SLAC-PUB-13436

Supporting Material for Dynamic quantum clustering: a tool for exploration of structures in data

October 13, 2008

Published in the Proceedings of the National Academy of Sciences

Work supported in part by US Department of Energy contract DE-AC02-76SF00515

1 Ripley's Crab data analysis in five dimensions

We begin by presenting a discussion of the full five-dimensional Crab dataset [2]. This is interesting because we see how varying the mass parameter, m, allows us to explore some features of the topology of local minima of the potential function in five dimensions. Figure 1 shows the before and after plots of the data points for the five-dimensional case. Since each data point is now a five-tuple, $(x^1, x^2, x^3, x^4, x^5)$, we display the data in two three-dimensional plots. The points in the top left plot show the (x^1, x^2, x^3) coordinates of the original data, the points in the top right plot exhibit the (x^3, x^4, x^5) coordinates of the same data. The lower plots display the same information for the DQC evolved data.

It is obvious from the first two plots that it is not as simple to separate the clusters for the five-dimensional data projected onto the five dimensional unit sphere. Although the separation between (red,blue) and (orange,green) is still fairly clear, in dimensions four and five we see that the (red,blue) and (orange,green) points are very intermingled. After DQC evolution we see that the original clusters are separated in the first three principal components, however things are still not as clean in dimensions four and five. Here we see that while the blues separate from the reds to some degree, the clusters are still intermingled. The same is true for the orange and green points.

We can attempt to improve the clustering further by following our previous scheme and iterating the DQC evolution one more time, starting from the enriched configuration. The results are shown in Figure 1.

Indeed, we see that the configurations have tightened up and clustering is much more clear. However we see that the red and blue, as well as the orange and green clusters seem to have developed smaller subclusters. If, as we would expect, this subclustering is due to the fact that the five-dimensional potential function has nearby local minima inside, then we would expect that significantly lowering the mass, m, should reduce the subclustering. In fact, this is exactly what happens. Figure 1 shows the five-dimensional data after carrying out the DQC evolution with a smaller mass value, $m = 0.00001$. In this case we see that the clustering in dimensions (x^1, x^2, x^3) remains unchanged, but now the subclustering seen in the fourth and fifth dimension has disappeared. This change in the pattern from the larger mass case is further enhanced in the second iteration of DQC evolution. Note, however, the existence of a small number of outliers and change in the clustering of a small number of points. Since we know the true classification of the data we

