformation. Combination with classical principal component mapping allows the examination of data structures in various views.

**2. Maximum mean square density gradient mapping**

Consider a set of  $n$ -dimensional real vectors  $x_j \in \mathbb{R}^n$ . Let  $d$  be an  $n$ -dimensional vector, and  $y_j = d^T x_j$  be the projection of  $x_j$  onto  $d$ . Let  $p(y)$  be the probability density function of the projections  $y_j$ . We want the rows of the mapping matrix to be orthogonal vectors for which the average gradient of  $p(y)$  is maximal. The motivation for that can be explained as follows. It is probable that cluster centers coincide with modes of  $p(y)$ . The higher the concentration of projections around cluster centers, the narrower and higher are the relevant modes in  $p(y)$ , and therefore the higher is the average gradient of  $p(y)$ . The gradient of a density function was applied for cluster analysis by Fukunaga and Hosteler (1975) and discriminant analysis by Huan Zhen-hua et al. (1984). We are stimulated by these works to use the gradient of a density function for mapping multidimensional data in the case of unsupervised learning.

The proposed mapping is based on the normalized gradient  $\nabla p(x)/p(x)$  of the density function  $p(x)$  of the samples  $x_i, i = 1, 2, \dots, N$ . It can be expressed in the form

$$\frac{\nabla p(x)}{p(x)} = \nabla \ln p(x). \tag{2}$$

Let the mapping be spanned on the directional vectors  $d_1$  and  $d_2$ . The transformation matrix  $D$  that produces the mapping is

$$D = [d_1 \ d_2]^T. \tag{3}$$

The projections of the normalized gradient  $\nabla \ln p(x)$  of the samples  $x_i, i = 1, 2, \dots, N$ , onto the lines defined by  $d_1$  and  $d_2$  are

$$z_{1i} = d_1^T \nabla \ln p(x_i), \tag{4}$$

$$z_{2i} = d_2^T \nabla \ln p(x_i), \quad i = 1, 2, \dots, N. \tag{5}$$

The mean squares of the projection values are

$$\begin{aligned} J_1 &= \frac{1}{N} \sum_{i=1}^N z_{1i}^2 \\ &= d_1^T \left[ \frac{1}{N} \sum_{i=1}^N \nabla \ln p(x_i) \nabla^T \ln p(x_i) \right] d_1 \\ &= d_1^T S^* d_1 \end{aligned} \tag{6}$$

and

$$J_2 = \frac{1}{N} \sum_{i=1}^N z_{2i}^2 = d_2^T S^* d_2, \tag{7}$$

where

$$S^* = \frac{1}{N} \sum_{i=1}^N \nabla \ln p(x_i) \nabla^T \ln p(x_i). \tag{8}$$

The total mean square value is

$$J = J_1 + J_2 = \text{Tr}[D^T S^* D]. \tag{9}$$

Assuming that we need to maximize  $J$  under the constraint  $d_1^T d_2 = 0$ , the problem can be converted to a conventional eigenvalue problem. The mapping that maximizes the mean square normalized gradient (9) is spanned on the eigenvectors which correspond to the two largest eigenvalues of the matrix  $S^*$  (8).

In order to explain the features of the mapping, we discuss this problem for the normal density function

$$\begin{aligned} p(x) &= \frac{1}{(2\pi)^{n/2} |A|^{1/2}} \\ &\quad \times \exp[-\frac{1}{2}(x-m)^T A^{-1}(x-m)], \end{aligned} \tag{10}$$

where  $m$  is the mean vector of the samples and estimated as

$$m = \frac{1}{N} \sum_{i=1}^N x_i, \tag{11}$$

$A$  is the covariance matrix of the samples which is estimated as

$$A = \frac{1}{N} \sum_{i=1}^N (x_i - m)(x_i - m)^T. \tag{12}$$

The normalized gradient of the density function (10) is

$$\nabla \ln p(x) = \frac{\nabla p(x)}{p(x)} = -C(x-m), \tag{13}$$

where  $C$  is the matrix

$$C = \frac{1}{(2\pi)^{n/2} |A|^{1/2}} A^{-1}. \tag{14}$$

Without loss of generality, we can assume that  $A$  and respectively  $A^{-1}$  are diagonal matrices

$$A = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n), \quad (15)$$

$$A^{-1} = \text{diag}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \dots, \frac{1}{\lambda_n}\right). \quad (16)$$

