Using a global parameter for Gaussian affinity matrices in spectral clustering

Sandrine Mouysset, Joseph Noailles and Daniel Ruiz

IRIT-ENSEEIHT, University of Toulouse, France
sandrine.mouysset@enseeiht.fr, jnoaille@enseeiht.fr and
daniel.ruiz@enseeiht.fr

Abstract. Clustering aims to partition a data set by bringing together similar elements in subsets. Spectral clustering consists in selecting dominant eigenvectors of a matrix called affinity matrix in order to define a low-dimensional data space in which data points are easy to cluster. The key is to design a *good* affinity matrix. If we consider the usual *Gaussian affinity matrix*, it depends on a scaling parameter which is difficult to select. Our goal is to grasp the influence of this parameter and to propose an expression with a reasonable computational cost.

1 Introduction

Clustering has many applications in a large variety of fields : biology, information retrieval, image segmentation, etc. Spectral clustering methods use eigenvalues and eigenvectors of a matrix, called affinity matrix, which is built from the raw data. The idea is to cluster data points in a low-dimensional space described by a small number of these eigenvectors. By far, it is commonly agreed that the design and normalization of this affinity matrix is the most critical part in the clustering process. We are concerned with the Gaussian affinity matrices because they are very largely used. The expression of the Gaussian affinity matrix depends on a parameter σ and the quality of the results drastically depends on the good choice of this parameter. As said by several authors [3],[6] and [4], the scaling parameter controls the similarity between data. We propose a new expression based on a geometrical interpretation which is a trade-off between computational cost and efficiency and test it with classical challenging problems. This definition integrates both dimension and density of data.

2 Algorithm *Ng, Jordan and Weiss* (NJW)

Let $x_1, ..x_m$ be a m points data set in a n-dimensional euclidean space. The aim is to cluster those m points in k clusters in order to have better within-cluster affinities and weaker affinities across clusters. We suppose that the number k of targeted clusters is given. The affinity between two points x_i and x_j could

be defined by $A_{ij} = exp(-\|x_i - x_j\|^2/\sigma^2)$ where $\|.\|$ is the euclidean norm.We consider the spectral clustering algorithm proposed by *NJW* [3] which is based on the extraction of dominant eigenvalues and their corresponding eigenvectors from the normalized affinity matrix A. This approach resumes in the following steps :

- Form the affinity matrix $A \in \mathbb{R}^{m \times m}$ defined by:

$$A_{ij} = \begin{cases} exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}) \text{ if } i \neq j, \\ 0 \text{ otherwise} \end{cases}$$

- Construct the normalized matrix : $L = D^{-1/2}AD^{-1/2}$ with $D_{i,i} = \sum_{j=1}^{m} A_{ij}$
- Construct the matrix $X = [x_1 x_2 ..x_k] \in \mathbb{R}^{m \times k}$ by stacking the eigenvectors associated with the k largest eigenvalues of L
- Form the matrix Y by normalizing each rows in the $m \times k$ matrix X

- Treat each row of Y as a point in $\mathbb{R}^k$, and group them in k clusters via the *K-means* method
- Assign the original point x_i to cluster j if and only if row i of matrix Y was assigned to cluster j.

NJW justify this algorithm by considering an ideal case with three well-separated clusters. With the assumption that the points are already indexed by clusters consecutively, the affinity matrix has a block-diagonal structure. Thus, the largest eigenvalue of the normalized affinity matrix is 1, with multiplicity of order 3. The normalized rows of the extracted dominant eigenvectors are piecewise constant. In the field of the rows of these largest eigenvectors, it is easy to identify the three well-separated points that correspond to these three piecewise constant eigenvectors, and then to define the clusters accordingly. As already said in [4], one crucial step is to select appropriately the parameter σ and, in that respect, we have to decide between a *global* parameter as in [3], [2] and [1] or a *local* parameter that depends on the points x_i and x_j as in [6].

