HAL open science

Quantum clustering analysis: Minima of the potential energy function

Aude Maignan, Tony Scott

To cite this version:

Aude Maignan, Tony Scott. Quantum clustering analysis: Minima of the potential energy function. International Conference on Machine Learning Techniques and Data Science (MLDS), Dec 2020, Sydney, Australia. hal-03169118

HAL Id: hal-03169118

https://hal.science/hal-03169118v1

Submitted on 15 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Quantum clustering analysis: Minima of the potential energy function

Aude Maignan[1] and Tony Scott[2]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.
aude.maignan@univ-grenoble-alpes.fr

[2] Institut für Physikalische Chemie, RWTH-Aachen University, 52056 Aachen, Germany

scotusts@yahoo.com

ABSTRACT

Quantum clustering (QC), is a data clustering algorithm based on quantum mechanics which is accomplished by substituting each point in a given dataset with a Gaussian. The width of the Gaussian is a σ value, a hyper-parameter which can be manually defined and manipulated to suit the application. Numerical methods are used to find all the minima of the quantum potential as they correspond to cluster centers. Herein, we investigate the mathematical task of expressing and finding *all* the roots of the exponential polynomial corresponding to the minima of a two-dimensional quantum potential. This is an outstanding task because normally such expressions are impossible to solve analytically. However, we prove that if the points are all included in a square region of size σ, there is only one minimum. This bound is not only useful in the number of solutions to look for, by numerical means, it allows to to propose a new numerical approach "per block". This technique decreases the number of particles (or samples) by approximating some groups of particles to weighted particles. These findings are not only useful to the quantum clustering problem but also for the exponential polynomials encountered in quantum chemistry, Solid-state Physics and other applications.

KEYWORDS

Data clustering, Quantum clustering, energy function, exponential polynomial, optimization.

1. INTRODUCTION

The primary motivation for this work stems from an important component of the area of information retrieval of the IT industry, namely *data clustering*. For any data of a scientific nature such as Particle Physics, pharmaceutical data, or data related to the internet, security or wireless communications, there is a growing need for data analysis and predictive analytics. Researchers regularly encounter limitations due to large datasets in complex simulations, in particular, biological and environmental research. One of the biggest problems of data analysis is data with no known *a priori* structure, the case of "unsupervised data" in the jargon of machine learning. This is especially germane to object or name disambiguation also called the "John Smith" problem [1]. Therefore *data clustering*, which seeks to find internal classes or structures within the data, is one of most difficult yet needed implementations.

It has been shown that the quantum clustering method (QC) [2,3] can naturally cluster data originating from a number of sources whether they be: scientific (natural), engineering and even text. In particular, it is more stable and is often more accurate than the standard data clustering method known as K-means [3]. This method requires isolating the minima of a quantum potential and is equivalent to finding the roots of its gradients i.e. an expression made of exponential polynomials. Finding all the clusters within the data means finding *all* the potential minima. The quantum clustering method can be viewed as "dual" or inverse operation of the machine learning process known as a nonlinear support vector machines when using Gaussian functions are used as its kernel function; this machine learning approach being the very inspiration of the quantum clustering method [4].

This is not the only problem in quantum mechanics requiring such solutions. The nodal lines of any given wave function characterize it with respect to internal symmetries and level of excitation. In general, if one arranges the eigenstates in the order of increasing energies, e.g. ϵ_1 ,ϵ_2 , ϵ_3, …the eigenfunctions likewise fall in the order of increasing number of nodes; the nth eigenfunction has $n-1$ nodes, between each of which the following eigenfunctions have at least one node [5]. In diffusion Monte-Carlo calculations for Molecules, a precise determination of the nodal structure of wave function yields greater accuracy for the energy eigenvalues [6,7,8]. Furthermore, solutions in terms of Gaussian functions involve the most developed mathematical "technology" of quantum chemistry (e.g. The Gaussian program [9]). This is not surprising for the following reasons:

1. In principle, we can get *all* the roots of polynomial systems. However, quantum mechanical systems need exponentials in order to ensure a square-integrable wave function over all space. About an atom, the angular components over a range $(0,2\pi)$ can be modeled in terms of polynomials of trigonometric quantities such as e.g. Legendre polynomials. However, the radial part extends over all space requiring exponential apodization.

2. Thanks to properties such as the Gaussian product theorem, Gaussian functions allow for exact analytical solutions of the molecular integrals of quantum chemistry [10,11,12].

3. In general, for small atoms and molecules, the nodal lines can be modeled as nodes of polynomial exponentials [13,14,15].

More recently, in the area of low temperature Physics (including superconductors), clustering within machine learning has been used in finding phases and separating the data into particular topological sectors [16,17,18]. High accuracy of the clustering is crucial in order to precisely identify transition points in terms of e.g. temperature or pressure.

To reiterate, any insight concerning the isolation of *all* the roots or nodal lines of polynomial exponentials is useful for quantum clustering and computational quantum chemistry and condensed matter Physics and data analysis. This has applications in all cases for any given function covering all space in principle but whose extrema and/or roots are in a finite local region of space.

1.1.Statement of the Problem

Consider a set of particles $(X_i)_{i=1..N}$, the quantum clustering is a process that detects the clusters of the distributed set $(X_i)_{i=1..N}$ by finding the cluster centers. Those centers are the minima of the potential energy function defined by [2,3]:

$$\frac{1}{2\sigma^2}\frac{1}{\sum_{i=1}^{N} e^{-\frac{(X-X_i)^2}{2\sigma^2}}}\sum_{i=1}^{N}(X-X_i)^2 e^{-\frac{(X-X_i)^2}{2\sigma^2}} \qquad (1)$$

such that $X \in R^2$. This function results from injecting a Parzen window into the Schrödinger wave equation [2,3] and balancing the resulting energy. Other methods based on energy variation may also be instructive [19]. The minima of this potential provides the cluster centers for a given standard deviation σ. As stated before, we limit ourselves to two dimensions. This method is more stable and precise than the standard K-means method [3].

Moreover, and in contradistinction to other data clustering methods, the determination of the parameter σ gives a number of extrema. The number of minima is not determined beforehand but obtained numerically.

One main difficulty is to determine the minima of the potential energy. Nowadays, the technique used to approach the minima is through the gradient descent or the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithms [3]. Some investigations have been made to improve the detection of clusters via the potential energy function. For instance, in 2018, Decheng *et al.* [20] improved the quantum clustering analysis by developing a new weighted distance once a minimum had been found. Improvements are needed to capture all the minima efficiently.