Omitting the constant factor in (14), we can obtain the matrix  $S^*$  (8)

$$\begin{aligned} S^* &= \frac{1}{N} \sum_{i=1}^N [A^{-1}(x_i - m)][A^{-1}(x_i - m)]^T \\ &= A^{-1} \left[ \frac{1}{N} \sum_{i=1}^N (x_i - m)(x_i - m)^T \right] A^{-1} \\ &= A^{-1} A A^{-1} = A^{-1}. \end{aligned} \quad (17)$$

Therefore, in the case of normal density  $p(x)$ , the maximum mean square gradient mapping is spanned on the eigenvectors which correspond to the two largest values of the matrix  $A^{-1}$  (16) or on the eigenvectors which correspond to the two smallest values of  $A$  (15). This result shows that the proposed mapping is opposite to the principal component mapping. It tends to compress the samples, while the classical principal component mapping provides the maximal scattering of the samples.

We want to apply the proposed mapping to cluster compression. In this case we assume the multimodal density function  $p(x)$ .  $p(x)$  can be estimated by

$$\hat{p}_N(x) = \frac{1}{Nh^n} \sum_{i=1}^N K\left(\frac{x - x_i}{h}\right), \quad (18)$$

where  $K(x)$  is a scalar function and  $h$  is a smoothing parameter which satisfies some requirements to guarantee asymptotic unbiasedness and consistency of the estimate (Fukunaga (1972)). The gradient of  $\hat{p}_N(x)$  (18) is

$$\nabla \hat{p}_N(x) = \frac{1}{Nh^{n+1}} \sum_{i=1}^N \nabla K\left(\frac{x - x_i}{h}\right). \quad (19)$$

Following Fukunaga and Hostetler (1975) we use the kernel function

$$K(x) = \begin{cases} c(1 - x^T x) & \text{if } x^T x \leq 1, \\ 0, & \text{if } x^T x > 1, \end{cases} \quad (20)$$

where

$$c = \pi^{-n/2} \left(\frac{n+2}{2}\right) \Gamma\left(\frac{n+2}{2}\right), \quad (21)$$

is the normalizing constant and  $\Gamma(\cdot)$  is the gamma function. This gives the gradient estimate

$$\nabla \hat{p}_N(x) = \hat{p}_N(x) \frac{n+2}{h^2} \frac{1}{k} \sum_{x_i \in S_h(x)} (x_i - x), \quad (22)$$

where

$$S_h(x) \equiv \{z: (z - x)^T(z - x) \leq h^2\} \quad (23)$$

and  $k$  is the number of observations falling within region  $S_h(x)$  and, therefore, the number in the sum in (22). The estimate of the normalized gradient is

$$\begin{aligned} \nabla \ln \hat{p}_N(x) &= \frac{\nabla \hat{p}_N(x)}{\hat{p}_N(x)} \\ &= -\frac{n+2}{h^2} (x - m_h(x)), \end{aligned} \quad (24)$$

where

$$m_h(x) = \frac{1}{k} \sum_{x_i \in S_h(x)} x_i \quad (25)$$

is the local mean of the observations in the region  $S_h(x)$ . The expression (24) shows that the gradient is opposite to the sample mean shift  $(x - m_h(x))$  in the region  $S_h(x)$ . Substituting (24) into (8) we obtain

$$S^* = \frac{1}{N} \sum_{i=1}^N (x_i - m_h(x_i))(x_i - m_h(x_i))^T. \quad (26)$$

The constant factor  $(n+2)/h^2$  is omitted in (26).

Therefore the mapping is spanned on the eigenvectors corresponding to the two largest eigenvalues of the matrix  $S^*$  (26).

The choice of the parameter  $h$  seems to depend upon the size of the clusters for which one is searching. This is due to the fact that  $h$  determines the amount of smoothing of the density and correspondingly the elimination of modes that are too narrow or too close to other modes. The choice of  $h$  is discussed by Woodroffe (1970).

In Figure 1 the effect of the proposed mapping is illustrated. It compresses the clusters onto the projection direction.

