The Method of Quantum Clustering

David Horn and Assaf Gottlieb
School of Physics and Astronomy
Raymond and Beverly Sackler Faculty of Exact Sciences
Tel Aviv University, Tel Aviv 69978, Israel

Abstract

We propose a novel clustering method that is an extension of ideas inherent to scale-space clustering and support-vector clustering. Like the latter, it associates every data point with a vector in Hilbert space, and like the former it puts emphasis on their total sum, that is equal to the scale-space probability function. The novelty of our approach is the study of an operator in Hilbert space, represented by the Schrödinger equation of which the probability function is a solution. This Schrödinger equation contains a potential function that can be derived analytically from the probability function. We associate minima of the potential with cluster centers. The method has one variable parameter, the scale of its Gaussian kernel. We demonstrate its applicability on known data sets. By limiting the evaluation of the Schrödinger potential to the locations of data points, we can apply this method to problems in high dimensions.

1 Introduction

Methods of data clustering are usually based on geometric or probabilistic considerations [1, 2, 3]. The problem of unsupervised learning of clusters based on locations of points in data-space, is in general ill defined. Hence intuition based on other fields of study may be useful in formulating new heuristic procedures. The example of [4] shows how intuition derived from statistical mechanics leads to successful results. Here we propose a model based on tools that are borrowed from quantum mechanics.

We start out with the scale-space algorithm of [5] that uses a Parzen-window estimator of the probability distribution based on the data. Using a Gaussian kernel, one generates from the N data points in a Euclidean space of dimension d a probability distribution given by, up to an overall normalization, the expression

$$\psi(\mathbf{x}) = \sum_i e^{-\frac{(\mathbf{x}-\mathbf{x}_i)^2}{2\sigma^2}} \tag{1}$$

where $\mathbf{x}_i$ are the data points. It seems quite natural [5] to associate maxima of this function with cluster centers.

The same kind of Gaussian kernel was the basis of another method, Support Vector Clustering (SVC) [6], associating the N data-points $\mathbf{x}_i$ with vectors in an abstract Hilbert space.

Here we will also consider a Hilbert space, but, in contradistinction with kernel methods where the Hilbert space is implicit, here we work with a Schrödinger equation that serves as the basic framework of the Hilbert space. Our method was introduced in [7] and is further expanded in this presentation. Its main emphasis is on the Schrödinger potential, whose minima will determine the cluster centers. This potential is part of the Schrödinger equation that ψ is a solution of.

2 The Schrödinger Potential

We define[7] the Schrödinger equation

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

for which $\psi(\mathbf{x})$ is a solution, or eigenstate.[1] The simplest case is that of a single Gaussian, when ψ represents a single point at $\mathbf{x}_1$. Then it turns out that $V = \frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{x}_1)^2$. This quadratic function, whose center lies at $\mathbf{x}_1$, is known as the harmonic potential in quantum mechanics (see, e.g., [8]). Its eigenvalue $E = d/2$ is the lowest possible eigenvalue of H, hence the Gaussian function is said to describe the ground state of H.

Conventionally, in quantum mechanics, one is given $V(\mathbf{x})$ and one searches for solutions, or eigenfunctions, $\psi(\mathbf{x})$. Here, we have already $\psi(\mathbf{x})$, as determined by the data points, we ask therefore for the $V(\mathbf{x})$ whose solution is the given $\psi(\mathbf{x})$. This can be easily obtained through

$$V(\mathbf{x}) = E + \frac{\frac{\sigma^2}{2}\nabla^2\psi}{\psi} = E - \frac{d}{2} + \frac{1}{2\sigma^2\psi}\sum_i(\mathbf{x}-\mathbf{x}_i)^2 e^{-\frac{(\mathbf{x}-\mathbf{x}_i)^2}{2\sigma^2}} \, . \tag{3}$$

