Novel clustering algorithm for microarray expression data in a truncated SVD space

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

October 30, 2002

ABSTRACT

Motivation: This paper introduces the application of a novel clustering method to microarray expression data. Its first stage involves compression of dimensions that can be achieved by applying SVD to the gene-sample matrix in microarray problems. Thus the data (samples or genes) can be represented by vectors in a truncated space of low dimensionality, 4 and 5 in the examples studied here. We find it preferable to project all vectors onto the unit sphere before applying a clustering algorithm. The clustering algorithm used here is the quantum clustering method that has one free scale parameter. Although the method is not hierarchical, it can be modified to allow hierarchy in terms of this scale parameter.

Results: We apply our method to three data sets. The results are very promising. On cancer cell data we obtain a dendrogram that reflects correct groupings of cells. In an AML/ALL data set we obtain very good clustering of samples into four classes of the data. Finally, in clustering of genes in yeast cell cycle data we obtain four groups in a problem that is estimated to contain five families.

Availability: http://neuron.tau.ac.il/~horn

Contact: horn@post.tau.ac.il

INTRODUCTION

Several authors have recently shown that, by using SVD, one can extract biologically interesting results from the gene/sample matrix X of microarray data. The relevant formalism is analyzed in detail by Alter *et al.* (2000) and will be outlined below. Similar approaches were used by Rayachudhuri *et al.* (2000) and Holter *et al.* (2000). The underlying idea is that the main features of the data presented by a huge association matrix, such as the gene/sample matrix in question, can be captured by a highly compressed form of the matrix. This compressed, or truncated, representation can be expressed in terms of eigengenes and eigenarrays (Alter *et al.* , 2000), that form eigenvectors (of X^TX and XX^T respectively) with leading eigenvalues. Thus, in the yeast cell-cycle data (Spellman *et al.* 1998), one can trace the correct temporal behavior in the relevant eigengenes.

We propose to use SVD in a different manner, as a preprocessing step of a clustering algorithm. The idea is that, rather than performing clustering on the initial data presented by the gene/sample matrix, we compress it first and perform clustering on the data as described in the truncated space. This approach is based on the experience gained in the community that studies texts using the latent semantic analysis (LSA) approach (Landauer *et al.* , 1998). The truncated space, known as concept space in the analysis of word/document association matrices, is known to be better suited for information retrieval purposes than the original matrix. Taking a similar attitude toward the gene/sample association matrix, it seems plausible that clustering after truncation should lead to meaningful answers. A similar approach has been recently proposed by Ding *et al.* (2002), who performed the clustering in the truncated space using the k-means algorithm. We perform this second stage of data analysis using the recently proposed method of Quantum Clustering (Horn and Gottlieb, 2002) and compare its results to k-means.

ALGORITHMS

Singular Value Decomposition (SVD)

Our study concerns an $m \times n$ gene/sample matrix X. Its columns may be interpreted as sample vectors defined in gene-space, and its rows are gene-vectors in sample space. This matrix of rank $k \leq \min(m,n)$ can be expanded into a sum of k unitary matrices of rank 1:

$$X = \sum_{\alpha=1}^{k} \sigma_\alpha \mathbf{u}_\alpha \mathbf{v}_\alpha^T \quad (1)$$

The two sets $\{\mathbf{u}_\alpha\}$ and $\{\mathbf{v}_\beta^T\}$ $\alpha,\beta = 1..k$, of column and row vectors, respectively, are orthonormal sets. This expression can be rewritten in the matrix representation

$$X = U\Sigma V^T \quad (2)$$

where Σ is a (non-square) diagonal matrix, and U, V are orthogonal matrices. Ordering the non-zero elements of Σ in descending order, we can get an approximation of a lower rank r to the matrix X by taking $\Sigma^r_{jj} = 0$ for $j > r$, leading to the matrix

$$Y = U\Sigma^r V^T. \tag{3}$$

This is the best approximation of rank r to X, i.e. it leads to the minimal sum of square deviations

