Probabilistic Quantum Clustering

Raúl Casaña-Eslava[a], Paulo J. G. Lisboa[a], Sandra Ortega-Martorell[a], Ian H. Jarman[a], José D. Martín-Guerrero[b]

[a] *Department of Applied Mathematics, Liverpool John Moores University, 3 Byrom Street, Liverpool, L3 3AF, UK.*
E-mail: raulcasana@gmail.com
[b] *Departament d'Enginyeria Electrònica - ETSE, Universitat de València, Av. Universitat, SN, 46100 Burjassot, València, Spain.*
*E-mail:*jose.d.martin@uv.es

Abstract

Quantum Clustering is a powerful method to detect clusters with complex shapes. However, it is very sensitive to a length parameter that controls the shape of the Gaussian kernel associated with a wave function, which is employed in the Schrödinger equation with the role of a density estimator. In addition, linking data points into clusters requires local estimates of covariance which requires further parameters. This paper proposes a Bayesian framework that provides an objective measure of goodness-of-fit to the data, to optimise the adjustable parameters. This also quantifies the probabilities of cluster membership, thus partitioning the data into a specific number of clusters, where each cluster probability is estimated through an aggregated density function composed of the data samples that generate the cluster, having each cluster an associated probability density function $P(K|X)$; this probability can be used as a measure of how well the clusters fit the data. Another main contribution of the work is the adaptation of the Schrödinger equation to deal with local length parameters for cluster discrimination by density. The proposed framework is tested on real and synthetic data sets, assessing its validity by measuring concordance with the Jaccard score.

Keywords: Quantum Clustering, Mixture of Gaussians, Probabilistic framework, Unsupervised assessment, Manifold Parzen window.

1. Introduction

Quantum Clustering (QC) is an appealing paradigm inspired by the Schrödinger equation [1] to identify and track connected regions, so clustering the data. The method is effective for modelling anisotropy and heteroscedasticity, since the use of gradient descent rather than distances for allocating points into clusters has the effect of linking together nearby points. However, the method is sensitive to the length scales that are inherent in the Schrödinger equation,

which requires parameter searches. Although some works have proposed methods like cluster consistency [2] for this purpose, they are intrinsically limited by the use of a single length scale to fit data that may have variable density.

The high sensitivity associated with the length scale, called σ, is initially due to the fact it appears explicitly in the Schrödinger equation, but above all it is because it controls the shape of the Gaussian kernel associated with a wave function, which is employed in the Schrödinger equation with the role of a density estimator. This dependency of the length scale will be amplified with the second derivatives involved in the computation of the potential function, needed to allocate points into clusters.

The original quantum clustering algorithm [1] generates a potential function $V(\mathbf{x})$ as the ground state of the time-independent Schrödinger equation:

$$H\Psi \equiv \left(-\frac{\sigma^2}{2}\nabla^2 + V(\mathbf{x})\right)\Psi(\mathbf{x}) = E\Psi(\mathbf{x}) \tag{1}$$

where H is the Hamiltonian, E the energy eigenvalue and the wave function Ψ acts as a density estimator. In the original formulation, the wave function was constructed as a Parzen estimator with a given length scale parameter, σ. The allocation of data points to clusters was determined by the application of gradient descent using the potential function. The wave function provides a parametrisation of local data density. This need not be Gaussian and may involve B-splines [3], Vector Quantisation [4] or the Epanechnikov kernel [5, 6]. However, exponential distributions are generally preferred due to their smoothness since the wave function has to be differentiable up to third order.

Clearly the length scale of the exponential functions, σ, is of critical importance as it determines the overlap between the wave function components from neighbouring observations. This has a critical impact on the shape and smoothness of the resulting potential function, affecting the number of local minima and, consequently, also the final number of clusters. The dependence on the bandwidth selection of the Parzen window has led to different approaches to estimate the local covariance, first using k-nearest neighbours (KNN) [2, 7], then with sample covariance estimators [8, 9, 10]. However the efficiency of KNN estimators varies considerably depending on the structure of the data [11].

Recent works have addressed the non-linear or non-spherical clustering problem from different perspectives, like using spectral clustering based on different similarity metrics instead of Euclidean distances or, considering semi-supervised learning. For instance [12, 13] propose a density-based algorithm similar to DBSCAN [14] but assuming local consistency —nearby points should have similar local density— and global consistency —high density regions should have the same structure or label— to define a density-adaptive metric based on the sensitivity of the local density. Another work [15] makes use of a supervised pairwise constraints to perform a spectral clustering based on Hidden Markov Random Fields (HMRF). Finally, [16] considers the Minimax distance based on minimum spanning tree clustering. Another clustering point of view in recent works

is through multi-task multi-view approach [17], where manifold learning techniques [18] are combined with graph-based methods [19] to get insights of the data structure. However, the main advantages of our approach with respect to these recent works stem from the fact that it is completely unsupervised and can identify the most appropriate hyperparameters automatically. It should be highlighted that while both classical and nature-inspired clustering algorithms use projective methods, quantum clustering defines an energy surface, which is the potential function, and data points slide along this surface to converge together in clusters, thus being a completely alternative way of facing unsupervised learning problems.

This paper addresses the need for an objective function to guide the optimisation of the adjustable parameters. We cast the quantum clustering paradigm in a probabilistic framework that defines a log-likelihood function to measure goodness of fit to the data. This enables parameter optimisation to be carried out reliably and systematically without prior knowledge of the data structure. Experiments with real-world data sets and challenging synthetic data sets demonstrate the effectiveness of the proposed approach even in the presence of anisotropy and heterocedasticity.

The method has only one free parameter, which is the number of nearest neighbours used in local covariance estimation. This underlines Probabilistic Quantum Clustering (PQC) as a plausible unsupervised method for the detection of complex data structure in low dimensional data. The proposed approach also indicates the presence of hierarchical data structure, identified by local minima in the objective function of goodness of fit.

Re-casting the method in a probabilistic framework has the further advantage of quantifying the probability of cluster membership and identifying the presence of outliers, which offers potential for use in novelty detection.

Another relevant contribution of the paper is the use of local length scales to discriminate clusters using density information. Indeed, as the gradient is sensitive to data density, it is capable of modelling high density clusters even when located inside clusters with lower density.

The rest of the paper is structured as follows: section 2 introduces the original QC and extends it into the proposed probabilistic framework. Section 3 describes the data sets that we considered to evaluate our proposal, reporting and discussing the achieved experimental results in section 4. Section 5 analyses the complexity of the method, concluding the paper in Section 6 with a critical summary of PQC, the drawn conclusions and possible directions for further work.

2. Methods

2.1. Current implementations of Quantum Clustering

In the original formulation of Quantum Clustering, QC_σ, a wave function composed of radial kernels centred on the data points generates a convex potential function from the steady-state eigenfunction of the Schrödinger equation.

Cluster allocation consists in identifying local basins of attraction in the potential function, by gradient descent (GD).

The wave function is parameterised as a mixture of Gaussians:

$$\Psi(\mathbf{x}) = \sum_{i=1}^{n} \psi_i(\mathbf{x}) = \sum_{i=1}^{n} e^{-\frac{(\mathbf{x}-\mathbf{x}_i)^2}{2\sigma^2}} \tag{2}$$

where n is the sample size and σ a global length scale comprising a single hyperparameter to adjust. The Gaussian normalisation factor is redundant as it will cancel out in the calculation of the potential function. The expression for the potential function is:

$$V(\mathbf{x}) = S + \frac{\sigma^2}{2} \frac{\nabla^2 \Psi(\mathbf{x})}{\Psi(\mathbf{x})} \tag{3}$$

where S is an arbitrary constant offset.