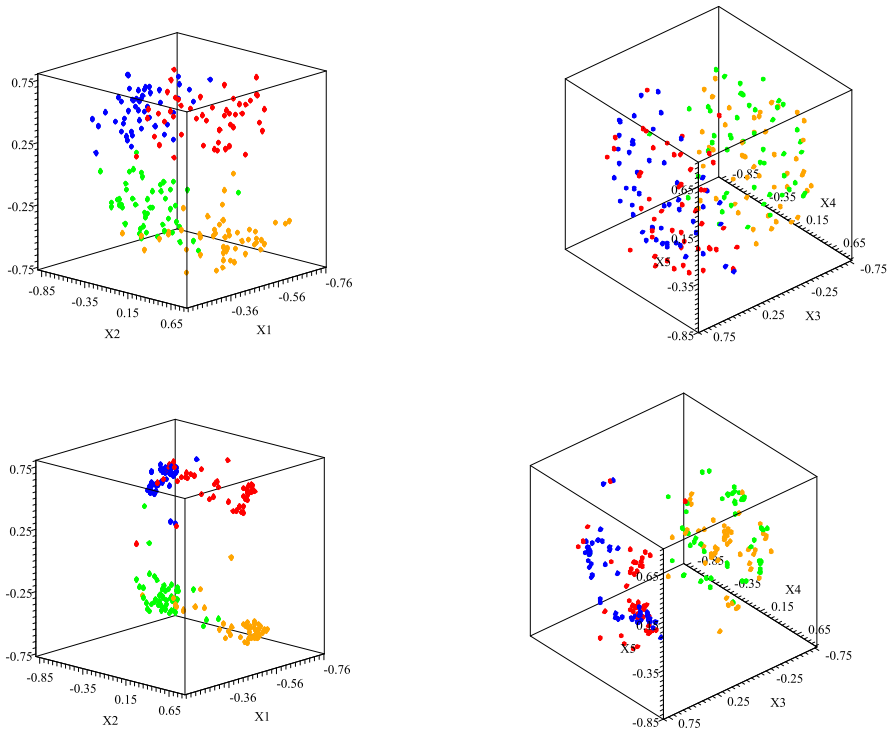

Figure S1 Plots of the five-dimensional data before beginning the quantum evolution and after completing the first step of the evolution. The upper two plots show the distribution of points in dimensions 1,2 and 3 and dimensions 3, 4 and 5 respecively. The values of σ,m and ϵ used to construct the Hamiltonian and evolution operator are: $\sigma = 0.11$, $m = 0.1$, and $\epsilon = 10^{-6}$.

see that the five-dimensional analysis did not improve the clustering based on the first three principal components.

2 Virus Data

Another example worth discussing is the case of the viruses dataset of S. Fauquet, 1988, discussed in a paper by Varshavsky et. al.[3]. In what follows we show the result of running DQC evolution of the same features selected for analysis by these authors (i.e., those three features which made the largest contribution to the SVD entropy). The dataset records 18 measurements

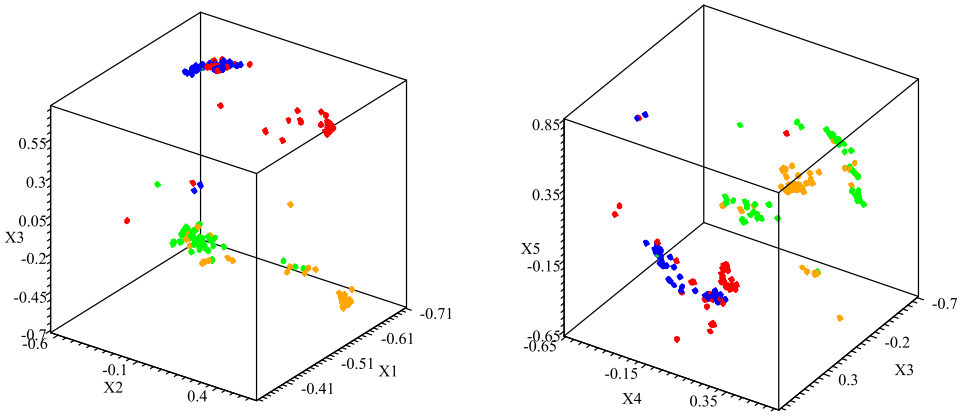

Figure S2 Plots of the five-dimensional data after the second iteration of the quantum evolution. The lefthand plot shows the distribution of points in dimensions 1,2 and 3. The lower two plots show the same distributions after the flow. The values of σ,m and ϵ used to construct the second Hamiltonian and evolution operator are as in the first case: $\sigma = 0.11$, $m = 0.1$, and $\epsilon = 10^{-6}$.

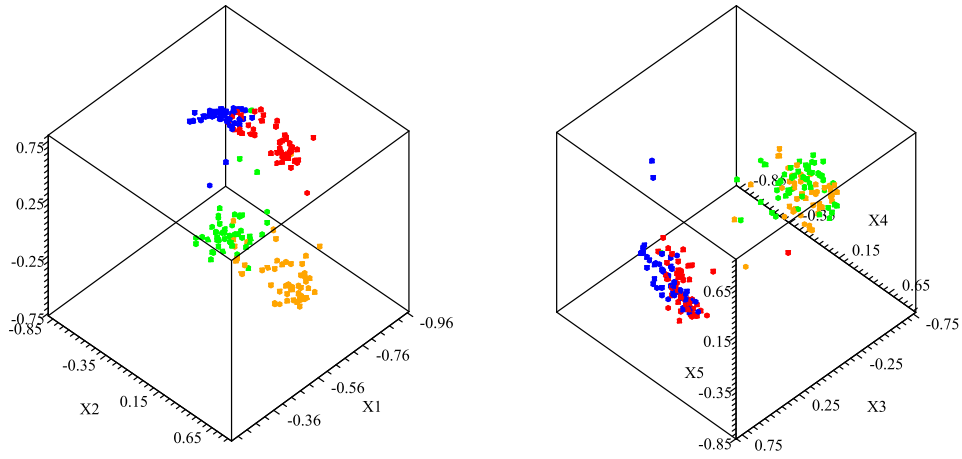

Figure S3 A plot of the five-dimensional data after flowing with a small mass, $m = 0.00001$. Clearly the separation of the points in the fourth and fifth dimensions, which is so apparent in the larger mass case, has disappeared.