The maximum mean square density gradient mapping expresses the inside structure of the clusters and tends to compress them. It is expected that the combination of the classical principal component mapping and the proposed mapping will improve the data structure examination. The mapping spanned on the principal component

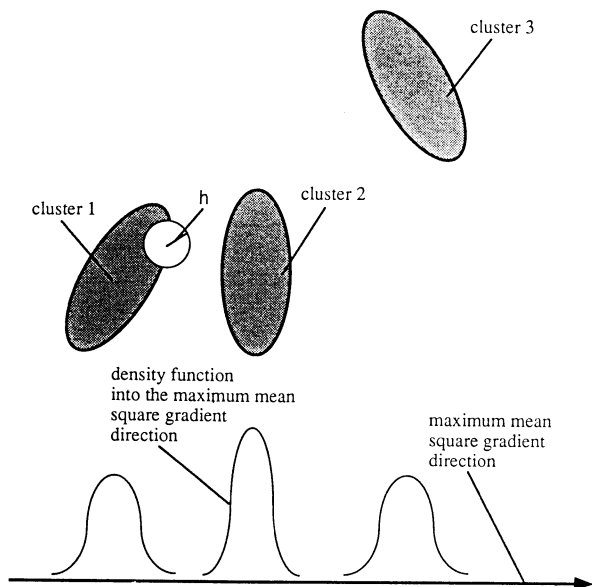


Figure 1. Maximum mean square gradient direction.

direction and maximal mean square gradient direction can be produced. It will express the separation of the clusters in the sense of the principal components and the cluster compression in the sense of the maximal mean square density gradient as well. These features are illustrated in Figure 2.

The maximum mean square gradient mapping is produced for a small value of  $h$  with respect to the sizes of the clusters and their distances apart. Letting the value of  $h$  be a large number we can obtain other features of the data structure. The following section describes this mapping.

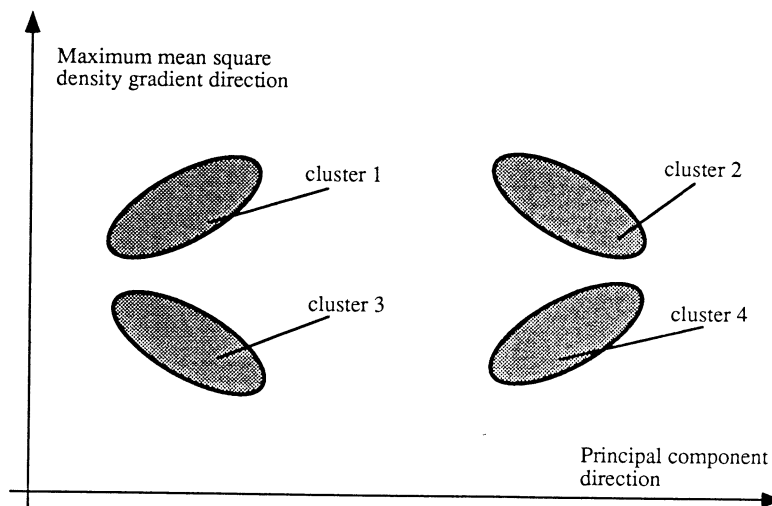


Figure 2. Combined principal component and maximum mean square gradient mapping.

### 3. Maximum local scatter mapping

The matrix  $S^*$  (26) is a generalization of the classical scatter matrix  $A$  (12).  $S^*$  reflects the global scattering of the samples around the global mean  $m$ , when  $h$  is large enough.  $S^*$  reduces to  $A$  when  $S_h(x)$  (23) contains all the samples  $x_i$ ,  $i = 1, 2, \dots, N$ . For the purpose of data structure examination, the scatter of clusters that are close to one another may be more important than the global scatter. Consider a case where two clusters are very near each other, and a third cluster is father away, as shown in Figure 3. Projecting the data onto the principal component direction yields two projected clusters because the projections of vectors belonging to the two close clusters are interleaved. Projecting the data onto a direction maximizing the local scatter conserves the three clusters of the projections. The value of  $h$  can be interactively modified until  $S_h(x)$  contains samples representing a local concentration of clusters. When this is achieved, the direction of the eigenvector corresponding to the largest eigenvalue of the respective  $S^*$  is the direction maximizing the average projection local scatter.

For the practical application of the mapping, the  $k$ -nearest-neighbor ( $k$ -NN) mean-shift estimate will be used. Letting  $h$  be replaced by the value of  $d_k(x)$ , the distance to the  $k$ -nearest-neighbor of  $x$ , and  $S_h(x)$  (23) be replaced by

