ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 173-184

Searching for the embedded manifolds in high-dimensional data, problems and unsolved questions

Jeanny Hérault[1], Anne Guérin-Dugué[2], Pierre Villemain[1],

[1] LIS, 46 Avenue Félix Viallet, 38031 Grenoble Cedex, France

[2] CLIPS, Bât. B 385 Rue de la Bibliothèque, 38041 Grenoble Cedex 9, France

{Jeanny.Herault,Pierre.Villemain}@inpg.fr, Anne.Guerin@imag.fr

Abstract. Starting from a recall of several classical – and less classical - remarks about high dimensional data spaces, this paper gives a bird's eye view over various techniques of data reduction, from linear multidimensional scaling to non-linear and non-parametric methods. Two kinds of approaches will be presented, the first one operating in the feature space, the second one operating in the dissimilarity space. A special attention will be devoted to the CCA algorithm, in a version which aims at capturing the mean manifold spanned by the data vectors. Some examples from artificial and real data are given.

1. Introduction

The general problem with high-dimensional data sets is, using the inherent redundancy of the data, first to obtain a reduction of dimension, second to obtain a representation of the intrinsic dimension of these data. That is to provide a "picture" which can be used to give a meaningful interpretation of the data. This can be done through various techniques such as Multidimensional Scaling (Shepard, 1962; Cox & Cox, 1995) or Non-Linear Mapping (Mardia *et al*. 1979; Borg, 1997).

These algorithms are based on the point mapping of *n*-dimensional vectors to a lower-dimensional space such that the inherent structure of the data is approximately preserved. The input data can be either vectors from a set of measurements (the input space is known), or an inter-point Euclidean distance matrix or a dissimilarity matrix (where the dimension of the input space is unknown). In some cases the data are non-metric: only the rank orders of the distances are known (Kruskal, 1964). For a similarity matrix $\{s_{ij}\}$, often found in psychometry, a conversion into a distance matrix $\{d_{ij}\}$ is required, for example by: $d_{ij} = \sqrt{2 - s_{ij} - s_{ji}}$, (D'Aubigny, 1989).

Notations. In the sequel, we will use the following notations:
A data set is composed of N observations (objects), which are considered as vectors $\mathbf{x_i}$ of n features ($\mathbf{x}_i \in \mathrm{R}^n, i = 1..N,$). According to some given norm, an inter-point distance (or dissimilarity) matrix D is defined by: $\mathbf{D} = \left\{d_{ij}\right\}, \mathbf{D} \in \mathrm{R}^{N \times N}$.

2. High-dimensional data spaces

A well known property of high dimensional spaces is the phenomenon of *empty space* (reviews in: Donoho, 2000; Verleysen, 2001; Landgrebe, 2002). This is illustrated by various considerations as follows.

- The Euclidean norm of the data tends to be constant as n increases: It has been demonstrated (Demartines, 1994) that for vectors with random iid components,

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 173-184

the mean of their Euclidean norms increases as $\sqrt{n}$ while their standard deviation tends to a constant.

- The volume of a hypersphere of radius r, compared to the volume of the hypercube of side $2r$, tends to be negligible. The ratio between the sphere's and the cube's volume, given by $R = \frac{1}{n2^{n-1}} \frac{\pi^{n/2}}{\Gamma(n/2)}$, decreases very rapidly to 0 as n increases (fig. 1), even for small values of n (10-20), it may become negligible. The consequence is that, for vector quantization of the space, the number of prototypes to be defined may become huge if the cost function is based on the Euclidean metrics.

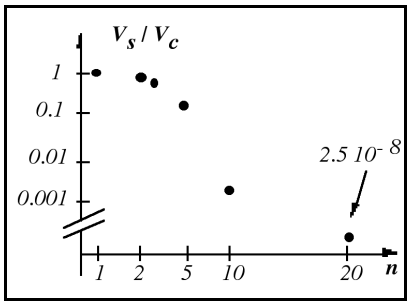

Figure 1. Sphere-to-cube volume ratio, according to the dimension n of the space.

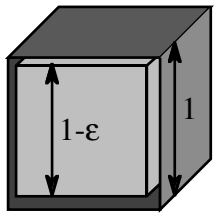

Figure 2. The external shell of a hypercube contains almost all the available volume (see text).

- The volume of a shell, defined (fig. 2) by the space between the unit hypercube and a centered hypercube of side (1-ε), is: $1-(1-\varepsilon)^n$. It tends to 0 as n increases: for n=50 and ε=5%, the shell represents 92.3% of the hypercube's volume! This fact is of particular importance if a Minkowski norm is considered (see later).

- Gaussian distribution: If the data are normally distributed, the probability density of finding a point at distance r from the center of the distribution is given by: $f = \frac{r^{n-1}}{2^{n/2-1}} \frac{e^{-r^2/2\lambda^2}}{\Gamma(n/2)}$ (fig. 3). It is maximum for $r/\lambda = \sqrt{n-1}$. This means that the space is relatively empty at the center, where the gaussian distribution is maximal!

