Applications of Spectral Clustering To Chemical Datasets

Rajarshi Guha and Peter C. Jurs
Department of Chemistry
The Pennsylvania State University

1

Clustering Chemical Datasets

- Clustering chemical datasets is a fundamental procedure in library design and selection
- Clustering can also be used as a data-partitioning method to focus on smaller subsets
- Clustering can be used as a classification technique

2

The Problem & Solutions

- Chemical datasets can be very large (> 10^5 observations)
- The feature space can be very large and highly correlated
- Clustering is *NP*-hard

- Throw more CPU power to the problem
- Use an approximation algorithm
- Avoid distance matrix calculations

3

SVD Clustering

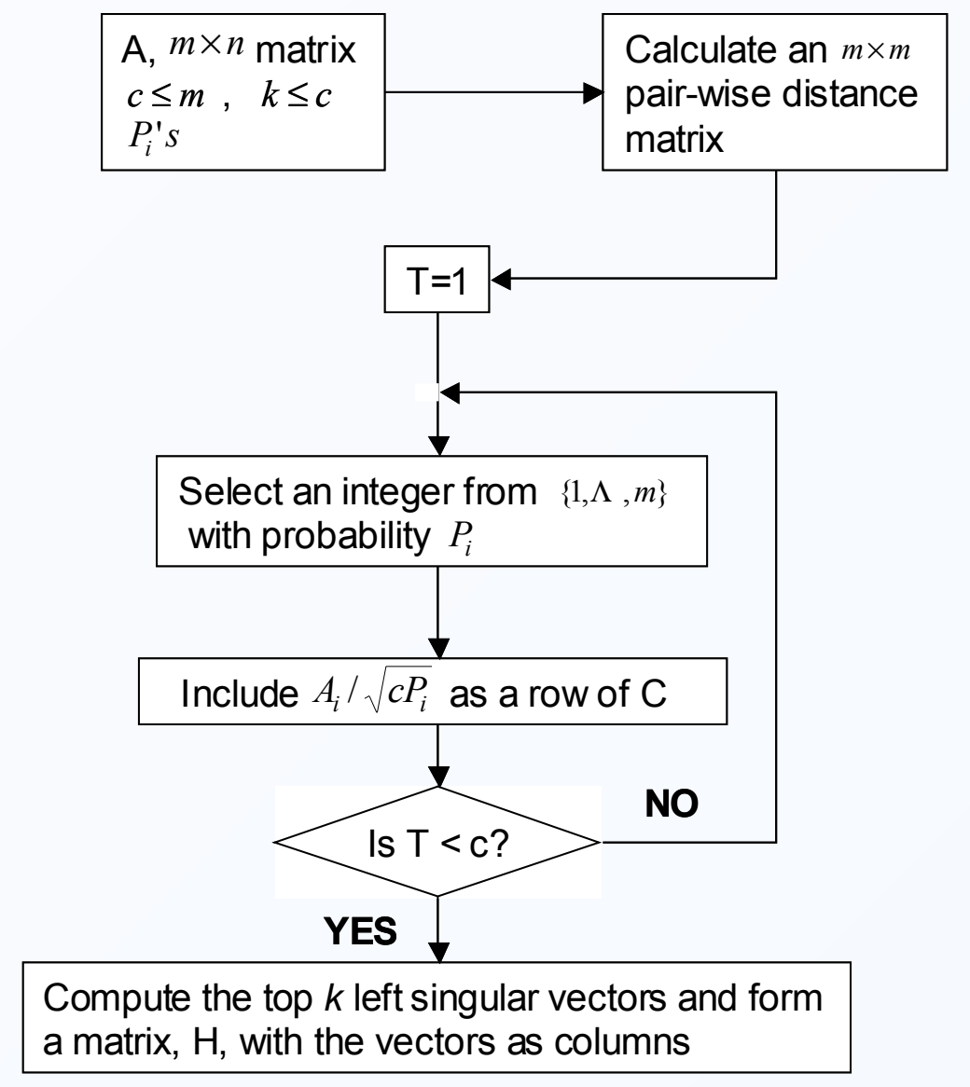
- Based on the Singular Value Decomposition
- Projects the original matrix onto a *k-D* subspace
- Clustering is performed in the reduced subspace
- The algorithm is a *polynomial* time approximation
- It has been shown that the SVD *itself* represents a clustering

Drineas, P.; Frieze, A.; Kannan, R.; Vempala, S; Vinay, V.., *Machine Learning*, 2004, **56**, 9-33

4

Why is *Fast* SVD a Better Way?