The present work consists, Subection 2.1, in simplifying the derivatives of the potential energy function such that the minima can be determined by some solution of a system of equations. Finding the extrema (minima, maxima and saddle points) of the function (1) is equivalent to solving a system

$$\begin{cases} M(x,y) = 0 \\ L(x,y) = 0 \end{cases} \qquad (2)$$

where $M(x,y)$ and $L(x,y)$ are bivariate exponential functions which can be expressed as polynomial in x, e^x, y and e^y. In this scenario, the degrees of M and L in x (respectively in y) are one. In Subsection 2.2, the implicit functions of $M = 0$ and $L = 0$ are investigated and the on going Crab example is presented Subsection 2.3. Section 3, A new block approach is presented. The aim of this new method is to reduce memory and computation costs. The main formal result is given Subsection 3.1. We prove that the function (1) has only one minimum if the set of particles $(X_i)_{i=1..N}$ are all included in a square of side σ. Then, we propose a method based on this result and a block approach to capture all the minima in a more efficient way. The presentation of benchmarks closed Section 3. Finally, we conclude Section 4.

2. PROBLEM REDUCTION AND FIRST ANALYSIS

In this section, we transform the minimization problem of the potential energy function (1) to the resolution of a system of two equations in two variables and $2N$ parameters, namely the particles coordinates $(X_i)_{i=1..N}$.

2.1. Problem reduction

It is known that the value of σ has a crucial role on the number of minima: the greater the value of σ, the smaller the number of minima. To simplify the potential energy function, we denote $Y = \frac{X}{\sqrt{2}\sigma}$. This variable change remove σ from the function. Discussion of σ will be presented at the end of this section.

We get

$$\frac{1}{2\sigma^2}\frac{1}{\sum_{i=1}^{N} e^{-\frac{(X-X_i)^2}{2\sigma^2}}}\sum_{i=1}^{N}(X-X_i)^2 e^{-\frac{(X-X_i)^2}{2\sigma^2}} = \frac{1}{\sum_{i=1}^{N} e^{-(Y-Y_i)^2}}\sum_{i=1}^{N}(Y-Y_i)^2 e^{-(Y-Y_i)^2} \qquad (3)$$

where for all i, $Y_i = \frac{X_i}{\sqrt{2}\sigma}$. We denote this equation $h(Y)$.

Theorem 1. *The extrema $Y = (x, y)$ of function*

$$h(x,y) = \frac{1}{\sum_{i=1}^{N} e^{-(Y-Y_i)^2}} \sum_{i=1}^{N} (Y - Y_i)^2 e^{-(Y-Y_i)^2}$$

satisfy the system of the following two bivariate functions: $\begin{cases} M(x,y) = 0 \\ L(x,y) = 0 \end{cases}$ *with $Y_i = (x_i, y_i)$ for all $i = 1..N$ and*

$$M(x,y) = \sum_{i=1}^{N} e^{-2x_i^2-2y_i^2}\, e^{4x_i x+4y_i y}(x - x_i) + \sum_{i=1}^{N}\sum_{j>i}^{N} e^{-x_i^2-y_i^2}\, e^{-x_j^2-y_j^2}\, e^{2(x_i+x_j)x+2(y_i+y_j)y} \quad (4)$$

and

$$L(x,y) = \sum_{i=1}^{N} e^{-2x_i^2-2y_i^2}\, e^{4x_i x+4y_i y}(y - y_i) + \sum_{i=1}^{N}\sum_{j>i}^{N} e^{-x_i^2-y_i^2}\, e^{-x_j^2-y_j^2}\, e^{2(x_i+x_j)x+2(y_i+y_j)y} \quad (5)$$

Remark: We will also use the shortest expression:

$$M(x,y) = \sum_{i=1}^{N} (x - x_i)K_i^2 + \sum_{i<j} c_{ij}\, K_i K_j \qquad (6)$$

and

$$L(x,y) = \sum_{i=1}^{N} (y - y_i)K_i^2 + \sum_{i<j} d_{ij}\, K_i K_j \qquad (7)$$

with for all i, $K_i = e^{-x_i^2-y_i^2} e^{2(x_i+x_j)x}$, and for all i, j, $i < j$,

$$c_{ij} = (2x - x_i - x_j)(1 - (x_i - x_j)^2) - (x_i - x_j)(y_i - y_j)(2y - y_i - y_j) \qquad (8)$$

and

$$d_{ij} = (2y - y_i - y_j)(1 - (y_i - y_j)^2) - (y_i - y_j)(x_i - x_j)(2x - x_i - x_j) \qquad (9)$$

Proof. $h(Y)$ is a fraction of two exponential polynomials, namely $h(Y) = \frac{f(Y)}{g(Y)}$ with $g(Y) = \sum_{i=1}^{N} e^{-(Y-Y_i)^2}$ and $f(Y) = \sum_{i=1}^{N} (Y - Y_i)^2 e^{-(Y-Y_i)^2}$.

Since $Y \in R^2$, Y is denoted $Y = (x, y)$, then f and g can also be written as

$$f(x,y) = \sum_{i=1}^{N} ((x - x_i)^2 + (y - y_i)^2) e^{-(x-x_i)^2-(y-y_i)^2} \qquad (10)$$

and

$$g(x,y) = \sum_{i=1}^{N} e^{-(x-x_i)^2-(y-y_i)^2} \qquad (11)$$

by denoting $Y_i = (x_i, y_i)$. The extrema of $h(x, y)$ satisfy the system $\begin{cases} \frac{\partial h(x,y)}{\partial x} = 0 \\ \frac{\partial h(x,y)}{\partial y} = 0 \end{cases}$ which is equivalent to:

$$\begin{cases} \frac{\partial f(x,y)}{\partial x} g(x,y) - \frac{\partial g(x,y)}{\partial x} f(x,y) = 0 \\ \frac{\partial f(x,y)}{\partial y} g(x,y) - \frac{\partial g(x,y)}{\partial y} f(x,y) = 0 \end{cases} \tag{12}$$

since $g(x, y) \neq 0$ everywhere.

The formal computation of the equations of the last system gives expressions which can be divided by $2e^{-x^2-y^2}$. We finally obtain Theorem 1. □

2.2. Cylindrical decomposition

For a given set of particles $(Y_i)_{i=1..N} = (x_i, y_i)_{i=1..N}$, the solutions of System (1) correspond to the intersection between the implicit functions of $M(x, y) = 0$ and those of $L(x, y) = 0$ (see Figure 1 for the example of crab with $N = 200$). An analysis on branches which will be detailed in a further work give the following result: Let us denote $ymax$ (resp. $xmax$) the index the greatest element of $(y_i)_{i=1..N}$ (resp. $(x_i)_{i=1..N}$) such that $\forall i \in \{1, \dots, N\} - \{ymax\}$ $y_{ymax} > y_i$. In the same way, we denote $ymin$ (resp. $xmin$) the index the smallest element of $(y_i)_{i=1..N}$ (resp. $(x_i)_{i=1..N}$) such that $\forall i \in \{1, \dots, N\} - \{ymin\}$ $y_{ymin} < y_i$.

