Algorithm for Data Clustering in Pattern Recognition Problems Based on Quantum Mechanics

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
(Received 16 July 2001; published 20 December 2001)

We propose a novel clustering method that is based on physical intuition derived from quantum mechanics. Starting with given data points, we construct a scale-space probability function. Viewing the latter as the lowest eigenstate of a Schrödinger equation, we use simple analytic operations to derive a potential function whose minima determine cluster centers. The method has one parameter, determining the scale over which cluster structures are searched. We demonstrate it on data analyzed in two dimensions (chosen from the eigenvectors of the correlation matrix). The method is applicable in higher dimensions by limiting the evaluation of the Schrödinger potential to the locations of data points.

DOI: 10.1103/PhysRevLett.88.018702 PACS numbers: 89.75.Kd, 02.70.–c, 03.65.Ge, 03.67.Lx

Clustering of data is a well-known problem of pattern recognition, covered in textbooks such as [1–3]. The problem we are looking at is defining clusters of data solely by the proximity of data points to one another. This problem is one of unsupervised learning, and is in general ill defined. Solutions to such problems can be based on intuition derived from physics. A good example of the latter is the algorithm by [4] that is based on associating points with Potts spins and formulating an appropriate model of statistical mechanics. We propose an alternative that is also based on physical intuition, this one being derived from quantum mechanics.

As an introduction to our approach we start with the scale-space algorithm by [5] who uses a Parzen-window estimator [3] of the probability distribution leading to the data at hand. The estimator is constructed by associating a Gaussian with each of the N data points in a Euclidean space of dimension d and summing over all of them. This can be represented, up to an overall normalization, by

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

where $\mathbf{x}_i$ are the data points. Roberts [5] views the maxima of this function as determining the locations of cluster centers.

An alternative, and somewhat related, method is support vector clustering (SVC) [6] that is based on a Hilbert-space analysis. In SVC, one defines a transformation from data space to vectors in an abstract Hilbert space. SVC proceeds to search for the minimal sphere surrounding these states in Hilbert space. We will also associate data points with states in Hilbert space. Such states may be represented by Gaussian wave functions, whose sum is $\psi(\mathbf{x})$. This is the starting point of our quantum clustering (QC) method. We will search for the Schrödinger potential for which $\psi(\mathbf{x})$ is a ground state. The minima of the potential define our cluster centers.

The Schrödinger potential.—We wish to view ψ as an eigenstate of the Schrödinger equation

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

Here we rescaled H and V of the conventional quantum mechanical equation to leave only one free parameter, σ. For comparison, the case of a single point at $\mathbf{x}_1$ corresponds to Eq. (2) with $V = \frac{1}{2\sigma^2}(\mathbf{x} - \mathbf{x}_1)^2$ and $E = d/2$, thus coinciding with the ground state of the harmonic oscillator in quantum mechanics.

Given ψ for any set of data points we can solve Eq. (2) for V:

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

Let us furthermore require that $\min V = 0$. This sets the value of

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

and determines $V(\mathbf{x})$ uniquely. E has to be positive since V is a non-negative function. Moreover, since the last term in Eq. (3) is positive definite, it follows that

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

We note that ψ is positive definite. Hence, being an eigenfunction of the operator H in Eq. (2), its eigenvalue E is the lowest eigenvalue of H, i.e., it describes the ground state. All higher eigenfunctions have nodes whose numbers increase as their energy eigenvalues increase. (In quantum mechanics, where one interprets $|\psi|^2$ as the probability distribution, all eigenfunctions of H have physical meaning. Although this approach could be adopted, we have chosen ψ as the probability distribution because of the simplicity of algebraic manipulations.)

Given a set of points defined within some region of space, we expect $V(\mathbf{x})$ to grow quadratically outside this

region, and to exhibit one or several local minima within the region. We identify these minima with cluster centers, which seems natural in view of the opposite roles of the two terms in Eq. (2): Given a potential function, it attracts the data distribution function ψ to its minima, while the Laplacian drives it away. The diffused character of the distribution is the balance of the two effects.

As an example we display results for the crab data set taken from Ripley's book [7]. These data, given in a five-dimensional parameter space, show nice separation of the four classes contained in them when displayed in two dimensions spanned by the second and third principal components [8] (eigenvectors) of the correlation matrix of the data. The information supplied to the clustering algorithm contains only the coordinates of the data points. We display the correct classification to allow for visual comparison of the clustering method with the data. Starting with $\sigma = 1/\sqrt{2}$ we see in Fig. 1 that the Parzen probability distribution, or the wave-function ψ, has only a single maximum. Nonetheless, the potential, displayed in Fig. 2, already shows four minima at the relevant locations. The overlap of the topographic map of the potential with the true classification is quite amazing. The minima are the centers of attraction of the potential, and they are clearly evident although the wave function does not display local maxima at these points. The fact that $V(\mathbf{x}) = E$ lies above the range where all valleys merge explains why $\psi(\mathbf{x})$ is smoothly distributed over the whole domain.