3 Towards the choice of a global parameter from a geometric point of view

As already said in introduction, the purpose is to build an affinity matrix that can integrate both the dimension of the problem as well as the density of points in the given n-th dimensional data set.

For the sake of efficiency, we shall investigate global parameters that can be used to derive the affinity matrix in the usual way, as a function of the distances between points in the data set. To this end, we first make the assumption that the n-th dimensional data set is isotropic enough, in the sense that there does not exist some privileged directions with very different magnitudes in the distances

between points along these directions. Let us denote by $S = \{x_i, 1 \leq i \leq m\}$ the data set of points, and by

$$D_{\max} = \max_{1 \leq i,j \leq m} \|x_i - x_j\| ,$$

the largest distance between all pairs of points in S. Under this first hypothesis, we can then state that the data set of points in essentially included in a n-th dimensional box with edge size bounded by $D_{\max}$.

If we expect to be able to identify some clusters within the set of points S, we must depart from the uniform distribution of m points in this enclosing n-th dimensional box. This uniform distribution is reached when dividing the box in m smaller boxes all of the same size, each with a volume of order $D_{\max}^n/m$, with a corresponding edge size that we shall denote as

$$\sigma = \frac{D_{\max}}{m^{\frac{1}{n}}} \tag{1}$$

What can be expected, indeed, is that if there exists some clusters, there must be at least some points that will be at a distance lower than a fraction of this edge size σ. Otherwise, the points should all be at a distance of order σ of each other, since we have made the assumption of isotropy and since all the points are included in the box of edge size $D_{\max}$.

Our proposal is thus to build the affinity matrix as a function of the ratio of the distances between points and the reference distance value $\frac{\sigma}{2}$. To incorporate the dimension n of the problem of clustering, we also propose to consider the control volumes around points instead of the square of the distances as commonly used. If we consider for instance the usual affinity matrix made of the exponential of these distances, we then propose to build the following matrix

$$A_{ij} = \{\exp\left(\frac{-\|x_i - x_j\|_2}{(\sigma/2)}\right)^n\} , \tag{2}$$

where $1 \leq i \leq m$ correspond to the row indexes and $1 \leq j \leq m$ to the column indexes in A, and to zero the diagonal in the usual way to get the affinity matrix to be used in the spectral embedding technique.

We first point out that this model relies upon the fact that the n-th dimensional box can be divided into smaller bricks in all directions. In other words, this means that the value $m^{\frac{1}{n}}$ is close to some integer and at least larger than 2. We shall come back later on this point, which will take some importance when the dimension n of the problem becomes large, in which case the above model might be weakened of slight modifications. This will be addressed in more details in the experiments.

Under the hypothesis that the n-dimensional data set is still isotropic enough, but when there exists some directions with varying amplitudes in the data, we can adapt slightly the computation of σ by considering that the set of points is included in a rectangular n-dimensional box. To approximate the volume of this

non square box, we compute the largest distances between all pairs of points along each direction to define the size of the edges :

$$\rho_k = \max_{1 \leq i \leq n} x_{ik} - \min_{1 \leq j \leq n} x_{jk}, k \in \{1, \ldots, m\}\,.$$

The vector ρ incorporates the sizes of the intervals in which each variable is included separately and, in this case, we shall consider that the enclosing rectangular box has the same aspect ratio as the one defined by the intervals lengths given in ρ, and with maximum edge size given by $D_{\max}$. Then, we can take

$$\sigma = \frac{D_{\max}\sqrt{n}}{\|\rho\|_2} \left(\frac{\prod_{i=1}^{n} \rho_i}{m} \right)^{\frac{1}{n}}\,, \tag{3}$$

which resumes to equation (1) when ρ is all constant and the box is square. A more general way, which is the basis of the Mahalanobis distance, would be to compute the spectral orientation of the dispersion of the data to fix the axes and compute the amplitudes along these axes. But this is more computationally demanding and assumes that the original data are linked together in some particular way. In this case, we can expect some preprocessing must be done to prepare the data appropriately.