$$S = \sum_i^m \sum_j^n (X_{ij} - Y_{ij})^2. \tag{4}$$

Once we apply SVD to a given matrix X we automatically define two spaces dual to each other. The matrix U has orthogonal columns (eigensamples) that serve as axes for representing all genes (rows of U), while the matrix V has orthogonal columns (eigengenes) that serve as axes of a space representing all samples (rows of V or columns of V^T). Truncating these representations to dimension r, the gene-vectors (truncated rows of U) and the sample-vectors (truncated columns of V^T) do not have equal norms. This leads to a problem for the clustering algorithm that is applied in these spaces since many vectors accumulate around the origin. We employ therefore rescaling of all vectors to unit length. In other words, we project these vectors onto the unit sphere in r-space. This approach is also consistent with the standard application of LSA (Landauer *et al.* 1998), where similarity between vectors in the truncated space is defined in terms of the (cosine of the) angle, rather than the proximity, between vectors.

In the following we provide two examples of clustering in sample space and one example in gene space. In both cases r is quite small, of order 4 to 5. This agrees with the observation of Ding *et al.* (2002) who concluded that the optimal r value should be of the order of the expected number of clusters.

Quantum Clustering (QC)

The clustering algorithm that we are going to use has been recently suggested by Horn and Gottlieb (2002). It starts out with a Parzen window approach, assigning to each data-point a Gaussian of width σ thus constructing

$$\psi(\mathbf{x}) = \sum_i e^{-\frac{(\mathbf{x}-\mathbf{x}_i)^2}{2\sigma^2}} \tag{5}$$

that can serve (but for an overall normalization) as a probability density generating the data. One then proceeds to construct a potential function

$$V(\mathbf{x}) = E + \frac{\frac{\sigma^2}{2}\nabla^2\psi}{\psi} \tag{6}$$

where

$$E = -\min\frac{\frac{\sigma^2}{2}\nabla^2\psi}{\psi} \tag{7}$$

thus rendering V positive definite. In fact V has a global minimum at zero, and grows as a polynomial of second order outside the domain over which the data points are defined. Within this domain, V develops minima that are identified with cluster centers.

The intuition behind this approach is that this choice of V is the correct one for the Schrödinger equation

$$H\psi \equiv (-\frac{\sigma^2}{2}\nabla^2 + V(\mathbf{x}))\psi = E\psi \tag{8}$$

whose solution (lowest eigenstate) is the probability density $\psi(\mathbf{x})$. In this equation, the potential function $V(\mathbf{x})$ can be regarded as the source of attraction, whereas the first Lagrangian term is the source of diffusion of the distribution, governed by the parameter σ.

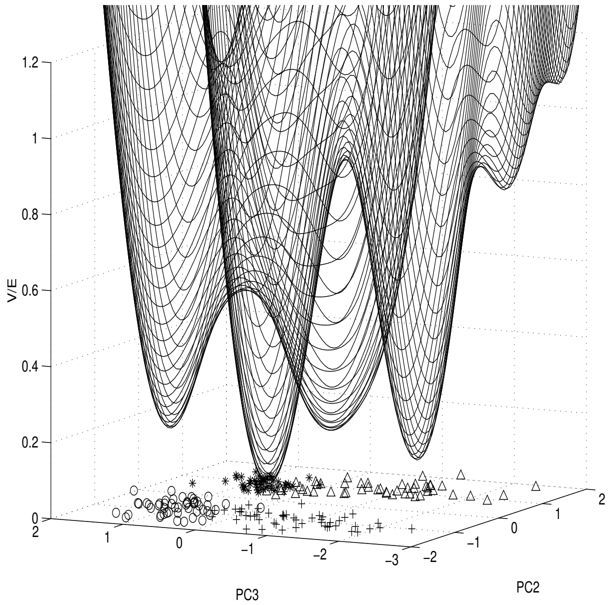

Figure 1: The potential function V generated from a set of 200 data points using the parameter value $\sigma = 1/2$. In this non-optimal case, two spurious local minima are generated in an otherwise well clustered potential that separates into four valleys corresponding to the four classes in the data. These classes are designated by different symbols in the two-dimensional plane to guide the eye.