- The infinite branches of the implicit functions of $M(x, y)$ tend to x_{ymin} at $-\infty$ and x_{ymax} at $+\infty$
- The infinite branches of the implicit functions of $L(x, y)$ tend to y_{ymin} at $-\infty$ and y_{ymax} at $+\infty$

2.3. Crab example

To illustrate our results, we use the crab data clustering example [3] using the dataset from Refs. [21,22]. This two dimensional case has been presented in Refs. [2,3]. This example is composed of four classes at 50 samples each, making a total of 200 samples i.e. particles and by taking $\sigma = 0.05$, we obtain, after the variable changes described in Section 2, a set of particles for which the x and y coordinates $(x_i)_{i=1..200}$ and $(y_i)_{i=1..200}$ satisfy $xmin = 150$, $xmax = 65$, $y_{xmax} = -0.3190$, $y_{xmin} = 0.3640$, $ymin = 35$, $ymax = 105$, $x_{ymax} = 0.0038$, $x_{ymin} = -0.7941$.

The curve $M(x, y) = 0$ is shown in red and the curve $L(x, y) = 0$ is shown in green. The intersection between the red and the green curves corresponds to the extrema of h.

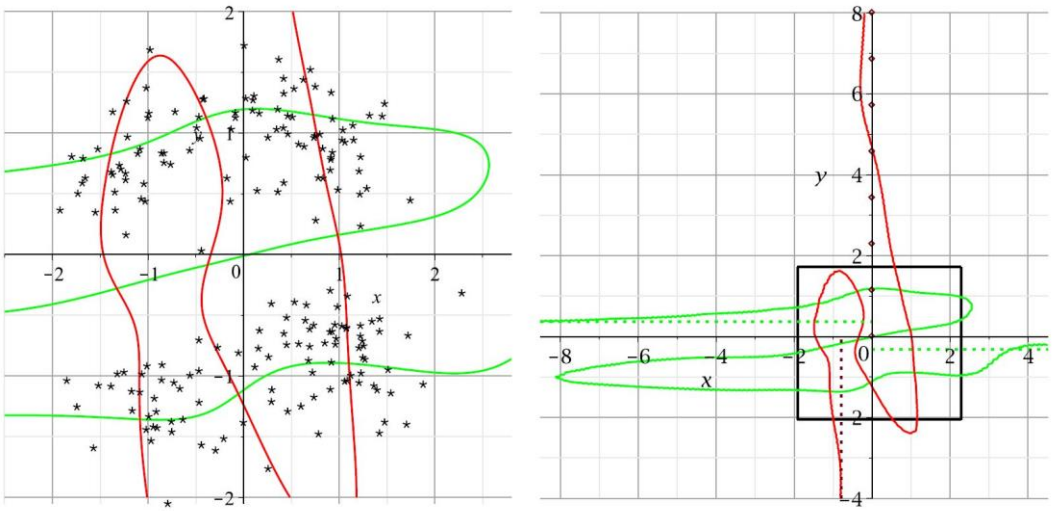


Figure 1: Crab example with $\sigma = 0.05$: (a) the set of points $(x, y)_i$, and the implicit curves of $M(x, y) = 0$ and $L(x, y) = 0$. (b) the limits of implicit curves.

Using the Maple computer algebra system [23], we obtain one maximum, four minima and four saddle points. Table 1 gives the approximation of the solutions in the Y base and in the X base which is the initial base.

solution	Y variables	$X = Y \times \sigma\sqrt{2}$ variables
minima	$(-1.390402, 0.8278737)$	$(-0.09831631, 0.05853951)$
	$(0.7257303, 1.150738)$	$(0.05131688, 0.08136950)$
	$(-1.084091, -1.357692)$	$(-0.07665683, -0.09600335)$
	$(1.099672, -0.8980522)$	$(0.07775858, -0.06350188)$
saddle points	$(-.3931250, 1.127632)$	$(-0.02779813, 0.07973564)$
	$(-0.05590183, -1.149097)$	$(-0.003952856, -0.08125344)$
	$(0.9828291, 0.1599075)$	$(0.069496, 0.01130716)$
	$(-1.326287, -0.2802702)$	$(-0.09378271, -0.01981809)$
maximum	$(-.3766093, -0.08563777$	$(-0.02663030, -0.006055504)$

Table 1: Extrema of the potential energy function (1) for $\sigma = 0.05$

Numerically, this variable change gives the advantages of a normalization of the values.

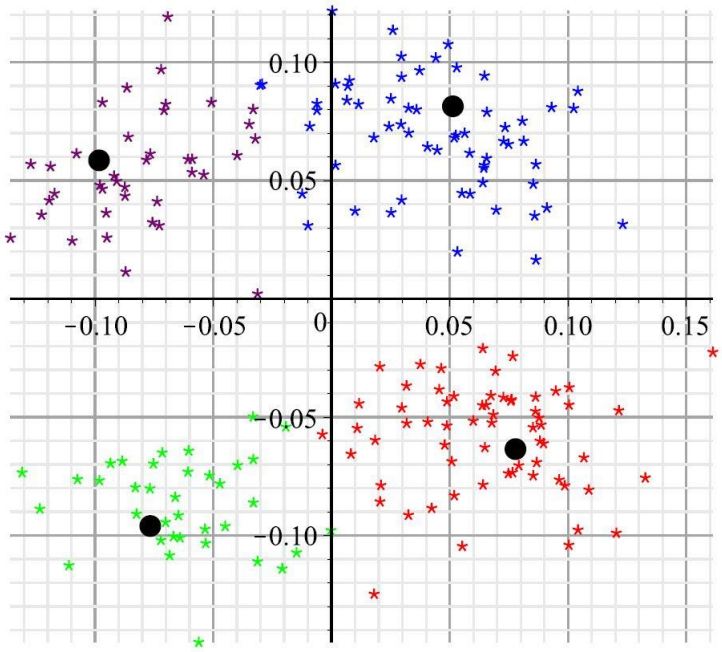


Figure 2: Crab clusters produced by using the minimum Euclidean distance from the minima ($\sigma = 0.05$)

The clusters produced by using the minimum Euclidean distance from the minima are shown Figure 2. For larger σ, the number of solutions decreases and hence, a coarser clustering is found. In Figure 3, two different values of σ are given. For $\sigma = 0.075$, there are two minima, whereas for $\sigma = 0.1$, only one solution exists which corresponds to a minimum. Table 2 gives the values of the corresponding minima.