4 Measures of Clustering

Ng and Weiss [3] suggest to make many tests with several values of σ and to select the ones with least distortion in the resulting spectral embedding. In some cases, the choice of σ is not very sensitive and good results can be obtained easily. Still, there exists many examples where this choice is rather tight, as for example in cases with geometrical figures plus background noise. In the following of this section, we introduce two *measures of quality* that can be used to identify the interval of appropriate values for the choice of σ.

4.1 Ratio of Frobenius Norms

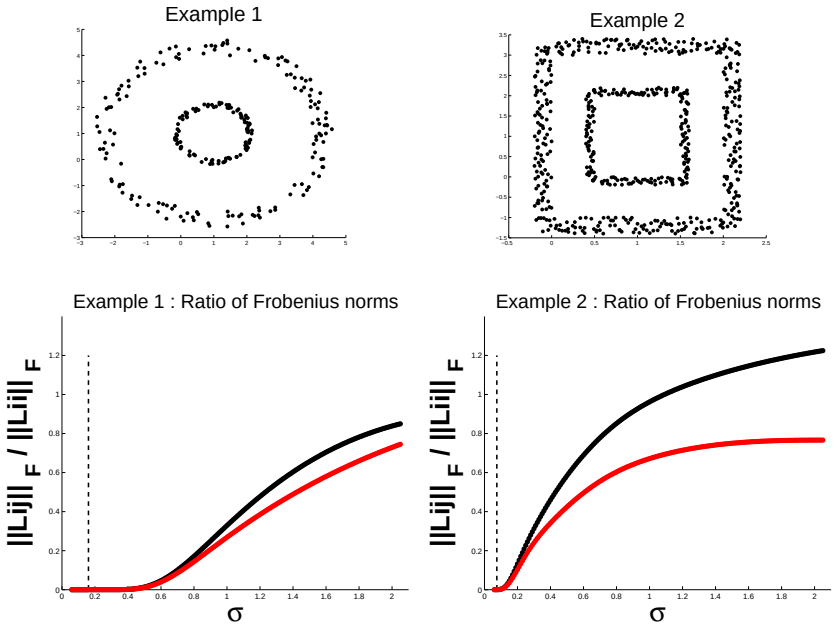
In general cases, the off-diagonal blocks in the normalized affinity matrix L are all non-zero and, for example, with $k = 3$, we can write :

$$\hat{L} = \begin{bmatrix} L^{(11)} & L^{(12)} & L^{(13)} \\ L^{(21)} & L^{(22)} & L^{(23)} \\ L^{(31)} & L^{(32)} & L^{(33)} \end{bmatrix}$$

We can then evaluate the ratios between the Frobenius norm of the off-diagonal-blocks and that of the diagonal ones.

$$r_{ij} = \frac{\|L^{(ij)}\|_F}{\|L^{(ii)}\|_F}$$

with $i \neq j$ and $i, j \in 1, .., k$ If the mean (or the max) of these values r_{ij} is close to 0, the affinity matrix has a near block diagonal structure. For example, in the following figure, we plot the value of these ratios in the case of two examples with two geometric clusters of points each.



From the behavior of these measures, we can see that there exists some interval in which the affinity matrix appears to be near block diagonal. This interval depends of course on the nature of the problem, and can be very different. For instance, in these examples, the length of this interval is of 0.4 in the first case and of 0.1 in the second one. The dash-dot line indicates the value of the heuristic (1) given in the previous section for the computation of σ. We can observe that this heuristic value falls in the corresponding intervals.