Introducing the following notation for the expected value of function F at discrete data points i with respect to the wave function ψ_i:

$$\langle F_i \rangle_\Psi \equiv \frac{\sum_i F_i \psi_i}{\sum_i \psi_i} \tag{4}$$

then the QC_σ potential simplifies to the second-order moment of the distance from the data points:

$$V(\mathbf{x}) = S + \frac{\sigma^2}{2} \frac{\nabla^2 \Psi(\mathbf{x})}{\Psi(\mathbf{x})} = S - \frac{d}{2} + \left\langle \frac{(\mathbf{x}-\mathbf{x}_i)^2}{2\sigma^2} \right\rangle_\Psi \tag{5}$$

where d is the input space dimension and with gradient given by:

$$\begin{aligned} \nabla V(\mathbf{x}) = \left\langle \frac{(\mathbf{x}-\mathbf{x}_i)}{\sigma^2} \right\rangle_\Psi \left(1 + \left\langle \frac{(\mathbf{x}-\mathbf{x}_i)^2}{2\sigma^2} \right\rangle_\Psi \right) - \\ \left\langle \frac{(\mathbf{x}-\mathbf{x}_i)(\mathbf{x}-\mathbf{x}_i)^2}{2\sigma^4} \right\rangle_\Psi \end{aligned} \tag{6}$$

Data points are allocated to clusters by sliding from their initial positions into basins of attraction corresponding to local minima in the potential function, by gradient descent:

$$\mathbf{y}_i(t + \Delta t) = \mathbf{y}_i(t) - \eta(t) \nabla V(\mathbf{y}_i(t)) \tag{7}$$

where $\eta(t)$ is an adjustable gain parameter.

We considered ADAM, a variant of Stochastic Gradient Descent (SGD) with an adaptive momentum term [20], which is suitable for sparse gradients that commonly occur with sparse data. Nowadays, there are many variants of SGD [21]. This work does not claim that ADAM is the most suitable choice for the QC problem, maybe other SGD variants could achieve a similar performance. The reason to eventually choose ADAM is twofold: first, the learning rates of each parameter —in our case, space dimension— are adapted as a function of the squared gradients, which speeds up the convergence in sparse regions where the potential gradient is too small; second, ADAM makes use of an exponential moving average of the past gradients —computing the mean and the variance— to update the gradient term, something similar to a momentum term that helps to avoid local minima.

The stopping criteria are thresholds for minimum change in either position or the value of the potential function. The smaller the value of the length scale, the higher the number of clusters detected. The application of this method is illustrated with a synthetic data set with four two-dimensional clusters that combine anisotropy and heteroscedasticity (artificial data set #1). Each cluster has 100 data points.

Figure 1a shows the cluster allocation by SGD with a length scale of $\sigma_{20\%}$, which corresponds with the mean value of all the possible values of σ, considering K-nearest neighbours with $K = 20\%N$, being N the sample size. The corresponding gradient directions and contour lines are shown in figure 1b; QC_σ with a single length scale results in a too large value to accurately capture the high density cluster and in a too small value for the sparse cluster at the bottom of the plot which breaks up into multiple local minima.

Information about local density can be included by defining σ as a function of the KNNs (QC_{knn}) instead of having a unique length scale. In particular, it can be typically expressed as a percentage of the total sample size: $K = \%\text{KNN}$ [2, 7, 11], i.e. the number of observations to estimate σ is parameterised as a percentage of K-nearest neighbours:

$$\sigma_i \equiv \frac{1}{K} \sum_{j \,\in\, knn(\mathbf{x}_i)}^{K} dist(\mathbf{x}_i, \mathbf{x}_j) \tag{8}$$

Note that the length scale of the original model (QC_σ) is the mean of these length scales: $\sigma = \frac{1}{N}\sum_{i=1}^{N} \sigma_i$. The variable length scale also allows to decouple the terms of the Schrödinger equation, being the kinetic term:

$$T_i = \frac{\sigma_i^2}{2}\nabla^2\psi_i = \left(\frac{(\mathbf{x}-\mathbf{x}_i)^2}{2\sigma_i^2} - \frac{d}{2}\right)\psi_i \tag{9}$$

The new potential and gradient terms become:

$$V(\mathbf{x}) = E + \frac{\sum_i \frac{\sigma_i^2}{2} \nabla^2 \psi_i}{\sum_i \psi_i} = E - \frac{d}{2} + \left\langle \frac{(\mathbf{x} - \mathbf{x}_i)^2}{2\sigma_i^2} \right\rangle_\Psi \tag{10}$$

$$\nabla V(\mathbf{x}) = \left\langle \frac{(\mathbf{x} - \mathbf{x}_i)}{\sigma_i^2} \right\rangle_\Psi \left(1 + \left\langle \frac{(\mathbf{x} - \mathbf{x}_i)^2}{2\sigma_i^2} \right\rangle_\Psi \right) - \left\langle \frac{(\mathbf{x} - \mathbf{x}_i)\,(\mathbf{x} - \mathbf{x}_i)^2}{2\sigma_i^4} \right\rangle_\Psi \tag{11}$$

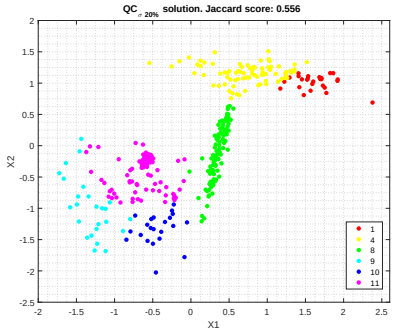
In contrast to QC_σ, the estimation of the length scale from nearest neighbours results in a wave function with a very pronounced peak in the high density region. The shape of QC_{knn} potential is more complex than that obtained by QC_σ, as it is now smooth in sparse regions and steep in dense areas, as required. Figure 1c shows the cluster allocation by SGD over this potential with accurate discrimination of the high density cluster against the surrounding sparse cluster. The potential also adapts to the local density changes, creating a sharp sink around the highest density peak; this region will be isolated in the clustering allocation by SGD, allowing a cluster discrimination by local densities. In this example, 20% is an appropriate parameter value; if σ were much smaller it would produce an over-fitted potential, generating too many sub-clusters. The value of the Jaccard Score (JS) with QC_{knn} (0.862) is much better than for QC_σ (0.556).

Adjusting the length scale via nearest neighbours is clearly effective for detecting clusters with very different densities and also to accommodate outliers with smooth and flat gradients that do not lead to an unnecessary fragmentation in low density regions. It partially solves the problem of heteroscedasticity but the amount of neighbours considered in the model is still a hyperparameter to be determined. There is a trade-off between too few neighbours resulting in an over-fitted density function with too many clusters, and too large a neighbourhood leading to a biased density function with too few clusters.

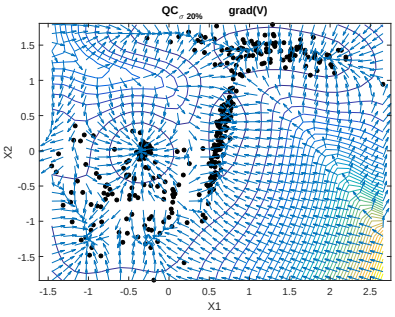
A natural extension to QC_{knn} is to consider the local non-spherical covariance matrices to set the value of the variable length scale, thus producing a wave function and a derived potential that could fit better the density distribution QC_{cov}:

$$\Sigma_i = \frac{1}{N_k - 1} \sum_{j \,\in\, knn}^{N_k} (\mathbf{x_j} - \mathbf{x}_i)^T (\mathbf{x_j} - \mathbf{x}_i) \tag{12}$$

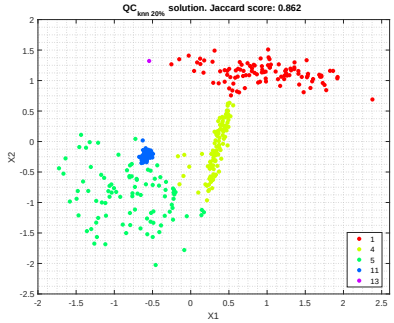
where, in reference to the notation of equations 12 and 13, Σ_i refers to the covariance matrix of $(sample)_i$, not to be confused with the summation symbol $\sum_i^n$.