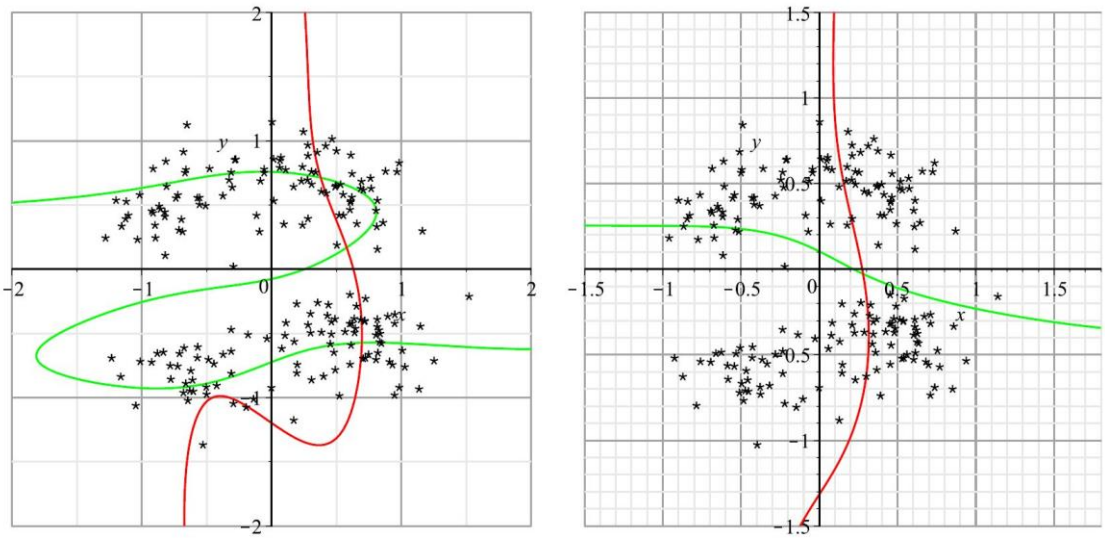


Figure 3: Crab example: t set of points $(x, y)_i$. and the implicit curves. (a) For $\sigma = 0.075$ (b) For $\sigma = 0.1$

σ	solution	Y variables	$X = Y \times \sigma\sqrt{2}$ variables
0.075	minima	$(0.3780743, 0.7125292)$ $(0.6945461, -0.5733816)$	$(0.04010084, 0.07557513)$ $(0.07366774, -0.06081630)$
0.1	minima	$(0.2794885, -0.02745293)$	$(0.03952565, -0.003882430)$

Table 2: minima of the potential energy function (1) with respect to σ

σ range	$0.085 \leq$	$[0.074, 0.084]$	$[0.071, 0.073]$	$[0.025, 0.070]$	$0.02 \geq$
clusters number	1	2	3	4	≥ 5

Table 3: range of σ with respect to the number of clusters

A deeper analysis is provided by Table 3. It gives for some σ ranges the resulting number of clusters. It shows that the non trivial number of clusters is more likely 4 because the corresponding σ range is the widest.

This first example of 200 samples can be fully solved numerically but the corresponding function $M(x, y)$ and $L(x, y)$ are sums of 20100 monomials in x, y, e^x and e^y. The size of M and L is an issue and the aim of the following section is to reduce the size of M and L while maintaining a good approximation of minima.

3. THE BLOCK APPROACH

In this section, we present a new numerical approach per block. First, we present the algebraic property needed to develop the new algorithm presented theoretically in the second subsection and algorithmically in the third subsection. Finally the Crab example is revisited and some other benchmarks are presented.

3.1. σ estimations

We have seen (Table 3) that the σ value is of crucial importance to the number of minima. The greater σ is, the smaller the number of minima. But obviously the number of minima also depends on the data. In this subsection, we link the value of σ with the values of the initial data in order to obtain a bound from which the number of minima is one.

Theorem 2. *Consider a set of particles $(X_i)_{i=1..N}$ where for all $i = 1..N$, $X_i = (v_i, w_i)$, the potential energy function $\frac{1}{2\sigma^2}\frac{1}{\sum_{i=1}^{N} e^{-\frac{(X-X_i)^2}{2\sigma^2}}}\sum_{i=1}^{N}(X - X_i)^2 e^{-\frac{(X-X_i)^2}{2\sigma^2}}$ has only one minimum for $\sigma = max(v_{max} - v_{min}, w_{max} - w_{min})$.*

To complete this proof, we use the variable changes proposed in Section 2 and we prove the equivalent property: System (2) $\begin{cases} M(x, y) = 0 \\ L(x, y) = 0 \end{cases}$ has only one solution if the set of points $(x_i, y_i)_{i=1..N}$ lies in a square of side $\frac{1}{\sqrt{2}}$. The proof is technical and the general idea is as follows: We first normalized and centralized System (2) into $\begin{cases} M_C(\alpha, \beta) = 0 \\ L_C(\alpha, \beta) = 0 \end{cases}$. Secondly, we prove that this last system has at most one minimum. Then we prove that at least one implicit curve of $M_c = 0$ (resp. $L_c = 0$) lies in the normalized square. Finally we conclude to the unicity of the minimum.

For instance, in the crab example, $max(v_{max} - v_{min}, w_{max} - w_{min}) = 0.297$ and without any computation, we know that if $\sigma \geq 0.297$, the function (1) has exactly one minimum.

To serve our new block method presented next subsection, we give another formulation of Theorem 2 as a corollary.

Corollary 1. *The bivariate function* $\frac{1}{2\sigma^2}\frac{1}{\sum_{i=1}^{N} e^{-\frac{(X-X_i)^2}{2\sigma^2}}}\sum_{i=1}^{N}(X - X_i)^2 e^{-\frac{(X-X_i)^2}{2\sigma^2}}$ *has only one minimum if the set of points* $(X_i)_{i=1..N}$ *are all included in a square of side* σ.

3.2. System approximation construction

In the general case of N particles, the functions $M(x, y)$ and $L(x, y)$ are sums of $\frac{N(N+1)}{2}$ exponential polynomials of the form $(x - x_i)K_i^2$, $c_{ij}K_iK_j$ or $d_{ij}K_iK_j$. We recall System (2):

$$\begin{cases} M(x, y) = 0 \\ L(x, y) = 0 \end{cases}$$

such that

$$M(x, y) = \sum_{i=1}^{N}(x - x_i)K_i^2 + \sum_{i<j} c_{ij}\, K_iK_j \quad (13)$$

and

$$L(x, y) = \sum_{i=1}^{N}(y - y_i)K_i^2 + \sum_{i<j} d_{ij}\, K_iK_j \quad (14)$$

where $K_i = e^{-(x-x_i)^2-(y-y_i)^2+x^2+y^2}$.

When N is large, we need a strategy to decrease the length of $M(x, y)$ and $L(x, y)$ while maintaining the main property of System (2) which is to define the cluster centers.