E is still left undefined. For this purpose we require V to be positive definite, i.e. min$V$=0. This sets the value of

$$E = -\min\frac{\frac{\sigma^2}{2}\nabla^2\psi}{\psi} \tag{4}$$

and determines $V(\mathbf{x})$ uniquely. Using Eq. 3 it is easy to prove that

$$0 < E \leq \frac{d}{2} \, . \tag{5}$$

3 2D Examples

3.1 Crab Data

To show the power of our new method we discuss the crab data set taken from Ripley's book [9]. This data set is defined over a five-dimensional parameter space. When analyzed in terms of the 2nd and 3rd principal components of the correlation matrix one observes a nice separation of the 200 instances into their four classes. We start therefore with this problem as our first test case. In Fig. 1 we show the data as well as the Parzen probability distribution $\psi(\mathbf{x})$ using the width parameter $\sigma = 1/\sqrt{2}$. It is quite obvious that this width is not small enough to deduce the correct clustering according to the approach of [5]. Nonetheless, the potential displayed in Fig. 2 shows the required four minima for the same width parameter. Thus we conclude that the necessary information is already available. One needs, however, the quantum clustering approach, to bring it out.

[1] H (the Hamiltonian) and V (potential energy) are conventional quantum mechanical operators, rescaled so that H depends on one parameter, σ. E is a (rescaled) energy eigenvalue in quantum mechanics.

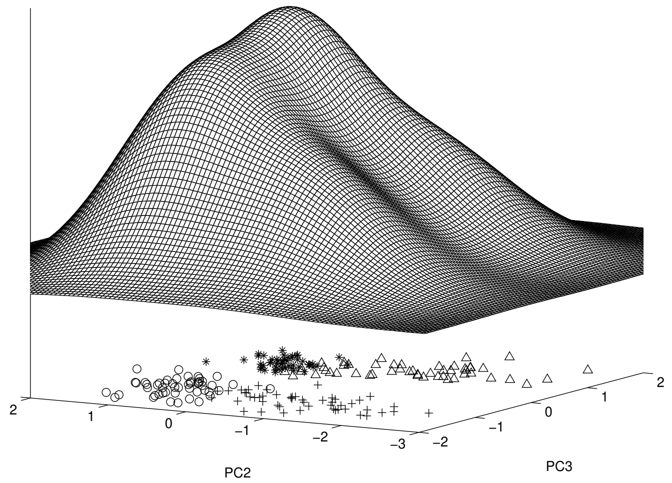

Figure 1: A plot of Roberts' probability distribution for Ripley's crab data [9] as defined over the 2nd and 3rd principal components of the correlation matrix. Using a Gaussian width of $\sigma = 1/\sqrt{2}$ we observe only one maximum. Different symbols label the four classes of data.

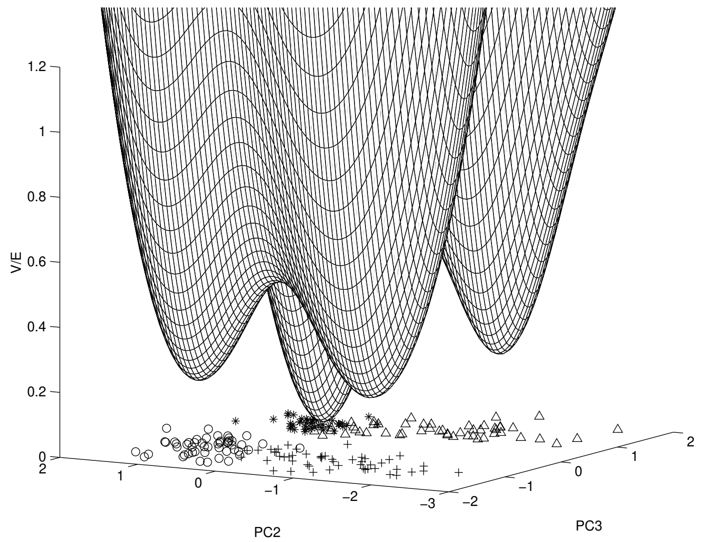

Figure 2: A plot of the Schrödinger potential for the same problem as Fig. 1. Here we clearly see the required four minima. The potential is plotted in units of E.

Note in Fig. 2 that the potential grows quadratically outside the domain over which the data are located. This is a general property of Eq. 3. E sets the relevant scale over which one may look for structure of the potential. If the width is decreased more structure is to be expected. Thus, for $\sigma = 1/2$, two more minima appear, as seen in Fig. 3. Nonetheless, they lie high and contain only a few data points. The major minima are the same as in Fig. 2.

3.2 Iris Data

Our second example consists of the iris data set [10], which is a standard benchmark obtainable from the UCI repository [11]. Here we use the first two principal components to define the two dimensions in which we apply our method. Fig. 4, which shows the case for $\sigma = 0.25$, provides an almost perfect separation of the 150 instances into the three classes into which they should belong.