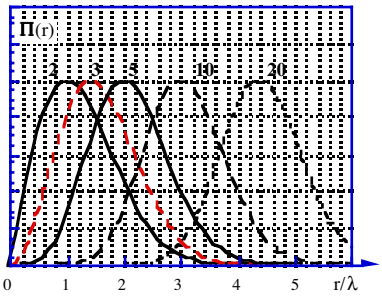

Figure 3. The probability density of a point at distance r from the center is, maximal at $\sqrt{n-1}$

- The number of training samples, in supervised classification, is related to the dimension. It has been shown (Fukunaga, 1990; Lee & Landgrebe, 1993) that the number of samples should be proportional to the dimension for linear classifiers, and to the square of the dimension for quadratic classifiers. However, for non-parametric classifiers, it may grow exponentially with the dimension: $N \approx K^n$.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 173-184

- Projections: according to the central limit theorem, low dimensional linear projections tend to be normally distributed as the dimension increases. This would mean that little information could be extracted in such cases (projection pursuit framework).

- The scalar product of a unit vector on the diagonal with any axis gives an angle defined by: $\cos(\theta) = 1/\sqrt{n} \xrightarrow[n\to\infty]{} 0$. Be careful, this does not mean that the diagonal is really orthogonal to the axes! Figure 4 shows that the angle is around 80°-95° for a large scale of dimensions.

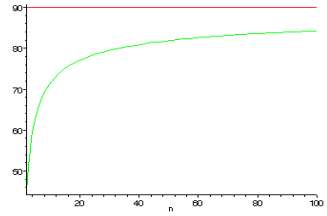

Fig. 4. Angle between the diagonal and the axes in dimension n.

However, fortunately the observed data features are not fully independent. This means that if the features are linked, the samples may span some subspace or some (even non-linear) manifold of reduced dimension. Then, it is mandatory to find this *intrinsic* dimension of the data set. Embedding it into a suitable subspace or manifold of low dimension is a means to provide a simpler representation of the data, useful for further classification tasks.

3- Finding the subspace or the manifold spanned by the data.

There are many methods for such a purpose, ranking from linear- to non-linear ones, metric or non-metric, parametric or non-parametric (Shepard, 1962, 1965; Sammon, 1969; Pekalska et al., 2001). Here are some classical examples of these methods.

3.1. Classical MDS

Let us consider a *dissimilarity matrix* $\mathbf{D}$ issued from N observations, not related to any known space (i.e. as obtained from a psychological experiment). Under the hypothesis that the underlying space is Euclidean ($\mathbf{X} \in \mathrm{R}^{N\times n}$: N vectors $\mathbf{x}_i$ in n dimensions), the classical Multi-Dimensional Scaling (Mardia & al., 1979) provides a representation of this unknown space.

If the distance matrix $\mathbf{D} \in \mathrm{R}^{N\times N}$ is Euclidean, we consider the (positive definite) matrix of its squared elements: $D^{(2)} = \left\{d_{ij}^2\right\}$. By using the centering matrix $\mathbf{J} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}\right)$, the covariance matrix of the data is obtained by: $\mathbf{B} = -\frac{1}{2}\mathbf{J}\mathbf{D}^{(2)}\mathbf{J}$.

The factorization of $\mathbf{B}$ by its eigendecomposition gives $\mathbf{X}\mathbf{X}^{\mathrm{t}} = \mathbf{B} = \mathbf{Q}\Lambda\mathbf{Q}^{\mathrm{t}}$, the number of non-zero eigenvalues gives the dimension k of the representation space and the (Nxk) data matrix is obtained by: $\mathbf{X} = \mathbf{Q}_k \Lambda_k^{1/2}$.

In the non-Euclidean case ($\mathbf{D}$ and $\mathbf{B}$ are not positive definite), various approximations may be obtained, either by neglecting the negative eigenvalues, or by transforming $\mathbf{D}^{(2)}$ as: $\mathbf{D}^{(2)} \to \mathbf{D}^{(2)} + c\left(11^{\mathrm{T}} - \mathbf{I}\right)$, with $c > |\lambda_m|$, λ_m being the smallest (negative) eigenvalue of $\mathbf{B}$. Another method consists of defining a *pseudo-Euclidean* space (Goldfarb, 1984) such as:

$\mathbf{B} = \mathbf{X}\mathbf{M}\mathbf{X}^{\mathrm{T}}$, with $\mathbf{M} = \begin{bmatrix} \mathbf{I}_{p\times p} & 0 \\ 0 & -\mathbf{I}_{q\times q} \end{bmatrix}$, p and q being respectively the numbers of positive and negative eigenvalues of $\mathbf{B}$.