$$S_{d_k}(x) \equiv \{z: \|z - x\| \leq d_k\} \tag{27}$$

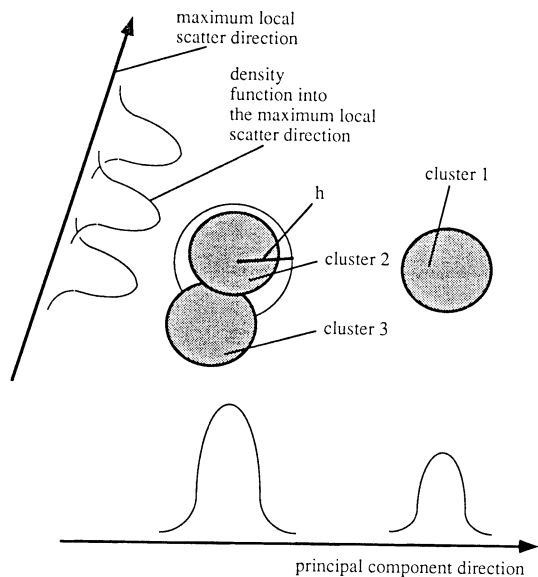


Figure 3. Maximum local scatter direction.

we obtain the  $k$ -NN estimate of the matrix  $S^*$

$$S_k^* = \frac{1}{N} \sum_{i=1}^N (x_i - m_k(x_i))(x_i - m_k(x_i))^T, \quad (28)$$

$$m_k(x_j) = \frac{1}{k} \sum_{n_i(x_j) \in S_{k_i}(x_j)} n_i(x_j) = \frac{1}{k} \sum_{i=1}^k n_i(x_j), \quad (29)$$

where  $n_i(x_j)$ ,  $i = 1, 2, \dots, k$ , are  $k$  nearest neighbors to  $x_j$  with respect to some metric in  $n$ -dimensional space  $\mathbb{R}^n$ . In  $S_k^*$  the variation of the value  $d_k(x)$  is neglected.

Based on the  $k$ -NN scatter matrix  $S_k^*$  (28) the following mapping algorithm is suggested:

- S-1. Set initial value of the number of the nearest neighbors,  $k$ .  
It is recommended that the trials be started with the value  $k = N/m$ , where  $m$  is the expected number of the clusters in the analyzed data set.
- S-2. Compute scatter matrix  $S_k^*$  (28).
- S-3. Find the eigenvectors  $r_{1k}$ ,  $r_{2k}$  of  $S_k^*$  corresponding to the two largest eigenvalues.
- S-4. Display the scatter plot of the sample projections onto the space spanned on  $r_{1k}$  and  $r_{2k}$ .
- S-5. Analyze the obtained scatter plot and make the following decisions:
  - Change the value of the number  $k$  of the nearest neighbors. Continue with S-1.
  - Stop analysis.

The proposed algorithm is an interactive mapping algorithm. The user analyzes the mapping

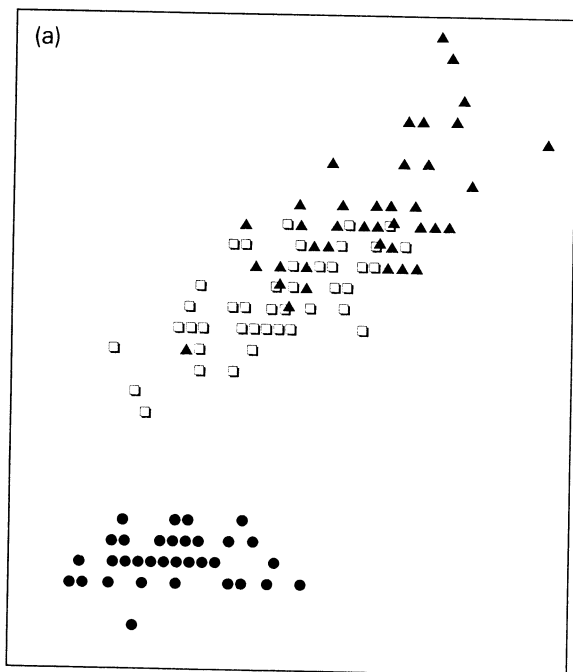


Figure 4a. Local data structure mapping for  $k = 5$ .

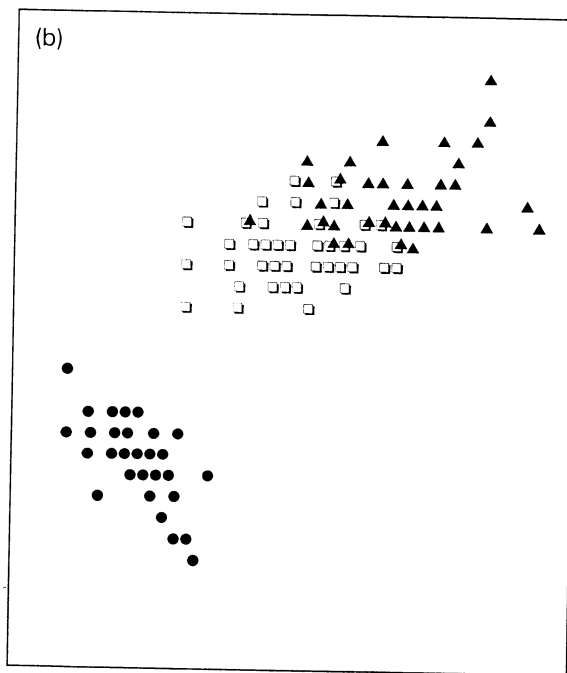


Figure 4b. Local data structure mapping for  $k = 20$ .