In Fig.1 we present an example for such a potential function generated from a set of data. The data come from the crab data set of Ripley (1996) presented in a two-dimensional subset of its principal components. These data were used for demonstration of the QC algorithm by Horn and Gottlieb (2002). The data are composed of four groups of 50 points each. The potential is plotted here for the (suboptimal) choice of $\sigma = 1/2$, for which four major valleys develop and two minor ones. For $\sigma = 1$ the minor valleys disappear.

Once the minima of $V(\mathbf{x})$ are defined as cluster centers, the assignment of data points to clusters can proceed through a gradient descent algorithm (Press *et al.* 1992) allowing auxiliary point variables $\mathbf{y}_i(0) = \mathbf{x}_i$ to follow dynamics of

$$\mathbf{y}_i(t + \Delta t) = \mathbf{y}_i(t) - \eta(t)\nabla V(\mathbf{y}_i(t)) \tag{9}$$

that lead to asymptotic fixed points $\mathbf{y}_i(t) \to \mathbf{z}_i$ coinciding with the cluster centers. In the example of Fig. 1 three of the 200 points fall into spurious local minima, whereas all other points will be classified into relevant clusters with a large degree of success. We refer to Horn and Gottlieb (2002) for further explanations. It should be emphasized that although a search is carried out here for the minima of a continuous function $V(\mathbf{x})$, which may be a complex problem in high dimensions, it can in fact be simplified by evaluating this function only at the data points (and their gradient descendants $V(\mathbf{y}_i)$) which is sufficient to carry out the algorithm of clustering.

Hierarchical QC

The QC algorithm has a free parameter σ that characterizes the length scale over which we search for cluster structures. Varying it from low to high values, we can get anywhere from N clusters (where N is the number of data points) to one cluster. The algorithm has to be applied judiciously, e.g. by limiting oneself to a small number of clusters that stays stable over a range of σ . It is however important to realize that this algorithm does not guarantee hierarchy, i.e. the assignments of data points to clusters does not follow a tree, or dendrogram representation, as σ is being varied.

We find it useful to define a modified version that produces a hierarchical formulation in an agglomerative manner. Starting out with very low σ, such that each datapoint is a cluster of its own, we have the first trivial clustering $\mathbf{z}_i^1 = \mathbf{x}_i$. Then we increase σ by some amount obtaining, after the QC gradient descent algorithm, new clustering centers $\mathbf{z}_i^2$. Although there are N values specified here, there should now be several coinciding with one another, thus describing small clusters with a few points in each. Let us now use $\mathbf{z}_i^2$ as the data points in our next stage of QC, after once again increasing σ. This leads to a new set of cluster values $\mathbf{z}_i^3$. This procedure is continued until large σ values are reached with only one cluster. On the way it defines a dendrogram whose clustering quality we may compare to biological sample data. We call this method hierarchical quantum clustering (HQC). It will be applied to the first of our three examples. The second and third examples are analyzed using QC with s judicial choice of σ to be explained below.

RESULTS

We test our method on data from 3 different microarray experiments in which the gene/sample classification is known. In the first two cases we will discuss clustering of cells, and in the third clustering of genes. In all three we compare our clustering results with known classifications.

Cancer cells

The NCI60 set is a gene expression profile of 60 human cancer cells using 9,703 cDNAs representing approximately 8000 unique genes. NCI60 includes cell lines derived from cancers of colorectal, renal, ovarian, breast, prostate, lung and central nervous system, as well as leukemias and melanomas. After application of selective filters Scherf *et al.* (2000) reduced the number of gene spots to a 1,376 subset. We applied HQC to the data points in an $r = 5$ eigengenes space. The obtained dendrogram shows that most samples cluster according to the cell/cancer type of the sample.