3.2. Non-parametric methods.

The main idea is the following: for every couple of distinct points (i,j), take every inter-point distance $X_{ij} = \|x_i - x_j\|$ in the input space and find the corresponding inter-point distance $Y_{ij} = \|y_i - y_j\|$ in a lower-dimensional output representation space. This can be done in different manners. One of them is basically to minimize the following quadratic form: $E = \sum_{i;j}\left(X_{ij} - Y_{ij}\right)^2$, for example by means of some gradient descent algorithm. If the dimensions of the input and output spaces are the same, the cost function E can be made null. Some normalized versions have been proposed as: $E = \frac{1}{\sum_{i,j} X_{ij}^2} \sum_{i,j}\left(X_{ij} - Y_{ij}\right)^2$, or as $E = \frac{1}{\sum_{i,j} 1/Y_{ij}^2} \sum_{i,j} X_{ij}^2 / Y_{ij}^4$, (Shepard, 1964), the last one is computationally very demanding and the equilibrium point is difficult to find. A better cost function, automatically normalized, has been designed (Sammon, 1969) to favor the mapping of small distances in the input space, thus assuring a "locally correct" topographic mapping: $E_s = \sum_{ij} \frac{\left(X_{ij} - Y_{ij}\right)^2}{X_{ij}} \Big/ \sum_{ij} X_{ij}$ through the "unfolding" of the data manifold.

However, this unfolding may be difficult or impossible to obtain in the case of strongly folded data, simply because the favoring of small distance X_{ij} must sometimes be relaxed: in a U-shaped input distribution, the extreme points which are not very distant should not contribute to the mapping cost function. We will see later a more interesting cost function, in which the terms are weighted by short distances *in the output space*: the CCA algorithm (Demartines & Hérault, 1997).

There are many other approaches to ascertain the structure of a data set. Apart from the well known Self-Organizing Maps (Kohonen, 1989), which acts more as a vector quantization under a topological constraint, there is the recent ISOMAP algorithm (Tenenbaum et al. 2000). The input distances are based on a Dijkstra algorithm on the graph obtained by considering the nearest neighbors. Then, a MDS procedure is applied to find the representation space.

4- Metrics

There are many possible metrics which may be used to represent the data space, though the Euclidean one is very common. We will give some emphasis to the Minkowski metrics for which the distance $d_{ij} = \left[\sum_k \left(x_{ik} - x_{jk}\right)^p\right]^{1/p}$ has often been considered as relevant for psychometric dissimilarities (Tversky & Krantz, 1970).

- The L^p (or Minkowski) norm is defined by $\|\mathbf{x}\|_p = \left[\sum_k x_k^p\right]^{1/p}$. If the data are L^p-normalized ($\|\mathbf{x}\|_p = 1$), they tend to map into the n-1 dimensional volume of

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 173-184

the unit cube's external shell as p increases (fig. 5): this complies with the property illustrated in figure 2.

- Filling the space with Minkowski hyperspheres. The volume ratio of the unit Minkowski sphere to the unit L^p cube is $R = \Gamma(1+1/p)^n / \Gamma(1+n/p)$. Fig 6 shows that choosing p according to n helps to pave the space better than with Euclidean spheres.

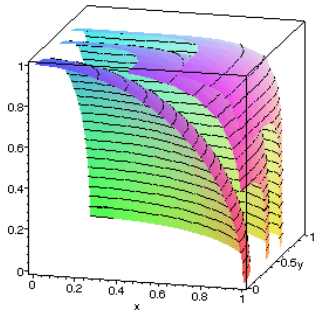

Figure 5. L^p normalized data in 3 dimensions, with p=2, 4, 8.

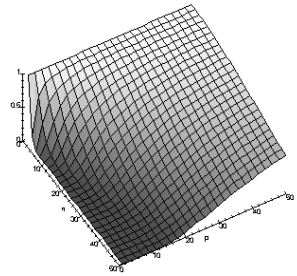

Figure 6. Filling the unit cube by Minkowski spheres.

- The rank r of the matrix $\mathbf{D}^{(p)} = \left\{ d_{ij}^p \right\} = \left\{ \Sigma_k \left(x_{ik} - x_{jk} \right)^p \right\}$ has a maximum value of $r = n(p-1) + 2$. Hints for proof: write that line number i of the matrix is a linear combination of r other lines. Then, every element is expressed as: $d_{ij}^p = \sum_{s=1}^{r} \lambda_{is} \, d_{sj}^p, \ \forall j \in [1,N]$, or: $\sum_{k=1}^{n} \left(x_{ik} - x_{jk} \right)^p = \sum_{s=1}^{r} \lambda_{is} \sum_{k=1}^{n} \left(x_{sk} - x_{jk} \right)^p$. It gives a set of equations where, after having developped the polynoms of degree p, it is possible to estimate the number of independent equations. This number is r, equal to the number of coefficients λ, that is, the rank of the matrix.