As σ is being decreased more minima will appear in $V(\mathbf{x})$. For the crab data, we find two new minima as σ is decreased to one-half. Nonetheless, the previous minima become deeper and still dominate the scene. The new minima are insignificant, in the sense that they lie at high values (of order E). Classifying data points to clusters according to their topographic location on the surface of $V(x)$, roughly the same clustering assignment is expected for a range of σ values. One important advantage of quantum clustering is that E sets the scale on which minima are observed. Thus, we learn from Fig. 2 that the cores of all 4 clusters can be found at V values below $0.4E$. In comparison, the additional maxima of ψ, which start to appear at lower values of σ, may lie much lower than the leading maximum and may be hard to locate numerically.

Principal component analysis (PCA).— In our example, data were given in some high-dimensional space and we analyzed them after defining a projection and a metric, using the PCA approach. The latter defines a metric that is intrinsic to the data, determined by second order statistics. But, even then, several possibilities exist, leading to nonequivalent results.

Principal component decomposition can be applied both to the correlation matrix $C_{\alpha\beta} = \langle x_\alpha x_\beta \rangle$ and to the covariance matrix

$$C_{\alpha\beta} = \langle\langle (x_\alpha - \langle x \rangle_\alpha)(x_\beta - \langle x \rangle_\beta) \rangle\rangle = C_{\alpha\beta} - \langle x \rangle_\alpha \langle x \rangle_\beta \,. \tag{6}$$

In both cases averaging is performed over all data points, and the indices indicate spatial coordinates from 1 to d. The principal components are the eigenvectors of these matrices. Thus we have two natural bases in which to represent the data. Moreover, one often renormalizes the eigenvector projections, dividing them by the square roots of their eigenvalues. This procedure is known as "whitening," leading to a renormalized correlation or covariance matrix of unity. This is a scale-free representation that would naturally lead one to start with $\sigma = 1$ in the search for (higher order) structure of the data.

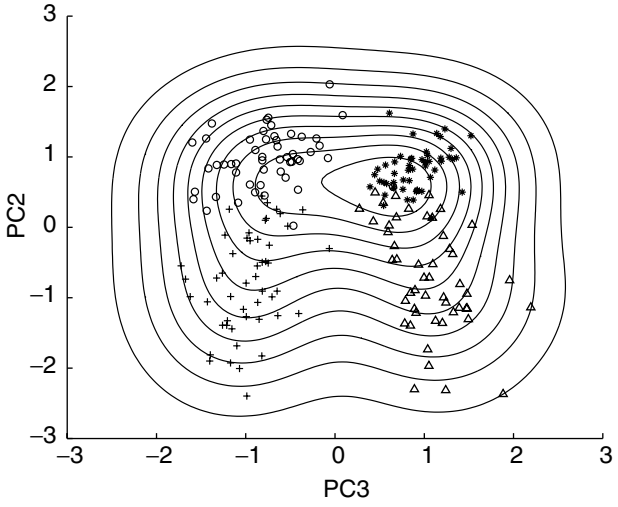

FIG. 1. Ripley's crab data [7] displayed on a plot of their second and third principal components with a superimposed topographic map of Roberts' probability distribution for $\sigma = 1/\sqrt{2}$.

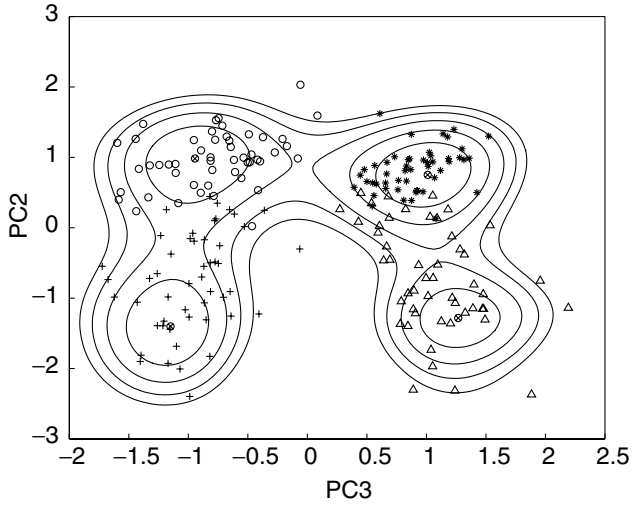

FIG. 2. A topographic map of the potential for the crab data with $\sigma = 1/\sqrt{2}$, displaying four minima (denoted by crossed circles) that are interpreted as cluster centers. The contours of the topographic map are set at values of $V(\mathbf{x})/E = 0.2, 0.4, 0.6, 0.8, 1$.