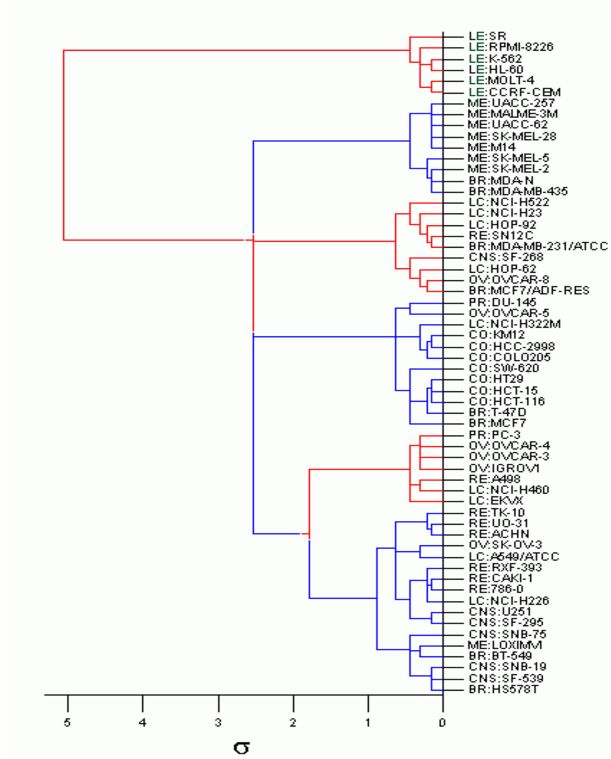

Figure 2: Dendrogram of 60 cancer cell samples. Clustering was performed on a truncated 5 dimensional space. The first 2 letters in each sample represent the tissue/cancer type.

As can be seen in Fig. 2, at $\sigma = 0.2$ one obtains many clusters, some including just one sample, others having 2-4 samples. From this point on we increased the width by dividing $1/\sigma^2$ by a factor of 2 at each step of HQC. Around $\sigma = 0.5$ one finds clustering into roughly the groups described by the first letters designating the cancer classes.

Let us use this example to explain the effect of the projection onto the sphere in r-space, after applying SVD to the data. We display in Fig. 3 data of four classes of cancer cells as they appear in two of the $r = 5$ truncated dimensions before (open circles) and after normalization to unit length. It is quite evident that this projection onto the sphere is an important preprocessing step for any clustering algorithm.

AML *vs* ALL

The second data set is taken from 72 leukemia patients (Golub *et al.* 1999) with 2 types of leukemia called ALL and AML. The ALL set is further divided into T-cell leukemia and B-cell leukemia and the AML set is divided into patients who have undergone treatment (with an anthracycline-cytarabine regimen) and those who did

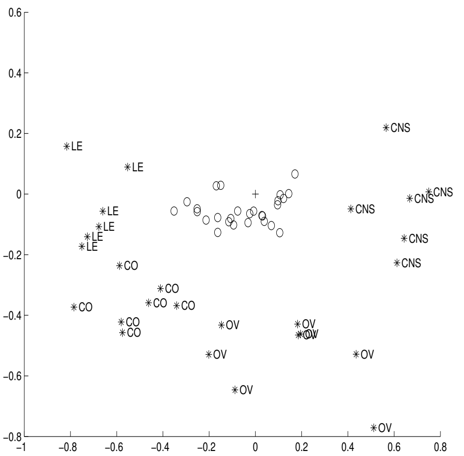

Figure 3: Representation of data of four classes of cancer cells on two dimensions of the truncated space. These data points (denoted by star and by the relevant letters) are shown after the normalization of each data point in r-space. The circles denote the locations of the data points before this normalization was applied.

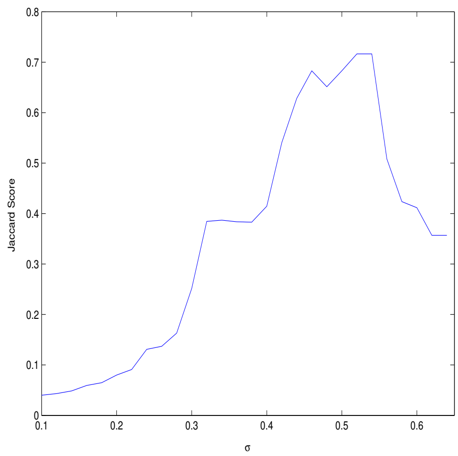

Figure 4: The Jaccard measure for the AML/ALL problem as function of σ.

not. We obtain good clustering results, conforming to the four classes of this problem. Here we applied QC directly, without HQC, because we were not interested in any dendrogram representation.