This result has several interesting consequences:

1. If the components x_k are redundant (linear combinations or not), the number of independent equations is reduced, and the dimension of the data space is $n'<n$.
2. If $r=N$, the number of observations may not be sufficient.
3. A data point is entirely defined by its r distances to other particular points. The data may be represented equivalently in the space of distances (see after).
4. A constant L^p norm in the input space X is represented by a hyperplan in the r dimensional space of distances $D^{(p)}$: this result may be of interest for kernel classifiers which are built in the space of distance (Pekalska et al., 2001).
5. The space of L^p distances offers more degrees of freedom than the Euclidean distance space because the rank of the distance matrix is higher.
6. If a class of points spanns a manifold of reduced dimension, the rank of its submatrix is lower than that of the whole data set.

Another interesting approach recently proposed, consists of a supervised learning of the suitable metrics, using information about external relevant data (Kaski, 2001). It has been used for financial or genetic data, and for text retrieval.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 173-184

5- Dissimilarity based Discrimination

Up to now, we have considered the task of representing the high dimensional data. Here we consider a *discrimination task*, from a specific point of view on discrimination analysis from dissimilarity measures.

Classifying objects according to their proximity is a fundamental task of pattern recognition and arises as a classification problem or discriminant analysis in experimental sciences (Fukunaga, 1990). In many application cases (biology, genetics, psychophysics, signal, image processing…), it is very hard to deal with an explicit feature-space and metrics representation. In such cases, implicit representations can be captured from the dissimilarity measures between observations provided by the acquisition process. When the feature space is not available, classical methods of Pattern Recognition cannot be directly used. As far as we know, four approaches dealing with dissimilarity-based pattern recognition have been proposed, starting from a dissimilarity matrix. (Pekalska et al, 2001) have named this approach "featureless pattern learning".

1. Use a ranking-based method, like the "K Nearest Neighbours rule".
2. Use an extended version of the "Support Vector Machines" directly with dissimilarity kernels, like "Support Vector Classifiers" (Vapnik, 1995; Schölkopf, 1997).
3. Use Multi-Dimensional Scaling techniques (Borg & Groenen, 1997) or other non-linear techniques such as CCA to embed dissimilarities into an Euclidean feature space, then use features-based pattern recognition classifiers.
4. For each observation, consider the list of dissimilarities with all other observations as a new features vector in a high dimensional space (one observation equal to one dimension). Then use feature-based classifiers, for example linear or quadratic classifier (Pekalska et al, 2001).

We have proposed an alternative approach to 1 and 4 (Guérin-Dugué & Celeux, 2000). Techniques for case 3, and especially those using CCA, are described in this paper. To these constraints, we add another important one: the dissimilarity matrix may be sparse. It is a common situation in practice, as the trend is to handle larger and larger databases where not all the dissimilarity values are known. The direct application of techniques based on the approach 4 is difficult in this case. But starting from these dissimilarity- based features, we can learn statistical moments extracted from the dissimilarity values between one observation and the remaining others. To do this, it is well known, in the Euclidean context, that if we consider only the first order statistics, the derived decision rule leads to a simple linear classifier (classification upon the class position). Let us consider a discrimination task into K classes (N_k observations in each learning class). In this framework, the linear decision rule on a new observation is: $class(e)=\arg\left[\underset{k}{Min}\left(\overline{d_k^2(e)}-I_k\right)\right]$, with

$$I_k = \frac{1}{2N_k^2}\sum_{i,j\in\omega_k} d^2(i,j) = \text{Inertia}(\omega_k),$$

and $\overline{d_k^2(e)}=\frac{1}{N_k}\sum_{i\in\omega_k} d^2(e,i)$.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 173-184

Non-linearity is introduced by means of the second order statistics with the variances on the dissimilarity distributions. In an Euclidean context, these quantities take into account the "shape" and the intrinsic dimension of each class. Even if these simple geometric interpretations are no longer valid with dissimilarity matrices, these approaches can be advantageously applied to provide new discrimination tools in such a context.

In addition, data-driven learning procedures are defined (Guérin-Dugué & Celeux, 2000). We introduce some adaptation to a specific class through a "shape coefficient" (variance / squared mean). It is defined from the classical coefficient of variation (standard deviation / mean) for each observation $class(e)=\arg\left[\underset{k}{Min}\left(\beta_k Cshape_k(e)\right)\right]$, with β_k a learning parameter optimised by cross validation, and the "shape coefficient" defined by $Cshape_k(e)=\left(\overline{d_k^2(e)}-I_k\right)\Big/Variance(d_k^2(e))$.

Finally, the advantages of the proposed method are, (i) a data driven versatility (adaptive parameters to learn the "shape" and the intrinsic dimension of each category/class), and (ii) adaptation to incomplete dissimilarity data by estimating statistics over all the available dissimilarity values.

