SVC, QC, SVD,QUI
class May 22, 2004

David Horn
School of Physics and Astronomy
Tel Aviv University

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

Topics to be covered

Support Vector Clustering (SVC)

Ben-Hur, Horn, Siegelmann and Vapnik, JMLR 2001

Quantum Clustering (QC)

Horn and Gottlieb, Phys. Rev. Lett. 2002

Singular Value Decomposition (SVD)

preprocessing for biological applications

Horn and Axel, Bioinformatics 2002

QUI

GUI for SVD and clustering, including QC

Cluster Analysis

The clustering problem:

partition a data-set into a number of groups such that points within each group are alike in some sense.

Approaches:

Hierarchical algorithms - cut the dendrogram at a certain level.

Partitional algorithms.

- Parametric methods.
  - K-means.
  - Mixture models.
- Nonparametric methods.
  - Graph theoretic methods.
  - Information theoretic.
  - Algorithms based on statistical physics.
  - Density estimation methods.

Support Vector Clustering

Given points x in data space, define images in Hilbert space.

Require all images to be enclosed by a minimal sphere in Hilbert space.

Reflection of this sphere in data space defines cluster boundaries.

Two parameters: width of Gaussian kernel and fraction of outliers

An enclosing sphere is defined by:

$$||\Phi(\mathbf{x}_j) - \mathbf{a}||^2 \leq R^2$$

Φ - map into feature space

$\mathbf{a}$: center of the sphere.

- Goal: minimize R^2 over all choices of $\mathbf{a}$ using the Largragian:

$$L = R^2 - \sum_j (R^2 - ||\Phi(\mathbf{x}_j) - \mathbf{a}||^2)\beta_j$$

β_j Lagrange multiplier

Derivatives with respect to R and $\mathbf{a}$:

$$\sum_j \beta_j = 1$$

$$\mathbf{a} = \sum_j \beta_j \Phi(\mathbf{x}_j)$$

The KKT complementarity conditions:

$$(R^2 - ||\Phi(\mathbf{x}_j) - \mathbf{a}||^2)\beta_j = 0$$

- $\beta_j \neq 0 \;\Rightarrow\; R^2 - ||\Phi(\mathbf{x}_j) - \mathbf{a}||^2 = 0$

Points with $\beta_j \neq 0$ are on the surface of the sphere (support vectors).

- Wolfe dual form:

$$W = \sum_j \Phi(\mathbf{x}_j)^2 \beta_j - \sum_{i,j} \beta_i \beta_j \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$$

with the constraints $\sum \beta_j = 1$

- The SV trick: represent the dot products by a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$

Lagrangian now becomes:

$$W = \sum_j K(\mathbf{x}_j, \mathbf{x}_j)\beta_j - \sum_{i,j} \beta_i \beta_j K(\mathbf{x}_i, \mathbf{x}_j).$$

- No need to know the specific form of Φ.

$$R = \{R(\mathbf{x}_i) \mid \mathbf{x}_i \text{ is a support vector}\}$$

The enclosing contour: $\{\mathbf{x} \mid R(\mathbf{x}) = R\}$

The shape of contour governed by the kernel parameter:

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-q||\mathbf{x}_i - \mathbf{x}_j||^2}$$

- As q increases the contour becomes a tighter fit; for certain values of q observe splitting.
- Need to identify the different components.
- Complexity: $O(N^2 D)$

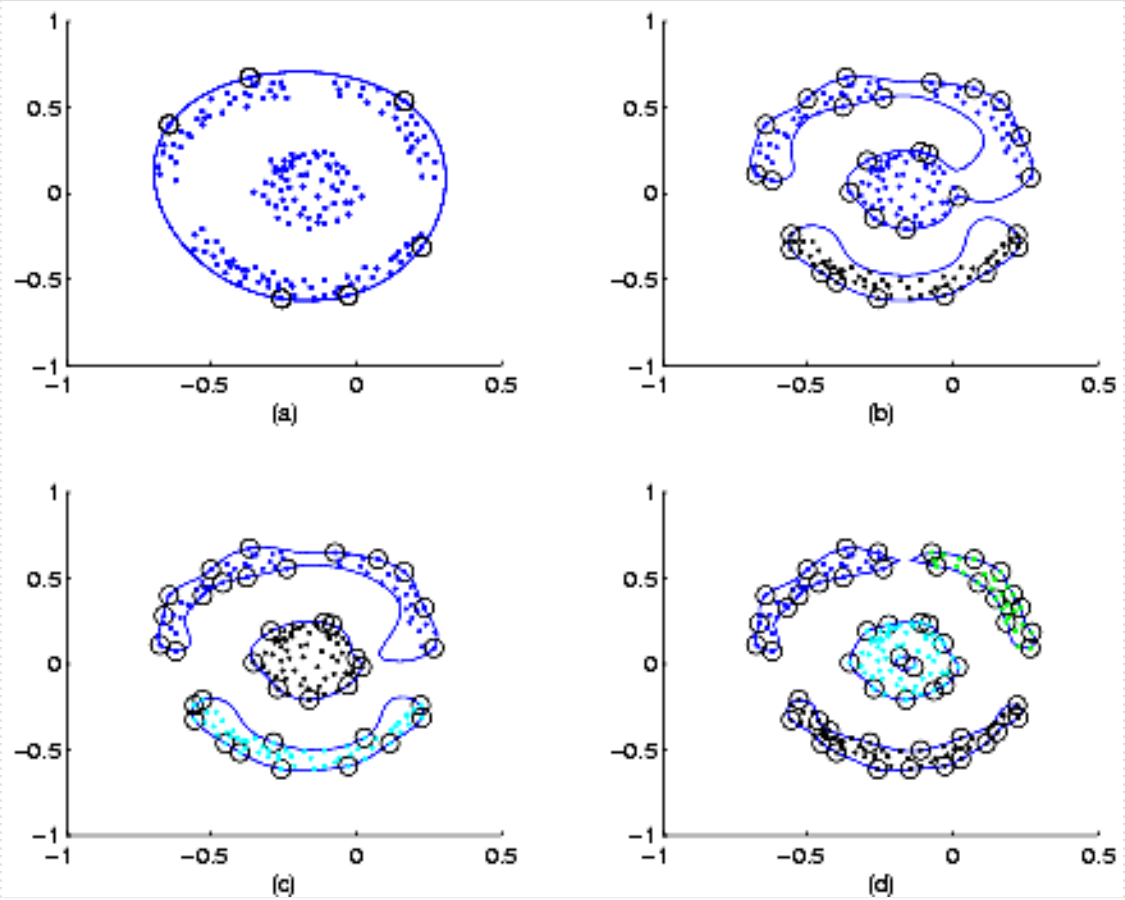
Variation of q allows for clustering solutions on various scales

q=1,

20,

24,

48



$$R = \{R(\mathbf{x}_i) \mid \mathbf{x}_i \text{ is a support vector}\}$$

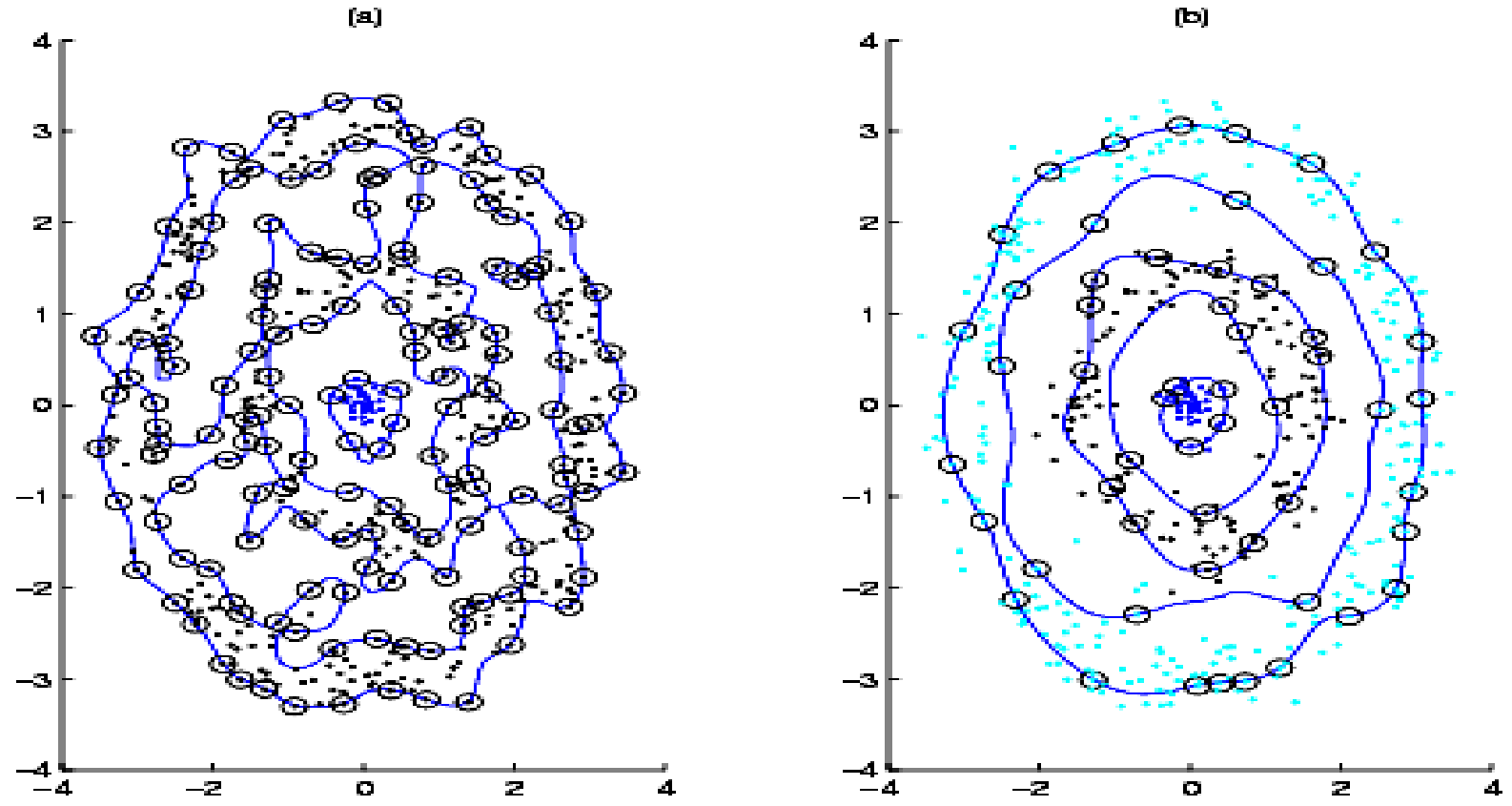
The enclosing contour: $\{\mathbf{x} \mid R(\mathbf{x}) = R\}$

The shape of contour governed by the kernel parameter:

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-q||\mathbf{x}_i - \mathbf{x}_j||^2}$$

- As q increases the contour becomes a tighter fit; for certain values of q observe splitting.
- Need to identify the different components.
- Complexity: $O(N^2 D)$

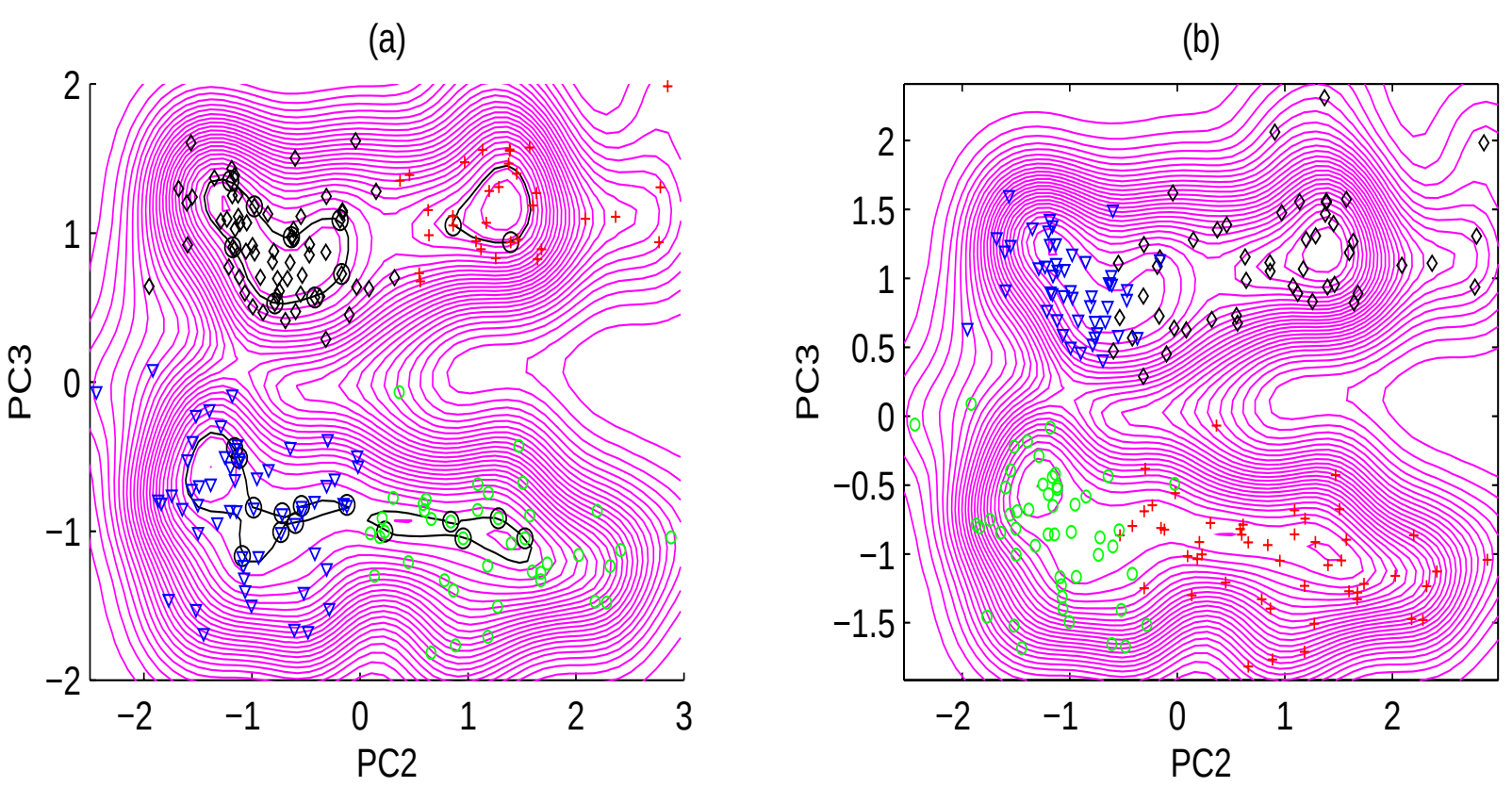
Example that allows for SVclustering only in presence of outliers. Procedure: limit $\beta < C = 1/pN$, where p=fraction of assumed outliers in the data.



q=3.5 p=0

q=1 p=0.3

Similarity to scale space approach for high values of q and p. Probability distribution obtained from R(x).



q=4.8 p=0.7

Wave-function Representation of Hilbert Space

The vectors $\Phi(\mathbf{x}_j)$ can be represented by wave-functions

$$\Phi(\mathbf{x}_j) \equiv ce^{-q(\mathbf{x}-\mathbf{x}_j)^2} \tag{1}$$

such that

$$\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \equiv c^2 \int e^{-q(\mathbf{x}-\mathbf{x}_j)^2} e^{-q(\mathbf{x}-\mathbf{x}_i)^2} d\mathbf{x} = e^{-q(\mathbf{x}_i-\mathbf{x}_j)^2} \tag{2}$$

The center of the SVC sphere $\mathbf{a}$ becomes

$$\mathbf{a} = \sum_j \beta_j \Phi(\mathbf{x}_j) \equiv c \sum_j \beta_j e^{-q(\mathbf{x}-\mathbf{x}_j)^2}. \tag{3}$$

This expression is the same as P_{svc}.