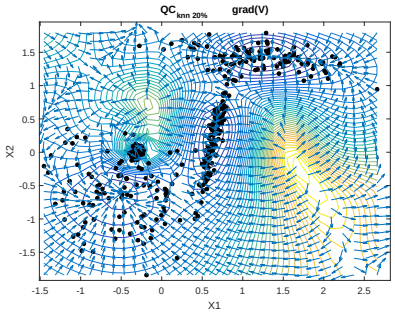
(a) SGD cluster allocation $QC_{\sigma\ 20\%}$



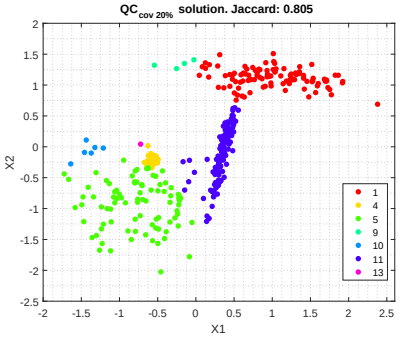
(b) Gradient $QC_{\sigma\ 20\%}$



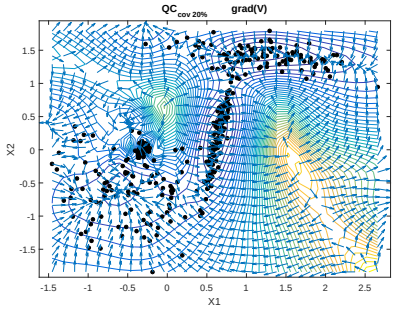
(c) SGD cluster allocation $QC_{knn\ 20\%}$



(d) Gradient $QC_{knn\ 20\%}$



(e) SGD cluster allocation $QC_{cov\ 20\%}$



(f) Gradient $QC_{cov\ 20\%}$

Figure 1: Cluster allocations by SGD (left) resulting from the gradients of the potential function (right) for artificial data set #1. The rows correspond to QC_σ, QC_{knn} and QC_{cov}, respectively. In all cases the length scales have been computed using a quantile of 20%. These solutions have Jaccard scores of 0.556, 0.862 and 0.805, respectively.

Now the kernels are normalised with multivariate Normal distributions:

$$\begin{aligned}\Psi(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^{n}\psi_i\left(\mathbf{x}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{|2\pi\Sigma_i|}}\exp\left(-\frac{1}{2}\left(\mathbf{x}-\mathbf{x}_i\right)^T\Sigma_i^{-1}\left(\mathbf{x}-\mathbf{x}_i\right)\right)\end{aligned} \tag{13}$$

Obviously, this approach may result in degenerate covariance matrices causing singularities in the covariance-inverse estimation. Moreover, if data are very anisotropic, the positive effect of superposition in the wave function is considerably reduced, resulting in a wave function that is less smooth which creates an excessive number of local minima. These disadvantages can be mitigated if all the local covariance matrices are restricted to be diagonal and replacing diagonal elements whose value is close to zero by a given arbitrary small threshold. As a result of a less smooth potential that leads to more local minima, some spurious clusters appear (figure 1e), thus degrading JS up to 0.805. The problems shown by QC_{cov} will be addressed in section 2.2, with our proposed QC_{cov}^{prob}.

2.2. Probabilistic Quantum Clustering, QC_{cov}^{prob}

The probabilistic framework QC_{prob} aims at interpreting the normalised mixture of Gaussians shown in eq. (13) as a joint probability distribution for the co-occurrence of the test point x and the data point x_i. This represents a generative model with a Gaussian kernel over each data point. The purpose of the PQC algorithm is to link together joint distributions of neighbouring points to form clusters.

Assigning to all observations an equal prior, once data have been assigned to clusters, the joint probability of observation of a test point x in a particular cluster k is given by:

$$P\left(k,\mathbf{x}\right) = \frac{1}{n}\sum_{i\in k}^{\#k}\psi_i\left(\mathbf{x}\right) \tag{14}$$

where n is the sample size, which automatically fulfils the consistency requirement that:

$$\sum_{k=1}^{K}P(k,\mathbf{x}) = P(\mathbf{x}) \tag{15}$$

being K the total number of clusters, and $\#k$ the number of observations in cluster k.

A key difference with respect to current implementations of quantum clustering is that there are two steps involved in cluster allocation:

a) Application of gradient descent to allocate individual observations into clusters, which partition the data by setting the indices i in eq. (14);

therefore, the probability of cluster membership for each data point is given by $P(k, \mathbf{x})$.

b) Reallocation of the observations using the maximum value of the probability of cluster membership, i.e., selecting the cluster index k for which $P(k|\mathbf{x})$ is maximal.

The probability of k follows by marginalizing the joint probability over $\mathbb{R}$:

$$\begin{aligned} P(k) &= \int_{\mathbb{R}} P(k, \mathbf{x}) d\mathbf{x} = \int_{\mathbb{R}} \frac{\sum_{i \in k}^{\#k} \psi_i(\mathbf{x})}{n} d\mathbf{x} \\ &= \sum_{i \in k}^{\#k} \frac{\int_{\mathbb{R}} \psi_i(\mathbf{x}) \, d\mathbf{x}}{n} = \sum_{i \in k}^{\#k} \frac{1}{n} = \frac{\#k}{n} \end{aligned} \tag{16}$$

Once the joint probability is defined, the Bayes' rule is applied to obtain the conditional probabilities used above:

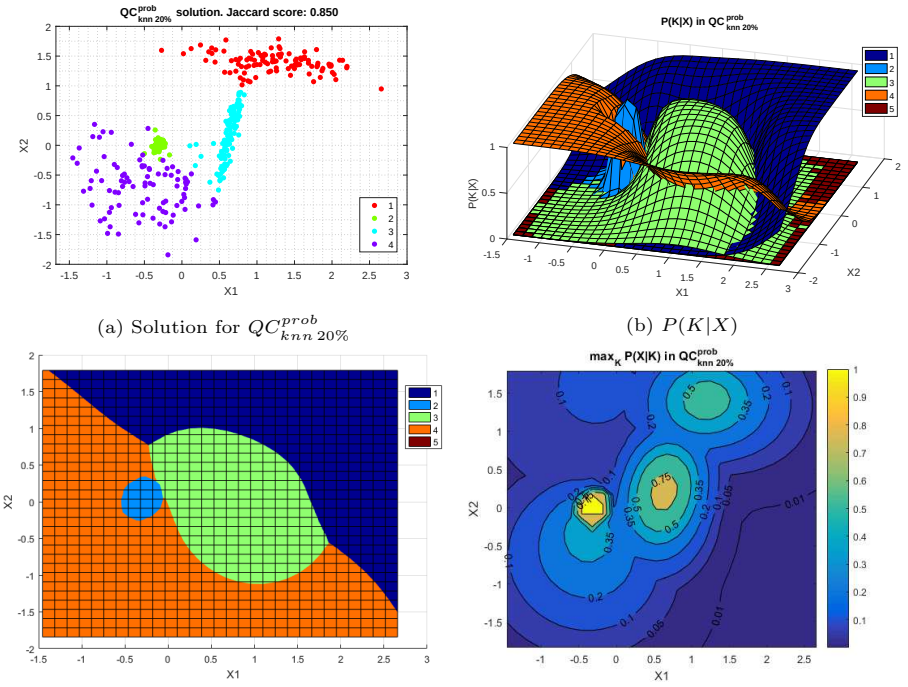
$$P(k|\mathbf{x}) = \frac{P(k, \mathbf{x})}{P(\mathbf{x})} = \frac{\sum_{i \in k}^{\#k} \psi_i(\mathbf{x})}{\sum_{k=1}^{K} \sum_{i \in k}^{\#k} \psi_i(\mathbf{x})} \tag{17}$$

$$P(\mathbf{x}|k) = \frac{P(k, \mathbf{x})}{P(k)} = \frac{\sum_{i \in k}^{\#k} \psi_i(\mathbf{x})}{\frac{\#k}{n}} \tag{18}$$

Note that the probabilistic framework allocates clusters to data from anywhere in input space, enabling the model to be used for unseen data. It is possible that following the second step fewer clusters will be allocated than the initial number identified by gradient descent. This will be dictated by the maximal values for the probabilities of cluster membership.

The gradient descent step is only applied once to set the probabilistic functions which then generalise to training and test data sets; hence, the probabilistic cluster allocation draws a probability map that defines the boundaries between clusters, which may have complex shapes. The aim is to connect together regions of similar data density, separating the clusters with values of different density, which may be less where data are sparse, or higher where heteroscedasticity means that one cluster is spatially entirely contained within another.

The experimental results show only a small effect on the cluster allocation, with a difference in JS of less than 2% when comparing QC_{knn} with its probabilistic counterpart. Figures 2b and 2c depict the probability maps using the four clusters detected in the QC_{knn} solution shown in figure 2a. Differences are greater in the case of QC_{cov}, with the probabilistic cluster allocation closer to the true labels than the SGD approach, with JS=0.882, i.e., the highest JS among all the experiments carried out, and with the selection of only four of the seven clusters detected in figure 1e, as desired.