4.2 Confusion Matrix

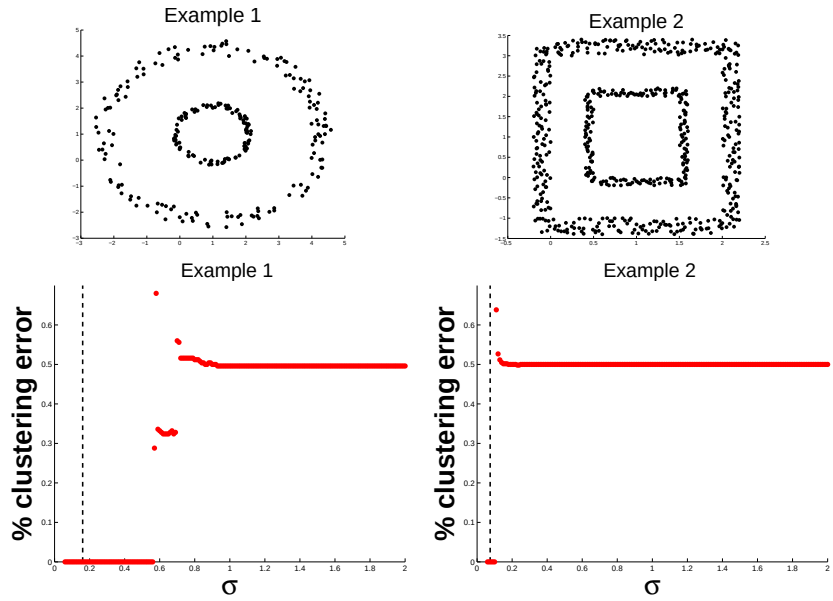
We now introduce an evaluation of the true error in clustering in the sense of the number of mis-assigned points within clusters. Let $C \in \mathcal{M}_{k,k}(\mathbb{R})$ be the so-called *confusion matrix* :

$$C = \begin{bmatrix} C^{(11)} & C^{(12)} & C^{(13)} \\ C^{(21)} & C^{(22)} & C^{(23)} \\ C^{(31)} & C^{(32)} & C^{(33)} \end{bmatrix}$$

(as for example in the case of three clusters) where $C^{(ij)}$ is the number of points that were assigned in cluster j instead of cluster i for $i \neq j$, and C^{ii} the number

of well-assigned points for each cluster i.
We define the *percentage of mis-clustered points* by :

$$p = \frac{\sum_{i \neq j}^{k} C^{(ij)}}{m}$$



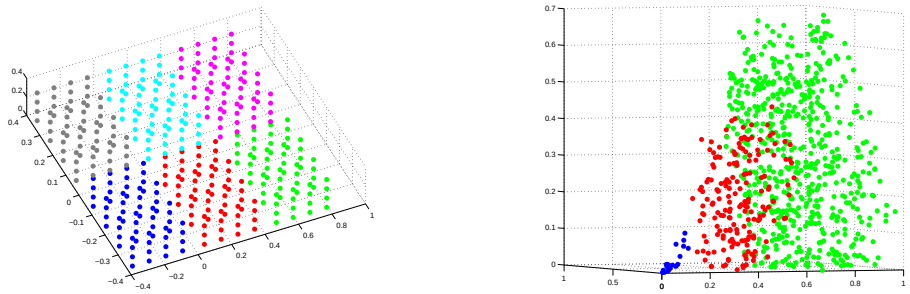
This matrix gives an estimate of the real error in the clustering method. The results from the previous two examples show that the interval value for the appropriate choice of parameter σ is approximatively the same as that observed with the ratios of Frobenius norms. We note that the clustering percentage of error varies almost instantaneously when σ just exits the appropriate interval. Again, we can observe that the value of the heuristic (1) corresponds to a value of σ with almost no clustering error.