4 Application of Quantum Clustering

The examples displayed in the previous section show that, if the spatial representation of the data allows for meaningful clustering using geometric information, quantum clustering (QC) will do the job. There remain, however, several technical questions to be answered: What is the preferred choice of σ? How can QC be applied in high dimensions? How does one choose the appropriate space, or metric, in which to perform the analysis? We will confront these issues in this section.

4.1 Varying σ

In the crabs-data we find that as σ is decreased to $\frac{1}{2}$, the previous minima of $V(\mathbf{x})$ get deeper and two new minima are formed. However the latter are insignificant, in the sense that they lie at high values (of order E), as shown in Fig. 3. Thus, if we classify data-points to clusters according to their topographic location on the surface of $V(\mathbf{x})$, roughly the same clustering assignment is expected for $\sigma = \frac{1}{2}$ as for $\frac{1}{\sqrt{2}}$. By the way, the wave function acquires only one additional maximum at $\sigma = \frac{1}{2}$. As σ is being further decreased, more and more maxima are expected in ψ and an ever increasing number of minima (limited by N) in V.

The one parameter of our problem, σ, signifies the distance that we probe. Accordingly we expect to find clusters relevant to proximity information of the same order of magnitude. One may therefore vary σ continuously and look for stability of cluster solutions, or limit oneself to relatively high values of σ and decide to stop the search once a few clusters are being uncovered.

4.2 Higher Dimensions

In the iris problem we obtained excellent clustering results using the first two principal components, whereas in the crabs problem, clustering that depicts correctly the classification necessitates components 2 and 3. However, once this is realized, it does not harm to add the 1st component. This requires working in a 3-dimensional space, spanned by the three leading PCs. Calculating $V(\mathbf{x})$ on a fine computational grid becomes a heavy task in high dimensions. To cut down complexity, we propose using the analytic expression of Eq. 3 and evaluating the potential on data points only. This should be good enough to give a close estimate of where the minima lie, and it reduces the complexity to N^2 irrespective of dimension. In the gradient-descent algorithm described below, we will require further computations, also restricted to well defined locations in space.

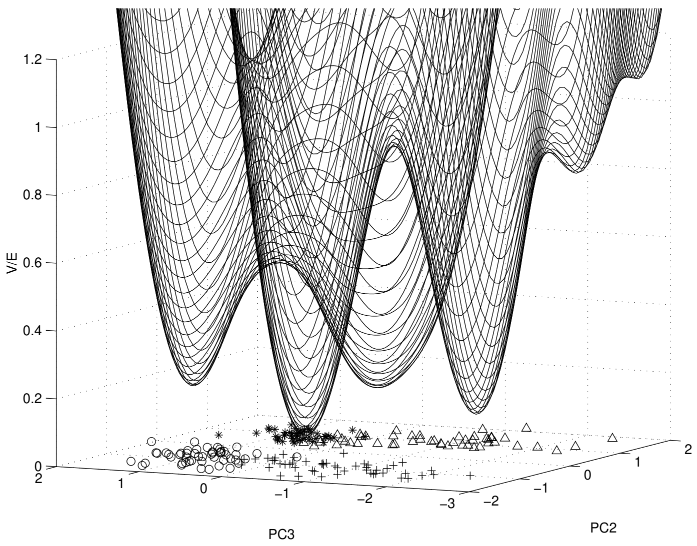

Figure 3: The potential for the crab data with $\sigma = 1/2$ displays two additional, but insignificant, minima. The four deep minima are roughly at the same locations as in Fig. 2.