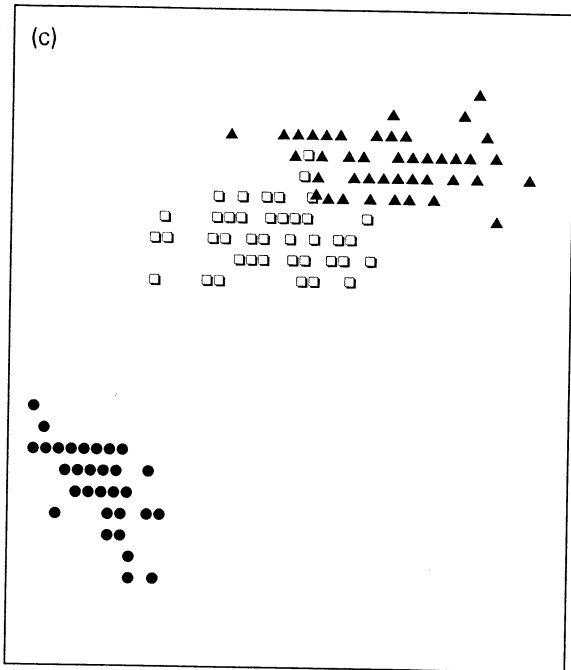


Figure 4c. Local data structure mapping for  $k=40$ .

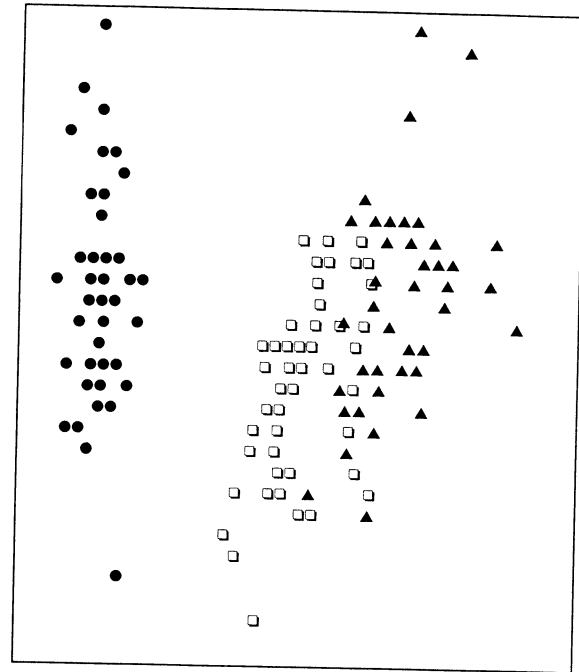


Figure 5. Principal component mapping.

results and makes the decision for a new sequence of trials.

#### 4. Experiment

##### 4.1. Data set

The classical Iris data (Kendall (1975)) were used in the experiment. The samples are four-dimensional. The data set includes 150 samples of three species of iris—Iris sesota, Iris versicolor and Iris virginica. 50 samples are taken from each population.

##### 4.2. Mappings obtained in the experiment

The proposed local scatter mapping (Section 3) was run. The values of numbers  $k$  of the nearest neighbors were varied as follows:

$$k=5, 20 \text{ and } 40.$$

In Figures 4a, 4b and 4c, the mappings obtained are presented. The samples on the mappings are labelled with '•' for Iris sesota, '□' for Iris versicolor and '▲' for Iris virginica.

For comparison, the following mappings are shown:

(1) In Figure 5, the principal component mapping is given.