From Scale-space to Quantum Clustering

Parzen window approach:

estimate the probability density by kernel functions (Gaussians) located at data points.

$$P(x) = c\sum_{i=1}^{N} f_i(x) = c\sum_{i=1}^{N} e^{-\frac{(x-x_i)^2}{2\sigma^2}}$$

$\sigma = 1/\sqrt{(2q)}$

Quantum Clustering

View P=Ψ as the solution of the Schrödinger equation:

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

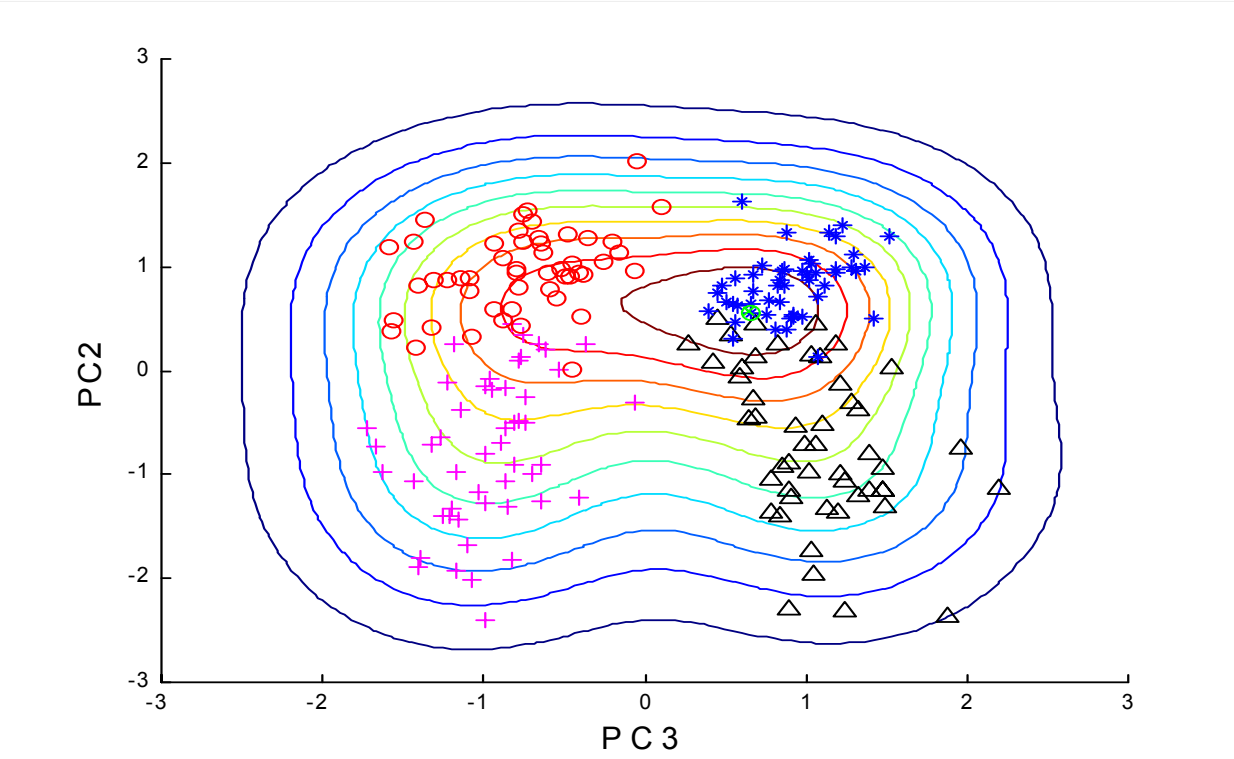
with the potential V(x) responsible for attraction to cluster centers and the Lagrangian causing the spread.

Find V(x):

$$V(x) = E + \frac{\frac{\sigma^2}{2}\nabla^2\Psi}{\Psi} = E - \frac{d}{2} + \frac{1}{2\sigma^2\Psi}\sum_i (x - x_i)^2 e^{-\frac{(x-x_i)^2}{2\sigma^2}}$$

Horn and Gottlieb, Phys. Rev. Lett. 88 (2002) 018702

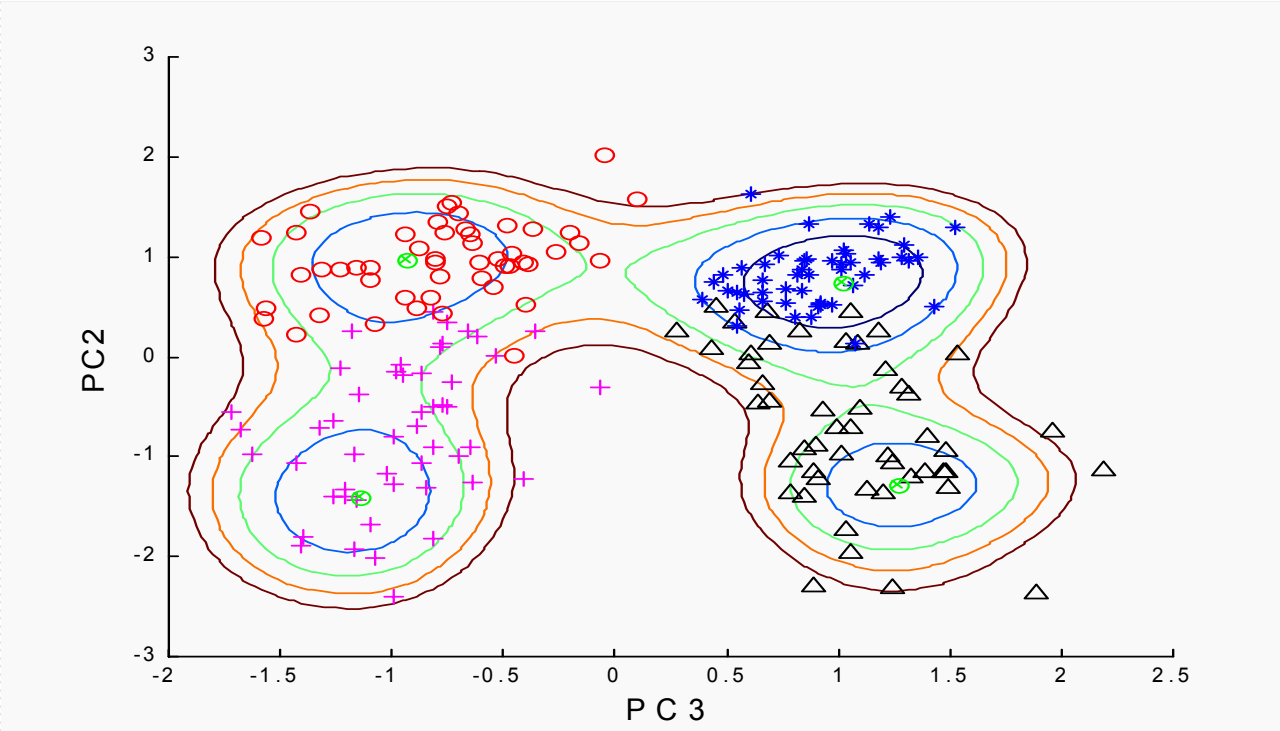
The Crabs Example (from Ripley's textbook)
4 classes, 50 samples each, d=5



A topographic map of the probability distribution for the crab data set with $\sigma=1/\sqrt{2}$ using principal components 2 and 3. There exists only one maximum.

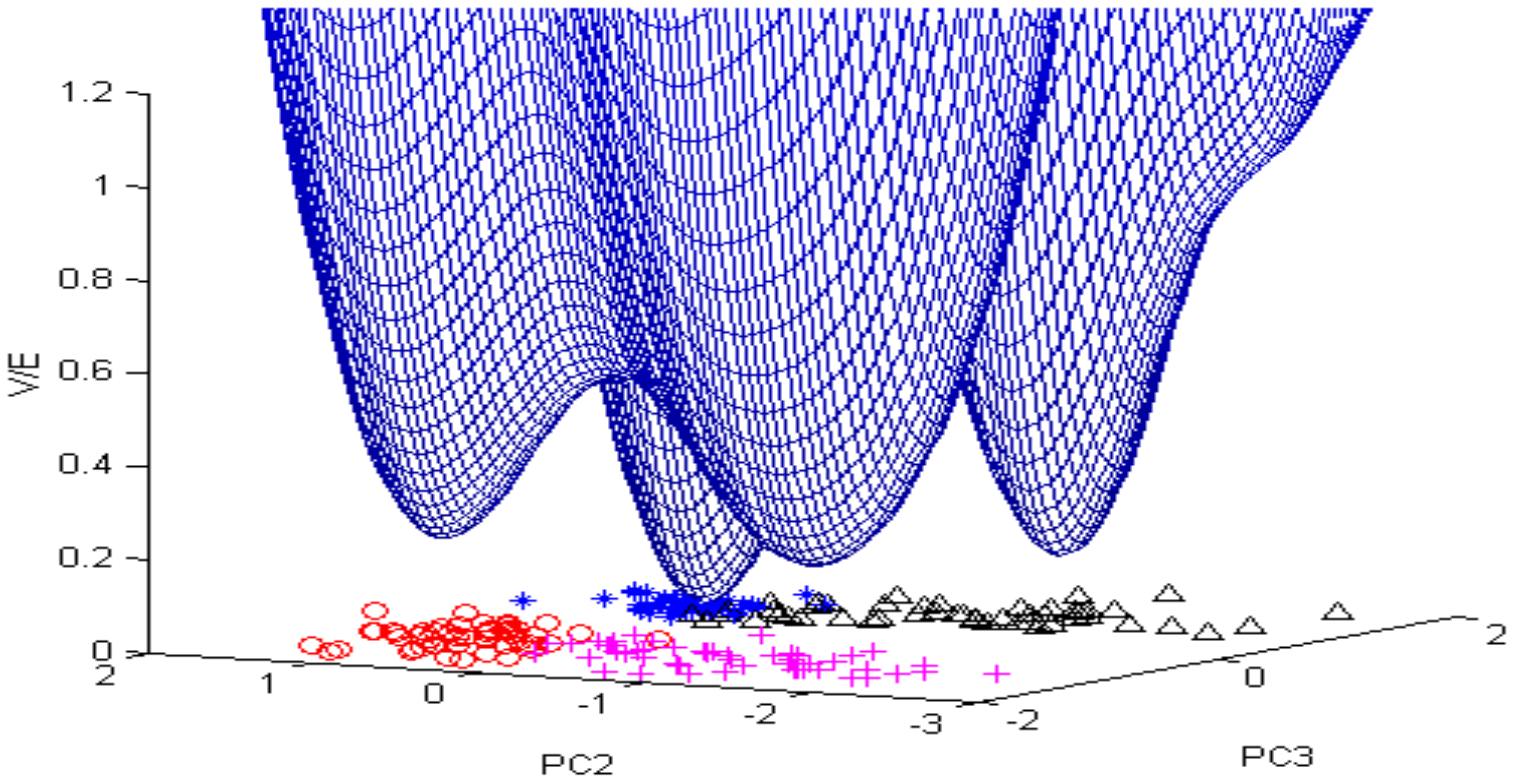
The Crabs Example

QC potential exhibits four minima identified with cluster centers



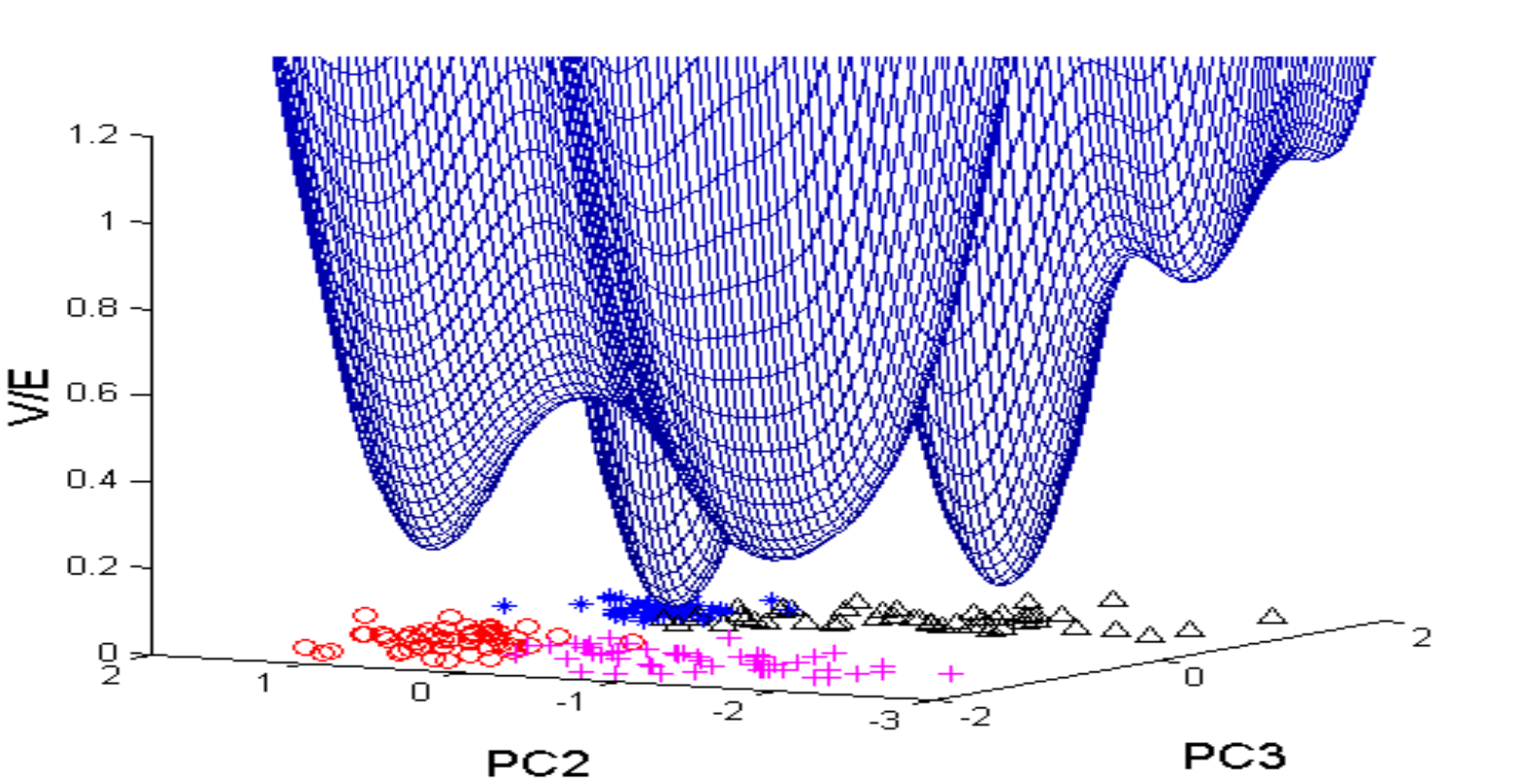
A topographic map of the potential for the crab data set with $\sigma=1/\sqrt{2}$ using principal components 2 and 3 . The four minima are denoted by crossed circles. The contours are set at values $V=cE$ for $c=0.2,\ldots,1$.

The Crabs Example - Contd.



A three dimensional plot of the potential for the crab data set with $\sigma=1/\sqrt{3}$ using principal components 2 and 3

The Crabs Example - Contd.



A three dimensional plot of the potential for the crab data set with $\sigma=1/2$ using principal components 2 and 3

Properties of V and E

E is chosen so that min(V)=0.

E sets the scale of structure observed in V(x).