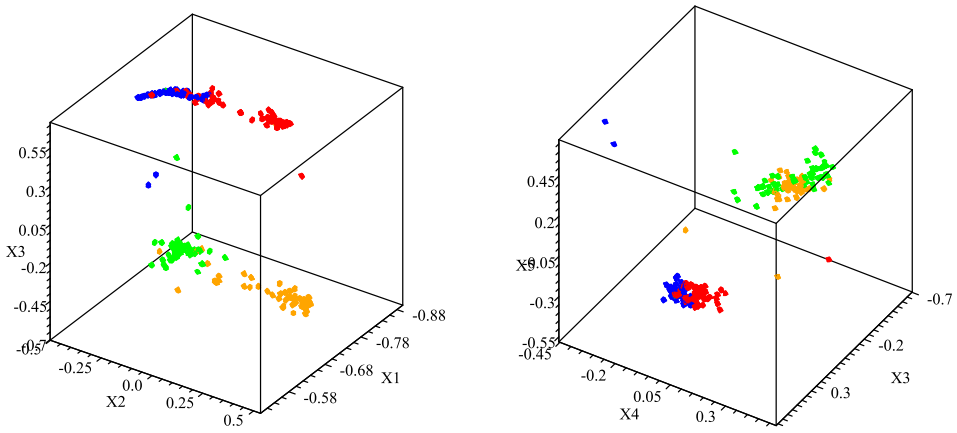

Figure S4 A plot of the second iteration of the DQC evolution of the five-dimensional data with a small mass, $m = 10^{-5}$.

of amino acid compositions for the coat proteins of 61 rod-shaped viruses which affect various crops. These viruses are known to fall into four classes; three Hordeviruses, six Tobraviruses, thirty-nine Tobamoviruses and thirteen Furoviruses.

Following the approach used in the feature selection based upon SVD entropy[3], we begin with doing an SVD decomposition on the 61×3 dimensional matrix, M, constructed selecting the second, sixth and sixteenth column from the original data. The data before DQC evolution is shown in the left hand plot on the top row of Figure 2. The color coding is that the Hordeviruses are shown in red, the Tobraviruses in blue, the Tobamoviruses in orange and the Furoviruses in green. The result of the first stage of DQC evolution is shown in the figure to the right. Once again, since the data oscillates about the minima, we stop the first stage of evolution when the clusters first come together. We then restart the flow from this configuration. As is evident from the plots, initially subclusters come together, then by the third iteration the data has organized itself into four well defined clusters. The content of these clusters is given in Table 1. We should note that the extraction of four tight clusters is a consequence of the feature extraction. If we consider the intermediate states of the evolution shown in Figure 2 we see that at intermediate stages the Tobamoviruses in orange and the Furoviruses in green first form well defined sub-clusters, and then these sub-clusters fuse to the final clusters. The question which should be addressed at this point is if there is any significance to this subclustering.