6- CCA Curvilinear Component Analysis

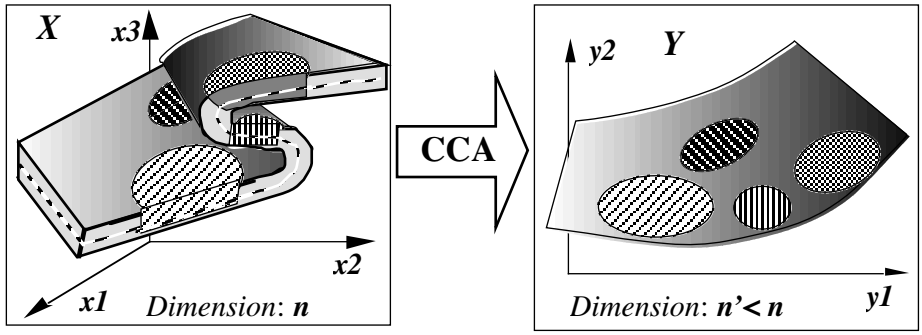

Figure 7. Principle of the CCA algorithm. After a possible vector quantization of the input data space (X) in n dimensions, the local topology of the input average manifold is mapped into an output representation space (Y) of dimension $n' < n$.

2.1. Principle

The input consists of N samples belonging to some theoretically n' dimensional manifold, embedded in an n-dimensional input space $\mathbf{X}=\{x_{ik}\}$, $i=1..N$, $k=1..n$. But the manifold may have some "thickness", therefore being possibly of higher dimension. The goal is to find the dimension of the average manifold of the data and to map it onto a representation space $\mathbf{Y}$ of lower dimension: to do this, we proceed to a Global Unfolding together with a Local Projection onto the average manifold (fig. 7). We use N neurons with n-dimensional input weights and n' dimensional output weights.

The CCA algorithm makes the neurons themselves find their neighborhood in the output space by adapting their output weights to the local topology of the input samples, according to some cost function.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 173-184

2.2. Choice of the cost function

Let us come back to the basic cost function (section 3.2), without normalization for sake of clarity: $E = \sum_{ij} E_{ij}$, with $E_{ij} = \left(X_{ij} - Y_{ij}\right)^2$. The input inter-point distances $X_{ij} = \|x_i - x_j\|$ being given, we start from a random configuration of points y_j. Then, for every point y_i in the output space, we move the points y_j so that the terms E_{ij} are minimized, for example by means of a gradient descent algorithm.

In order to map the average manifold of the data, two cases are to be considered (see figure 8): first, we need a *global unfolding* of the average manifold of the data, and second, we need a *local projection* of the data onto the average manifold.

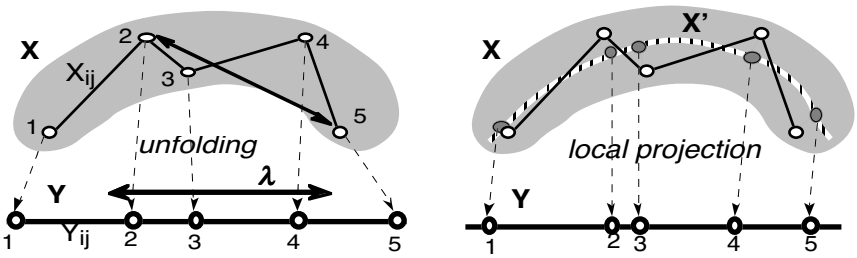

Figure 8. Illustration of the problem of data representation, in two cases: either only an unfolding is desired, or also a local projection is desired (see text).

Unfolding: Let us consider the first case alone (fig. 8.a). In order to unfold the data, only some E_{ij} terms of the basic cost function need to be minimized: those for which the distance Y_{ij} is smaller than some pre-defined distance λ. Thus, allowing the matching for only short distances is a way to respect the local topology. It has been proved that this condition (*applied on the output distances*) ensures a global unfolding much better than other mapping techniques, which apply it to the input distances (Demartines, 1994). In this case, the general term to be minimized becomes:

$$E_{ij}^{u} = \left(X_{ij} - Y_{ij}\right)^2 F_\lambda\left(Y_{ij}\right), \text{with } F_\lambda(\cdot) = 1 \text{ for } Y_{ij} < \lambda \text{ and } F_\lambda(\cdot) = 0 \text{ for } Y_{ij} > \lambda .$$

The choice of λ strongly depends on the data structure (e. g. curvature of the average manifold, spreading of the data around this manifold). As the data structure is unknown, λ decreases as the number of iterations increases, like in SOM's.

We should remark that, apart from the desired global unfolding, there is also some tendency to make a local projection. Look at the input distribution in figure 8.a: because we ask the mapping of X_{14} simultaneously with the mapping of X_{12}, X_{23} and X_{34}, the resulting compromise will lead to $Y_{12} < X_{12}$, $Y_{23} < X_{23}$ and $Y_{34} < X_{34}$, which makes an approximate projection. This property will be useful hereafter.