To describe the quality of the results we calculate, at each stage of σ, the Jaccard score

$$J = \frac{n_{11}}{n_{11} + n_{10} + n_{01}} \tag{10}$$

where n_{11} is the number of pairs of samples that appear in the same cluster both according to the cell type and according to our clustering algorithm, whereas $n_{10} + n_{01}$ is the number of pairs that appear together in one classification and not in the other. This score should be 1 for perfect clustering and decrease as the clustering quality decreases. The results of an $r = 5$ analysis, displayed in Fig. 4, show that the best performance is obtained around $\sigma = 0.5$, which is where four clusters are the preferred solutions of the QC algorithm. The clustering itself is presented in Fig. 5. The first two clusters are the ALL B-cells and T-cells, where we have only 2 (out of 47) misclassifications. We compare here the QC results with a k-means analysis, which turns out to be worse. The Jaccard scores are 0.72 for the best QC result (varying over σ) and 0.48 for the best k-means (varying over k and averaging over initial conditions). It can be seen in Fig. 5 that the k=4 k-means analysis has one quite empty cluster. Indeed, the best k-means results were obtained for k=3. This set of data has been analyzed by Lin *et al.* (2002) who obtained perfect separation into ALL T-cells *vs* B-cells. In our analysis, while allowing a less than perfect separation into these two categories, we have obtained an interesting separation into the four classes in these data.

Yeast Cell Cycle

A famous benchmark is that of the yeast data of Spellman *et al.* (1998), a case studied by several groups who have applied SVD, as explained in the Introduction. Here we wish to test clustering of genes, whose classification into groups was investigated by Spellman *et al.* (1998). The gene/sample matrix that we start from has dimensions of 798×72. We truncate it to $r = 4$ and obtain, once again, our best results for $\sigma = 0.5$ where four clusters follow from the QC algorithm. The original data were classified by Spellman *et al.* (1998) into five classes, whereas we obtained four. The resulting Jaccard score is 0.5. When we group two of the five classes into one, the score increases to 0.54. In other words, the clustering and classification have only partial overlap. In Fig. 6 we display the distribution of the five classes of data, as projected onto two of the truncated $r = 4$ dimensions. Once again this demonstrates how the processing via SVD and our sphere-normalization allows for meaningful clustering of data that are given in a high number (72) of dimensions. It also gives some feeling as to the mixture between the classes.

The quality of clustering can be judged from Fig. 7. We see four cluster assignments of the genes that are presented in an order that preserves their original classification into five groups. The fourth and fifth classes are strongly mixed by QC. The same kind of mixing appears also in k-means analysis. We obtained Jaccard scored of 0.5 for QC and 0.46 for k-means with k=4.

References

[1] Alter,O., Brown,P.O., and Botstein,D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. Proceedings of the National Academy of Sciences 97 (18), pp. 10101-10106.

[2] Rayachudhuri,S., Stuart,J.M. and Altman,R.B. (2000). Principle component analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.* 455-466.

[3] Holter,N.S., Mitra,M., Maritan,A., Cieplak,M., Banavar, J. R., and Fedoroff, N.V. (2000). Fundamental