Let us denote $R = [x_{min}, x_{max}] \times [y_{min}, y_{max}]$ the rectangle containing all the points $(Y_i)_{i=1..N}$. The basic idea is to partition R into squares and approximate the minimum locally by considering for each square, only its particles. These new points will correspond to a weighted approximation of the particles in the square. They will therefore correspond to the weighted particles of the approximate system.

The block construction consists of subdividing R into k^2 square blocks of length

$$\frac{1}{k} max(x_{max} - x_{min}, y_{max} - y_{min}) \quad (15)$$

Since the particles are numbered from 1 to N, we denote $B(i)$ the block containing the particle i. i is named a representative of the block and we have: $B(i) = B(j)$ if i and j belong to the same square. We denote R a set containing exactly one representative of each non empty block.

Let $\alpha \in R$, the function M is reduced to the particles of the block $B(\alpha)$ which is denoted $M_{B(\alpha)}$ and

$$M_{B(\alpha)}(x, y) = \sum_{i \in B(\alpha)}(x - x_i)K_i^2 + \sum_{i<j, i \in B(k), j \in B(\alpha)} c_{ij}\, K_iK_j \quad (16)$$

Similarly,

$$L_{B(\alpha)}(x,y) = \sum_{i\in B(\alpha)} (y - y_i)K_i^2 + \sum_{i<j, i\in B(k), j\in B(\alpha)} d_{ij}\, K_i K_j \qquad (17)$$

By setting $\sigma = \frac{1}{k} max(x_{max} - x_{min}, y_{max} - y_{min})$, Theorem 2 guarantees that the system governed by

$$\begin{cases} M_{B(\alpha)}(x,y) = 0 \\ L_{B(\alpha)}(x,y) = 0 \end{cases} \qquad (18)$$

has exactly one minimum $(x_{B(\alpha)}, y_{B(\alpha)})$.

Therefore, $M(x,y) = \sum_{\alpha\in R} M_{B(\alpha)} + \sum_{i<j, j\notin B(i)} c_{ij}\, K_i K_j$ and we approximate $M(x,y)$ by

$$M_{Bls}(x,y) = \sum_{\alpha} p_{B(\alpha)} \left(x - x_{B(\alpha)}\right) K_{B(\alpha)}^2 + \sum_{k\in R, l\in R, \alpha<\beta} p_{B(\alpha)}\, p_{B(\beta)} c_{B(\alpha)B(\beta)} K_{B(\alpha)} K_{B(\beta)} \qquad (19)$$

where $p_{B(\alpha)}$ corresponds to the number of particles inside $B(\alpha)$. Equivalently, we approximate $L(x,y)$ by $L_{Bls}(x,y)$ to obtain the block system

$$\begin{cases} M_{Bls} = 0 \\ L_{Bls} = 0 \end{cases} \qquad (20)$$

M_{Bls} and L_{Bls} are now sums of at most $\frac{k^2(k^2+1)}{2}$ exponential polynomials and $k^2 << N$.

Remark (Limit preservation): the minima of System (2) are usually in the domain R. Nevertheless, the limit preservation of the approximate system is important. To do so, and according to Section 2, the four extrema (x_{minx}, y_{minx}), (x_{miny}, y_{miny}), (x_{maxx}, y_{maxx}) and (x_{maxy}, y_{maxy}) are usually not integrated into blocks and appear without any modification in System (20).

3.3. Algorithm

The main steps of the algorithm are as follows:

- Input: the list of particles $L = ((x_i, y_i))_{i=1..N}$ and k
- Compute $\sigma = \frac{1}{k} max(x_{max} - x_{min}, y_{max} - y_{min})$
- For all $(i,j) \in \{1..k\}^2$
  - Compute $B = [x_{min} + i\sigma, x_{min} + (i+1)\sigma] \times [y_{min} + j\sigma, y_{min} + (j+1)\sigma]$,
  - Find the list L_B of all the particles belonging to S,
  - If $L_B \neq \emptyset$ compute the minimum m_B of the block-system $\begin{cases} M_B = 0 \\ L_B = 0 \end{cases}$ involving only the particles of L_B,
  - The weight p_B of this minimum corresponds to the number of particles inside the square. $p_B = card(L_B)$.

- Consider the list L_m of all the minima with their corresponding weight. Compute the minima of the corresponding block system $\begin{cases} M_{Bls} = 0 \\ L_{Bls} = 0 \end{cases}$ involving L_m.

Remark: With regards to the third item: we have proved, thanks to Corollary 1, that $\begin{cases} M_B = 0 \\ L_B = 0 \end{cases}$ has exactly one minimum m_B. Indeed, the size of the block B is σ and the construction of the function M_B and L_B involves only the particles in the block B. This minimum is often close to the mass center of the cluster. Finding this minimum using a Newton-Raphson method with the mass center as a starting point has fast convergence. Moreover, one can consider a variation of our approach where σ depends on an additional parameter $l \geq 1$: $\sigma = \frac{l}{k} max(x_{max} - x_{min}, y_{max} - y_{min})$. In this variation, Theorem 2 holds since $l \geq 1$ and σ and k can be chosen independently such that $\frac{\sigma}{kmax(x_{max}-x_{min}, y_{max}-y_{min})} \geq 1$. Therefore we can consider an approximation involving more blocks without changing σ.