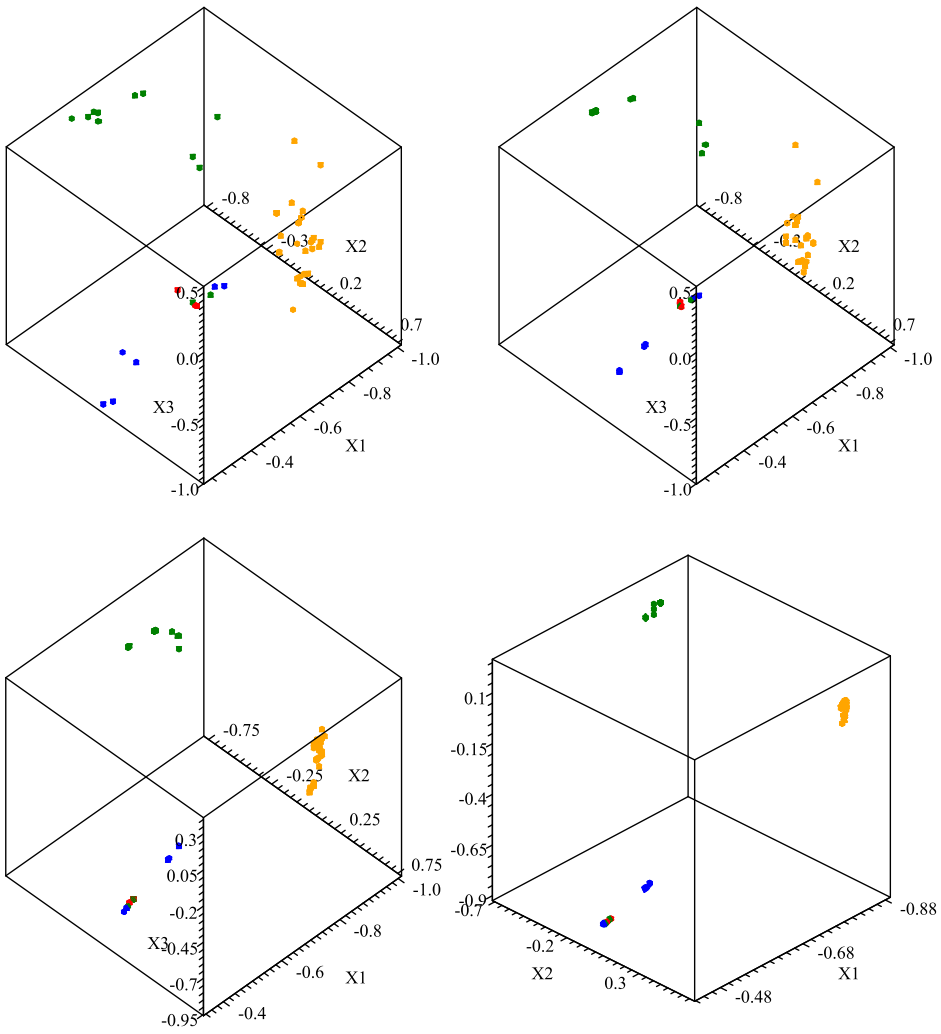

Figure S5 A plot of the feature selected virus data at the beginning, after the first, second and third iteration of DQC evolution. The values used for the evolution are $\sigma = 0.15$, $m = 0.001$ and $\epsilon = 10^{-5}$. The results are not very sensitive to these choices.

To address this question we can simply do an SVD decomposition on the full dataset and then restrict ourselves to the first three principal components. In this case the initial data looks like the first plot in Figure 2. Now, it is evident from DQC before evolution, that the Tobamoviruses in orange and the Furoviruses in green are sub-clustered in much the same way

Cluster	Row Number	Colors
1	4,5,8,9	4 Blue
2	10,11,12,13,14,15,16,17,18,19,20, 21,22,23,24,25,26,27,28,29,30,31, 32,33,34,35,36,37,38,39,40,41,42, 43,44,45,46,47,48	39 Orange
3	1,2,3,6,7,55,56	3 Red, 2 Blue, 2 Green
4	49,50,51,52,53,54,57,58,59,60,61	11 Green

Table S1 The content of the four final virus clusters

as in the feature selected data at intermediate stages. The difference is that now these sub-clusters do not merge with further evolution. Table 2 gives the members of the six final clusters obtained from the data without feature selection. The important thing to note about these two cases is that DQC, as we already noted, enhances the clustering already present in the selected data. It does not seem to produce dramatic artifacts on its own.

Cluster	Row Number	Colors
1	4,5,6,7,8,9,55,56	6 Blue, 2 Green
2	20,21,35,36,37	5 Orange
3	17	1 Orange
4	10,11,12,13,14,15,16,18,19,22,23, 24,25,26,27,28,29,30,31,32,33,34, 38,39,40,41,42,43,44,45,46,47,48	33 Orange
5	1,2,3,57	3 Red, 1 Green
6	49,50,51,52,53,54	6 Green
7	58,59,60,61	4 Green

Table S2 The content of the six final virus clusters plus one singleton in the analysis based on the first three principal components.