(a) Solution for $QC^{prob}_{knn\,20\%}$

(b) $P(K|X)$

(c) Projection of $P(K|X)$

(d) $max_K\ P(X|K)$

Figure 2: Top left figure shows the probabilistic cluster allocation with $QC^{prob}_{knn\,20\%}$ (JS=0.850). Top right figure shows its probability map of cluster membership, $P(K|X)$. A top-down projection can be observed in the bottom left figure, where only the highest cluster membership regions are observed. The bottom right figure depicts $max_K\ P(X|K)$, which is useful for outlier detection.

A further advantage of the probabilistic approach is the identification of outliers, by simply thresholding $\max_k P(k|\mathbf{x})$. This is potentially of value to defend against unwanted extrapolation of the clustering structure, as well as providing a probabilistic framework for novelty detection.

2.3. Performance assessment

Optimisation of length scales and other adjustable parameters in quantum clustering currently lacks a systematic methodology based on an objective measurement of the goodness-of-fit of the clusters to the data, a common limitation of unsupervised learning.

We propose the use of the probabilistic framework not only for cluster allocation but also to define the fit to the data, by maximising the log-likelihood for the probability of cluster membership. This measure will then be used to optimise the value of the only hyperparameter in the model, which is the proportion of neighbours used to estimate the local covariance matrix at each data point. This is parameterised as the ratio %KNN.

As shown in eq. (17), each observation is allocated to the cluster k_w with the highest probability $P(k_w|\mathbf{x}_i)$:

$$P(k_w|\mathbf{x}) = \frac{P(k_w, \mathbf{x})}{P(\mathbf{x})} = \frac{P(k_w, \mathbf{x})}{\sum_k P(k, \mathbf{x})} \tag{19}$$

Aggregating over the complete data set, the overall likelihood of cluster membership:

$$\mathrm{LL}(K|\mathbf{X}) = log\left(\prod_i^n P(k_w|\mathbf{x}_i)\right) = \sum_i^n log\left(P(k_w|\mathbf{x}_i)\right) \tag{20}$$

To normalise the score in the range $[0, 1]$, the Average negative Log-Likelihood (ALL) is used:

$$\mathrm{ALL}(K|\mathbf{X}) = \frac{-\sum_i^n log\left(P(k_w|\mathbf{x}_i)\right)}{N} \tag{21}$$

Its value clearly depends on the length scale parameter, %KNN, because the length scale controls the number of clusters and the smoothness of the potential function. The lower the ALL, the better the model fitting with the exception of the trivial solution corresponding with a too large value of the length scale that leads to a single cluster covering all of the data (ALL = 0).

The ALL provides an unsupervised figure of merit which is highly correlated with the supervised JS. Therefore, it can be used as a measure of the clustering performance without the need of prior information about the number of clusters or their composition. Figure 3 shows ALL and JS for different length scales in QC^{prob}, to illustrate their correlation. It also reveals the hierarchical structure of the data, where an abrupt change in ALL means a significant change in the data structure. The bottom plot of figure 3 shows how the number of clusters depends on the length scale, although the QC^{prob} considerably cushions the fluctuation compared with the original QC.

2.4. Extended ALL score with Energy threshold

The extended ALL score improves ALL by setting a threshold E_{th} to merge clusters according to the maximum potential difference between their centroids. By default, the ALL score uses a fixed E_{th} that depends on the SGD convergence criteria in the last iteration:

$$E_{th_{(default)}} = max\Big(\epsilon_V, max\big(\Delta V(\mathbf{x}_{iter\,max})\big)\Big) \tag{22}$$

This E_{th} takes the maximum value between the minimum SGD precision, ϵ_V, and the last SGD update. Figure 4 shows the extended representation of ALL, including its relationship with variations of E_{th} that is no longer a fixed value;

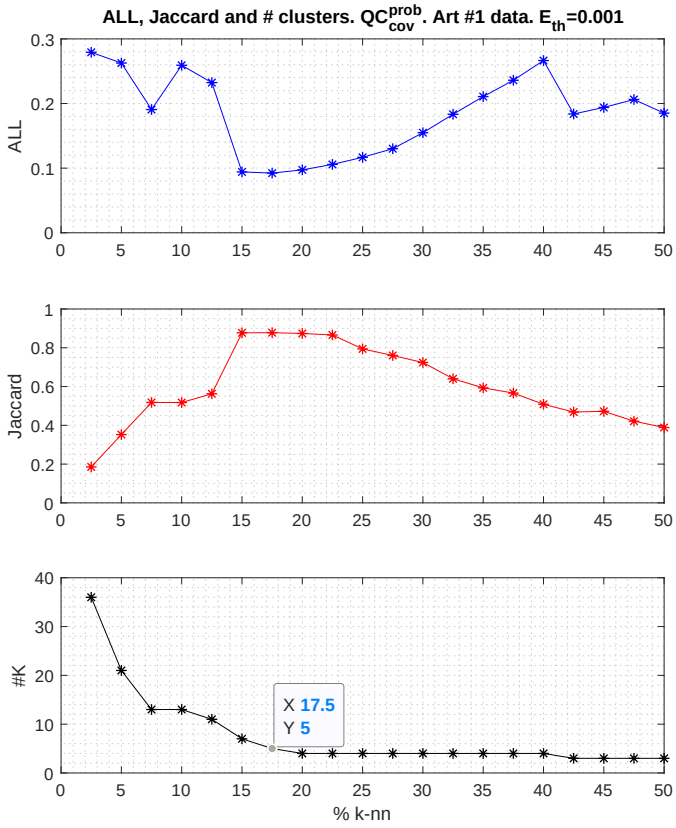


Figure 3: Comparative plot of ALL (top) and Jaccard score (middle) versus %KNN, using QC_{cov}^{prob} in artificial data set #1. The lower values of ALL coincide with the higher values of Jaccard score. Also ALL points out how the structure of the data is changing when %KNN varies. The bottom figure depicts the number of clusters per length scale solution.

for low E_{th} values, being $E_{th} = 0.001$ the default value, the figure 4 presents the same pattern of ALL observed in figure 3. To avoid confusion with non-trivial solutions, scores associated with a trivial solution are assigned to the highest ALL score. The interpretation of the ALL plots is partly subjective, as the plots give an indication of the clustering structure in the data which may be multi-level when the data are hierarchical. Therefore, one can consider that ALL score is a useful tool for finding the hyperparameters associated with good solutions of the problem. The following steps must be taken to identify such solutions (the procedure is also described in more detail in algorithm 1):

a) Starting with the lowest E_{th} value, like default $E_{th} = 0.001$, look for a local minimum in the direction of %KNN axis giving priority to the lowest

values of %KNN.

b) Local minima must have a stable valley in the direction of E_{th}, any solution within this valley is a good solution.

c) Repeat the process if there are more local minima in the direction of %KNN, in ascending order.

d) Looking at the E_{th} direction, if there is a stable region of low ALL values, wide enough to cover several values of E_{th} and %KNN, that region contains a meaningful solution. Solutions with a high E_{th} must be taken with caution because they might correspond to solutions with a few number of clusters, produced after merging clusters hierarchically.