3.4. Crab Example Revisited

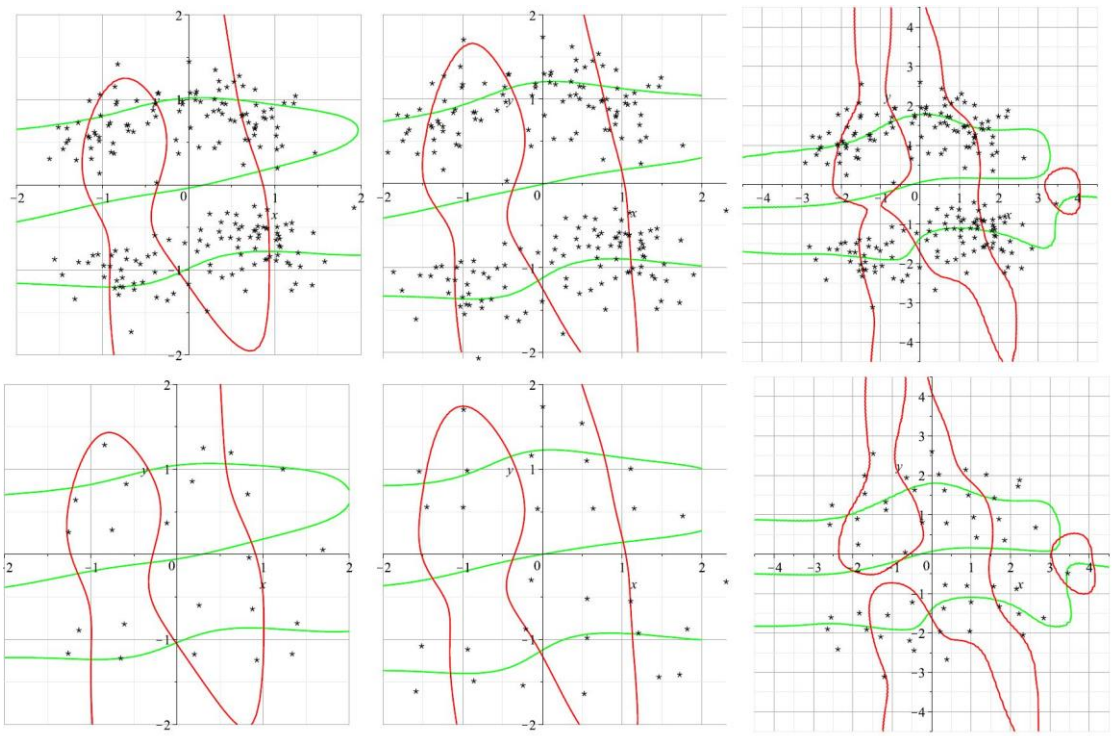


Figure 4: The first three plots (above) represent the implicit function of M in red and L in green and the set of particles (Original problem) with various $\sigma = 0.0594$, $\sigma = 0.0495$ and $\sigma = 0.0330$. The three remaining plots (below) represent the implicit function of MBls in red and LBls in green and the set Lm of the block minima for respectively k = 5 k = 6 and k = 9

The block algorithm has been tested on the crab example [3,21,22] with varying values of k. For $k = 5$, we have reduced the minimizing problem on 200 particles to a minimizing problem on 23 weighted particles. These new 23 particles correspond to minima of a sub-problem reduced to blocks. Table 4 shows for various k, the number of non empty blocks it produces (column two) and the value of $\sigma = \frac{1}{k} max(x_{max} - x_{min}, y_{max} - y_{min})$ (column 3). It also shows, in the fourth column, the approximation of the minima of the block System (20) in the X variable, whereas the sixth column shows the approximation of the minima of the original System (2). In the fifth and

seventh column, the number of particles per clusters is given. The clusters are obtained by computing the Euclidean distance between a particle and the four minima namely m_1, m_2, m_3 and m_4. A particle p belongs to the cluster i if $|pm_i| = min(|pm_1|, |pm_2|, |pm_3|, |pm_4|)$.

k *and* σ	blocks numb.	minima $\begin{cases} \mathcal{M}_{Bls} = 0 \\ \mathcal{L}_{Bls} = 0 \end{cases}$	clust. size	minima $\begin{cases} \mathcal{M} = 0 \\ \mathcal{L} = 0 \end{cases}$	clust. size
5 and 0.0594	23	$[-0.102270, 0.0658670]$	40	$[-0.09612, 0.06528]$	41
		$[-0.083935, -0.103749]$	36	$[-0.07728, -0.10097]$	36
		$[0.0474319, 0.0895055]$	60	$[0.04818, 0.08389]$	59
		$[0.084879, -0.073041]$	64	$[0.07861, -0.06591]$	64
6 and 0.0495	30	$[-0.104058, 0.0584710]$	41	$[-0.09830, 0.05817]$	41
		$[-0.080777, -0.097700]$	36	$[-0.07653, -0.09568]$	36
		$[0.0540185, 0.0816837]$	59	$[0.05145, 0.08114]$	59
		$[0.078933, -0.065498]$	64	$[0.07766, -0.06330]$	64
9 and 0.0330	53	$[-0.100231, 0.0463999]$	41	$[-0.09671, 0.04727]$	41
		$[-0.072107, -0.086702]$	37	$[-0.07653, -0.08632]$	37
		$[0.062279, 0.0685270]$	59	$[0.06196, 0.06852]$	59
		$[0.076056, -0.054503]$	63	$[0.07562, -0.05440]$	63

Table 4: Comparison of the minima and the clusters using the block method and the direct method

We have compared the clusters produced by the direct method with $\sigma = 5$ and those produced by the block method with $k = 5$, we observe that the result is the same except for one particle. For $k = 6$ or $k = 9$, we obtain the same clusters from both methods.

3.5. Benchmarks

The block method can be tested on larger set of particles. In this subsection, we propose two other examples:

- Clustering of Exoplanet data [3]. This is data from the "Extrasolar Planets Encyclopedia" [3,24] or more specifically Tahir Yaqoob [25]. Figure 5 is a plot of mass in Earth units versus the period in Astronomical Units (AU) on a log base 10 scale. The number of particles is $N = 1093$. It shows some very complex behavior, but three rather well-defined groups of data can be discerned as revealed by the quantum clustering method. The block method with $k = 14$ and $l = 1.5$ gives a σ value of 0.74 and the three following minima at $(-1.784092251, 1.149209809)$, $(0.07545310832, 0.4043352565)$ and $(0.4394030237, 3.008141919)$.

  The data cluster on the lower right-hand side corresponds to the massive, short-period hot Jupiters that have been discovered.

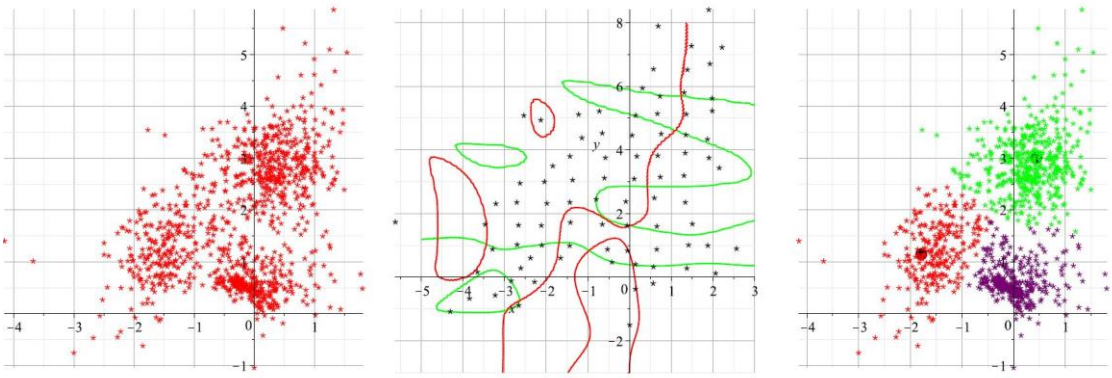


Figure 5: Exoplanets example. From left to right : the original data, the main characteristics of the corresponding block method of parameters $k = 4$ and $l = 1.5$, the clustering

The next two examples are known to be difficult examples and the clustering outcome is usually imperfect.