The single point case corresponds to the harmonic potential

$$V = \frac{1}{2\sigma^2}(x - x_i)^2 \qquad E = d / 2$$

In general

$$0 < E \leq d / 2$$

Identifying Clusters

Local minima of the potential are identified with cluster centers.

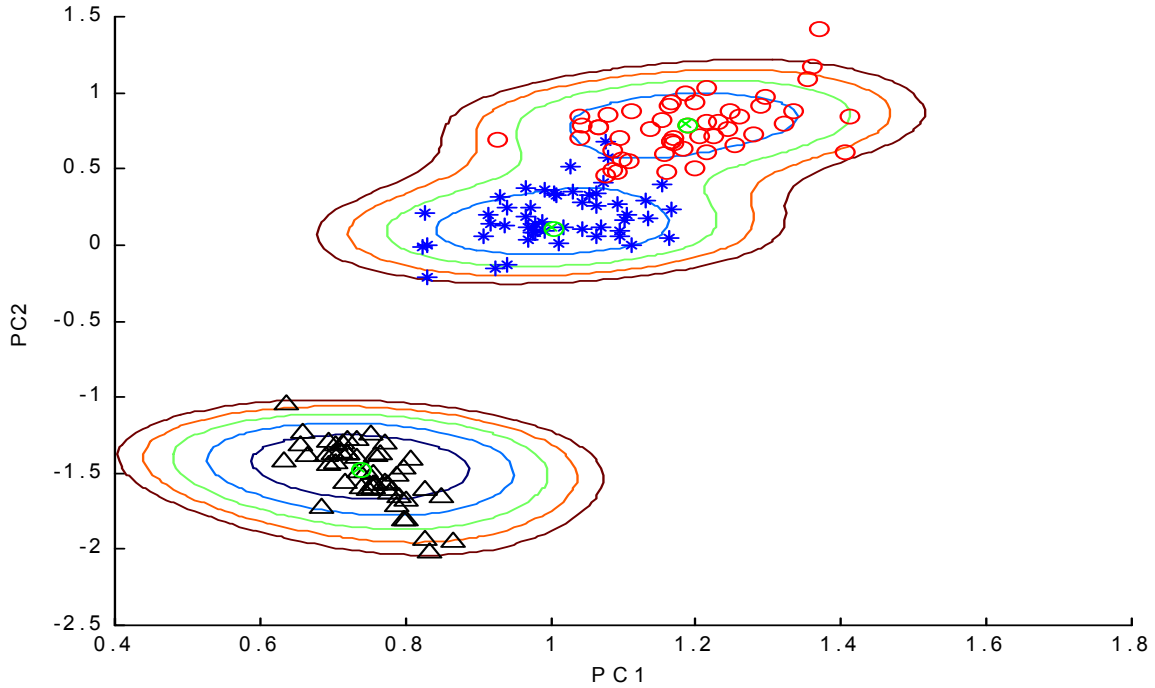
Data points are assigned to clusters according to:

-minimal distance from centers, or,

-sliding points down the slopes of the potential with gradient descent until they reach the centers.

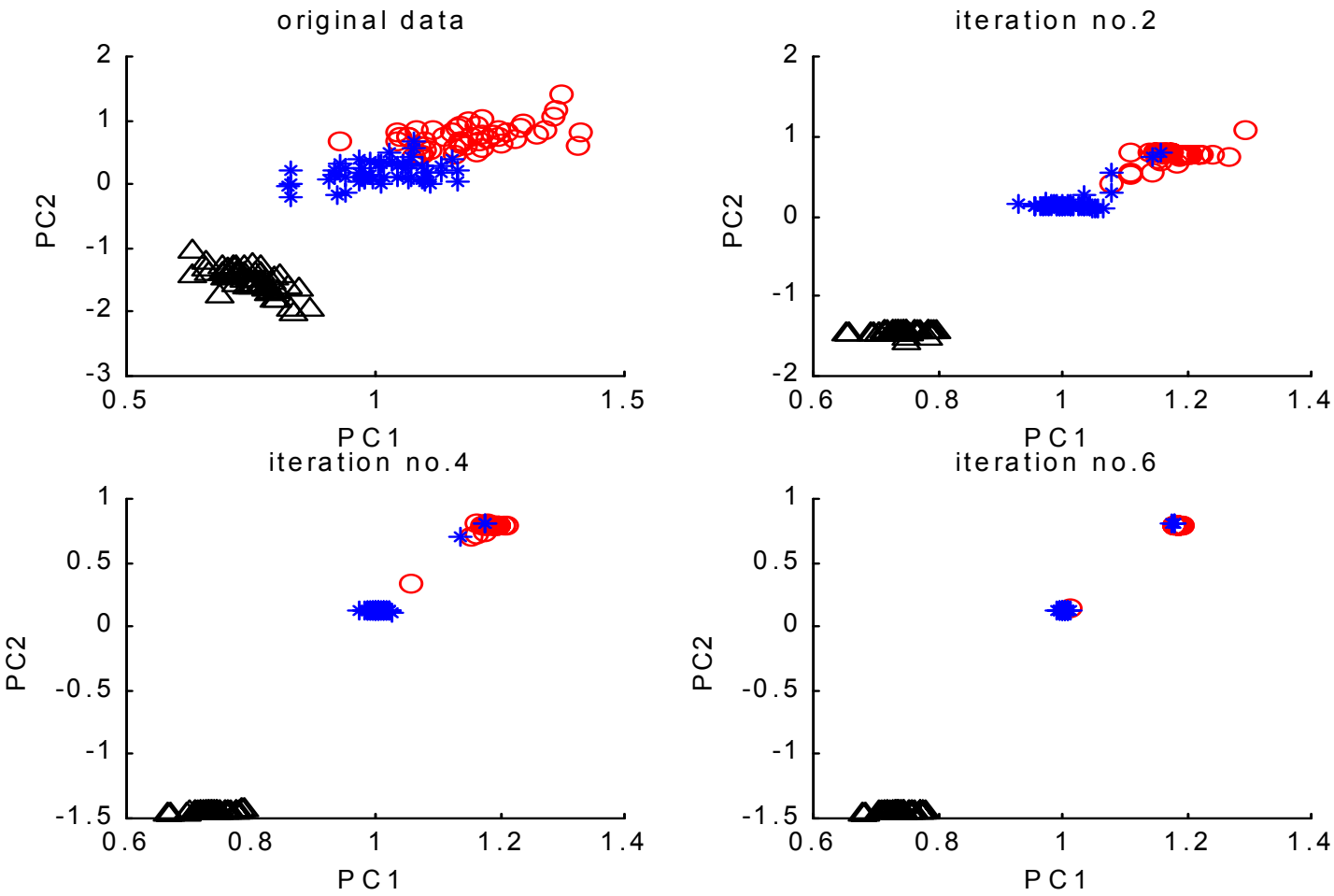
The Iris Example

3 classes, each containing 50 samples, d=4



A topographic map of the potential for the iris data set with σ=0.25 using principal components 1 and 2. The three minima are denoted by crossed circles. The contours are set at values V=cE for c=0.2,…,1.

The Iris Example - Gradient Descent Dynamics



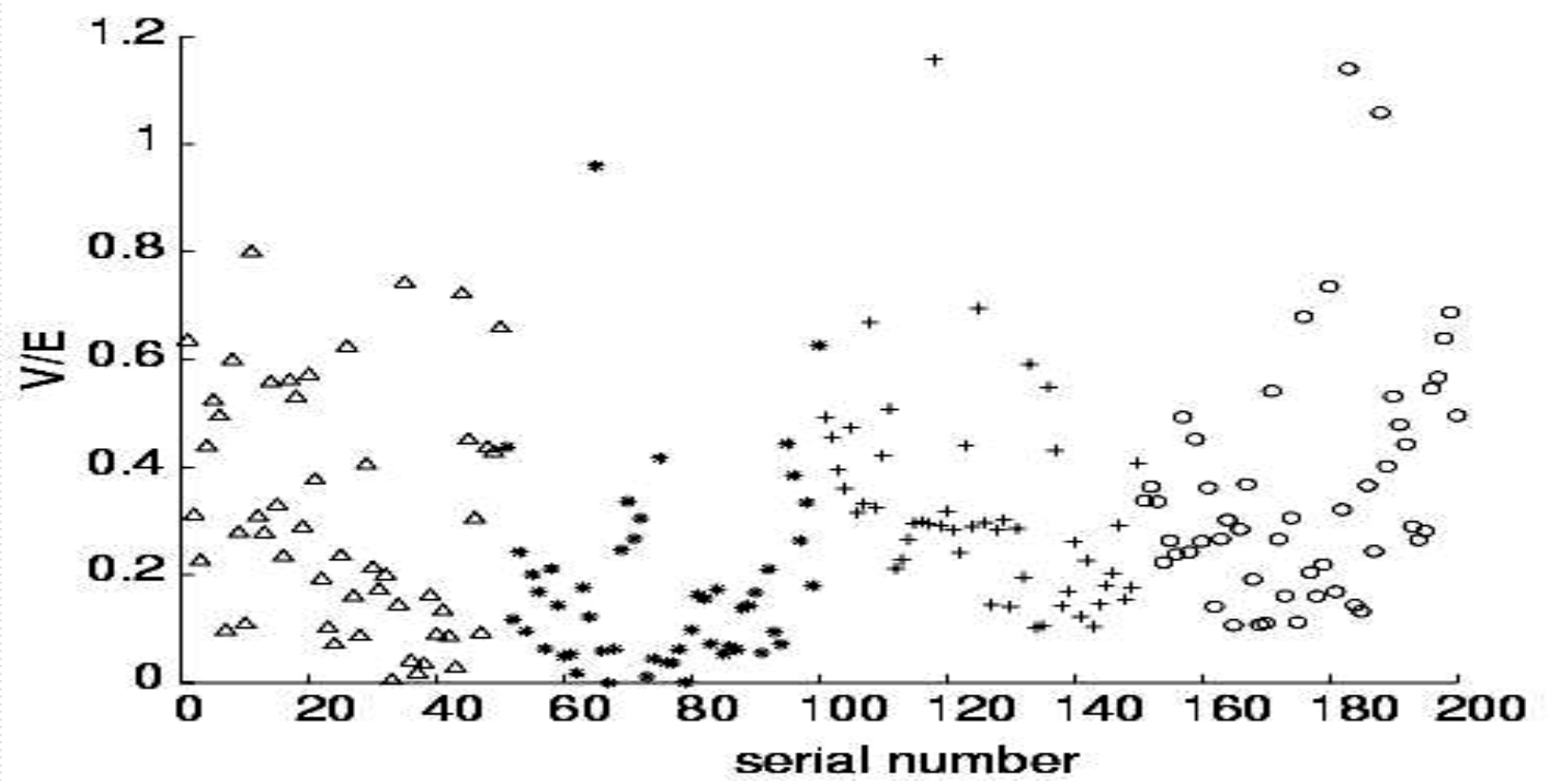
Application in High Dimensions.

Evaluate the potential at data points only:

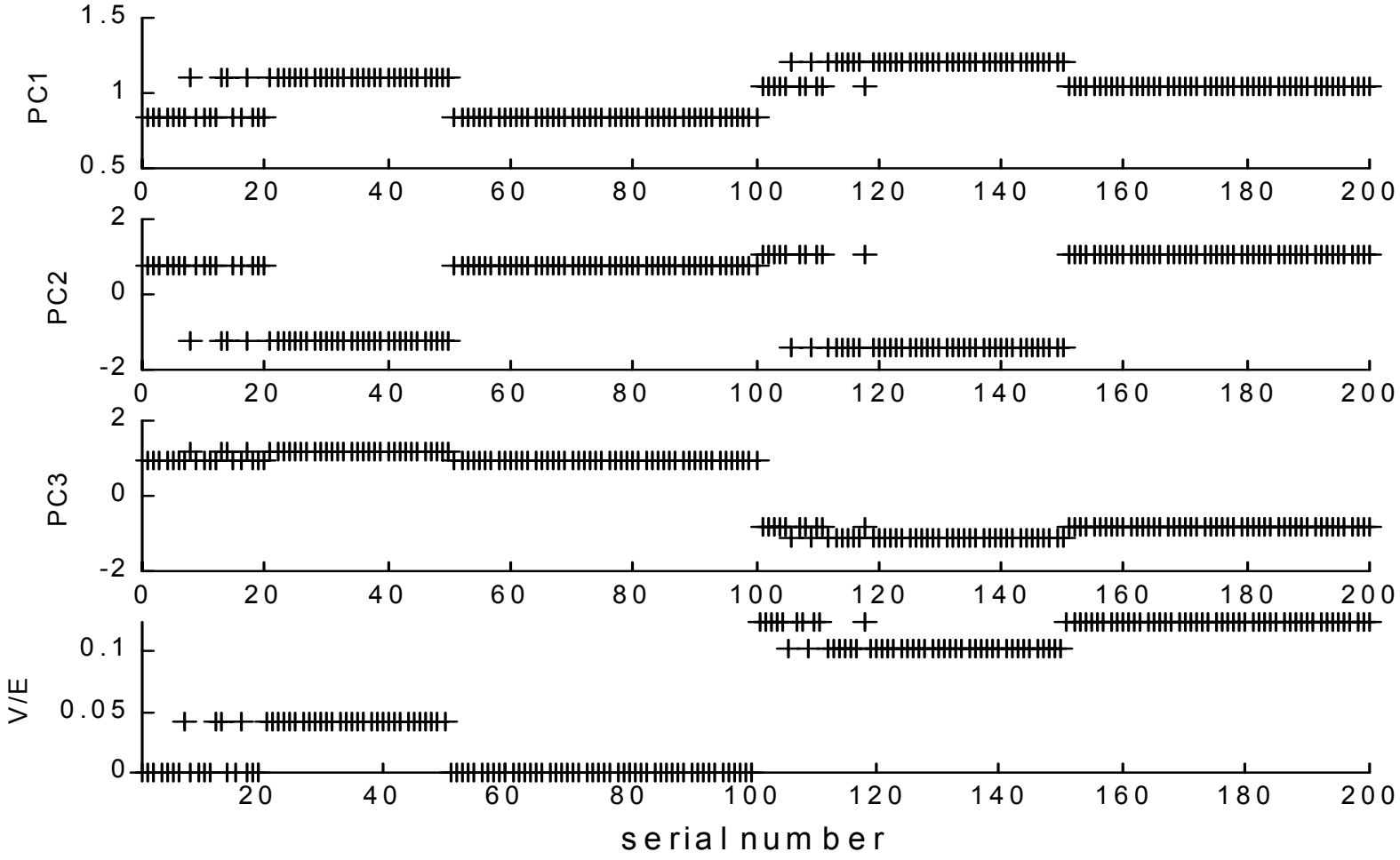
$$V(x_i) = V_i = E - \frac{d}{2} + \frac{1}{2\sigma^2 \Psi_i} \sum_j (x_i - x_j)^2 e^{-\frac{(x_i - x_j)^2}{2\sigma^2}}$$

Since minima are expected to lie close to data points, it suffices to use this discrete approach. Moreover, this formulation allows one to apply QC to a situation when only **distances** (rather than space coordinates) are known.