3 Iterations of the analysis of the Leukemia dataset of Golub et al.

In the case of the leukemia data [1] it is a-priori expected that many of the quantities measured by the Affymetrix GeneChip will have little or nothing

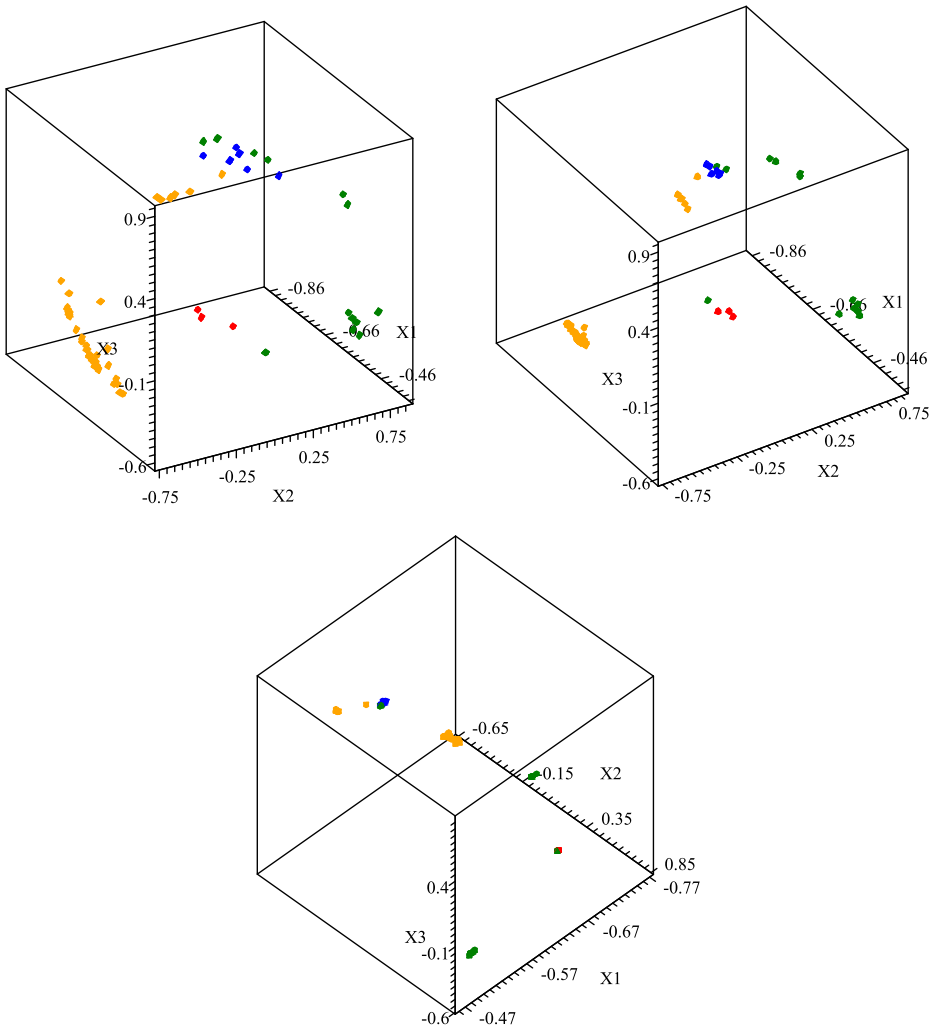

Figure S6 A plot of the virus data for the first three principal components at the beginning, after the first and second iteration of DQC evolution.

to do with the fact that the cell we are looking at is a cancer cell. Clearly not every gene is significant in this classification. Moreover, the measurements of gene expression have a statistical uncertainty which is not always well understood. Hence, it is important to filter out particularly noisy information, even though we don't know its origin. This is where SVD-entropy based filtering plays a role. Note, when asking these questions, the colors assigned to points are no longer there for pedagogical purposes. Now they

are a visual way of quickly checking whether feature selection and DQC evolution are producing the desired result. In other words, this can serve as a supervised feature selection method.