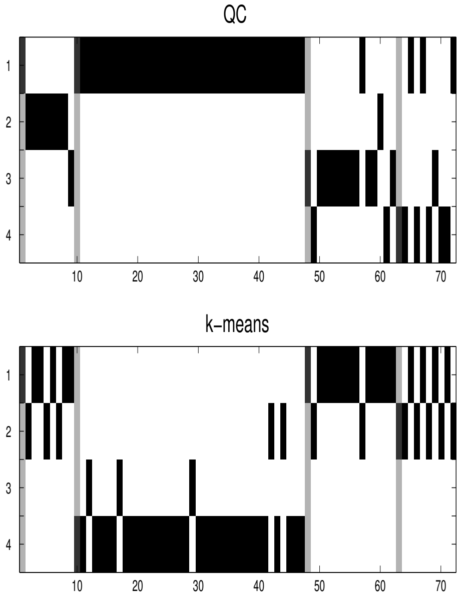

Figure 5: Clustering solutions for the AML/ALL problem using QC with $\sigma = 0.54$ (upper frame) and k-means with k=4 (lower frame). The samples are ordered on the x-axis according to the classification into four groups. Each new group starts with a grey line.

patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl. Acad. Sci. USA* 97, 8409-8414.

[4] Spellman, P. T. , Sherlock, G. , Zhang, M. Q. , Iyer, V. R. , Anders, K. , Eisen, M. B. , Brown, P. O. , Botstein, D. & Futcher, B. (1998) Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Mol. Biol. Cell* 9, 3273-3297.

[5] Landauer, T. K., Foltz, P. W., and Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes,* 25, 259-284.

[6] Ding, H., He, X., Zha, H. and Simon, H. D. (2002). Adaptive dimension reduction for clustering high dimensional data. *Proc. 2nd IEEE Intl. Conf. Data Mining*, Maebashi Japan.

[7] Horn, D. and Gottlieb, A. 2002. Algorithm for Data Clustering in Pattern Recognition Problems Based on Quantum Mechanics. *Phys. Rev. Lett.* 88, 018702.

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

[9] 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.

[10] Horn, D. and Gottlieb, A. 2002. Algorithm for Data Clustering in Pattern Recognition Problems Based on Quantum Mechanics. *Phys. Rev. Lett.* 88, 018702.

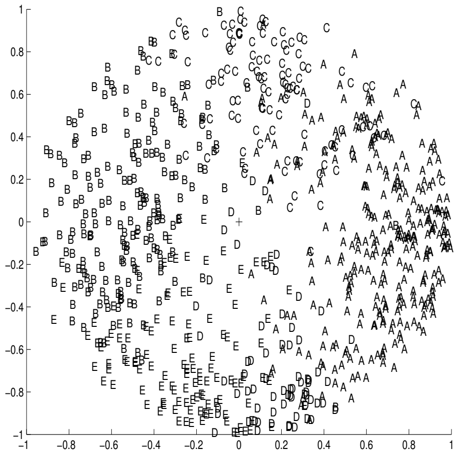

Figure 6: The five gene families as represented in two coordinates of our $r = 4$ dimensional truncated space.

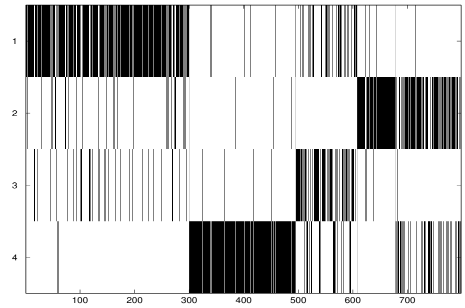

Figure 7: Cluster assignments of genes for QC with $\sigma = 0.46$ as compared to the classification by Spellman *et al.* (1998) shown by the ordering of genes on the x-axis.

[11] Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.L., Kohn, K.W., Myers, T.G., Eisen, M.B., Reinhold, W.C., Andrews, D.T., Scudiero, D.A., Pommier, Y., Botstein, D., Brown, P.O. and Weinstein, J.N. 2000. A gene expression database for the molecular pharmacology of cancer. *Nature Genetics* 24(3): 227-234.

[12] Golub,T., Slonim,D., Tamayo,P., Huard,C., Gaasenbeek,M. Mesirov,J., Coller,H., Loh,M., Downing,J., Caligiuri,M. Bloomfield,C., Lander,E. 1999. Molecular classification of cancer: class discovery and class prediction by geneexpression monitoring. *Science*, 286:531–537.