- The original matrix is randomly sampled
- The sub-matrix is then decomposed
- Though the *fast SVD* algorithm utilizes the SVD, the matrix being decomposed can be significantly smaller


A, $m \times n$ matrix $c \le m$, $k \le c$ P_i's

Calculate an $m \times m$ pair-wise distance matrix

T=1

Select an integer from $\{1, \Lambda, m\}$ with probability P_i

Include $A_i / \sqrt{cP_i}$ as a row of C

Is T < c?

NO

YES

Compute the top *k* left singular vectors and form a matrix, H, with the vectors as columns

5

Datasets & Descriptors

- AMES Test
- 4337 compounds
- Categorial

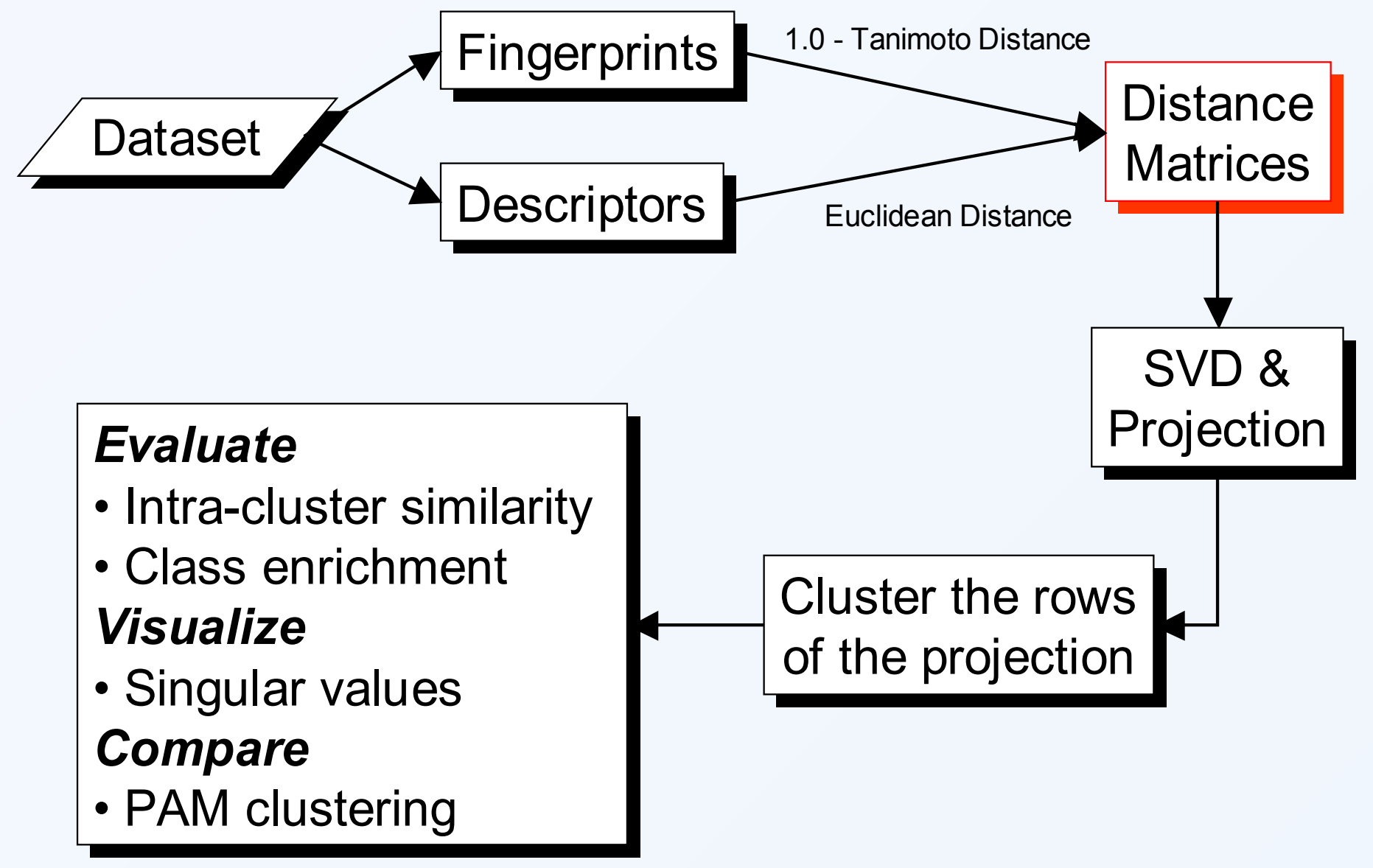
- Aqueous Solubility
- 1256 compounds
- Real valued