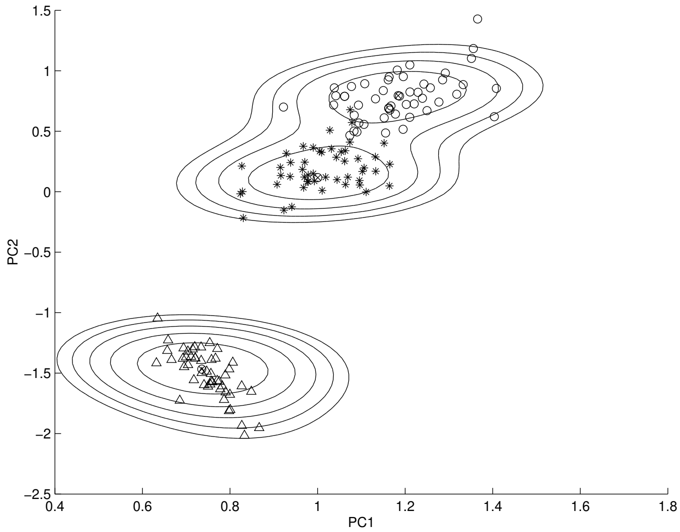

Figure 4: Quantum clustering of the iris data for $\sigma = 0.25$ in a space spanned by the first two principal components. Different symbols represent the three classes. Equipotential lines are drawn at $V(\mathbf{x})/E = .2, .4, .6, .8, 1$.

When restricted to the locations of data points, we evaluate V on a discrete set of points $V(\mathbf{x}_i) = V_i$. We can then express V in terms of the distance matrix $D_{ij} = |\mathbf{x}_i - \mathbf{x}_j|$ as

$$V_i = E - \frac{d}{2} + \frac{1}{2\sigma^2} \frac{\sum_j D_{ij}^2 e^{-\frac{D_{ij}^2}{2\sigma^2}}}{\sum_j e^{-\frac{D_{ij}^2}{2\sigma^2}}} \tag{6}$$

with E chosen appropriately so that minV_i=0.

All problems that we have used as examples were such that data were given in some space, and we have exercised our freedom to define a metric, using the PCA approach, as the basis for distance calculations. The previous analysis tells us that QC can also be applied to data for which only the distance information is known.

4.3 Principal Component Metrics

The QC algorithm starts from distance information. The question how the distances are calculated is another - very important - piece of the clustering procedure. The PCA approach defines a metric that is intrinsic to the data, determined by their second order statistics. But even then, several possibilities exist, leading to non-equivalent results.

Principal component decomposition can be applied both to the correlation matrix and to the covariance matrix. Moreover, whitening normalization may be applied. The PCA approach that we have used is based on a whitened correlation matrix. This turns out to lead to the good separation of crab-data in PC2-PC3 and of iris-data in PC1-PC2. Since our aim was to convince the reader that once a good metric is found, QC conveys the correct information, we have used the best preprocessing before testing QC.

5 The Gradient Descent Algorithm

After discovering the cluster centers we are faced with the problem of allocating the data points to the different clusters. We propose using a gradient descent algorithm for this purpose. Defining $\mathbf{y}_i(0) = \mathbf{x}_i$ we define the process

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

letting the points $\mathbf{y}_i$ reach an asymptotic fixed value coinciding with a cluster center. More sophisticated minimum search algorithms, as given in chapter 10 of [12], may be used for faster convergence.

To demonstrate the results of this algorithm, as well as the application of QC to higher dimensions, we analyze the iris data in 4 dimensions. We use the original data space with only one modification: all axes are normalized to lie within a unified range of variation. The results are displayed in Fig. 5. Shown here are different windows for the four different axes, within which we display the values of the points after descending the potential surface and reaching its minima, whose V values are shown in the fifth window. These results are very satisfactory, having only 5 misclassifications. Applying QC to data space without normalization of the different axes, leads to misclassifications of the order of 15 instances, similar to the clustering quality of [4].

6 Discussion

In the literature of image analysis one often looks for the curve on which the Laplacian of the Gaussian filter of an image vanishes[13]. This is known as zero-crossing and serves as

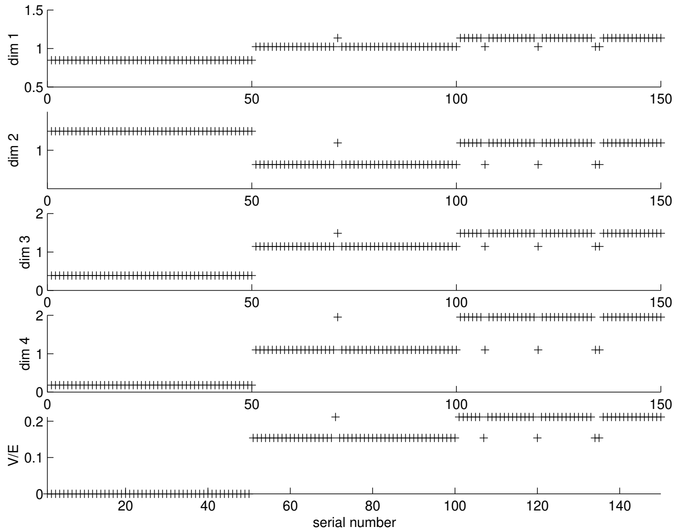

Figure 5: The fixed points of the four-dimensional iris problem following the gradient-descent algorithm. The results show almost perfect clustering into the three families of 50 instances each for $\sigma = 0.21$.