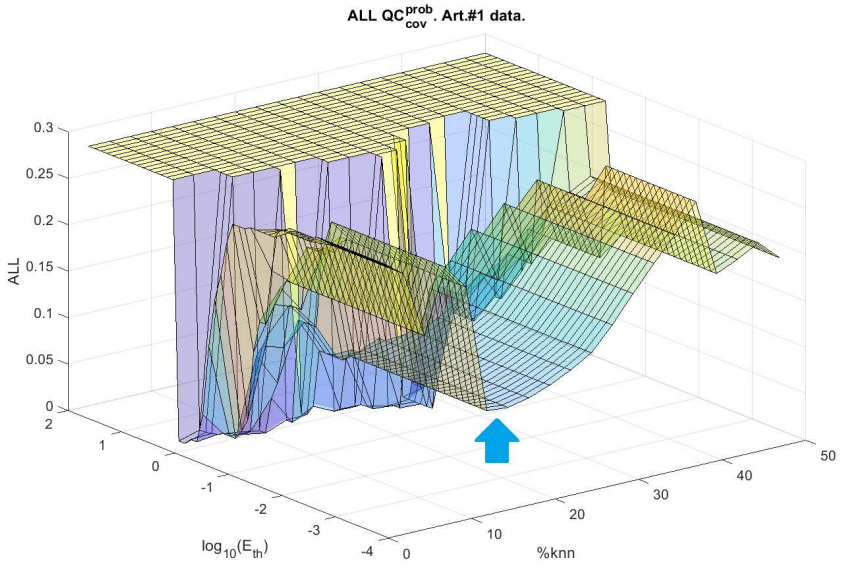


Figure 4: Extended ALL score versus %KNN and E_{th} for data set #1. The ALL scores always diminish when E_{th} increases, because it is an implicit reduction of the number of clusters. That explains why the ALL score is less reliable when there are few clusters, because the trivial solution, with a unique cluster, always leads to ALL equal to zero. To avoid confusion with non-trivial solutions, scores associated with a trivial solution are assigned with the highest ALL score.

3. Data sets

Two challenging artificial data sets and two real-world data sets were employed to test the theoretical hypotheses and evaluate the clustering performance.

Algorithm 1 Procedure of PQC hyperparameter selection from ALL plot

Inputs: collection of PQC models fitted to $\mathbf{X}$ and characterised by their hyperparameters $\%knn$ and E_{th}
ALL $\leftarrow function\,(\%knn, E_{th})$ ▷ Goodness of fit score for each model
Plot ALL as a function of $\%knn$ and E_{th}

procedure LocalMinimaKnn(ALL, $\%knn, E_{th}$)
    $E'_{th} \leftarrow min(E_{th})$
    $\Delta E'_{th} \leftarrow$ Small E_{th} variation
    $Parameters_1 \leftarrow$ Empty list
    for $\%knn' \leftarrow min(\%knn), max(\%knn)$ **do**
        if ALL $(\%knn', E'_{th})$ is local minimum **then**
            if ALL $(\%knn', E'_{th} + \Delta E'_{th}) \approx$ ALL $(\%knn', E'_{th})$ **then**
                Model with $(\%knn', E'_{th})$ is a meaningful solution
                $Parameters_1 \leftarrow$ Append $(\%knn', E'_{th})$
            end if
        end if
    end for
    return $Parameters_1$
end procedure

procedure LocalMinimaEth(ALL, $\%knn, E_{th}$)
    $\Delta E'_{th} \leftarrow$ Small E_{th} variation
    $\Delta\%knn' \leftarrow$ Small $\%knn'$ variation
    $Parameters_2 \leftarrow$ Empty list
    for $\%knn' \leftarrow min(\%knn), max(\%knn)$ **do**
        for $E'_{th} \leftarrow min(E_{th}), max(E_{th})$ **do**
            if ALL $(\%knn' \pm \Delta\%knn', E'_{th} \pm \Delta E'_{th})$ is $\approx$ absolute minimum **then**
                Model with $(\%knn', E'_{th})$ is a meaningful hierarchical solution
                $Parameters_2 \leftarrow$ Append $(\%knn', E'_{th})$
            end if
        end for
    end for
    return $Parameters_2$
end procedure

$SelectedParameters \leftarrow Parameters_1 + Parameters_2$ ▷ Solutions to check

3.1. Data set #1 (artificial): Local densities

This data set, already employed in section 2 to illustrate the characteristics of the different methods presented in the paper, has two main characteristics which challenge clustering algorithms: first, there are two clusters with cigar shapes; second, there are two clusters partially overlapped but with different local densities. The original QC was able to detect anisotropic clusters, but it is less able to discriminate clusters with different local densities. The data set is two-dimensional to aid visualisation and comprises four clusters with 100 observations each.

3.2. Data set #2 (artificial): Two spirals

This is a two-dimensional spiral data set with standard deviation in the first spiral of 0.1 and 0.025 in the second spiral. Each cluster has 200 observations.

3.3. Data set #3 (real): Crabs

This well-known data set was used in the original QC paper [22]. It describes five morphological measurements on 50 crabs of each of two colour forms and both sexes, of the species Leptograpsus variegatus collected at Fremantle, W. Australia. Therefore, there are 200 observations and four different labels, two for gender and two for each species. To compare the results with the original paper, principal component analysis (PCA) has been applied, selecting only the two first principal components (PCs).

3.4. Data set #4 (real): Italian olive oil

This data set consists of 572 observations and 10 variables [23]. Eight variables describe the percentage composition of fatty acids found in the lipid fraction of these oils, which is used to determine their authenticity. The remaining two variables contain information about the classes, which are of two kinds: three "super-classes" at country level: North, South, and the island of Sardinia; and nine collection area classes: three from the Northern region (Umbria, East and West Liguria), four from the South (North and South Apulia, Calabria, and Sicily), and two from the island of Sardinia (inland and coastal Sardinia). The hierarchical structure of this data set makes it especially appealing for testing clustering algorithms.

4. Results

This section evaluates the extent to which the ALL score can determine the most suitable %KNN to maximise the JS, highlighting the peculiarities of each data set and particularly comparing the results of QC_{knn}^{prob} and QC_{cov}^{prob}. As ALL tends to be smaller as the number of clusters decreases, when several local minima appear in ALL, the ones associated with lower %KNN values should have priority over the ones with higher %KNN values, as shown in algorithm 1.

The tables of results include the following information:

- Column 1: data set number and QC model.
- Column 2: score employed to select the quantile (%KNN), firstly the supervised choice according to the best JS, then the unsupervised option based on the local minima found in ALL, and finally checking if the extended ALL has a stable region increasing the E_{th} parameter.
- Column 3: the E_{th} parameter; by default is used $E_{th} = 0.001$, but then the extended ALL plot is analysed to find stable ALL regions with solutions of higher hierarchical order.
- Column 4: length scale parameter in quantiles (%KNN)
- Column 5: number of clusters (#K)
- Column 6: ALL score
- Column 7: Jaccard score - for the Olive oil data set, there are two possible classifications, with three or nine regions.
- Column 8: Cramer's V score - for the Olive oil data set, there are two possible classifications, as above. The Cramer's V-index (C_v) is a normalised version of the standard chi-square test for contingency tables; C_v measures the concordance between different cluster allocations.
- Column 9: Pearson's linear correlation coefficient between ALL score with $E_{th} = 0.001$ and the Jaccard score.
- Column 10: The p-values associated to the correlation coefficient.

4.1. Data set #1: Local densities

Table 1 shows that both models, QC_{knn}^{prob} and QC_{cov}^{prob}, perform similarly for this data set. QC_{knn}^{prob} has the correct number of clusters, four, with a $JS = 0.85$, however QC_{cov}^{prob} with five clusters has a slightly better value, $JS = 0.88$. In both cases, the ALL corresponds with the JS. Besides, there is not a stable region of low ALL with high E_{th} values, so no hierarchical solution was considered.

4.2. Data set #2: Two spirals

Results are shown in table 2 and figures 5–8. Figure 5 shows that JS is quite low. Actually, JS is not a good metric for this data set as it does not give any relevance to the fact that the spirals are not mixed, it only measures similarity with the true labels. To address this issue, C_v was used ($C_v < 1$ when the spirals were mixed). C_v shows that the spirals are not mixed until 25% KNN for QC_{cov}^{prob}, but they are fragmented into sub-clusters. Length scales greater than 25% KNN make the potential too smooth thus making potential wells mix the spirals.

If guided only by the ALL score in figure 5, two local minima would be selected, the first one at 7.5%KNN and the second one at 35%KNN, keeping E_{th} with the default value (0.001). Both solutions are illustrated in figure 6.

Table 1: Data set #1: Local densities. The supervised solution with best JS matches the unsupervised solution proposed by ALL.