- Gionis *et al.* [26] propose a method consisting of an aggregation of other approaches including single linkage, complete linkage, average linkage, K-means and Ward's clustering. The dataset proposed in [26] has $N = 788$ particles and contains narrow bridges between clusters and uneven-sized clusters that are known to create difficulties for the clustering algorithms. The aggregation method gives seven clusters.

  Our quantum block method (with $k = 9$, $\sigma = 3.6889$) gives also seven minima and thus seven clusters. Figure 6 Shows 6 drawing : The first drawing is the initial data. In the second one, the black dots corresponds to the new set of weighted particles obtained by using the block method with parameters $k = 9$ and $l = 1$ (Consequently, σ becomes $\sigma = 3.6889$). The red and green curves correspond to the implicit functions of M_{Bls} and L_{Bls} (The scale has been modified here following the variable changes proposed in Section 2 The determination of the clusters is done here from the minima using the Euclidean distance. Unfortunately, it faces some difficulties and some improvements could be done by using spectral clustering. Here, we use a ϵ-neighborhood graph to produce the spectral clustering as shown in the second line of Figure 6. The MATLAB algorithm used needs as input the data *and* the number of clusters. First, we see the level lines and the clusters of the block data. The last drawing gives the rebuilding of the clustering on the initial data. It shows that the quality of the clustering is similar to the one of the aggregation of five different clustering approaches (see [26]).

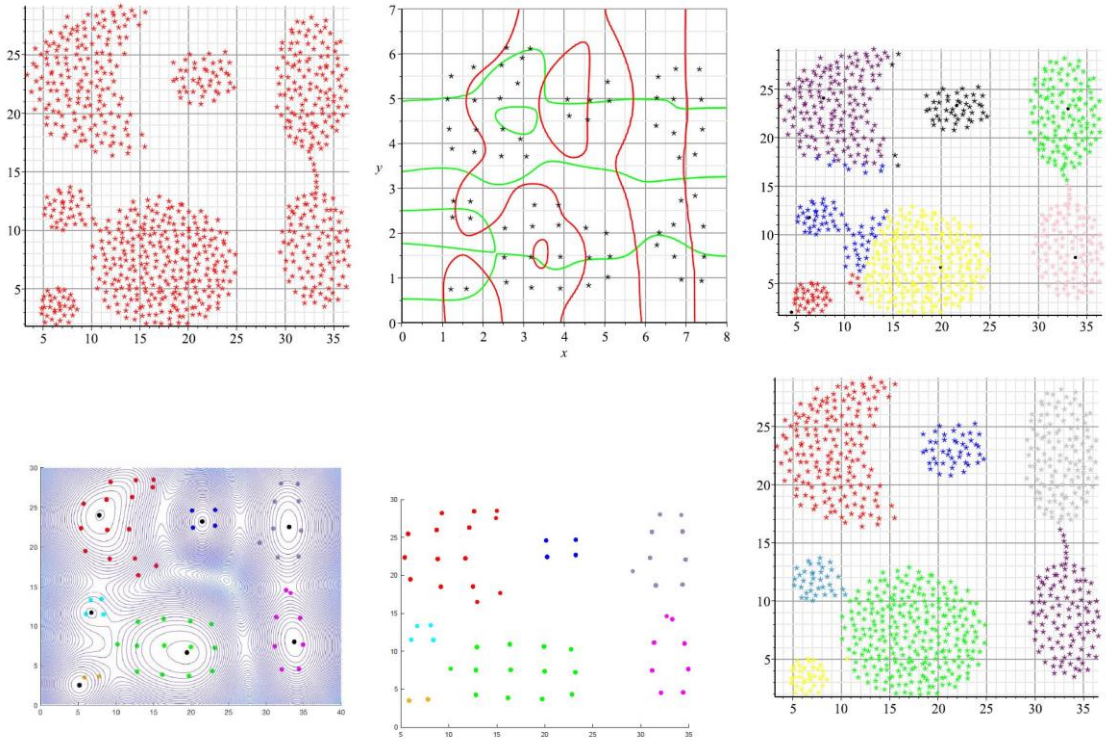


Figure 6: Example from [26] of $N = 788$ particles. From left to right, the initial data, the main characteristics of the corresponding block method of parameters $k = 9$ and $l = 1$, the clustering using the Euclidean distance from the computed center (in black). Second line: The level line of the block quantum equation, new clustering based on the spectral clustering method on the block data, reconstruction of the clustering on the initial data.

Unfortunately, some specific shapes such as ring-shaped or spiral-shaped clusters are challenging for numerous clustering methods including our QC block method. To overcome this issue, an approach based on optimization of an objective function, is proposed in [27] to detect specifically elliptic ring-shaped clusters. However, this approach is not appropriate when different kind of shapes coexist as for example in the case of Zahn's compounds [28]. It also requires a skilled operator to visualize the clusters. It will be a great challenge to improve the QC approach in order to detect such shapes.

3.6. Perspectives

In spite of claims to the contrary [29], even with extensions, K-means is no longer state-of-the-art. A means of finding *all* the potential minima of the quantum potential and consequently the number of clusters for a given range of σ is an essential key feature for data clustering under program control without prior visualization whilst K-means and even MATLAB's spectral clustering require the number of cluster centers on input and thus skilled operators. The quantum clustering approach yields this number for a given range. Automatic Data clustering under program control allows the processing of much bigger and more complex mixed datasets potentially providing a more robust industrial standard. It would multiply the number of platforms with large data collection tools such as Hadoop or MongoDB and thus a greater realization of patents for name of object disambiguation [1].

4. CONCLUSIONS

Herein, we have made considerable progress in dealing with the outstanding problem of getting all the centers of the quantum clustering method, namely finding *all* the minima of the quantum potential of Equation (1) where σ is the standard deviation of the Gaussian distribution. The extrema of this potential are the roots of two coupled equations, which in principle are impossible to solve analytically. After simplifications, those equations become bivariate exponential polynomial equations and a number of useful properties have been proved. More precisely, limits of implicit function branches are given and the case of two particles is analytically solved. We also proved that the coupled equations have only one minimum if the data are included in a square of side σ. This bound is directly useful to propose a new approach "per block". This technique decreases the number of particles by approximating some groups of particles to weighted particles. The minima of the corresponding coupled equations are then given numerically by which the number of clusters is obtained. Those minima can be used as cluster centers. However, for some complex examples, other clustering approaches such that spectral clustering gives better visual results (though they still require the number of clusters on input). On such examples, the approach consisting in the use of the block method (for the number of clusters but also for the weighted particles) gives very good results. Example 3, from Gionis *et al.* shows that the quality of the clustering is similar to the one of the aggregation of five approaches.