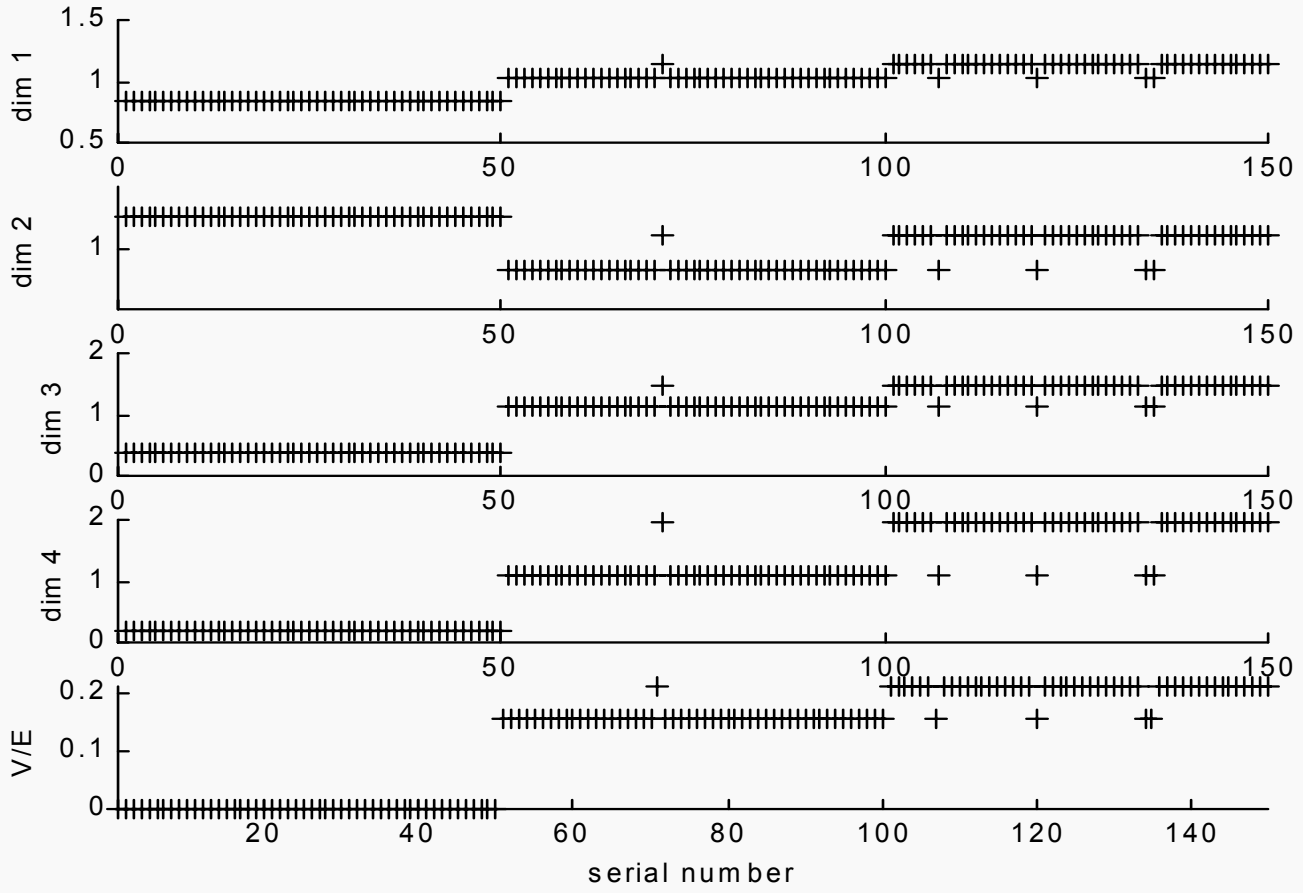
Crab data in 3D: PC1,2,3



Clusters of crab-3D after gradient descent

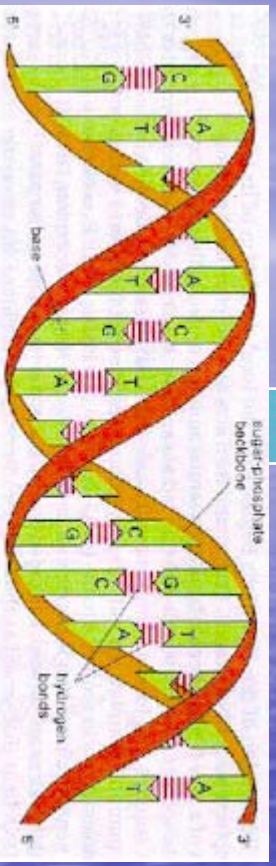


The Iris Example - Using Raw Data in 4D.



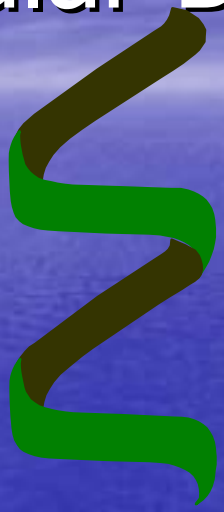
There are only 5 misclassifications. σ=0.21.

Central Dogma of Molecular Biology



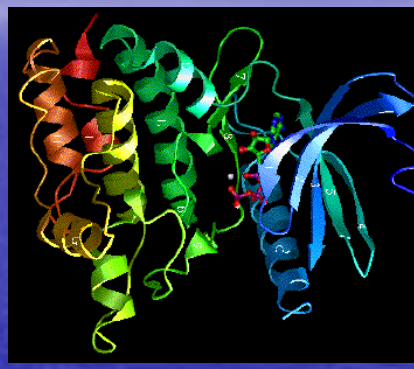
Gene (DNA)

Transcription



mRNA

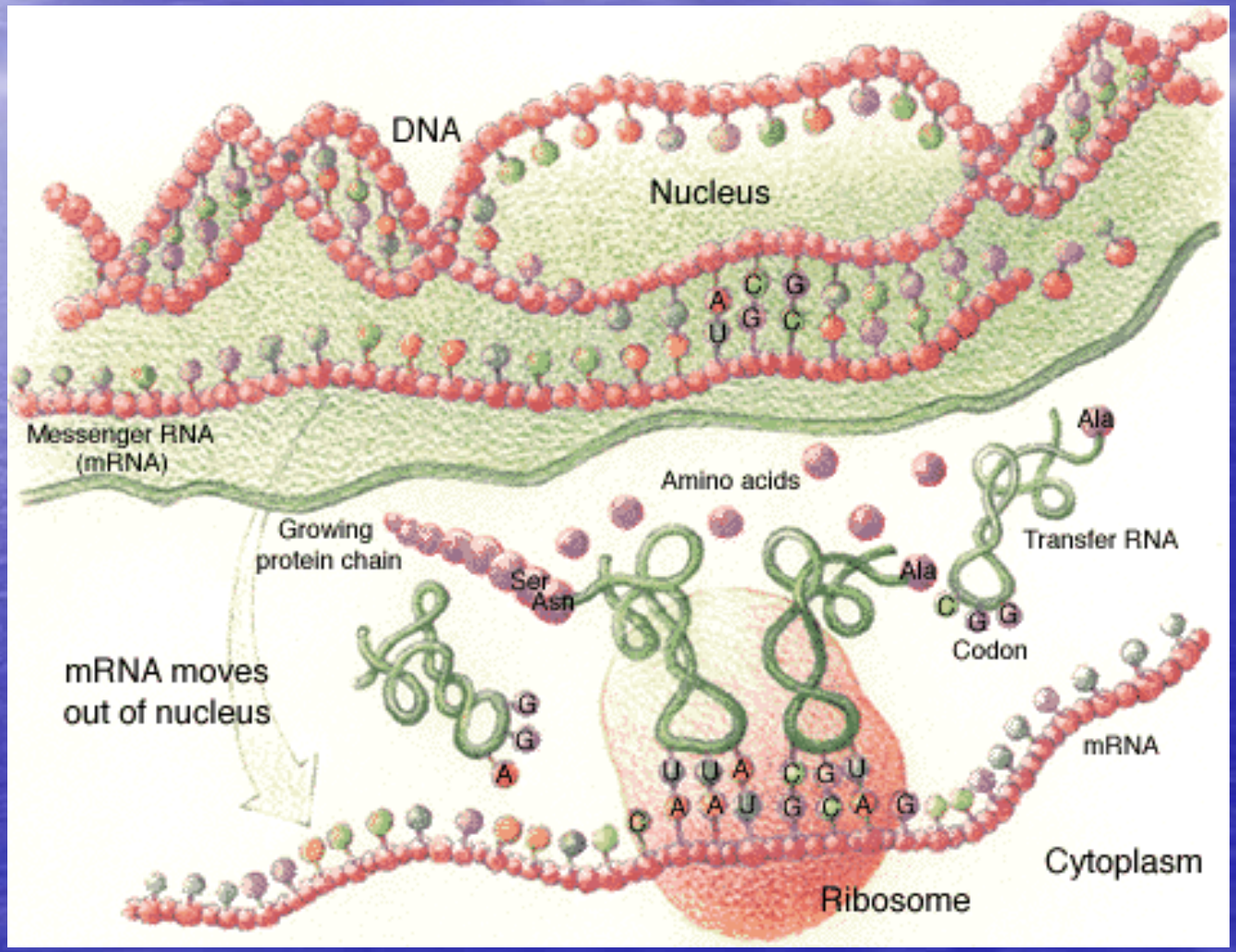
Translation



Protein

Cells express **different** subset of the genes in different tissues and under different conditions

From genes to proteins



Source: Oak Ridge National Laboratory

Affymetrix GeneChip® probe array

Image courtesy of Affymetrix

Hybridization of tagged probes

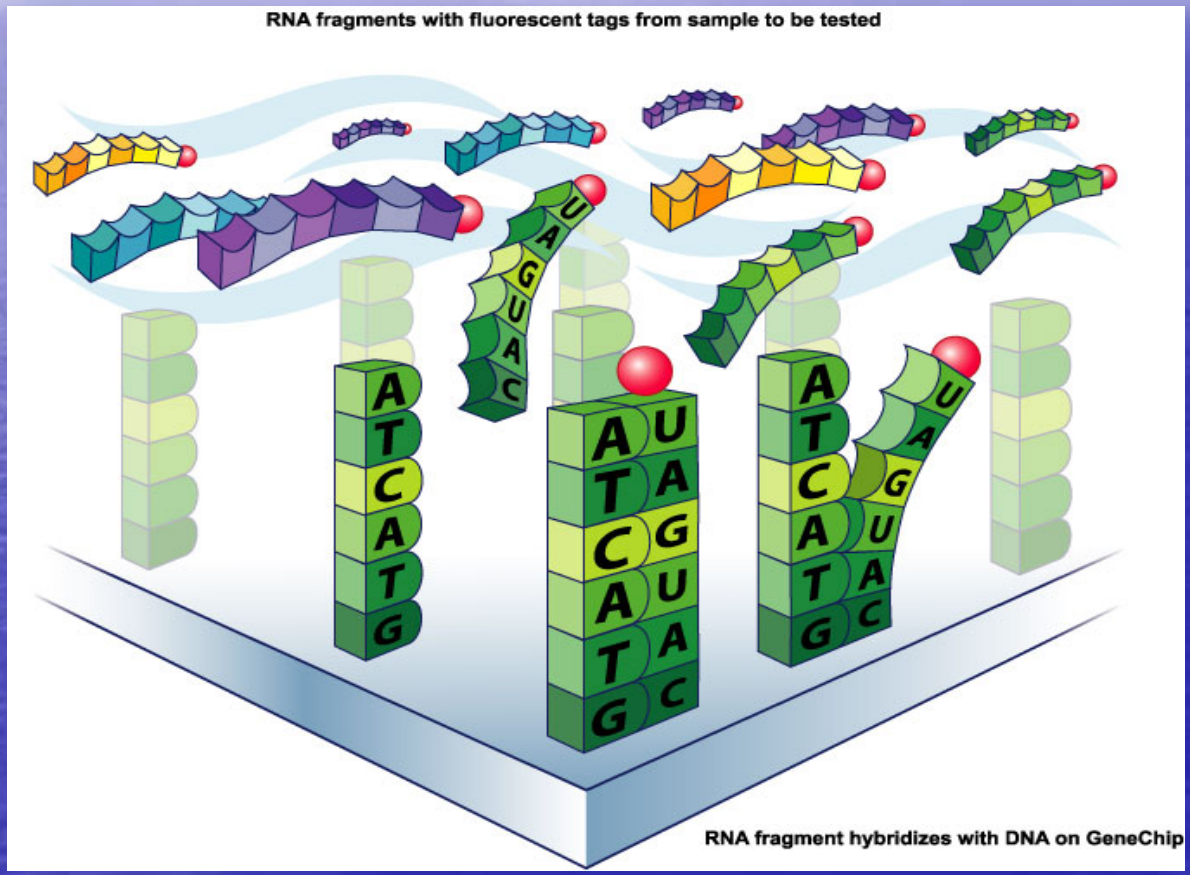
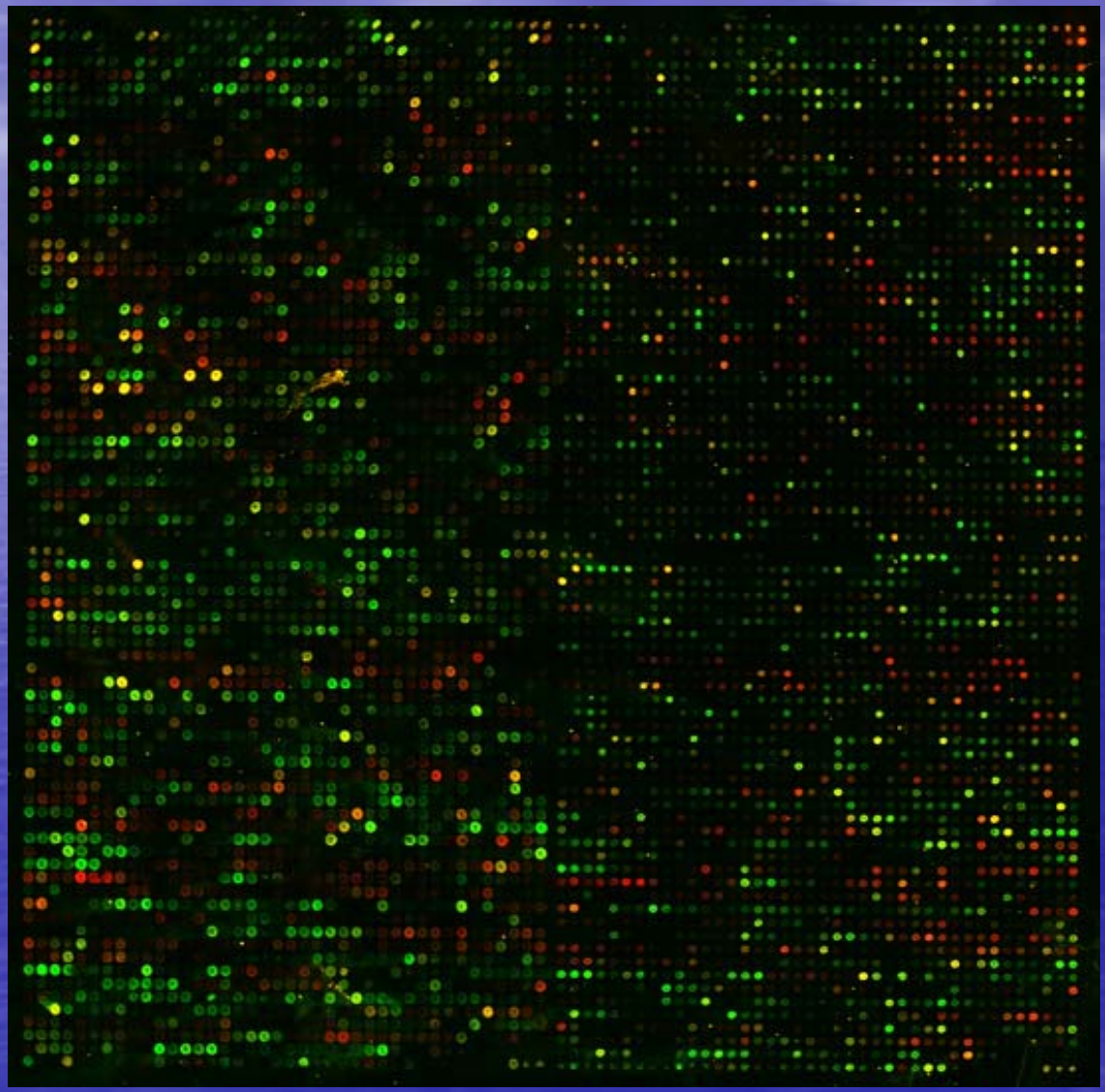