Data #1 Density	*Score*	E_{th}	*%KNN*	*#K*	*ALL*	*JS*	C_v	$\rho_{E_{th}}$	*p-val*
QC_{knn}^{prob}	Best JS	0.001	17.5	4	0.082	0.85	0.94	-0.81	1.5E-5
	Best ALL	0.001	17.5	4	0.082	0.85	0.94	-	-
	ALL stable at high Eth	No	-	-	-	-	-	-	-
QC_{cov}^{prob}	Best JS	0.001	17.5	5	0.092	0.88	0.96	-0.87	6.0E-7
	Best ALL	0.001	17.5	5	0.092	0.88	0.96	-	-
	ALL stable at high Eth	No	-	-	-	-	-	-	-

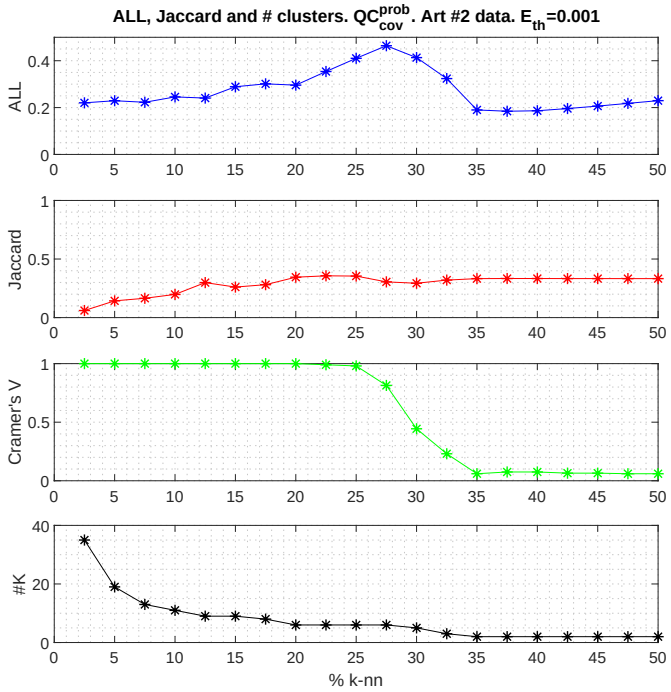
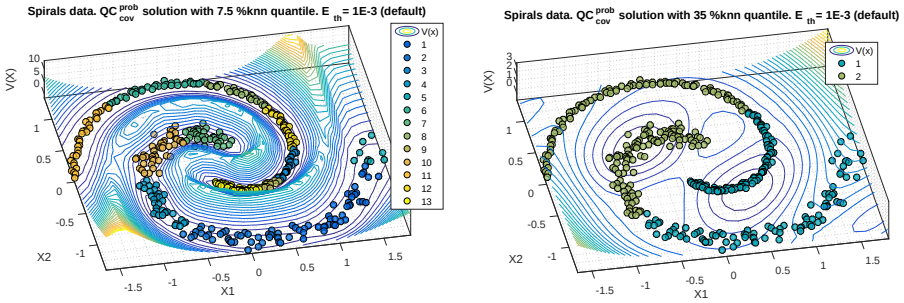


Figure 5: ALL, Jaccard score, C_v and number of clusters obtained by QC_{cov}^{prob} for data set #2. ALL splits the graph into two regions separated by a value of KNN equal to 22.5%; at the left side, the spirals are not mixed but broken up; while at the right side the spirals are mixed but there are only two clusters. Obviously, an external supervision would prefer non-mixed spirals.



(a) Solution for $QC^{prob}_{knn\ 20\%}$ (b) $P(K|X)$

Figure 6: QC^{prob}_{cov} solutions with $E_{th} = 0.001$ (default initial value equal to the precision of the SGD stopping criteria). The left figure uses a 7.5%KNN, here the spirals are not mixed but each one is fragmented into sub-clusters. The right figure uses 35%KNN, where the length scale is too big to preserve the spirals not mixed. There are two clusters but the spirals are mixed. These cases show the need of the extended ALL plots.

In order to find the optimal solution ($JS = 1$), where the spirals are neither mixed nor fragmented, the value of E_{th} should be increased until reaching a region of low ALL values, as shown in figure 7. The best solution depicted in figure 8, is achieved in regions with low values of %KNN (< 20) and high values of E_{th} ($\in [10^{-1}, 10^{0}]$).

Although ALL is not highly correlated with JS along the E_{th} axis direction, a stable region of low ALL with high E_{th} implies an underlying hierarchical structure that produces a good JS. Since the JS is not ideally suited for this data set, the expected inverse correlation with ALL is not present in Table 2. The stability region varies depending on the QC model, but can be inspected visually using the ALL plot.

Table 2: Data set #2: Two spirals. In this case, the supervised solution with best JS (only varying %KNN) has a poor performance without modifying the E_{th} parameter. ALL of the stability region proposes a solution with JS=1.

Data #2 Spirals	*Score*	E_{th}	*%KNN*	*#K*	*ALL*	*JS*	C_v	$\rho_{E_{th}}$	*p-val*
QC^{prob}_{knn}	Best JS	0.001	47.5	1	0.510	0.50	-	0.60	0.005
	Best ALL1	0.001	7.5	14	0.237	0.16	1.00	-	-
	Best ALL2	0.001	35.0	2	0.229	0.33	0.06	-	-
	ALL stable at high Eth	$[0.2, 0.8]$	$[2.5, 10]$	2	6.8E-5	**1.00**	1.00	-	-
QC^{prob}_{cov}	Best JS	0.001	22.5	6	0.354	0.36	0.99	0.19	0.412
	Best ALL1	0.001	7.5	13	0.223	0.17	1.00	-	-
	Best ALL2	0.001	35.0	2	0.190	0.33	0.06	-	-
	ALL stable at high Eth	$[0.5, 1.5]$	$[2.5, 20]$	2	1.0E-5	**1.00**	1.00	-	-

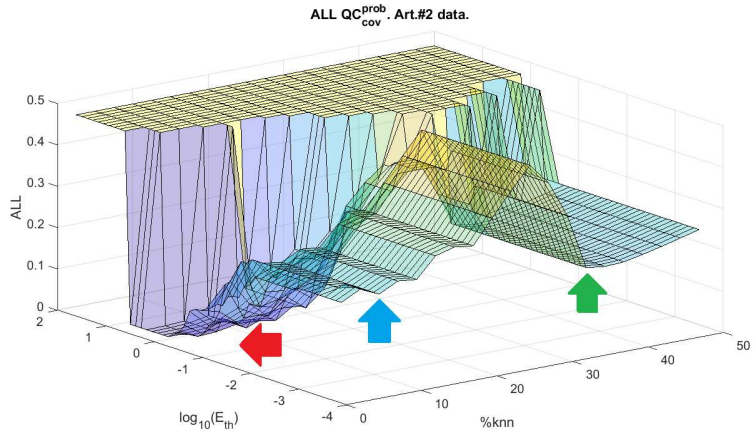


Figure 7: Extended ALL score showing the stability region for high E_{th} values. This region offers a solution based on low length scales where the sub-clusters are merged hierarchically to form the two spirals without being mixed. The ALL plot indicates three regions of interest: local minima with small length scale (blue arrow), local minima with higher length making a too smooth potential (green arrow), and the stable region of high E_{th} offering the most interesting solution (red arrow).

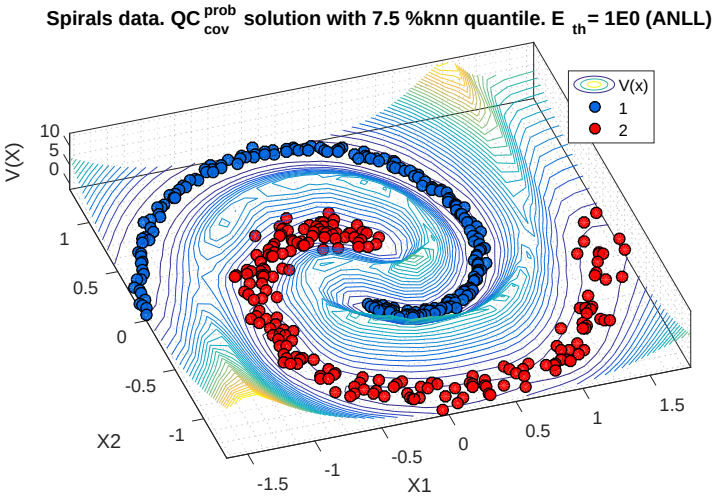


Figure 8: Spiral solution based on the stable region parameters in the extended ALL plot.