In Figure 4 of the article we see the result of applying this procedure to the original data, doing an SVD decomposition of the resulting matrix, then doing a DQC evolution of the three dimensional problem obtained by restricting attention to principal coordinates 2, 3 and 4. It should be readily apparent from the left hand plot that applying a single stage of SVD-entropy based filtering has a dramatic effect upon the clustering, even before DQC evolution. The right hand plot shows what happens after a single stage of DQC evolution. This shows the general features which will persist throughout the discussion. Now the fact that the blue points form a well separated cluster is quite obvious. As is the fact that the red points really want to divide into two subclusters. The green and orange clusters seem to be somewhat intermingled, although two obvious subclusters have begun to appear and there are a few outliers which don't get well incorporated into the main clusters. This last situation will resolve itself with further stages of DQC clustering, even at this first stage of filtering. However rather than pursue this further, we wish to demonstrate what happens if we iterate the filtering process, since the effects are quite dramatic.

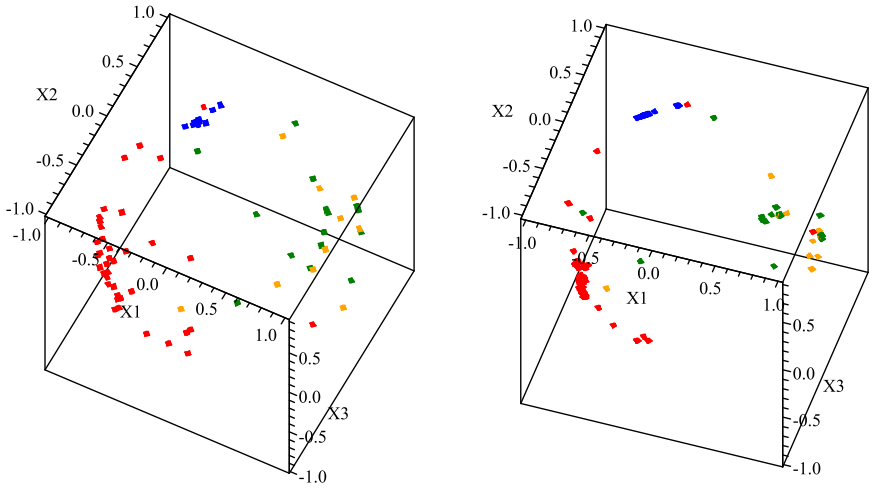

Figure S7 The left hand plot is the data after three stages of SVD-entropy based filtering, but before DQC evolution. The right hand plot is the same data after DQC evolution.

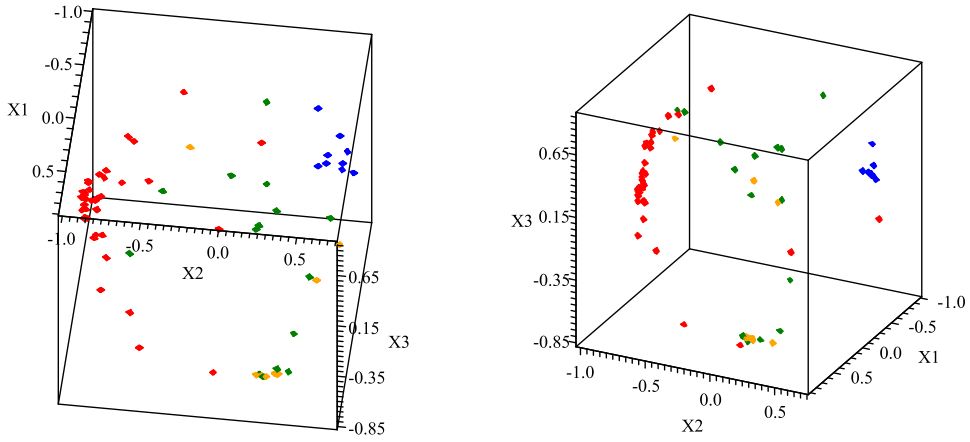

Figure S8 The left hand plot is the data after five stages of SVD-entropy based filtering, but before DQC evolution. The right hand plot is the same data after DQC evolution. Note the extremely well separated blue cluster.