Image courtesy of Affymetrix

Microarray Experiment Result



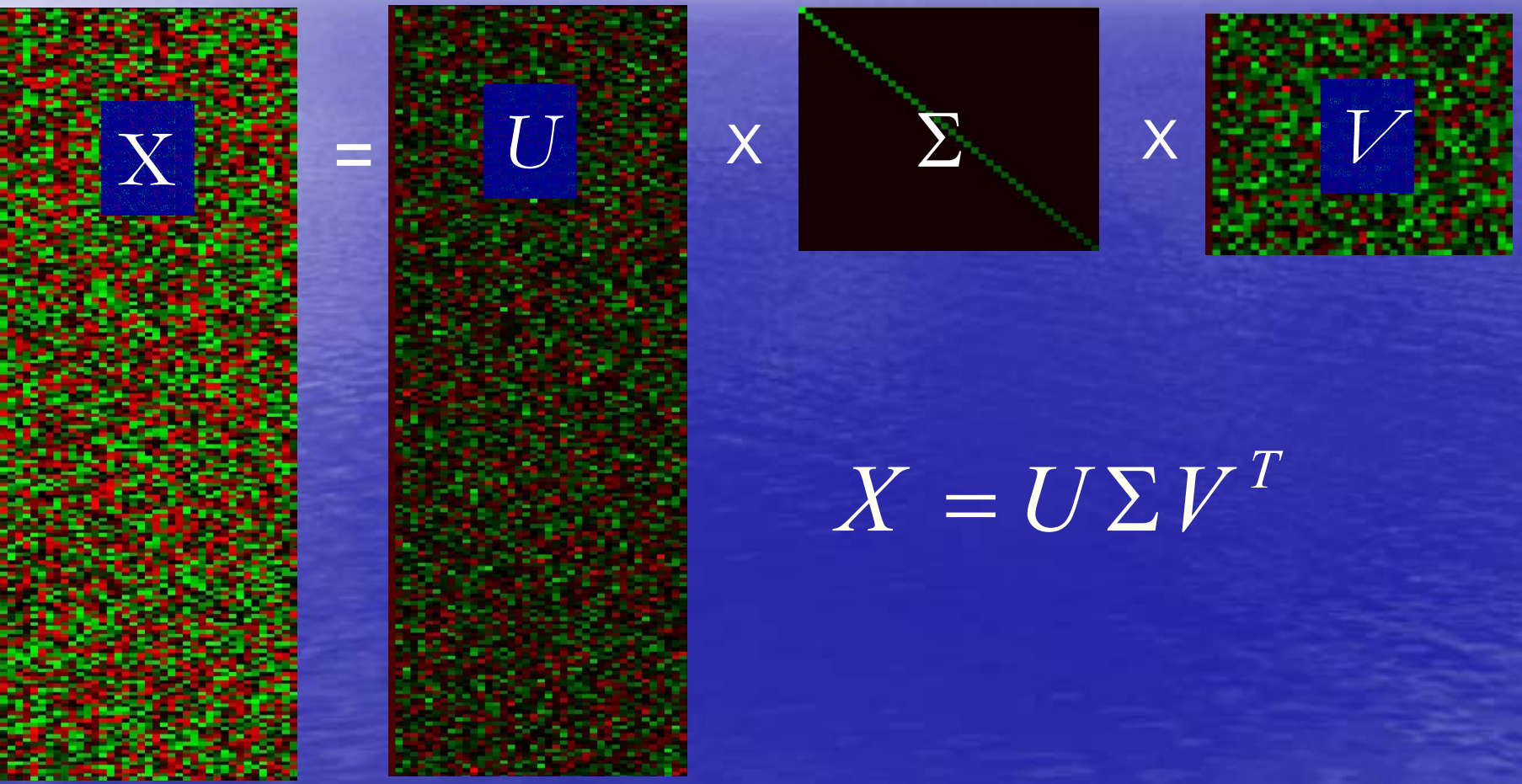
Microarrays

- Advantage
  - High throughput
- Disadvantages
  - The measure of expression level is not accurate
  - Experiments produces high level of noise

Most common analysis: clustering

- Input: N data points, $\mathbf{X}_i$, i=1,2,…,N in a d dimensional space.
- Output: classification of the data points into "natural" groups

Singular Value Decomposition



$$X = U \Sigma V^T$$

Singular Value Decomposition

X is any m by n matrix:

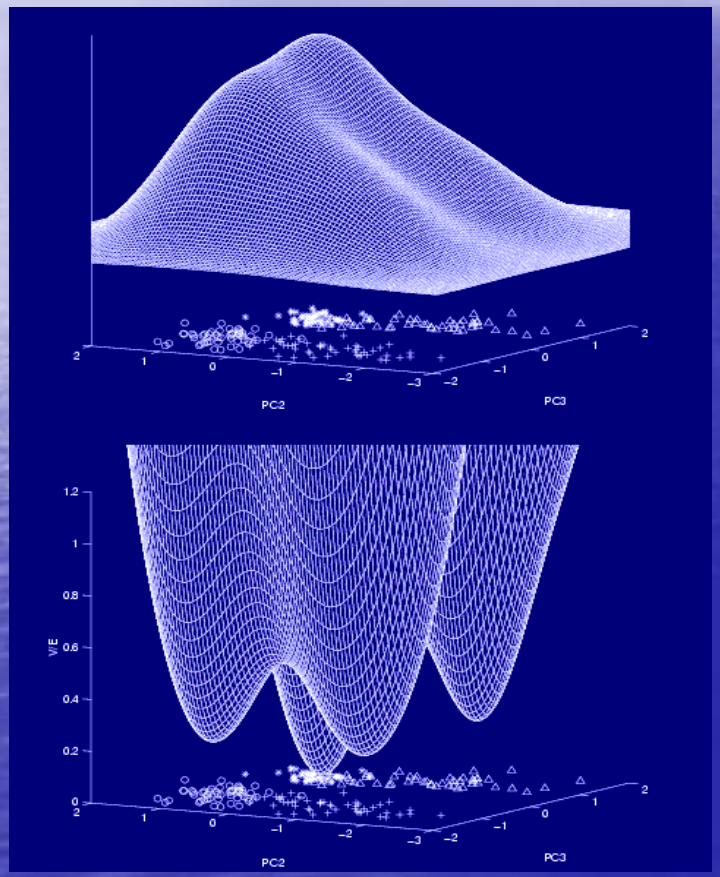
$$X = U\Sigma V^T$$

$$Y = U\Sigma^r V^T$$

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

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

Quantum Clustering



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

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

Hierarchical Quantum Clustering (HQC)

- Start with raw data matrix containing gene expression profiles of the samples.
- Apply SVD and truncate to r-space by selecting the first r significant eigenvectors
- Apply QC in r-dimensions starting at small scale σ, obtaining many clusters. Move data points to cluster centers and reiterate the process at higher σ. This produces hierarchical clustering that can be represented by a dendrogram.

Example 1– Clustering of human cancer cells

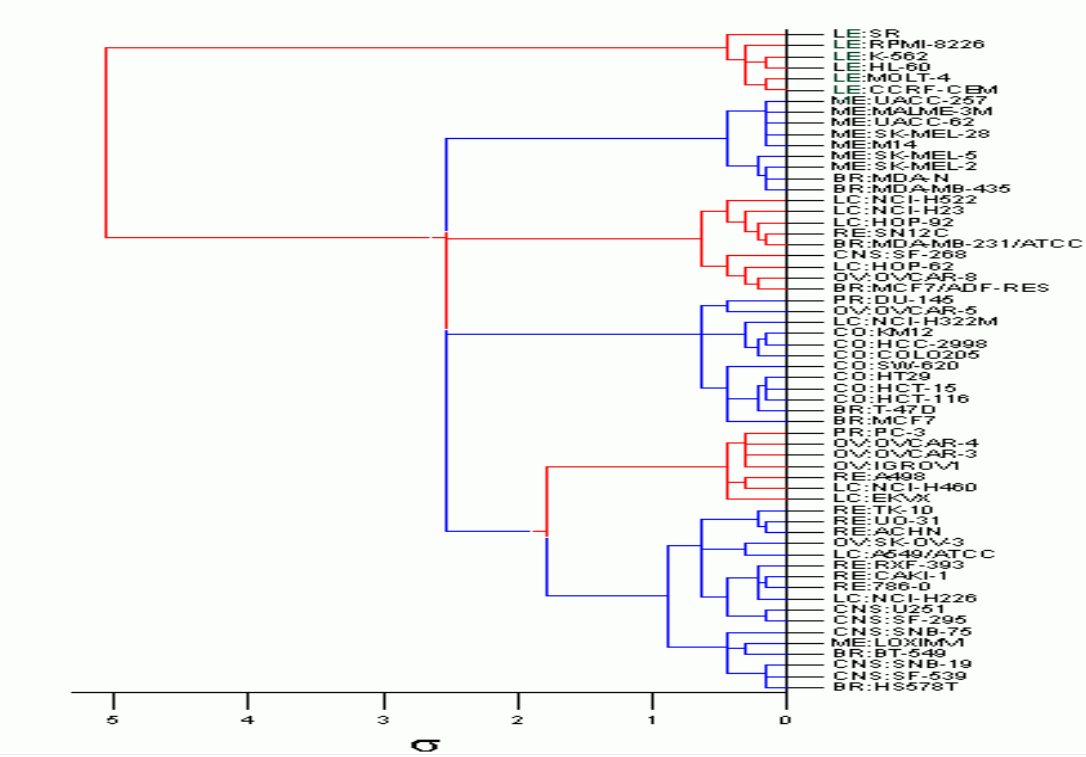
The NCI60 set is a gene expression profile of ~8000 genes in 60 human cancer cells.

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 the number of gene spots is reduced to 1,376 gene subset.
(Scherf et al. – Nature Genetics 24 , 2000)

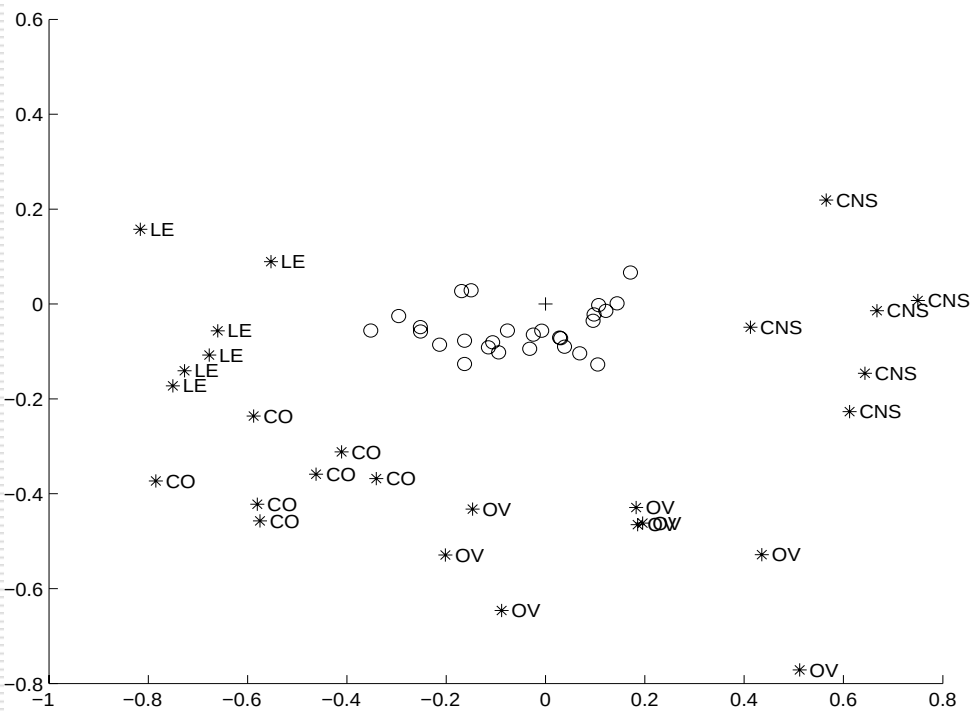
We applied HQC with r=5 dimension.

Example 1 – Clustering of human cancer cells



Dendrogram of 60 cancer cell samples. The clustering was done in 5 truncated dimensions. The first 2 letters in each sample represent the tissue/cancer type.

Example 1 - Projection onto the unit sphere



Representation of data of four classes of cancer cells on two dimensions of the truncated space. The circles denote the locations of the data points before this normalization was applied

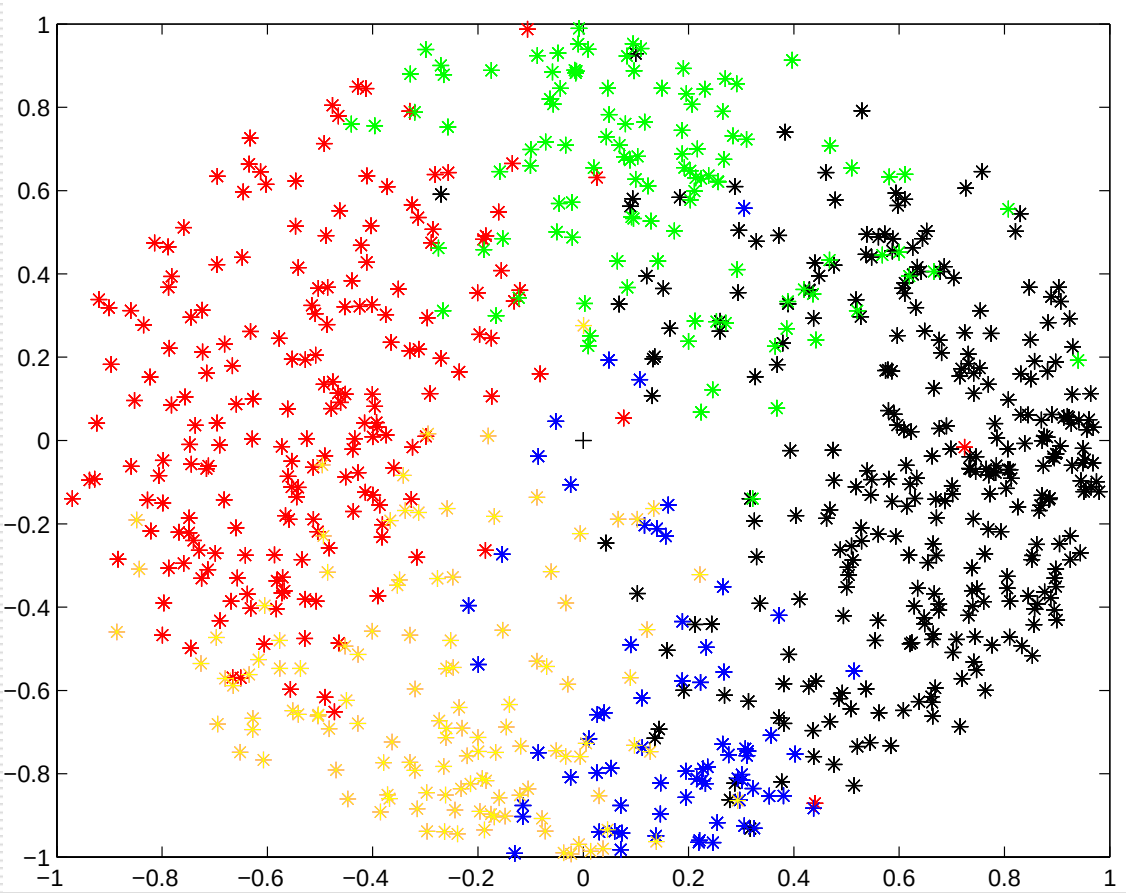
Example 2 – Yeast cell cycle

Yeast cell cycle data were studied by several groups who have applied SVD. *(Spellman et al. Molecular Biology of the Cell, 9, Dec. 2000)* We use it to test clustering of genes, whose classification into groups was investigated by Spellman et al.