4.3. Data set #3: Crabs

For the Crabs' data set, ALL also obtains the appropriate %KNN corresponding with the best JS. Table 3 shows that QC_{knn}^{prob} leads to $Js = 0.70$ and QC_{cov}^{prob} to $Js = 0.74$, respectively. In relation to the extended ALL score there are no stable hierarchical solutions.

Table 3: Data set #3: Crabs. The supervised solution with best JS matches with the unsupervised solution proposed by ALL.

Data #3 Crabs	*Score*	E_{th}	*%KNN*	*#K*	*ALL*	*JS*	C_v	$\rho_{E_{th}}$	*p-val*
QC_{knn}^{prob}	Best JS	0.001	17.5	4	0.110	0.74	0.90	-0.83	5.7E-6
	Best ALL	0.001	17.5	4	0.110	0.74	0.90	-	-
	ALL stable at high Eth	No	-	-	-	-	-	-	-
QC_{cov}^{prob}	Best JS	0.001	15.0	4	0.126	0.70	0.89	-0.88	2.8E-7
	Best ALL	0.001	15.0	4	0.126	0.70	0.89		
	ALL stable at high Eth	No	-	-	-	-	-	-	-

4.4. Data set #4: Olive oil

Table 4 shows the main results for this data set. For the QC_{knn}^{prob}, the first ALL local minimum is closer to the real classification of nine regions but ALL does not identify the best length scale available, as it proposes 7.5%KNN (JS=0.55) instead of 2.5%KNN (JS=0.73). The second ALL local minimum obtains a similar JS to the best possible one, although the length scale is quite different: 22.5%KNN instead of 12.5%KNN. Despite not matching exactly with the highest JS, the information provided by the two minima is of paramount relevance, as they point out the two underlying structures, namely, three and nine clusters. The ALL-JS correlation is quite poor, partly due to the fact that ALL is compared with two different JS curves.

Nonetheless, the QC_{cov}^{prob} clearly outperforms QC_{knn}^{prob}, ALL finds solutions with JS practically as good as the best JS ones, the ALL-JS correlation is better, and the number of clusters is closer to the real one (#K: 4 and 9).

A further detailed explanation can be obtained observing figure 9. The algorithm starts with many sub-clusters with the first KNN; it is important to take into account that dealing with more than 100 clusters is computationally very expensive during the cluster allocation because it has to check many ($100 \cdot 99 = 9900$) possible paths between potential wells (centroids). Then, the number of clusters decreases drastically until obtaining nine clusters in 15% KNN, and it is here where the first local minimum appears in ALL, matching with the highest JS for the structure of nine areas. Then, a subtle local minimum appears at 45% KNN, very close to the highest JS for the structure of three regions of Italy.

Table 4: Data set #4: Olive oil. JS in bold refer to the value that should be compared to the corresponding ANNL, depending on whether the model is a solution of the 3-class or 9-class problem. The supervised solutions with best JS match the unsupervised solutions proposed by ALL, excepting for the QC_{knn}^{prob} ALL1 in JS2.

Data #4 Olive	*Score*	E_{th}	*%KNN*	*#K*	*ALL*	*JS1 JS2*	C_V1 C_V2	$\rho_{E_{th}}$	*p-val*
QC_{knn}^{prob}	Best JS1 3 regions	0.001	12.5	5	0.241	0.77 —	0.83 —	0.08	7.5E-1
	Best JS2 9 regions	0.001	2.5	9	0.167	— 0.73	— 0.98	-0.33	1.5E-1
	Best ALL1	0.001	7.5	5	0.162	0.64 **0.55**	0.85 0.89	-	-
	Best ALL2	0.001	22.5	2	0.230	**0.74** 0.36	0.97 0.95	-	-
	ALL stable at high Eth	No	-	-	-	-	-	-	-
QC_{cov}^{prob}	Best JS1 3 regions	0.001	47.5	4	0.231	0.79 —	0.76 —	-0.67	1.4E-3
	Best JS2 9 regions	0.001	20.0	8	0.187	— 0.73	— 0.76	-0.52	1.9E-2
	Best ALL1	0.001	15.0	9	0.175	0.52 **0.72**	0.99 0.72	-	-
	Best ALL2	0.001	45.0	4	0.220	**0.78** 0.41	0.76 0.81	-	-
	ALL stable at high Eth	No	-	-	-	-	-	-	-

Lastly, there is another ALL minimum at 50% KNN; it is not a real solution but an effect of dealing with very few clusters. The best JS for three regions is $JS = 0.73$, and for nine areas is $JS = 0.79$.

5. Complexity analysis

This work is focused on proposing a new algorithm for cluster allocation with automatic hyperparameter selection, and hence, efficient computation has not been analysed deeply. In fact, possible future works could be related to parallelizing some tasks, like the calculation of the potential function per sample, in order to speed up the process. However, an estimation of the PQC runtime has been carried out. It depends on the following factors:

- m: Sample size.
- d: Space dimensionality.
- $iter_{SGD}$: Number of iterations until SGD convergence.
- K_0: Initial number of potential wells with at least one allocated sample; many of them can be considered as sub-clusters to be merged depending on their potential difference. The shortest path of all-pairs, that has a

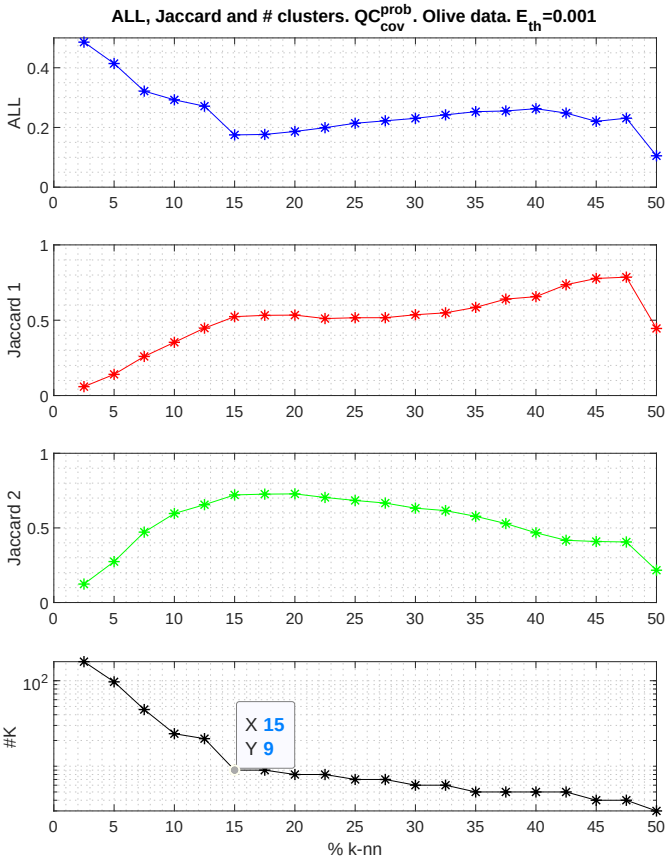


Figure 9: ALL, Jaccard score (three areas: Jaccard 1; nine areas: Jaccard 2) and number of clusters obtained by QC_{cov}^{prob} for the olive oil data set. ALL points out firstly 15% as the most suitable KNNs, and then 45%KNN.

quadratic dependence runtime with K_0, must be used. This factor depends on the σ value, being the greater K_0 the smaller σ. In QC_{cov}^{prob}, the initial K_0 might be quite large because the potential is less smooth and tends to create sub-clusters, although the sub-clusters eventually disappear with the probabilistic allocation; this effect may slow down the algorithm compared to QC_{knn}^{prob}.

In particular, the order of complexity can be described as:

$$\mathcal{O}\left(2 * m \cdot d \cdot iter_{SGD} + K_0^2\right) \tag{23}$$

Additionally, in order to build the ALL score map, that depends on the hyparameters $\%knn$ and E_{th}, the total runtime will be multiplied by the amount of different $\%knn$ values that are scanned. The runtime associated with the different hierarchical solutions based on E_{th} is very short, as all solutions share the same PQC model, which depends only on σ.

The runtime of the experiments shown in section 4 is presented in table 5. It shows the runtime as a function of the length scale; it also indicates the initial sub-clusters of the model, before being merged by having an energy difference $E \leq E_{th}$. This effect has a stronger runtime impact in the case of QC_{cov}^{prob} and small length scales, as one can observe for the dataset #4 between $7\%knn$ with $K_0 = 129$ and $14\%knn$ with $K_0 = 71$.