a measure of segmentation of the image. Its analogue in the scale-space approach is where $\nabla^2\psi(\mathbf{x}) = 0$. Clearly each such contour can also be viewed as surrounding maxima of the probability function, and therefore representing some kind of cluster boundary, although different from the conventional one [5]. It is known that the number of such boundaries [13] is a non-decreasing function of σ. Note that such contours can be read off Fig. 4. Comparison with Eq. 3 tells us that they are the $V(\mathbf{x}) = E$ contours on the periphery of this figure. Clearly they surround the data but do not give a satisfactory indication of where the clusters are. Cluster cores are better defined by $V = 0.4E$ curves in this figure. One may therefore speculate that equipotential levels of V may serve as alternatives to $\nabla^2\psi = 0$ curves in future applications to image analysis.

Image analysis is a 2-dimensional problem, in which differential operations have to be formulated and followed on a fine grid. Clustering is a problem that may occur in any number of dimensions. It is therefore important to develop a tool that can deal with it accordingly. Since the Schrödinger potential, the function that plays the major role in our analysis, has minima that lie in the neighborhood of data points, we find that it suffices to evaluate it at these points. This enables us to deal with clustering in high dimensional spaces. The results, such as the iris problem of Fig. 5, are very promising. They show that the basic idea, as well as the gradient-descent algorithm of data allocation to clusters, work well.

Quantum clustering does not presume any particular shape or any specific number of clusters. It can be used in conjunction with other clustering methods. Thus one may start with SVC to define outliers which will be excluded from the construction of the QC potential. This would be one example where not all points are given the same weight in the construction of the Parzen probability distribution.

It may seem strange to see the Schrödinger equation in the context of machine learning. Its usefulness here is due to the fact that the two different terms of Eq. 2 have opposite effects on the wave-function. The potential represents the attractive force that tries to concentrate the distribution around its minima. The Laplacian has the opposite effect of spreading the wave-function. In a clustering analysis we implicitly assume that two such effects exist. QC models them with the Schrödinger equation. Its success proves that this equation can serve as the basic tool of a clustering method.

References

[1] A.K. Jain and R.C. Dubes. *Algorithms for clustering data*. Prentice Hall, Englewood Cliffs, NJ, 1988.

[2] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego, CA, 1990.

[3] R.O. Duda, P.E. Hart and D.G. Stork. *Pattern Classification*. Wiley-Interscience, 2nd ed., 2001.

[4] M. Blat, S. Wiseman and E. Domany. Super-paramagnetic clustering of data. *Phys. Rev. Letters* 76:3251-3255, 1996.

[5] S.J. Roberts. Non-parametric unsupervised cluster analysis. *Pattern Recognition*, 30(2):261–272, 1997.

[6] A. Ben-Hur, D. Horn, H.T. Siegelmann, and V. Vapnik. A Support Vector Method for Clustering. in *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference* Todd K. Leen, Thomas G. Dietterich and Volker Tresp eds., MIT Press 2001, pp. 367–373.

[7] David Horn and Assaf Gottlieb. Algorithm for Data Clustering in Pattern Recognition Problems Based on Quantum Mechanics. Phys. Rev. Lett. 88 (2002) 018702.

[8] S. Gasiorowicz. *Quantum Physics*. Wiley 1996.

[9] B. D. Ripley *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge UK, 1996.

[10] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annual Eugenics*, 7:179–188, 1936.

[11] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.

[12] W. H. Press, S. A. Teuklosky, W. T. Vetterling and B. P. Flannery. *Numerical Recipes - The Art of Scientific Computing* 2nd ed. Cambridge Univ. Press, 1992.

[13] A. L. Yuille and T. A. Poggio. Scaling theorems for zero crossings. *IEEE Trans. Pattern Analysis and Machine Intelligence* PAMI-8, 15-25, 1986.