5 Results

In order to validate the geometrical approach detailed in section 3, we consider two n-dimensional benchmark examples, one with six n-th dimensional uniform blocks slightly separated from each other, and another one made of pieces of n-spheres in $\mathbb{R}^n$ (see figure 1). The affinity matrix is defined by :

$$A_{ij} = \{\exp\left(\frac{-\|x_i - x_j\|_2}{\sigma/2}\right)^d\} \tag{4}$$

where d will be alternatively set to the different integer values from 1 to 5, in order to verify experimentally the adequacy of the power d (usually taken as



(a) n=3, m=384 (b) n=3,m=1200

Fig. 1. Example 1 & 2 : six blocks and three pieces of n-spheres

$d = 2$) with respect to the dimension of the problem n, as suggested in section 3. To obtain the results given in the following tables, we have tried consecutive values of σ from 0.01 to 0.15 and computed the two error measures discussed in the previous section. This enabled us to determine approximately an interval of feasibility for the values of σ. The purpose of that was to verify if the heuristics (1) or (3) would belong to the appropriate intervals or not.

5.1 First example : six blocks

This geometrical example is made of n-th dimensional blocks with uniform distribution each, slightly separated from each other, and is in perfect agreement with the assumption of isotropy used in the developments of section 3. Each block is composed of p^n points with a step size of 0.1 in each direction, and with $p = 4$ in the case of $n = \{2, 3\}$ and $p = 3$ in the case of $n = 4$. Finally, the blocks are separated from each other by a step size of 0.13. This example corresponds to the basic configuration that the heuristic (3) for σ should address well by default, and is therefore a fundamental case study.

In table (1), we indicate the results obtained for three different values of the dimension n of the problem, and we also indicate in each case the values σ_1 and σ_2 corresponding to the heuristics (1) and (3) respectively. For each of these dimensions, we vary the power d for the computation of the affinity matrix as indicated in (4), and we compute in all of these cases the intervals of feasibility for the values of σ with respect to the quality measures introduced in section 4. To determine these intervals in the case of ratio of Frobenius norms between blocks, the quality has been taken as acceptable when the mean of these ratios was inferior or equal to 0.15.

The results in table (1) show that the two heuristics σ_1 and σ_2 fall into the appropriate intervals in almost all cases. This is in agreement with the expectations in the sense that the affinity matrix is able to separate well the data. We also mention that the lengths of the interval, specially with the first quality

measure, are larger for a value of d close to n, which is partly in favor of the consideration of the volumes instead of squared distances when building the affinity matrix in usual way.

5.2 Second example : three pieces of n-spheres with 1200 points

This second example is built in the same spirit as the first one, except that each cluster has a different volume, and the spherical shape on some of the boundaries prevents k-means like techniques to separate well the clusters from scratch.

As in the previous example, table (2) shows the results for the spectral clustering with the two quality measures in function of both d and n. We can observe again that the heuristics (1) and (3) are within the validity interval for both measures, and that for increasing values of the dimension n, the affinity matrix is better determined with the clusters when the power d is closer to n.

5.3 Image segmentation :

We consider now an example of image segmentation. In this case, we investigate two different approaches to define the affinity matrix :

- *as a 3-dimension rectangular box :* since the image data can be considered as isotropic enough because the steps between pixels and brightness are about the same magnitude, we can try to identify the image data as a 3-dimensional rectangular set and incorporate the heuristic (3) for σ in the affinity matrix given by (2).
- *as a product of a brightness similarity term and a spatial one :* the second possibility is to consider that the image data are composed of two distinct sets of variables, each one with specific amplitude and density. Indeed, the spatial distribution of the pixels is isotropic but the brightness is scattered into levels (256 maximum) and the brightness density cannot be derived from the number of points. Therefore, what is usually considered in papers dealing with image segmentation (see for instance [4, 5]) is the product of an affinity matrix for the spatial data with an affinity matrix for the brightness values, each with its specific σ parameter reflecting the local densities. We then propose to build the affinity matrix in the following way :

$$A_{ij} = exp(-\frac{\|x_i - x_j\|^2}{(\sigma_G/2)^2} - \frac{|I(i) - I(j)|}{(\sigma_B/2)}), \tag{5}$$

where $I(i)$ is the brightness value in $\mathbb{R}$ and x_i the coordinates of pixel i in $\mathbb{R}^2$. The parameter σ_G is given by (1) applied only to the spatial data, and σ_B is fixed to $(I_{\max}/\ell)$ with ℓ a characteristic number of brightness levels. For instance, in the following example, ℓ is equal to the number of threshold in the picture and σ_B will define the length of the intervals under which brightness values should be grouped together.
We point out that this way of doing is still in the spirit of the developments in section 3, because σ given by (1) reflects a clustering reference distance in the