Table 5: Influence of length scales in the algorithm runtime for the different datasets. Times t are expressed in seconds

Data	Subclusters (K_0)	Length scale ($knn\%$)					Size
	Runtime (s)	7%	14%	21%	28%	35%	Dimension
#1	$K_0\,(QC_{knn})$	15	8	7	5	4	400
	$K_0\,(QC_{cov})$	14	7	8	4	4	
	$t\,(QC_{knn})$	32.5	21.8	18.8	18.2	37.4	2
	$t\,(QC_{cov})$	24.2	33.3	16.1	14.7	17.6	
#2	$K_0\,(QC_{knn})$	14	8	6	4	2	400
	$K_0\,(QC_{cov})$	13	9	8	6	3	
	$t\,(QC_{knn})$	20,4	38,9	36,7	19,3	19,3	2
	$t\,(QC_{cov})$	21.6	25.1	41.6	18.0	15.5	
#3	$K_0\,(QC_{knn})$	12	5	4	4	4	200
	$K_0\,(QC_{cov})$	20	9	4	4	4	
	$t\,(QC_{knn})$	18.9	10.4	9.9	10.5	11.0	2
	$t\,(QC_{cov})$	6.4	4.6	4.2	3.8	3.7	
#4	$K_0\,(QC_{knn})$	8	5	5	3	2	572
	$K_0\,(QC_{cov})$	129	71	41	31	24	
	$t\,(QC_{knn})$	46.3	44.4	139.9	50.4	50.5	8
	$t\,(QC_{cov})$	2267.6	793.5	727.0	557.8	491.8	

Table 6 shows the runtime for the computation of the hierarchical solutions based on E_{th}; in particular, basic statistics of runtime associated with 20 E_{th}

Table 6: Influence of E_{th} in the algorithm runtime for the different datasets. Times are expressed in seconds

QC variant	Data	Runtime (s) of $E_{th} \in [10^{-4}, 10^{2}]$			
		min	max	mean	std
knn	#1	0.046	0.017	0.033	0.013
cov		0.029	0.013	0.023	0.006
knn	#2	0.041	0.016	0.030	0.010
cov		0.042	0.012	0.019	0.007
knn	#3	0.014	0.011	0.012	0.002
cov		0.012	0.005	0.009	0.003
knn	#4	0.108	0.057	0.086	0.024
cov		0.562	0.172	0.438	0.137

values log-spaced between $[10^{-4}, 10^{2}]$. In all cases, the runtime for this step is quite short since no new model is being computed in this process.

6. Conclusion

This paper has presented two main novel contributions within the paradigm of QC. Firstly, an adaptation of the Schrödinger equation to deal with independent local length scales, thus allowing cluster discrimination by density. Secondly, a probabilistic framework for QC to detect the underlying structure in data; it enables outlier detection as well as the delineation of Bayesian optimal cluster boundaries.

This framework leads to a merit function to measure goodness-of-fit in the form of ALL. This utilises a Bayesian framework to enable optimisation of a control parameter for the estimation of local length scales using set percentages of nearest neighbours. Local minima of ALL have empirically shown a high correlation with the highest values of JS. Therefore, we suggest that ALL can become a useful objective performance index for unsupervised learning. Furthermore, the ALL provides useful guidance and insight into QC solutions to detect hierarchical structures in the data.

Two new models for PQC with different levels of computational complexity have been proposed. Attending to its simplicity and versatility QC^{prob}_{knn} may outperform QC^{prob}_{cov} in general. However, QC^{prob}_{cov} may perform better than QC^{prob}_{knn} in data sets with challenging peculiarities.

The main limitation of QC^{prob}_{cov} stems from its less smooth potential functions as local-covariance kernels have less superposition effect than spherical kernels. As a consequence of this:

- QC^{prob}_{cov} needs more iterations than QC^{prob}_{knn} in order to achieve the same SGD convergence .
- QC^{prob}_{cov} tends to create more sub-clusters due to the presence of more local minima. This is not an inconvenience in itself because these sub-clusters

can fit better the data and can be later merged in the cluster allocation process. However, the computation time needed to check all the possible paths between all the centroids may be excessive.

QC methods are well-known to have poor performance for high-dimensional data. The proposed framework shares this inherent limitation, the root of which lies in the ultra-metric nature of Euclidean distances in high dimensions as well as sparsity which causes difficulties for local covariance estimation. This remains an area of further work.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness under project with reference number TIN2014-52033-R supported by European FEDER funds, and by the Spanish Ministry of Education, Culture and Sport under project with reference number PR2015-00217.

References

[1] D. Horn, A. Gottlieb, The method of quantum clustering, in: Advances in neural information processing systems, 2002, pp. 769–776.

[2] N. Nasios, A. G. Bors, Kernel-based classification using quantum mechanics, Pattern Recognition 40 (2006) 875–889.

[3] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, Bayesian data analysis, Vol. 2, Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[4] R. Gray, Vector quantization, IEEE Assp Magazine 1 (2) (1984) 4–29.

[5] B. W. Silverman, Density estimation for statistics and data analysis, Vol. 26, CRC press, 1986.

[6] D. Comaniciu, P. Meer, Distribution free decomposition of multivariate data, Pattern analysis & applications 2 (1) (1999) 22–30.

[7] N. Nasios, A. G. Bors, Finding the number of clusters for nonparametric segmentation, in: Computer Analysis of Images and Patterns, 11^{th} International Conference CAIP, LNCS 3691, 2005, pp. 213–221.

[8] L. Zelnik-Manor, P. Perona, Self-tuning spectral clustering, in: Advances in neural information processing systems, 2005, pp. 1601–1608.

[9] Y. Li, Y. Wang, Y. Wang, L. Jiao, Y. Liu, Quantum clustering using kernel entropy component analysis, Neurocomputing 202 (2016) 36–48.

[10] P. Vincent, Y. Bengio, Manifold parzen windows, in: Advances in Neural Information Processing Systems, 2003, pp. 849–856.

[11] R. V. Casaña-Eslava, I. H. Jarman, P. J. Lisboa, J. D. Martín-Guerrero, Quantum clustering in non-spherical data distributions: Finding a suitable number of clusters, Neurocomputing 268 (2017) 127–141.

[12] M. Du, S. Ding, Y. Xue, Z. Shi, A novel density peaks clustering with sensitivity of local density and density-adaptive metric, Knowledge and Information Systems 59 (2) (2019) 285–309.

[13] X. Xu, S. Ding, Z. Shi, An improved density peaks clustering algorithm with fast finding cluster centers, Knowledge-Based Systems 158 (2018) 65–74.

[14] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the 2nd international conference on Knowledge Discovery and Data mining (KDD'96), AAAI Press, 1996, pp. 226–231.

[15] S. Ding, H. Jia, M. Du, Y. Xue, A semi-supervised approximate spectral clustering algorithm based on hmrf model, Information Sciences 429 (2018) 215–228.

[16] Q. Liu, R. Zhang, R. Hu, G. Wang, Z. Wang, Z. Zhao, An improved path-based clustering algorithm, Knowledge-Based Systems 163 (2019) 69–81.

[17] Y. Zhang, Y. Yang, T. Li, H. Fujita, A multitask multiview clustering algorithm in heterogeneous situations based on lle and le, Knowledge-Based Systems 163 (2019) 776–786.

[18] T. Deng, D. Ye, R. Ma, H. Fujita, L. Xiong, Low-rank local tangent space embedding for subspace clustering, Information Sciences 508 (2020) 1–21.

[19] H. Wang, Y. Yang, B. Liu, H. Fujita, A study of graph-based system for multi-view clustering, Knowledge-Based Systems 163 (2019) 1009–1019.

[20] D. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

[21] S. Ruder, An overview of gradient descent optimization algorithms, arXiv preprint arXiv:1609.04747.

[22] D. Horn, A. Gottlieb, The method of quantum clustering, in: T. G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14, MIT Press, 2001, pp. 769–776.

[23] M. Forina, C. Armanino, S. Lanteri, E. Tiscornia, Food Research and Data Analysis, Applied Science Publishers, 1983.