Projection (fig. 8.b). If the data were projected onto their average manifold, the inter-point distances X'_{ij} of the projected data would *locally* minimize the quadratic error: $(X_{ij}^2 - X'^2_{ij})$. Then, the output vectors should map this local projection. That is, translated into a cost function problem, they should minimize: $E_{ij}^{p} = \left(X_{ij}^2 - Y_{ij}^2\right)^2$.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 173-184

This applies only if $Y_{ij} \leq X_{ij}$, a situation which is initiated by the above-mentioned tendency to make a local projection. Conversely, if $Y_{ij} \geq X_{ij}$, we are in the condition of unfolding. Hence, the two situations (unfolding or projection) do not overlap, and the global cost function can merge both E_{ij}^{u} and E_{ij}^{p}, provided that the gradient continuity between them be assured when $Y_{ij} = X_{ij}$ (Herault et al., 1999). Notice that the input distance X_{ij} can be chosen of any type (Euclidean, Minkowski...).

2. 3. Monitoring unfolding and projection

The quality of the mapping is monitored during the gradient descent by the joint distribution of input and output inter-point distances: dx/dy (fig. 9).

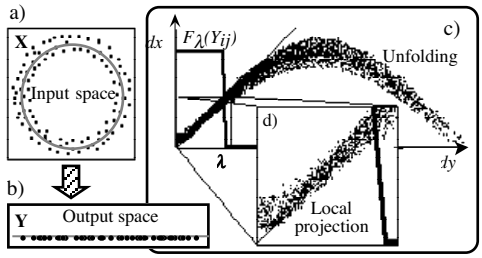

Figure 9. Evaluation of the quality of the mapping. a) Example of a 2-dimensional data space with a 1-dimensional average manifold. b) 1-dimensional output representation. c) The dx/dy joint distribution showing the regions where unfolding and local projection occur.

If the dimension of the output space is lower than that of the input space the joint distribution dx/dy presents two aspects. In the case of unfolding, the points lie on the dy>dx side of the first diagonal and, in the case of projection, they lie on the dy<dx side. A "good" mapping is obtained when there is an unfolding for large dy values and a projection for small values (fig. 9). Driving the value of λ by hand often increases the quality of the mapping.

2. 4. Example of a difficult mapping

Figure 10. Mapping of a 3-D data set of two interlaced rings onto a 2-D representation space. a) input space, b) output space where the two rings are broken in order to satisfy at best the 2-D representation, c) the result of the Sammon mapping.

In the case of a complex input data structure (fig. 10), CCA is able to find the "best" solution for the mapping, possibly at the expense of breaking the local topologic constraint when necessary. In this example, the inherent 2D structure is respected everywhere but in two points, a feature outstripping many other mapping techniques.

7- Application examples

CCA has proved to be efficient in many applications of various kinds (Guérin-Dugué et al., 1999), ranging from audio-visual speech recognition (Teissier et al., 1998) to nuclear detectors (Vigneron et al., 1997).

7.1- Scene classification by CCA

Another difficult problem that has been approached is the one of scene categorization from spatial statistics (mean value) of the energy distribution of an image in various frequency bands and orientations (Hérault J. *et al* 1997). An image is analyzed by a bank of spatial filters, according to four orientations and five 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, and each image is a 20D vector in this space. By CCA (see figure 11 left), we have found that a 2-dimensional representation was possible and that, in this space, the organization of the data was surprisingly in accordance with some semantic meaning: Natural/Artificial scenes for each side of the grey line, and Open/Closed landscapes along this line.

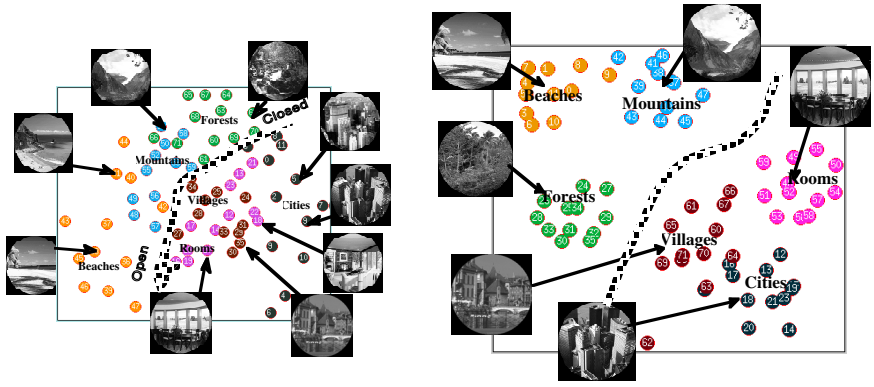

Fig. 11. Two-dimensional representation by CCA of a set of 72 images. (Left) Images are in a 20-dimensional data space obtained by the mean energies in four orientations and five spatial frequency bands. (Right) Organization obtained with an input psychological distance matrix collected on the same dataset with 20 human subjects.