Figures 3 and 3 show the results of three and five iterations of SVD-entropy, before and after DQC evolution. These plots, especially the after DQC pictures, show dramatic clustering, especially for the blue points. With each stage of filtering we see that the blue points cluster better and better, in that the single red outlier separates from the cluster and the cluster separates more and more from the other points. The blue points are what we will refer to as an *obviously robust cluster* which has been identified in early stages of filtering. If one continues iterating past the fifth stage, however, the clear separation of the blue points from the others begins to diminish. Thus we see that the SVD-entropy based filtering, in trying to enhance the clumping of the red points, starts throwing away those features which make the blue cluster distinct. Since this effect is quite pronounced we would say that features that are important to distinguishing the blue cluster from the others begin to be removed at the sixth and higher iterations of filtering. This is, of course, just what we are looking for, a way of identifying those features which are important to the existing biological clustering. Out of the original 7129 features, we have reduced ourselves to 2766 features by the fifth iteration. In going from step five to step six this gets further reduced to 2488 features, so we could begin searching among the 278 eliminated features

to isolate those most responsible for the separation of the blue cluster from the others. Instead, we will take another track and, since it is so robust and easily identified, remove the blue cluster from the original data and repeat the same process without this cluster. The idea here is that now the SVD-entropy based filtering will not be pulled by the blue cluster and so it will do a better job of sorting out the red, green and orange clusters. As we will see, this is in fact the case.

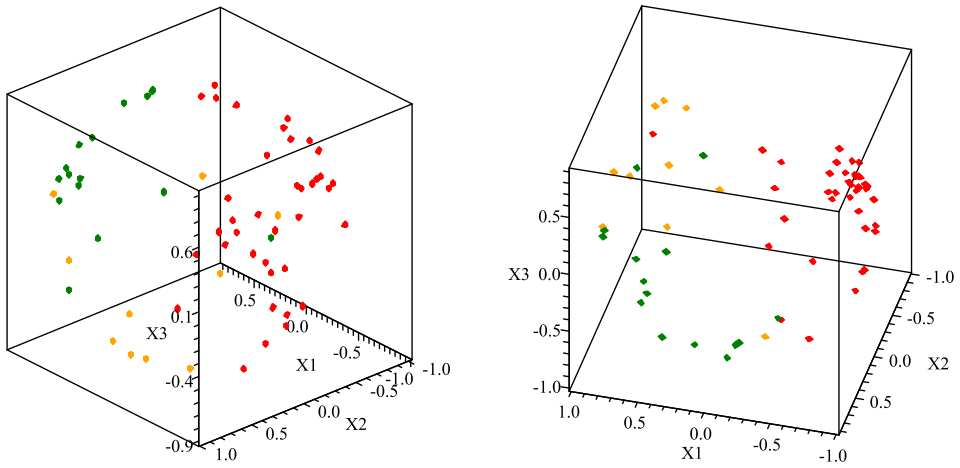

Figure S9 The left hand plot is what the starting data looks like if one first removes the blue points and does one stage of SVD-entropy based filtering. The right hand plot is what the starting data looks like after three stages of filtering.

In Figure 3 we see a plot of what the starting configurations look like if one takes the original data, removes the identified blue cluster and re-sorts the reduced data set according to the SVD-entropy based filtering rules. The left hand plot is what happens if one filters a single time, removing those features, i, whose one-left-out comparison, CE_i, is less than or equal to zero. The right hand plot shows what happens if one repeats this procedure two more times, each time removing features for which $CE_i \leq 0$. There is no problem seeing that each iteration of the SVD-entropy based filtering step improves the separation of the starting clusters. By the time we have done five SVD-entropy based filtering steps the red, green and orange clusters are distinct, if not obviously separated.