(2) In Figure 6, the mapping obtained by the

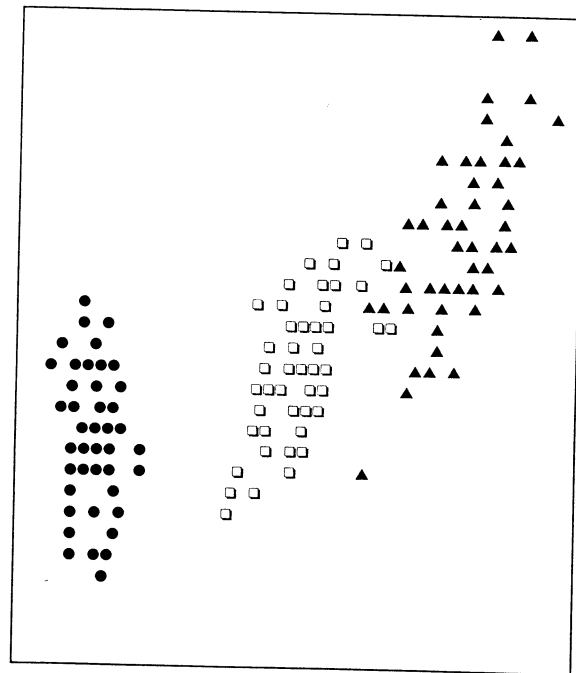


Figure 6. Multidimensional scaling mapping.

multidimensional scaling (Sammon (1969)) is presented.

### 5. Discussion and concluding remarks

The experimental results confirm the efficiency of the proposed local scatter mapping (Section 3). It (Figure 4c) separates the clusters of the various species better than the principal component mapping (Figure 5). The separation obtained is comparable with the cluster separation of the multidimensional scaling (Figure 6).

The principal advantage of the proposed method versus multidimensional scaling is the analytical expression of the mapping projection.

The analysis of the mappings obtained (Figures 4a, 4b and 4c) shows that the accuracy of cluster separation depends strongly on the value  $k$  of the number of nearest neighbors. The clusters of Iris versicolor ('-') and Iris virginica ('▲') are not formed for the values of  $k=5$  and 20 (Figures 4a and 4b). The appropriate mapping is obtained for the value of the control parameter  $k$  which is close to the number of the samples in the clusters (Figure 4c).

The mapping algorithms are intended for interactive data analysis. An important feature of the interactive procedures is the computation time. The principal component mapping and the proposed mappings are based on the solving of the symmetric eigenproblem. Using contemporary software (IMSL, NAG), the eigenvectors and eigenvalues can be obtained very efficiently. The time complexity for these computations is of the order of  $n^3$  ( $O(n^3)$ ), where  $n$  is the dimension of the data. For the principal component mapping, the computation of the scatter matrix  $A$  is ( $O(n^2 \times N)$ ), where  $N$  is the number of the data samples. The computation complexity required for the matrix  $S_k^*$  is higher than that required for  $A$ . The distances

between all pairs of data samples must be calculated and sorted before the local means and the matrix can be computed. The computation complexity depends on  $N^2$ , making the local structure mapping substantially more computational expensive compared to the principal component transform. However, the computation complexity of the nonlinear mapping proposed by Sammon (1969) is  $O(N^3)$ . As demonstrated in the previous section, the results of the local structure mapping are very similar to those of the multidimensional scaling mapping, and both are by far superior to the principal component mapping results. Therefore, the local structure mapping is an attractive tool for interactive data structure analysis.

### References

- Aladjem, M. (1991a). Parametric and nonparametric linear mappings of multidimensional data. *Pattern Recognition* 24, 543-553.
- Aladjem, M. (1991b). PNM: A program for parametric and nonparametric mapping of multidimensional data. *Computers in Biology and Medicine* 21, 321-343.
- Fukunaga, K. (1972). *Introduction to Statistical Pattern Recognition*. Academic Press, New York.
- Fukunaga, K. and L.D. Hostetler (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inform. Theory* 21, 32-40.
- Fukunaga, K. and J.M. Mantock (1983). Nonparametric discriminant analysis. *IEEE Trans. Pattern Anal. Machine Intell.* 5, 671-678.
- Huan Zhen-hua, Li Ming-hong and N. Laveen (1984). A non-parametric feature extraction algorithm. *Proc. Int. Conf. Syst. Man Cybern.* 1, 591-595.
- Kendall, M.G. (1975). *Multivariate Analysis*. Charles Griffin Company, London.
- Sammon, J.W. Jr. (1969). A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* 18, 401-409.
- Siedlecki, W., K. Siedlecka and J. Sklansky (1988). An overview of mapping techniques for exploratory data analysis. *Pattern Recognition* 21, 411-429.
- Woodroffe, M. (1970). On choosing a delta-sequence. *Ann. Math. Statist.* 41, 1965-1971.