The gene/sample matrix that we start from has dimensions of 798x72, using the same selection as made by *(Shamir, R. and Sharan, R. 2002).*

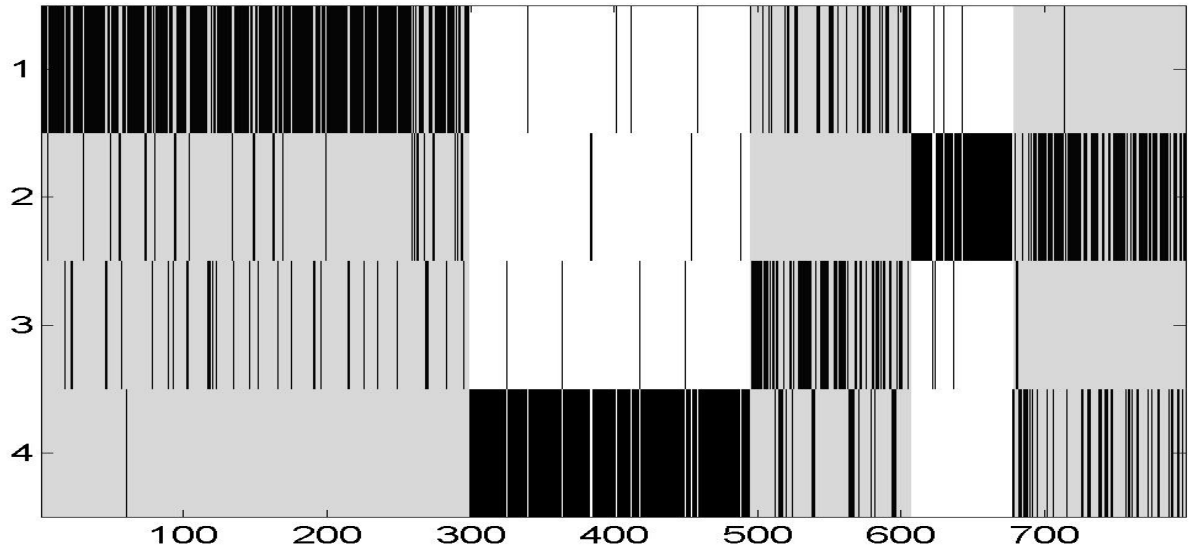
We truncate it to r=4 and obtain, once again, our best results for σ=0.5, where four clusters follow from the QC algorithm.

Example 2 – Yeast cell cycle



The five gene families as represented in two coordinates of our r=4 dimensional space.

Example 2 – Yeast cell cycle



Cluster assignments of genes for QC with s=0.46 , as compared to the classification by Spellman into five classes, shown as alternating gray and white areas .

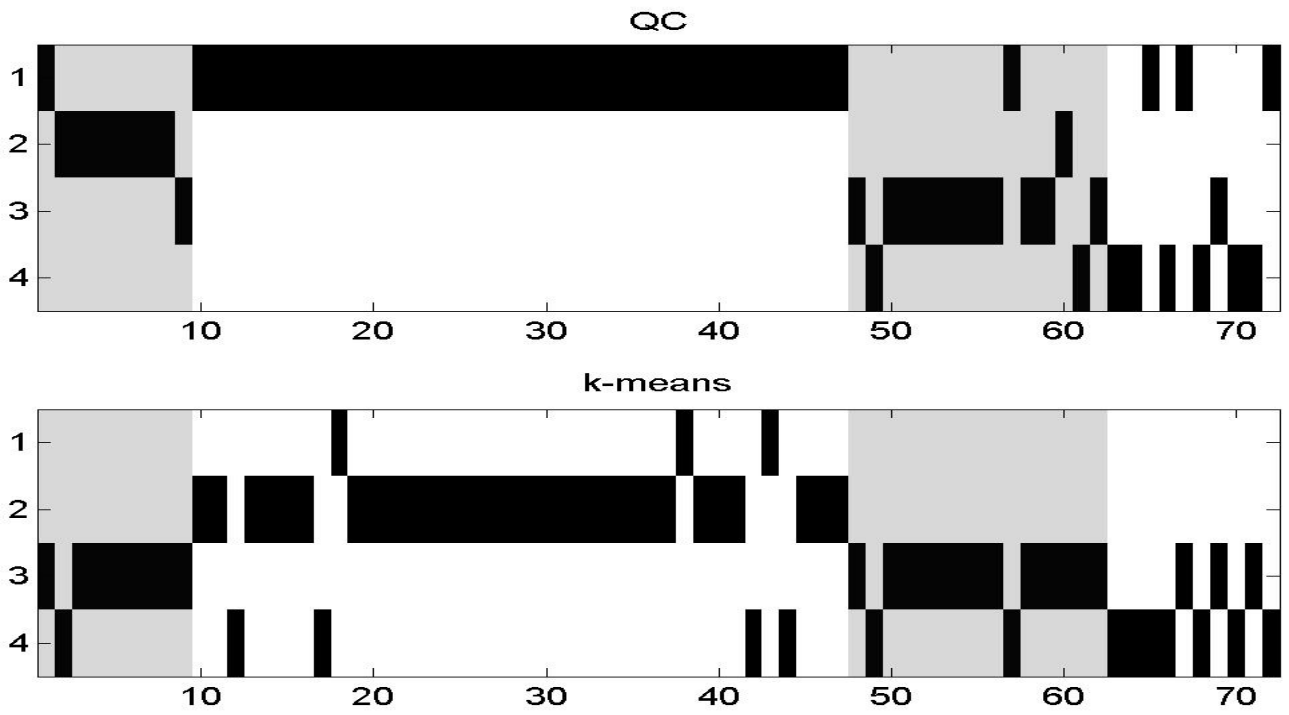
Example 3 – AML/ALL

This data set is taken from 72 leukemia patients with 2 types of leukemia called ALL and AML. *(Golub,T. et al. 1999. Science, 286)*

The ALL set is further divided into T-lineage leukemia and B-lineage leukemia and the AML set is divided into patients who have undergone treatment (with an anthracycline-cytarabine regimen) and those who have not.

The microarray data correspond to 72 samples tested on 7129 genes. Using SVD we truncate it down to r=5, where we obtain clustering results, conforming to the four classes that exist in this data set.

Example 3 – AML/ALL



Clustering solutions for the AML/ALL problem using . The samples are ordered on the x-axis according to the classification into four groups indicated by alternative gray and white areas

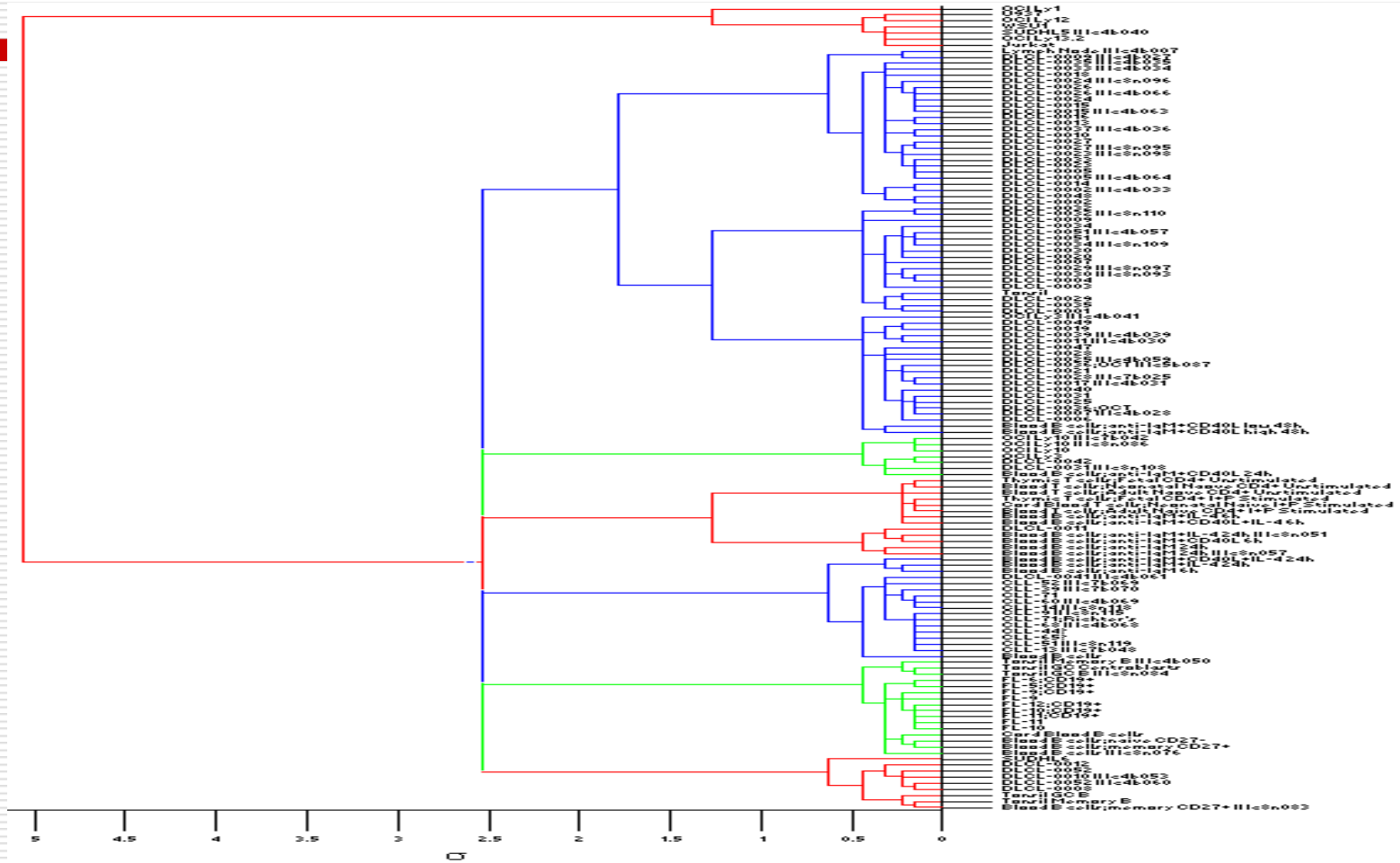
Example 4 – B-cell lymphoma

Measurements of gene expression were made using 128 microarrays and 96 samples of normal and malignant lymphocytes.

The microarrays used were specialized ones (`Lymphochip‘), designed by selecting genes that are preferentially expressed in lymphoid cells and genes with known or suspected roles in processes important in immunology or cancer. *(Alizadeh AA, et al. Nature 403 , 2000)*

We applied HQC with r=6 dimensions and obtained a Jaccard score of 0.85 when compared to given classification of samples.

Example 4 – B-cell lymphoma



Dendrogram of 124 samples of normal and malignant lymphocytes. Clustering was performed in an r=5 eigengene space.

Example 5 – Lung cancer

- We applied QC in 3 dimensions in order to separate between 28 healthy and cancerous tissues – data was taken from an experiment done by Jossi Hillel's lab
- The result was 1 misclassification when using ~5000 out of 22,000 genes
- Attempt to further separate the samples into smokers and non-smokers has failed at this point

Measuring Cluster Quality

Given the true solution T for a clustering problem we can compare a proposed solution S in a way called Jaccard score:

$$Jaccard = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

n_{11} - The Number of pairs of elements that are in the same cluster in both S and T.

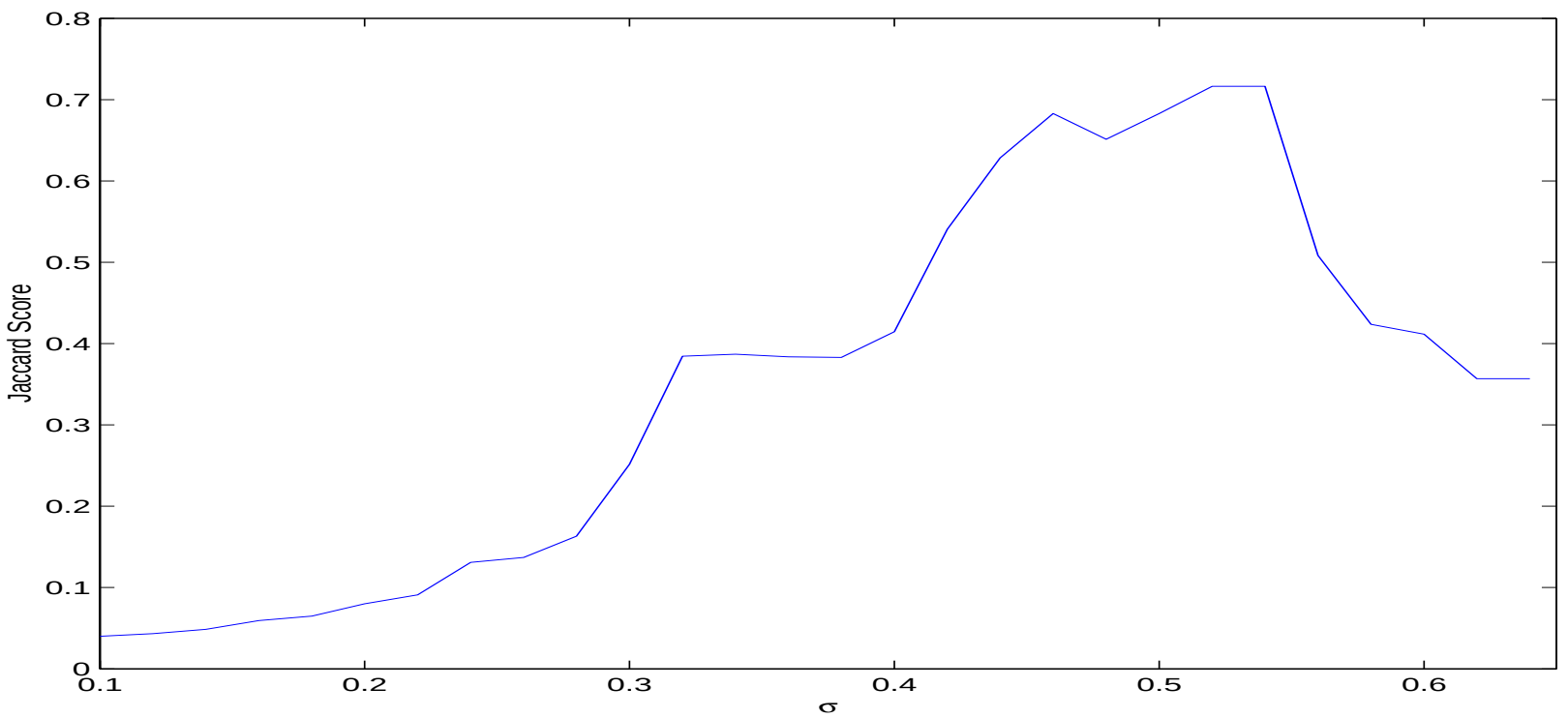
n_{01} - The Number of pairs that are in the same cluster only in S

n_{10} - The Number of pairs that are in the same cluster only in T.