Finally, to complete our discussion, we show Figure 3. This figure shows

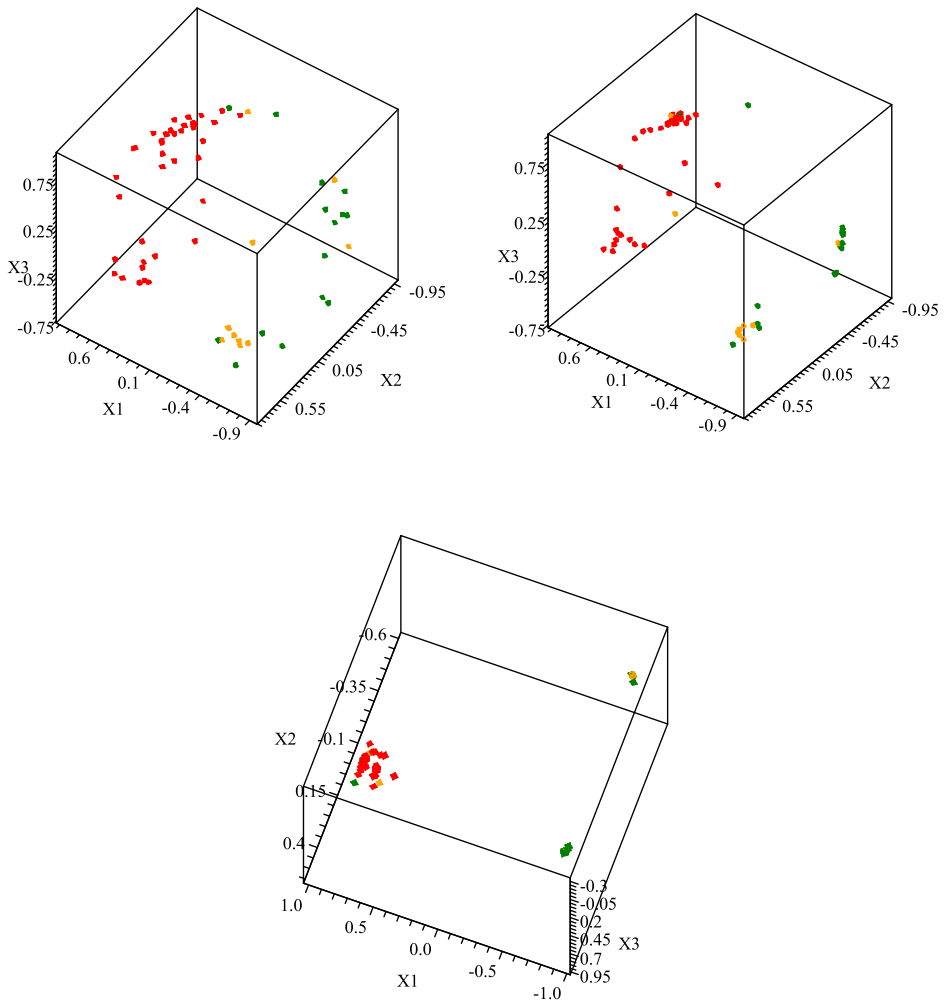

Figure S10 The left hand plot is what the starting data looks like if one first removes the blue points and does five stages of SVD-entropy based filtering. The right hand plot is what happens after one stage of DQC evolution. The bottom plot is the final result after iterating the DQC evolution step two more times. At this point the clusters are trivially extracted.

the results of doing five iterations of the SVD-entropy based filtering and following that with three stages of DQC evolution. The dramatic clustering accomplished by DQC evolution makes it easy to extract clusters. Note however, that in the second plot we see what we have seen throughout, that

the red points first form two distinct sub-clusters which only merge after two more stages of DQC evolution. This constant repetition of the same feature, which is only made more apparent by SVD-entropy based filtering, is certainly a real feature of the data. It presumably says that what appears to be a sample of a single type of cell at the biological level is in reality two somewhat different types of cell when one looks at gene expression. Table 3 shows the content of the clusters as shown in the last picture in Figure 3 augmented by the previously removed pure cluster of nine blue points. The Jaccard score [1] for this result is 0.762, higher than the value 0.707 obtained by [3].

Cluster	Colors
1	9 Blue
2	38 Red, 1 Green, 2 Orange
3	4 Green, 6 Orange
4	9 Green, 2 Orange
5	1 Green

Table S3 The content of the four final cancer cell clusters plus one singleton, as extracted from the data after including the robust blue cluster and the results of filtering the reduced dataset five times and running DQC three times. The Jaccard score for this result is 0.762.

References

[1] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA et al: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science **286** 531 (1999).

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

[1] The Jaccard score is evaluated by considering all pairs of data points, and asking if they cluster together and if they fit in the same class, as judged by the expert. The Jaccard score is then defined by $J = \frac{tp}{tp+fp+fn}$ where tp, fp, fn, stand for true-positive, false-positive and false-negative, correspondingly.

[3] R. Varshavsky, A. Gottlieb, M. Linial and D. Horn. Novel Unsupervised Feature Filtering of Biological Data. Bioinformatics **22** no. 14 (2006), e507-e513.