Table 1. 6 n-dimensional blocks

(a) $n = 2$ and $\sigma_1 = 0.1398$ and $\sigma_2 = 0.1328$

d	1	2	3	4	5
Ratio of Frobenius norm < 0.15	[0.02;0.3]	[0.02;0.56]	[0.02;0.6]	[0.02;0.62]	[0.02;0.6]
Clustering error	[0.12;1.6]	[0.06;1.24]	[0.08;1.06]	[0.1;1.18]	[0.12;1.1]

(b) $n = 3$ and $\sigma_1 = 0.1930$ and $\sigma_2 = 0.1510$

d	1	2	3	4	5
Ratio of Frobenius norm < 0.15	[0.02;0.3]	[0.02;0.58]	[0.02;0.64]	[0.02;0.66]	[0.02;0.66]
Clustering error	[0.02;3]	[0.06;2.2]	[0.04;1.4]	[0.04;1.2]	[0.04;1.2]

(c) $n = 4$, $\sigma_1 = 0.2234$ and $\sigma_2 = 0.1566$

d	1	2	3	4	5
Ratio of Frobenius norm < 0.15	[0.02;0.3]	[0.02;0.22]	[0.02;0.26]	[0.02;0.28]	[0.02;0.28]
Clustering error	[0.02;1.2]	[0.04;0.78]	[0.02;0.62]	[0.12;0.52]	[0.14;0.48]

Table 2. 3 pieces of n-sphere

(a) $n = 2$ and $\sigma_1 = 0.0288$ and $\sigma_2 = 0.0278$

d	1	2	3	4	5
Ratio of Frobenius norm < 0.15	[0.04;0.14]	[0.04;0.18]	[0.6;0.2]	[0.06;0.2]	[0.06;0.18]
Clustering error	[0.04;0.08]	[0.02;0.14]	[0.06;0.18]	[0.6;0.18]	[0.06;0.18]

(b) $n = 3$ and $\sigma_1 = 0.1074$ and $\sigma_2 = 0.1044$

d	1	2	3	4	5
Ratio of Frobenius norm < 0.15	[0.02;0.12]	[0.02;0.3]	[0.04;0.5]	[0.06;0.5]	[0.08;0.5]
Clustering error	[0.02;0.12]	[0.04;0.16]	[0.04;0.16]	[0.08;0.18]	[0.08;0.18]

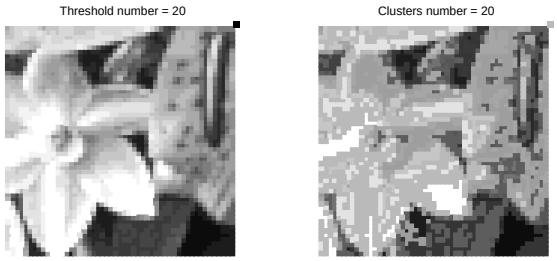
(c) $n = 4$, $\sigma_h = 0.1704$ and $\sigma_2 = 0.1658$

d	1	2	3	4	5
Ratio of Frobenius norm < 0.15	[0.04;0.04]	[0.02;0.04]	[0.04;0.08]	[0.06;0.1]	[0.1;0.16]
Clustering error	[0.02;0.02]	[0.04;0.1]	[0.06;0.16]	[0.08;0.17]	[0.1;0.2]