The approach used here is potentially useful for other types of exponential polynomials found in numerous Physical applications such as, for example, quantum mechanical diffusion Monte-Carlo calculations, where a precise knowledge of the nodal lines ensures accurate energy eigenvalues.

REFERENCES

[1] M. Fertik, T Scott and T Dignan, US Patent No. 2013/0086075 A1, Appl. No. 13/252,697 - Ref. US9020952B2, (2013)

[2] D. Horn and A. Gottlieb, Phys. Rev. Lett. **88**, 18702 (2002)

[3] T. C. Scott, M. Therani and X. M. Wang, Mathematics **5**, 1-17 (2017)

[4] A. Ben-Hur, D. Horn, H. T. Siegelmann and V. Vapnik, J. Mach. Learn. Res. **2**, 125-137 (2002)

[5] A. Messiah, Quantum Mechanics (Vol. I), English translation from French by G. M. Temmer, North Holland, John Wiley & Sons, Cf. chap. IV, section III. chap. 3, sec.12, 1966.

[6] A. Lüchow and T. C. Scott, J. Phys. B: At. Mol. Opt. Phys. **40**, 851-867 (2007)

[7] A. Lüchow, R. Petz R and T. C. Scott, J. Chem. Phys. **126**, 144110-144110 (2007)

[8] T. C. Scott, A. Lüchow, D. Bressanini and J.D. Morgan III, Phys. Rev. A (Rapid Communications) **75**, 060101 (2007)

[9] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, Montgomery, Jr., J. A., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, Gaussian Inc. , Wallingford CT (2009)

[10] T. C. Scott, I. P. Grant, M. B. Monagan and V. R. Saunders, Nucl. Instruments and Methods Phys. Research **389A**, 117-120 (1997)

[11] T. C. Scott, I. P. Grant, M. B. Monagan and V.R. Saunders, MapleTech **4**, 15-24 (1997)

[12] C. Gomez and T. C. Scott, Comput. Phys. Commun. **115**, 548-562 (1998)

[13] Achatz, M., McCallum, S., & Weispfenning, V. (2008). Deciding polynomial-exponential problems. In D. Jeffrey (Ed.), ISSAC'08: Proceedings of the 21st International Symposium on Symbolic and Algebraic Computation 2008 (pp. 215-222). New York: Association for Computing Machinery. https://doi.org/10.1145/1390768.1390799

[14] A. Maignan, Solving One and Two-dimensional Exponential Polynomial Systems, ISSAC98, acm press, pp 215-221.

[15] Scott McCallum, Volker Weispfenning, Deciding polynomial-transcendental problems, Journal of Symbolic Computation, Volume 47, Issue 1, 2012, Pages 16-31, ISSN 0747-7171, https://doi.org/10.1016/j.jsc.2011.08.004.

[16] J. F. Rodriguez-Nieva and M. S. Scheurer, Identifying topological order through unsupervised machine learning, Nature Physics, Nature, Physics **15**, 790 (2019)

[17] Eran Lustig, Or Yair, Ronen Talmon, and Mordechai Segev, Identifying Topological Phase Transitions in Experiments Using Manifold Learning, Phys. Rev. Lett. **125**, 127401 (2020)

[18] Jielin Wang, Wanzhou Zhang, Tian Hua and Tzu-Chieh Wei, Unsupervised learning of topological phase transitions using Calinski-Harabaz score, accepted by Physical Review Research, (2020)

[19] Shervan Fekri Ershad, Texture Classification Approach Based on Energy Variation IJMT **2**, 52-55 (2012)

[20] Fan Decheng, Song Jon, Cholho Pang, Wang Dong, CholJin Won, Improved quantum clustering analysis based on the weighted distance and its application, Heliyon, Volume 4, Issue 11, 2018, e00984, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2018.e00984.

[21] B. Ripley, Cambridge University Press, Cambridge, UK (1996)

[22] B. Ripley, Available online , http://www.stats.ox.ac.uk/pub/PRNN/ (accessed on 3 January 2017)

[23] L. Bernardin, P. Chin, P. DeMarco, K. O. Geddes, D. E. G. Hare, K. M. Heal, G. Labahn, J. P. May, J. McCarron, M. B. Monagan, D. Ohashi and S. M. Vorkoetter, MapleSoft , Toronto (2012)

[24] Exoplanet.eu-Extrasolar Planets Encyclopedia, Available online , http://exoplanet.eu/ Retrieved 16 November 2015 (accessed on 2 January 2017)

[25] T. Yaqoob, New Earth Labs (Education and Outreach) , Baltimore, MD (USA, 16 November 2011)

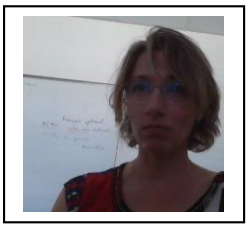
[26] A. Gionis, H. Mannila, and P. Tsaparas, Clustering aggregation. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007. 1(1): p. 1-30.

[27] Isak Gath and Dan Hoory, Fuzzy clustering of elliptic ring-shaped clusters, Pattern Recognition Letters", Vol. 16, 1995, p. 727-741, https://doi.org/10.1016/0167-8655(95)00030-K.

[28] http://cs.joensuu.fi/sipu/datasets/

[29] A. Ahmad and S. S. Khan, "Survey of State-of-the-Art Mixed Data Clustering Algorithms," in IEEE Access, vol. 7, pp. 31883-31902, 2019, doi: 10.1109/ACCESS.2019.2903568.

Aude Maignan



Aude Maignan received the Ph.D. degree in applied mathematics from Limoges University, France, in 2000. She is an Associate Professor at Université Grenoble Alpes, France. Her research interests include complex systems, generalized Lambert function and graph rewriting.

Tony Scott

Tony C. Scott graduated in 1991 with a Ph.D. in Theoretical Physics and was awarded the Pearson Medal for best Physics Doctoral thesis at the University of Waterloo (1991). His Master's thesis in Applied Mathematics (1986) is cited in the Wikipedia section on the Wheeler-Feynman absorber theory. Awarded with an N.S.E.R.C. postdoctoral scholarship, he subsequently did research in Mathematical Physics at the Harvard-Smithsonian in Cambridge MA USA(90-92). Afterwards, he did research in relativistic quantum chemistry and pioneered a mathematics course using computer algebra at the Mathematical institute in Oxford University in the UK ('93-'95). This was followed by further work at the University of Ben-Gurion in Israel ('96-'97), INRIA in France ('98-'99), the Forschungszentrum in Juelich Germany (2003) and eventually RWTH-Aachen University Germany (2003-2006) where he retains an affiliation. He worked for 7 years as a Data Scientist in Silicon Valley in the San Francisco bay area before returning as a professor in China. He is back in the private sector working in the area of Data Science.