Note that the Jaccard score is better when score is higher with best score at 1

Measuring Clustering Quality

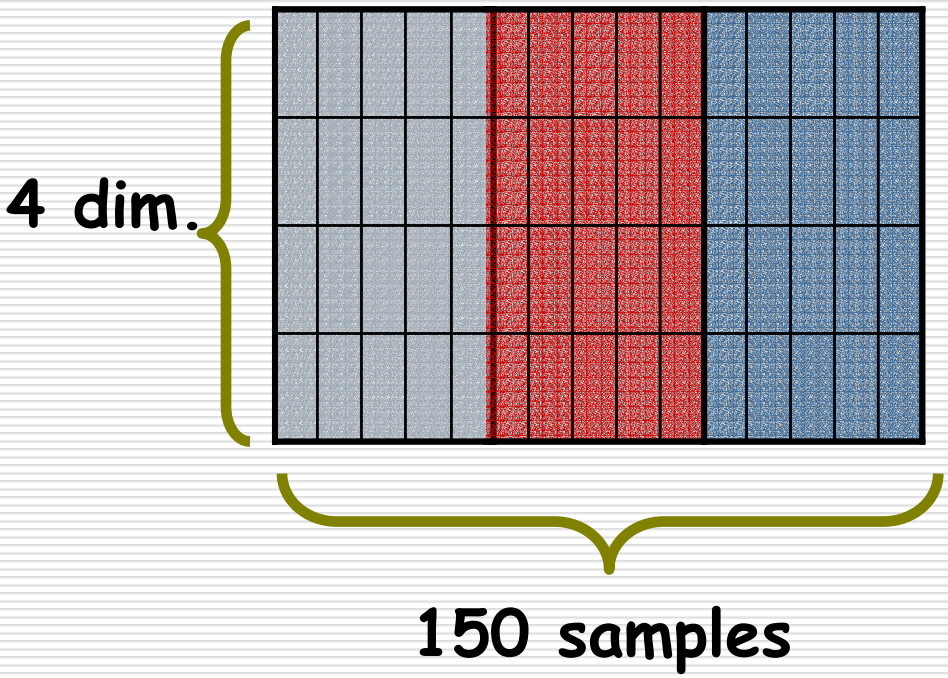
The Jaccard score behavior as a function of sigma



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

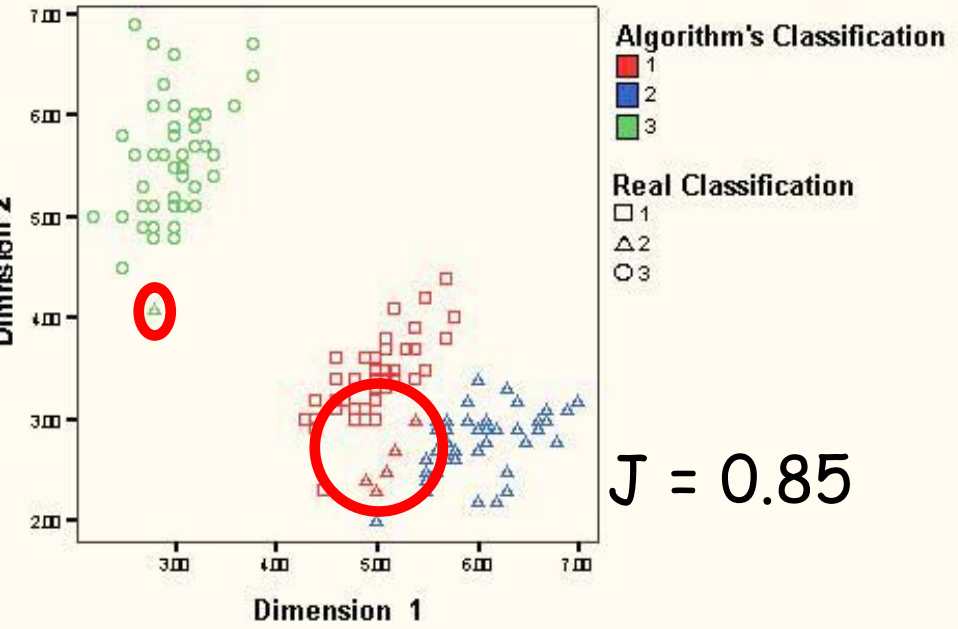
back to The Iris Flower Example

- Iris Flower
  - 150 flowers
  - 4 indexes
  - 3 groups



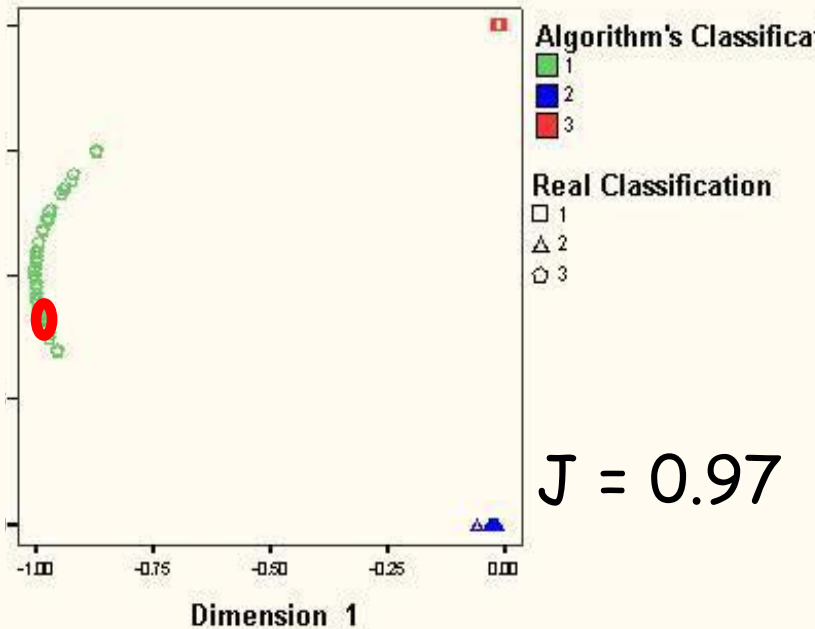
Iris Flower (Cont.)

Raw data (dims 1 & 2)



J = 0.85

Data After SVD (dims 1 & 3)



J = 0.97

QUI – Quantum clustering User Interface

- ***QUI*** is a GUI Matlab tool that enables an easy and intuitive way to compare some clustering methods.
- ***QUI*** is a five-step wizard that envelops some basic Matlab clustering methods and introduces the Quantum clustering algorithm. ***QUI*** provides a flexible and customizable interface for clustering data with high dimensionality.
- ***QUI*** allows both textual and graphical display for the clustering results

How to Install?

QUI is a self-extracting package. In order to install and run the QUI tool, follow these three easy steps

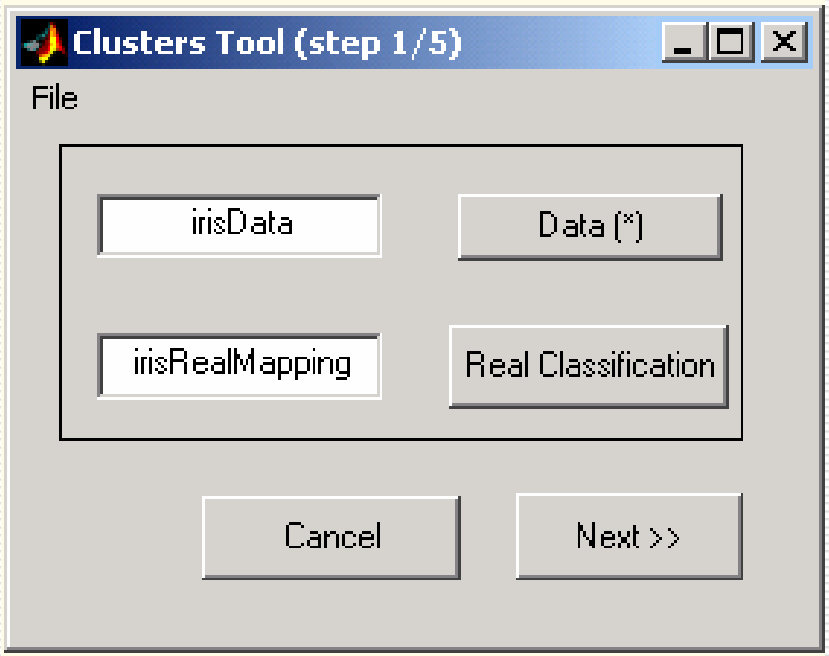
Download the QUI.zip packaged to your local drive.

Add the QUI destination directory to your Matlab path.

Within Matlab, type 'qui' at the command prompt.

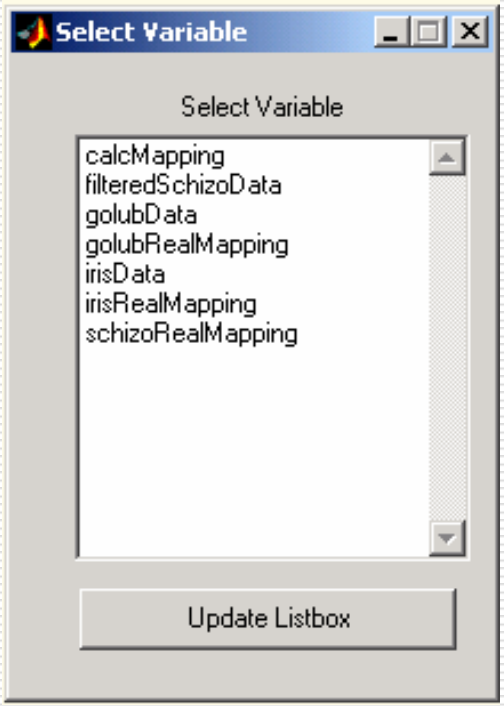
Steps – 1

Input parameters



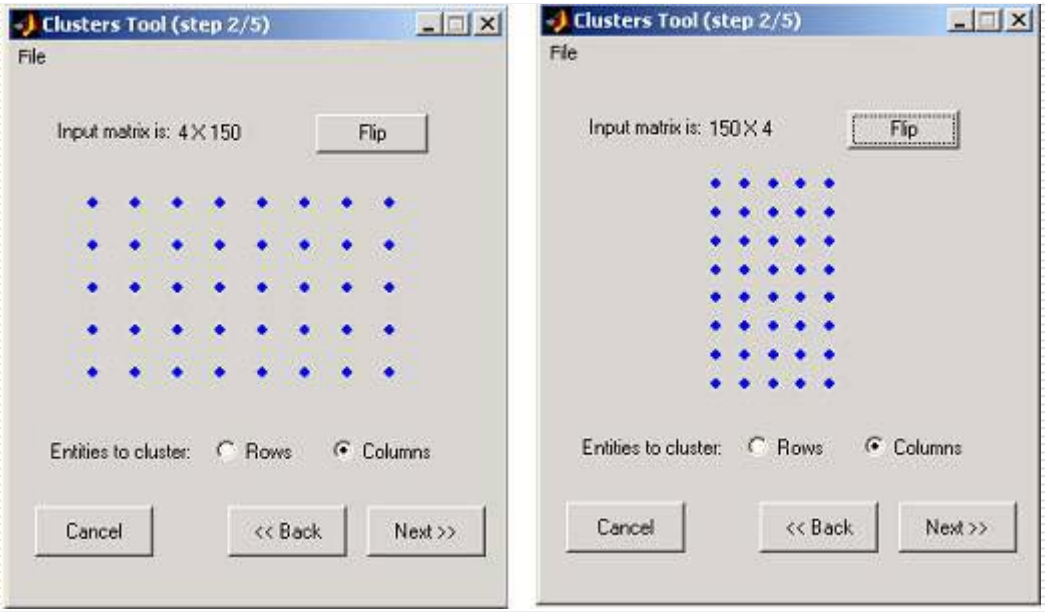
Steps – 1

Selecting variables



Steps – 2

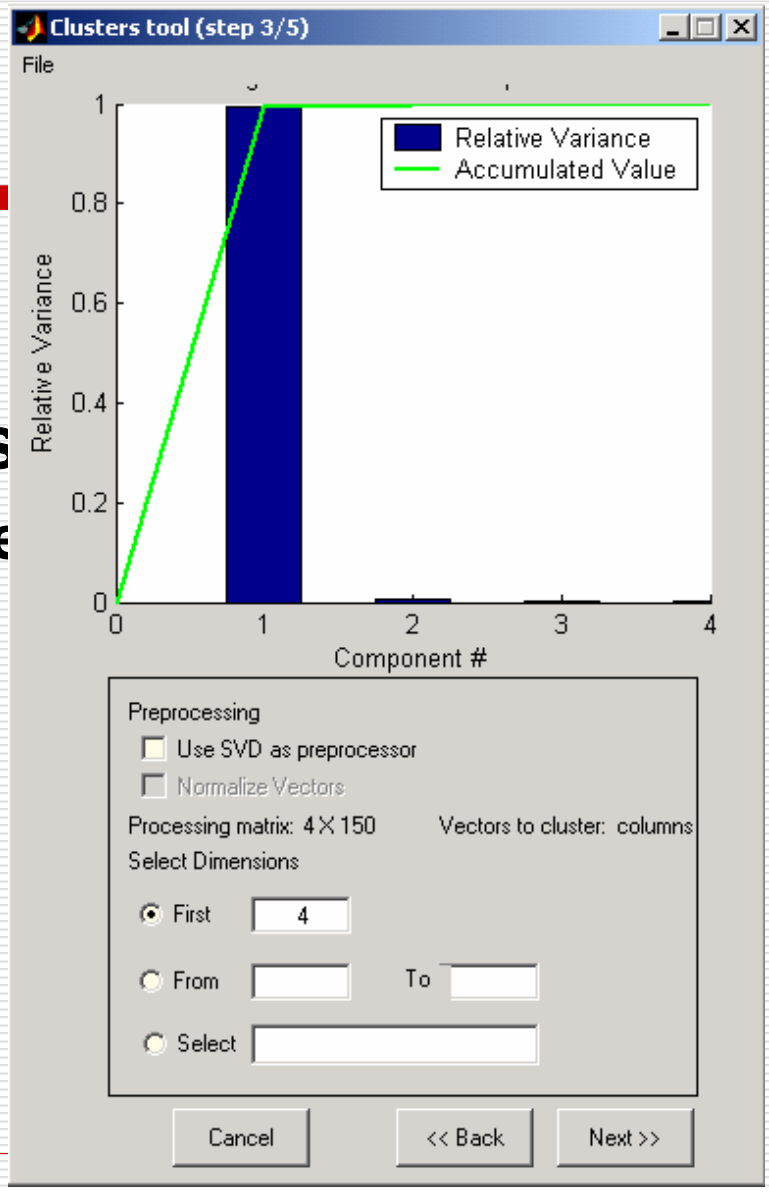
Determining the matrix shape and vectors to cluster