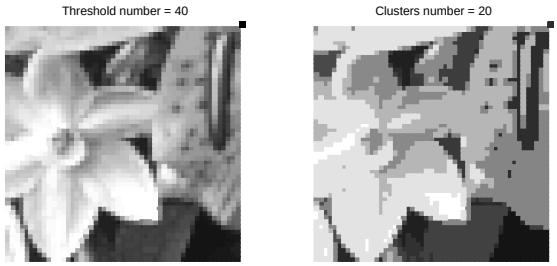
case of locally isotropic and scattered enough distribution of points. With 256 maximum brightness levels, the distribution cannot be considered anymore as locally scattered (lots of values are even equal to each other) and one must give a priori the characteristic distance under which brightness values can be clustered. We note also that the solution of taking $\sigma_B = I_{\max}/256$ would result in grouping the brightness values into clusters of length one approximately, and the segmentation of the image will require the analysis of a lot of clusters made of pixels close to each other and with about the same brightness level, equivalent to a very fine grain decomposition of the image.

In the following results, we test the approaches (2) and (5) for the computation of the affinity matrix on a 50×50 pixels picture. On the left, we show the original thresholded image and, on the right, the results obtained with either (2) in figure 2 or with (5) in figure 3.

In both cases, the results are visually acceptable. The 3-dimensional approach seems to provide nicer results than the 2D by 1D product, but we need more investigations to ensure which one of these two approaches is the best in general and to refine the results.

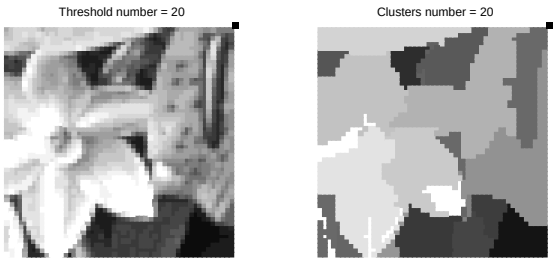


(a) 20 bright levels and 20 clusters

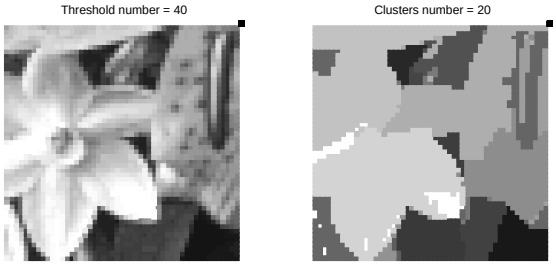


(b) 40 bright levels and 20 clusters

Fig. 2. Test of the 3-dimension rectangular affinity box on a flower



(a) 20 bright levels and 20 clusters



(b) 40 bright levels and 20 clusters

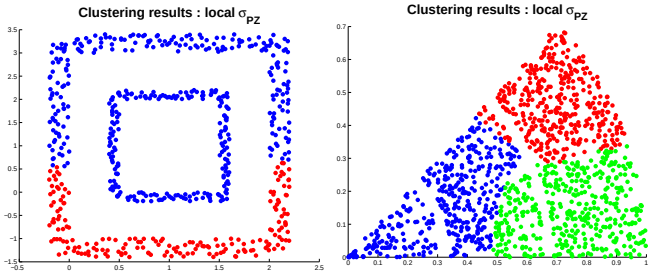
Fig. 3. Test of the product 2D by 1D affinity boxes on a flower

6 Remarks and Conclusions

The problematic of choosing an adequate parameter in order to improve the results has also been treated by some authors. Different points of view could be adopted.

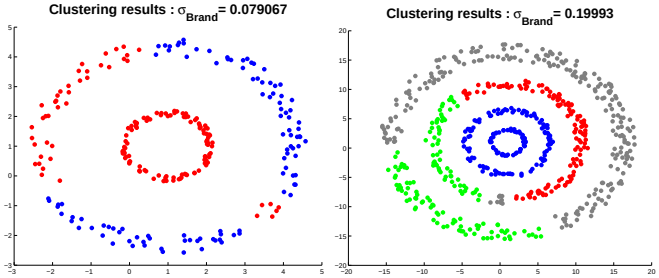
- *Perona and Zelnik-Manor* [6] propose a locally approach. They assign a different scaling parameter σ_i to each point x_i in the data set. σ_i is equal to the distance between x_i and its P-th neighbors. This method gives great results in some kind of problems where the effect of local analyze provides enough information to create the clusters : for example, recovering a tight cluster within background noise. But computing a value of σ for each point x_i can be costly and the value P must be fixed empirically (P=7).
- *Brand and Huang* [2] define a global scale parameter : the mean between each data point and its first neighbor. In many examples, we obtain well clustered data representations.