The PCA approach that we have used in our example was based on whitened correlation matrix projections. Had we used the covariance matrix instead, we would get similar, but slightly worse, separation of the crab data. Our example is meant to convince the reader that once a good metric is found, QC conveys the correct information. Hence we allowed ourselves to search first for the best geometric representation, and then apply QC.

QC in higher dimensions.—Increasing dimensionality means higher computational complexity, often limiting the applicability of a numerical method. Nonetheless, here we can overcome this "curse of dimensionality" by limiting ourselves to evaluating V at locations of data points only. Since we are interested in where the minima lie, and since invariably they lie near data points, no harm is done by this limitation. The results are depicted in Fig. 3. Here we analyzed the crab problem in a three-dimensional (3D) space, spanned by the first three PCs. Shown in this figure are V/E values as functions of the serial number of the data, using the same symbols as in Fig. 2 to allow for comparison. Using all data of $V < 0.3E$, one obtains cluster cores that are well separated in space, corresponding to the four classes that exist in the data. Only 9 of the 129 points that obey $V < 0.3E$ are misclassified by this procedure. Adding higher PCs, first component 4 and then component 5, leads to deterioration in clustering quality. In particular, lower cutoffs in V/E, including lower fractions of data, are required to define cluster cores that are well separated in their relevant spaces.

One may locate the cluster centers, and deduce the clustering allocation of the data, by following the dynamics of gradient descent into the potential minima. By defining $\mathbf{y}_i(0) = \mathbf{x}_i$, one follows the steps of $\mathbf{y}_i(t + \Delta t) = \mathbf{y}_i(t) - \eta(t)\nabla V(\mathbf{y}_i(t))$, letting the points $\mathbf{y}_i$ reach an asymptotic fixed value coinciding with a cluster center. More sophisticated minimum search algorithms (see, e.g., chapter 10 in [9]) can be applied to reach the fixed points faster. The results of a gradient-descent procedure, applied to the 3D analysis of the crab data shown in Fig. 3, are that the three classes of data points 51 to 200 are clustered correctly with only five misclassifications. The first class, data points 1–50, has 31 points forming a new cluster, with most of the rest joining the cluster of the second class. Only 3 points of the first class fall outside the 4 clusters.

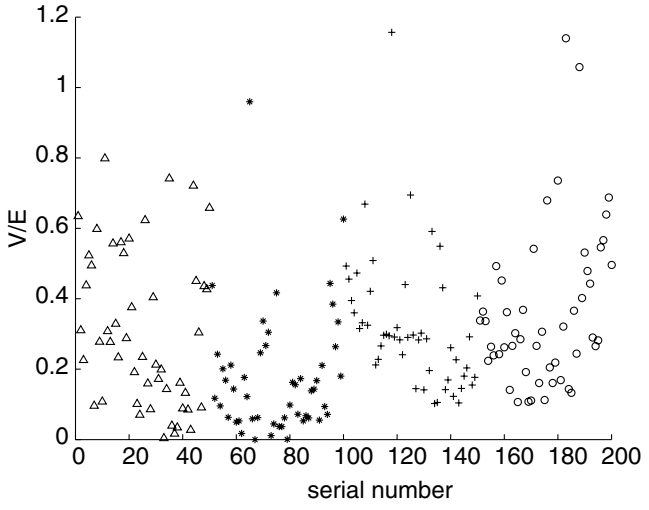

FIG. 3. Values of $V(\mathbf{x})/E$ are depicted in the crab problem with three leading PCs for $\sigma = 1/2$. They are presented as a function of the serial number of the data, using the same symbols of data employed previously. One observes low lying data of all four classes.

We also ran our method on the iris data set [10], which is a standard benchmark obtainable from the UC Irvine (UCI) repository [11]. The data set contains 150 instances, each composed of four measurements of an iris flower. There are three types of flowers, represented by 50 instances each. Clustering of these data in the space of the first two principal components, using $\sigma = 1/4$, has the amazing result of misclassification of 3 instances only. Quantum clustering can be applied to the raw data in four dimensions, leading to misclassifications of the order of 15 instances, similar to the clustering quality of [4].

Distance-based QC formulation.—Gradient descent calls for the calculation of V both on the original data points as well as on the trajectories they follow. An alternative approach can be to restrict oneself to the original values of V, as in the example displayed in Fig. 3, and follow a hybrid algorithm to be described below. Before turning to such an algorithm let us note that, in this case, 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

$$V_i = E - \frac{d}{2} + \frac{1}{2\sigma^2}\frac{\sum_j D_{ij}^2 e^{-D_{ij}^2/2\sigma^2}}{\sum_j e^{-D_{ij}^2/2\sigma^2}} \qquad (7)$$

with E chosen appropriately so that $\min V_i = 0$. This type of formulation is of particular importance if the original information is given in terms of distances between data points rather than their locations in space. In this case we have to proceed with distance information only.