Steps – 3

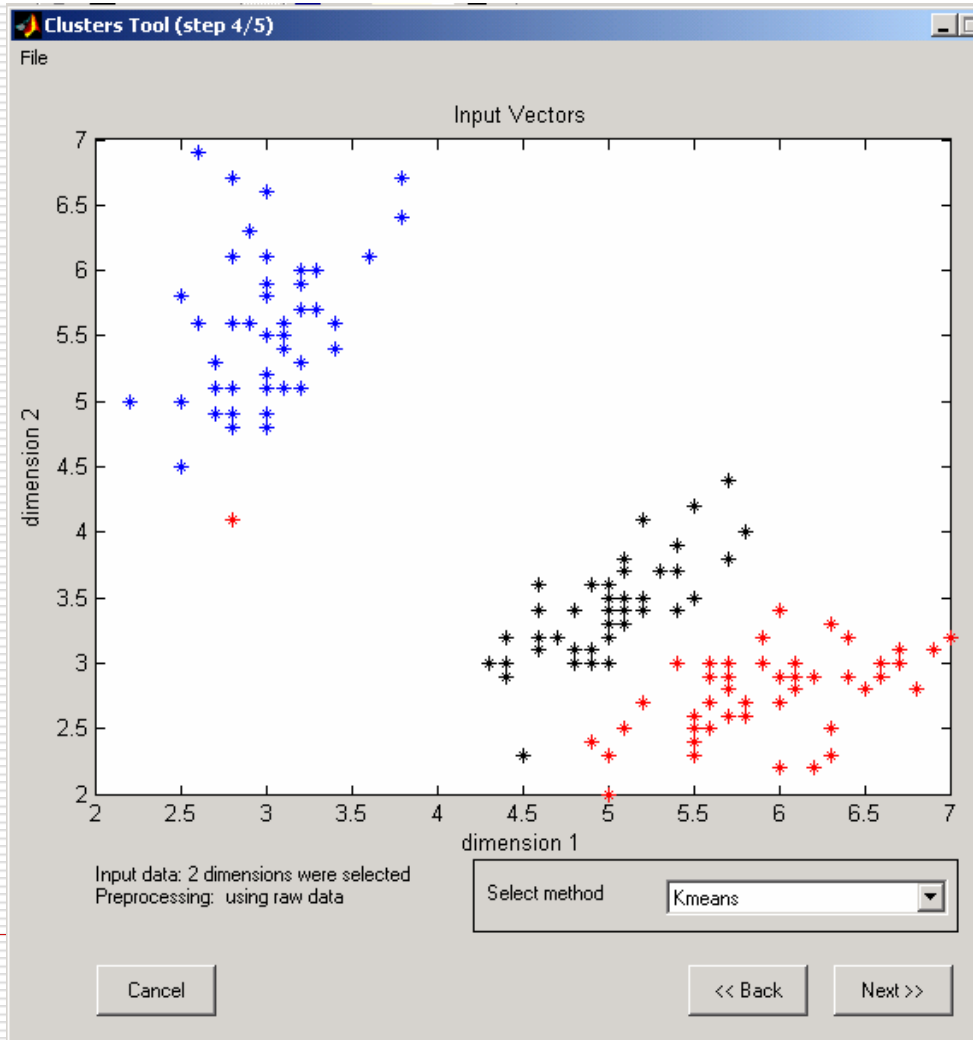
Preprocessing Procedures

- Components' variance graphs
- Preprocessing parameters



Steps – 4

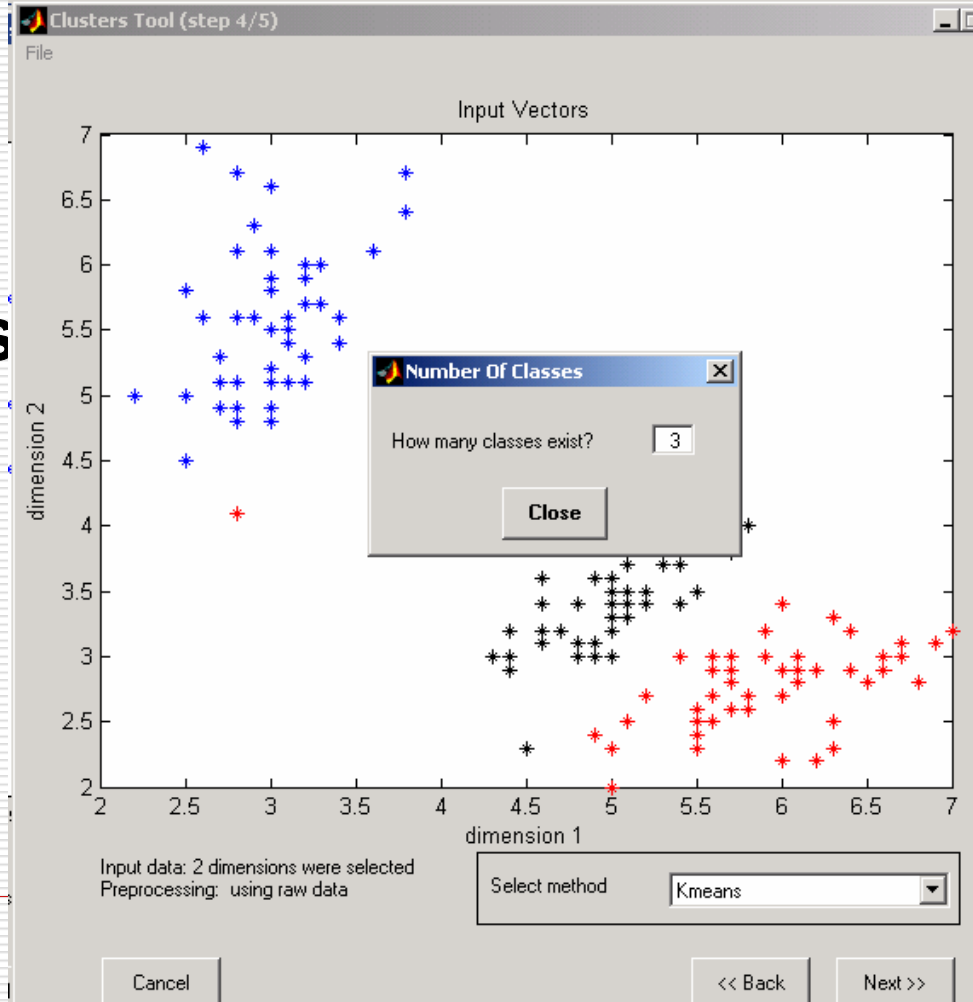
Points distribution preview and clustering method selection



Steps – 5

Parameters for clustering algorithms

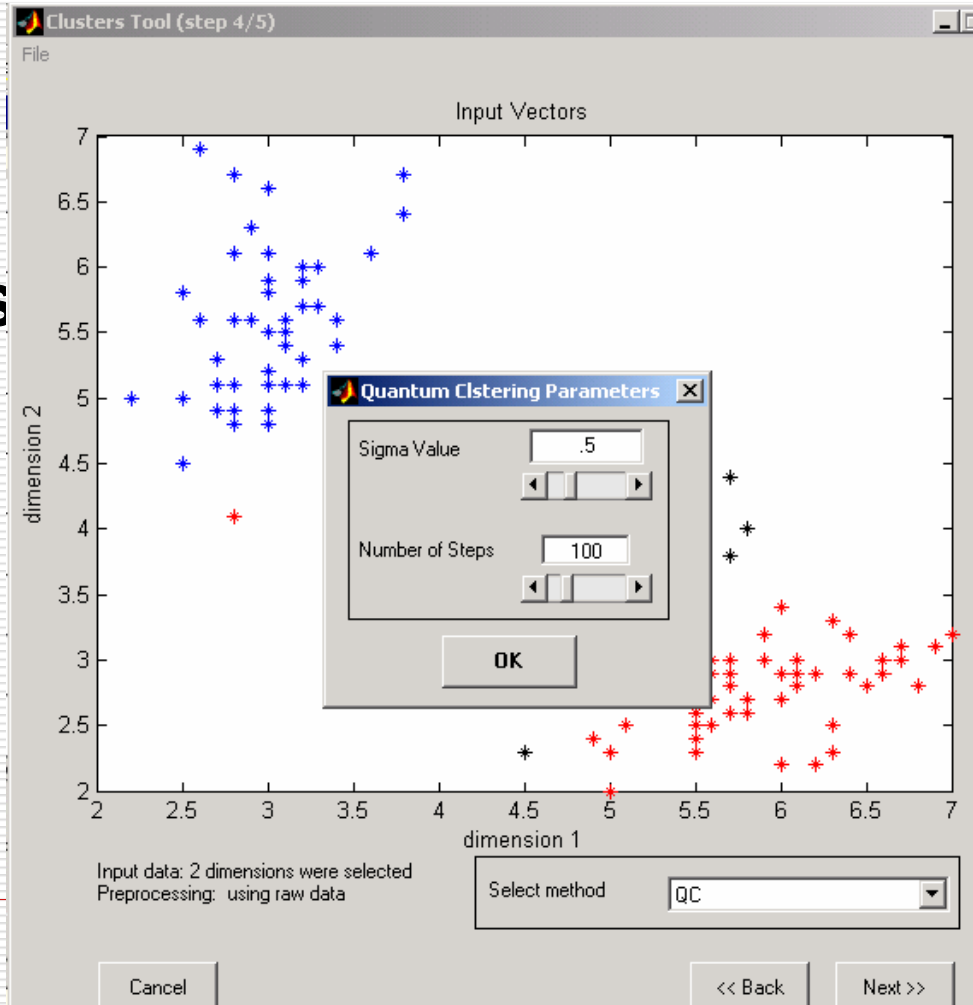
- Kmeans



Steps – 5

Parameters for clustering algorithms

- QC



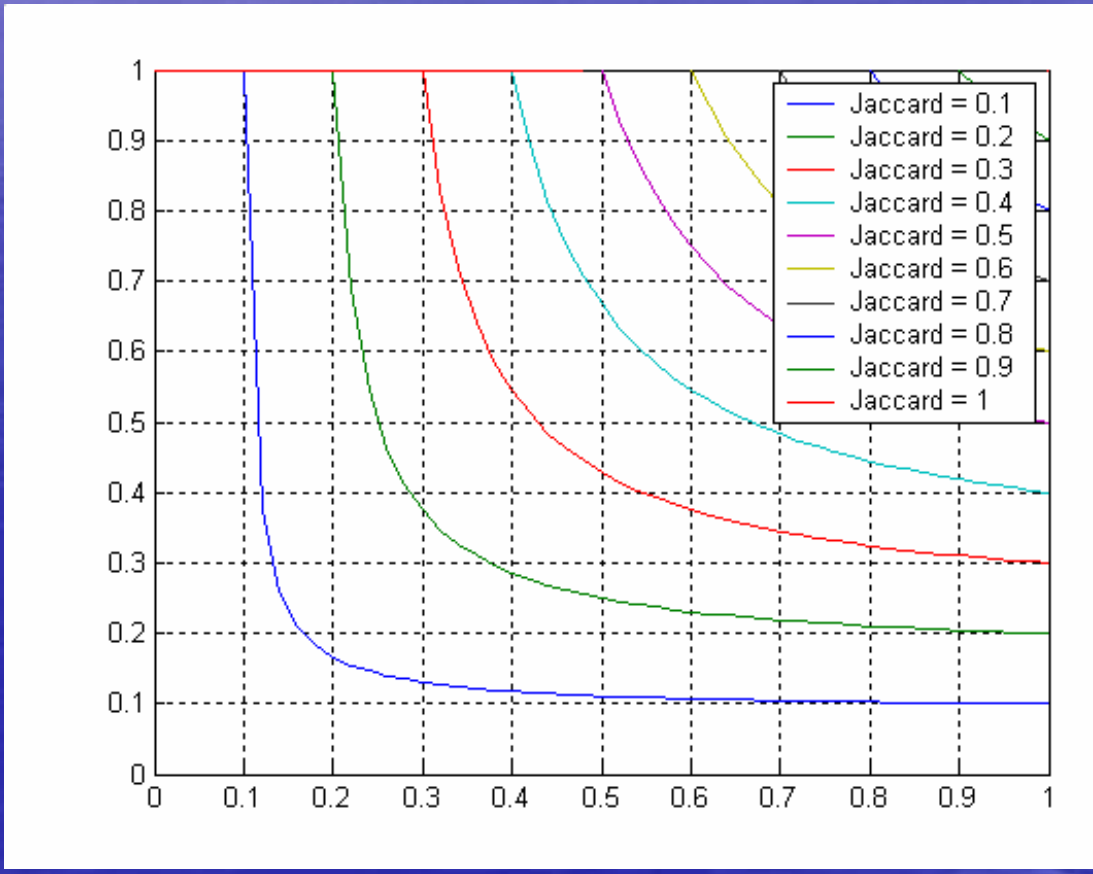
Steps – 6

QUI result

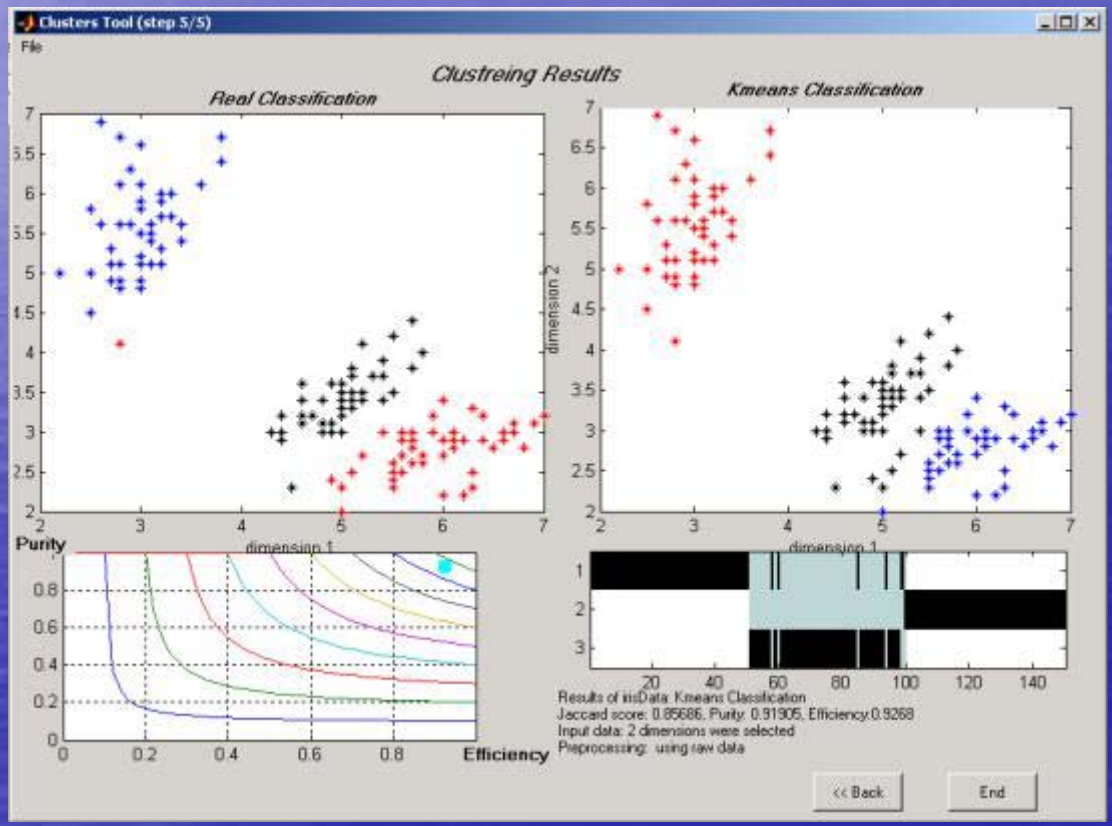
$$Efficiency = \frac{n_{11}}{n_{11} + n_{10}}$$

$$Purity = \frac{n_{11}}{n_{11} + n_{01}}$$

$$Jaccard = \frac{n_{11}}{n_{11} + n_{01} + n_{10}}$$



Steps – 6
Results



Summary - SVC

SVC is a kernel method employing the principle of enclosing images of data points in a minimal sphere in Hilbert space.

Working with a Gaussian kernel we obtain good clustering results varying the width parameter q.

For a problem with overlapping distributions of data we make use of the possibility of handling outliers in SVM methods. The fraction p of outliers becomes another parameter of our system.

SVC supplies cluster contours in data space.

For large p values it bears similarity to the state-scale (Parzen window) approach.

Summary - QC

The QC algorithm constructs a potential function from normalized second order moments of Parzen window distribution.

It identifies potential minima with cluster centers.

It has one free parameter, the scale $\sigma=1/\sqrt{(2q)}$.

QC can work with full spatial information in low dimensions (e.g. few PCs) or with distance information in any dimension.

Cluster assignment via gradient descent leads to very good results.

Summary – biological applications

We have employed a new form of SVD processing to microarray data along with QC as a clustering method.

Truncating SVD down to 4-5 dimensions, turns out to be a useful method for representing microarray data. Projection onto the unit sphere allows for separation between different classes.

QC, and its hierarchical version HQC, provide robust clustering tools that work well in multi-dimensional spaces.

In all test cases we have obtained very good clustering results for choices of σ around 0.5.