In the examples introduced in figure 5, the density of points varies within each cluster. These results illustrate the fact that without global density information,



(a) $m = 600$ (b) $m = 1200$

Fig. 4. Examples with σ proposed by *Perona*



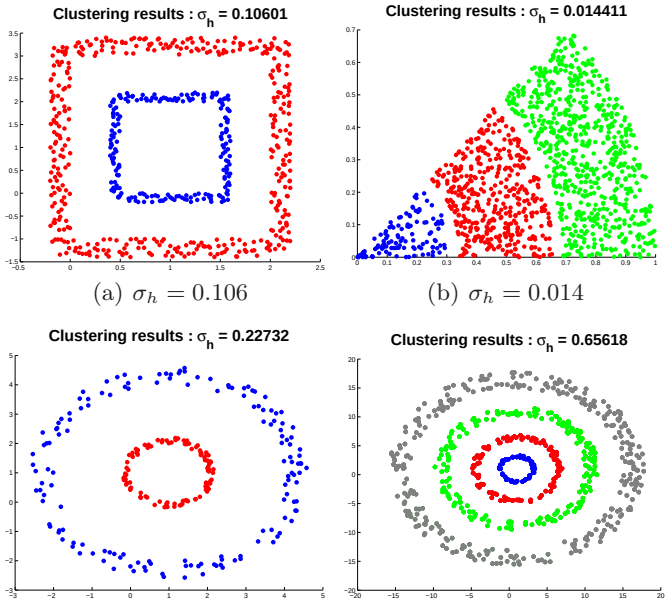
(a) $m = 250$ (b) $m = 650$

Fig. 5. Examples with σ proposed by *Brand*

it can be difficult to cluster well the data points in some cases. We test our definition (1) of global parameter for these examples in figure 6:

As recalled above, this global parameter gives good results in these four cases. We mention however that this heuristic parameter gives information about the spatial repartition of the data in a box of dimension $D_{\max}$. So when we have cases with an important noise density, the noise is difficult to separate from the existing clusters and can be assigned to its closest cluster. Only a local parameter can help to identify the noise from the cluster.

In conclusion, we have proposed a parameter for the construction of the affinity matrix used within spectral clustering techniques. This approach is adapted to n-dimensional cases, and based on a geometric point of view. With an isotropic assumption, this parameter represents the threshold of affinity between points within the same cluster. With quality measures such as ratio of Frobenius norms and confusion matrix, the rule of σ is observed and our definition is validated on a few n-dimensional geometrical examples. We have also tried a case of image segmentation, but we still need deeper investigations and larger sets of test



(a) $\sigma_h = 0.106$ (b) $\sigma_h = 0.014$

(c) $\sigma_h = 0.227$ (d) $\sigma_h = 0.656$

Fig. 6. Examples with the heuristic σ for $n = 2$

examples to ensure the validity as well as to determine the limitations of this approach. We plan also to test this general approach in a case of biologic topic.

References

1. M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems*, 14(3), 2002.
2. M. Brand and K. Huang. A unifying theorem for spectral embedding and clustering. *9th International Conference on Artificial Intelligence and Statistics*, 2002.
3. A. Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: analysis and an algorithm. *Proc.Adv.Neural Info.Processing Systems*, 2002.
4. Freeman W.T. Perona, P. A factorization approach to grouping. *European Conference on Computer Vision*, 1998.
5. Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
6. Perona P. Zelnik-Manor, L. Self-tuning spectral clustering. *NIPS*, 2004.