By applying QC we can reach results such as in Fig. 3 without invoking any explicit spatial distribution of the points in question. One may then analyze the results by choosing a cutoff, e.g., $V < 0.2E$, such that a fraction (e.g., one-third) of the data will be included. On this subset we select groups of points whose distances from one another are smaller than, e.g., 2σ, thus defining cores of clusters. Then we continue with higher values of V, e.g., $0.2E < V < 0.4E$, allocating points to previous clusters or forming new cores. Since the choice of distance cutoff in cluster allocation is quite arbitrary, this method cannot be guaranteed to work as well as the gradient-descent approach.

Generalization.—Our method can be easily generalized to allow for different weighting of different points, as in

$$\psi(\mathbf{x}) = \sum_i c_i e^{-(\mathbf{x}-\mathbf{x}_i)^2/2\sigma^2} \tag{8}$$

with $c_i \geq 0$. This is important if we have some prior information or some other means for emphasizing or deemphasizing the influence of data points. An example of the latter is using QC in conjunction with SVC [6]. SVC has the possibility of labeling points as outliers. This is done by applying quadratic maximization to the Lagrangian

$$W = 1 - \sum_{i,j} \beta_i \beta_j e^{-(\mathbf{x}_i-\mathbf{x}_j)/2\sigma^2} \tag{9}$$

over the space of all $0 \leq \beta_i \leq \frac{1}{pN}$ subject to the constraint $\sum_i \beta_i = 1$. The points for which the upper bound of β_i is reached are labeled as outliers. Their number is regulated by p, being limited by pN. Using for the QC analysis a choice of $c_i = \frac{1}{pN} - \beta_i$ will eliminate the outliers of SVC and emphasize the role of the points expected to lie within the clusters.

Discussion.—QC constructs a potential function $V(\mathbf{x})$ on the basis of data points, using one parameter, σ, that controls the width of the structures that we search for. The advantage of the potential V over the scale-space probability distribution is that the minima of the former are better defined (deep and robust) than the maxima of the latter. However, both of these methods put the emphasis on cluster centers, rather than, e.g., cluster boundaries. Since the equipotentials of V may take arbitrary shapes, the clusters are not spherical, as in the k-means approach. Nonethelss, spherical clusters appear more naturally than, e.g., ring-shaped or toroidal clusters, even if the data would accommodate them. If some global symmetry is to be expected, e.g., global spherical symmetry, it can be incorporated into the original Schrödinger equation defining the potential function.

QC can be applied in high dimensions by limiting the evaluation of the potential, given as an explicit analytic expression of Gaussian terms, to locations of data points only. Thus the complexity of evaluating V_i is of order N^2 independent of dimensionality.

Our algorithm has one free parameter, the scale σ. In all examples we confined ourselves to scales that are of order 1, because we have worked within whitened PCA spaces. If our method is applied to a different data space, the range of scales to be searched for could be determined by some other prior information.

Since the strength of our algorithm lies in the easy selection of cluster cores, it can be used as a first stage of a hybrid approach employing other techniques after the identification of cluster centers. The fact that we do not have to take care of feeble minima, but consider only robust deep minima, turns the identification of a core into an easy problem. Thus, an approach that drives its rationale from physical intuition in quantum mechanics can lead to interesting results in the field of pattern classification.

We thank B. Reznik for a helpful discussion.

[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, New York, 2001), 2nd ed.
[4] M. Blat, S. Wiseman, and E. Domany, Phys. Rev. Lett. **76**, 3251 (1996).
[5] S. J. Roberts, Pattern Recognit. **30**, 261 (1997).
[6] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, in *Proceedings of the Conference on Advances in Neural Information Processing Systems 13, 2000,* edited by Todd K. Leen, Thomas G. Dietterich, and Volker Tresp (MIT Press, Cambridge, MA, 2001), p. 367.
[7] B. D. Ripley, *Pattern Recognition and Neural Networks* (Cambridge University Press, Cambridge, UK, 1996).
[8] I. T. Jolliffe, *Principal Component Analysis* (Springer-Verlag, New York, 1986).
[9] W. H. Press, S. A. Teuklosky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes—The Art of Scientific Computing* (Cambridge University, Cambridge, UK, 1992), 2nd ed.
[10] R. A. Fisher, Ann. Eugenics **7**, 179 (1936).
[11] C. L. Blake and C. J. Merz, UCI repository of machine learning databases, 1998.