Figure 11-right shows the perceptual organization of the same grey level images set according to perceptual similarities. These perceptual estimates were captured from a psychophysical experiment where each subject was asked to give similarity marks (1 to 10) between images. The obtained map compares favorably with the preceding one. Here, the clusters are better isolated, maybe due to the use of semantic information by human subjects. As previously, a clear line separates artificial contexts (human-made) from natural contexts (non human-made). This approach with perceptual metrics (Mojsilovic & Rogowitz, 2001) is of highest importance for image retrieval.

7.2- Cortical flattening by CCA

Cortical unfolding and flattening becomes a very important investigation tool when studying the activity of human brain at relatively high spatial resolution (Tootel *et al*., 1996). The problem is as follows: given 3D grey level images of the cortex (see figure 6a), how is it possible to unfold and flatten a specific interface (surface between grey and white matter) inside the 3D data? The intrinsic dimension here is known (2D). The difficulty is in the intrinsic curvature of the cortical ribbon, the non-linearity because of highly folded gyri and sulci, and the noise due to the MRI acquisition and the segmentation process. By CCA, we have shown that a cortical flattening can be easily obtained with several advantages, (i) fast processing time (fast learning strategy of CCA), (ii) good quality of the resulting map by minimizing the

overall distortion. A modified version of CCA has been developed for this application (Guérin-Dugué *et al*, 2000) in which the input distance matrix has been processed by geodesic distances along the cortical ribbon. In order to speed up these estimates, the geodesic distances between all the points in the ribbon are only processed on few points called "anchors". For the other points, only approximate geodesic distances by Euclidean distances are computed inside a given neighborhood. The distances with the anchor points give the global structure of the mapping, and the others give the local structure.

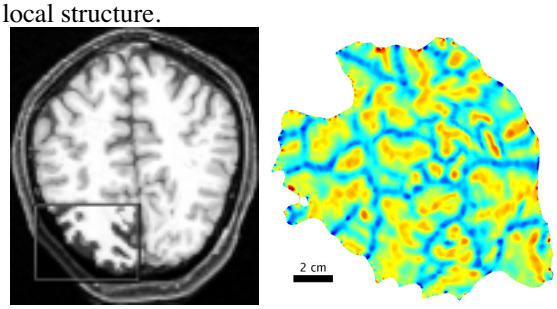

Fig. 12. Flattening the cortical surface. left bottom: segmented slice of anatomical MRI, right: 2D flattened cortical ribbon of the left hemisphere of visual cortex after CCA. The color information on the map represents the local curvature.

This process is currently used at INSERM-Grenoble for fMRI (Warnking, 2002). Figure 12 illustrates the flattening process. For the CCA algorithm, the number of nodes is 19809 of which only 10 are anchors. For the others points (19799), only the local distances are computed inside a 10 order neighborhood (for this example, the mean number of neighbors per point is 395). Though a very sparse distance matrix (only 2% of the distances are computed), both the global and local structure of the data is captured. The computing time is 140 sec on a SUN ULTRA 10 workstation for a spatial resolution of the anatomical MRI that gives an output map of 128 cm^2.

4- Conclusion

As it can be seen from the various problems presented in this mini-review, the world of high dimensional data is very complex. The schemes extrapolated from the Euclidean geometry in 2 or 3 D seem no longer suitable to describe such spaces. The first problem is that of defining a suitable metrics, either when the data space is known, or when it is not available (i.e. in the case of dissimilarity matrices obtained from human assessments). The second one is to find a "good" representation space of low dimension, which would be useful for visualization purposes (2 or 3 D) or for classification objectives. However, despite the important works devoted to this subject during the second part of the 20^{th} century, the fundamentals of the question remain mostly unsolved: when good results happen to be obtained, it often looks more due to chance than due to a reliable mathematical reasoning. Because of the increasing number of researchers who broach this topic, many interesting views become available, leading to a better and better insight of the question.

5- References

Borg I. and Groenen P. (1997). *Modern Multidimensional Scaling: Theory and Applications*. Springer Series in Statistics.

Cox, T.F. Cox, M.A.A. (1995). Multidimensional Scaling on the Sphere, *Comunications in Statistics*, 20, pp.2943-2953.

D'Aubigny G., *L'analyse Multidimensionnelle des Données de Dissimilarités*, Thèse d'état, Université Grenoble I, 1989.

Demartines P. (1994). Analyse de données par réseaux de neurones auto-organisés. Ph.D. thesis, INPG, Grenoble, France.

Demartines, P., Hérault J. (1997). Curvilinear Component Analysis: A Self-Organizing Neural Network for Nonlinear Mapping of Data Sets. *IEEE Trans. on Neural Networks*, 8, 1.

Donoho D.L. (2000). High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. *Am. Math. Soc. conf.* "*Math Challenges of the 21st Century*", Los Angeles.

Fukunaga, K. (1990). *Introduction to statistical Pattern Recognition*. Acad. Press, San Diego.

Goldfarb L. (1984). A unified approach to Pattern Recognition. *Patt. Rec*, 17: 575-82.

Guérin-Dugué A., Olympieff S., Gispert Lopez J.D., Rubin C., Segebarth C., Représentation Plane du Cortex Visuel en Imagerie Fonctionnelle à Résonance Magnétique, RFIA'2000, pp. II.29-II.38, Paris, fev. 2000.

Guérin-Dugué A., Celeux G., Discriminant Analysis on Dissimilarity Data : A New Fast Gaussian like Algorithm, AISTATS'2001, Florida, USA, january 2001

Guérin-Dugué A., Teissier P., Delso-Gafaro G. and Hérault J. (1999). Curvilinear Component Analysis for High-dimensional Data Representation: II. Examples of introducing additional mapping constraints for specific applications. IWANN'99, Alicante, Spain.

Hérault J., Jausions-Picaud C. and Guérin-Dugué A. (1999). Curvilinear Component Analysis for high dimensional data representation : I. Theoretical aspects and practical use in the presence of noise, IWANN'99, Alicante, Spain.

Hérault J., Oliva A., Guérin-Dugué A. (1997). Scene Categorisation by Curvilinear Component Analysis of Low Frequency Spectra. ESANN-97, Bruges, BE.

Kaski S. (2001). Learning metrics for exploratory data analysis. In *Neural Networks for Signal Processing XI* - Proceedings of IEEE Workshop.

Kohonen T. (1989). *Self-Organisation and Associative Memory*. Springer-Verlag, Berlin.

Kruskal J.B. (1964). Non-metric multidimensional scaling: a numerical method. Psychometrika, 29:115--129.

Landgrebe D. (2002). "Hyperspectral Image Data Analysis as a High Dimensional Signal Processing Problem," (Invited), *Special Issue of the IEEE Sig. Proc. Mag.*, 19, 1, pp. 17-28.

Lee C. & LandgrebeD.A. (1993). Analyzing High Dimensional Multispectral Data. *IEEE Trans. Geosc. And remote Sensing*, Vol. 31, No 4, 792-800.

Mardia K.V., Kent J.T., and Bibby J.M. (1979). *Multivariate Analysis*. Acad. Press, London.

Mojsilovic A. & Rogowitz, B.E. (2001), Capturing image semantics with low-level descriptors, ICIP'01, vol.I, pp.18-21, Thessaloniki, Greece, 7-10 october.

Pekalska E., Paclik P. & Duin P.W. (2001). A Generalized Kernel Approach to Dissimilarity-based Classification. *J. of Machine Learninng Res.*, 2, 175-211.

Sammon, J. W. (1969). A non-linear mapping algorithm for data structure analysis. *IEEE Trans. Computers,* C-18(5):401-409.

Schölkopf B., 1997, Support vector learning. Published by: R. Oldenbourg Verlag, Munich, PhD thesis.

Shepard R. N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. Psychometrica, vol. 27, pp.125-139.

Shepard R.N. (1965). Parametric representation of non-linear data structure. In Krishnaiah P.R., editor, *Int. Symposium on Multivariate Analysis,* pages 561-592. Academic Press.

Teissier P., Guérin-Dugué A., Schwartz J.L. (1998). Models for Audiovisual Fusion in a Noisy-Vowel Recognition Task, Journal of VLSI Signal Processing, vol 20, pp. 25-44.

Tenenbaum J.B., da Silva V. & Langford C. (2000). A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500):2319-2323.

Tootel R., Dale A., Sereno M., Malach R., New images from human visual cortex, Trends in Neurosciences, vol. 19, n° 11, pp. 481-489, 1996.

Tversky A. & Krantz D.H. (1970). The Dimensional Representation and the Metric Structure of Similarity Data. *J. of Math. Psychology*, Vol. 7, 572-597.

Vapnik V., 1995, The nature of Statistical learning, Springer, N.Y.

Verleysen M. (2001). Learning high-dimensional data. LFTNC 2001, NATO advanced reserach workshop on Limitations and Future Trends in Neural Computation, Siena, 22-24.

Vigneron V., Maiorov V., Berndt R. Sanz-Ortega J. and Schillebeeckx P. (1997). Neural network application to enrichment measurements with n*ai* detectors. *VC CSR Proc.*

Warnking J., Dojat M., Guérin-Dugué A., Delon-Martin C., Olympieff S., Richard N., Chehikian A., Segebarth C., (2002), fMRI Retinotopic Mapping : Step by Step, submitted to Neuroimage, march, 2002.