- 166 bit MACCS fingerprints
- Constitutional, geometric and topological descriptors
- Calculations and analysis performed with MOE and R

Huuskonen, *J. Chem. Inf. Comput. Sci.*, 2000, **45**, 123-456
Kazius, J.; McGuire, R.; Bursi, R., *J. Med. Chem.*, 2005, **48**, 312-320

6

Methodology


Dataset

Fingerprints

Descriptors

1.0 - Tanimoto Distance

Euclidean Distance

Distance Matrices

SVD & Projection

Cluster the rows of the projection

Evaluate
- Intra-cluster similarity
- Class enrichment

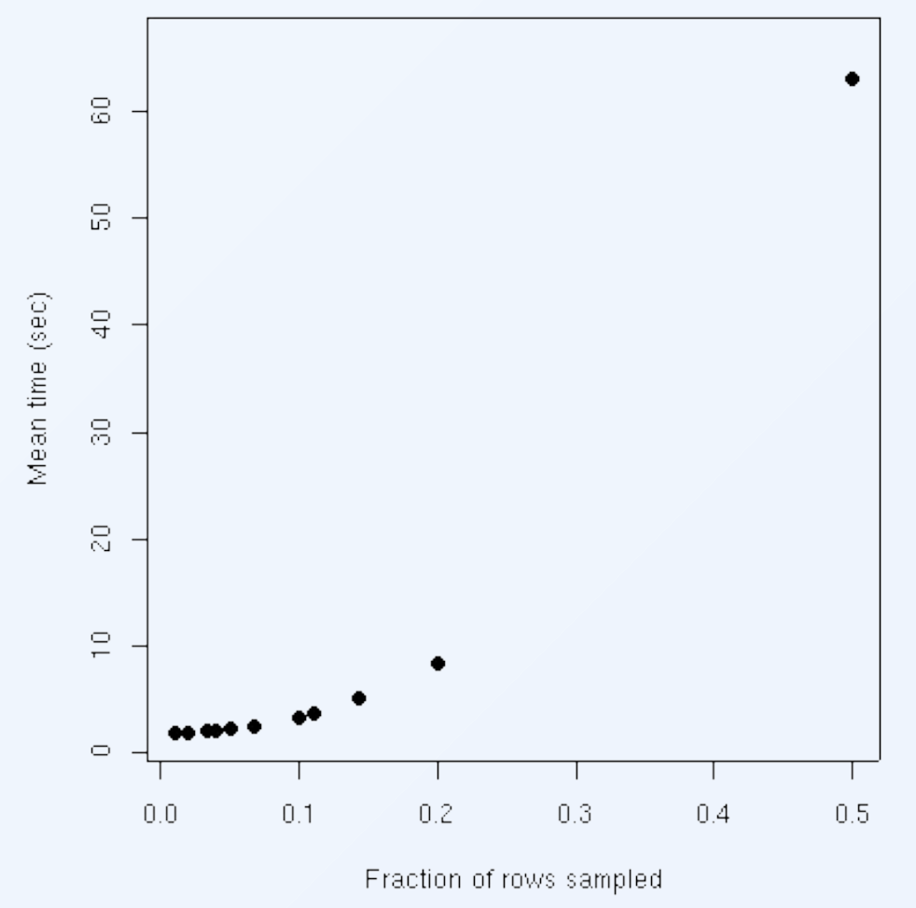
Visualize
- Singular values

Compare
- PAM clustering

7

How Fast is the Fast SVD?

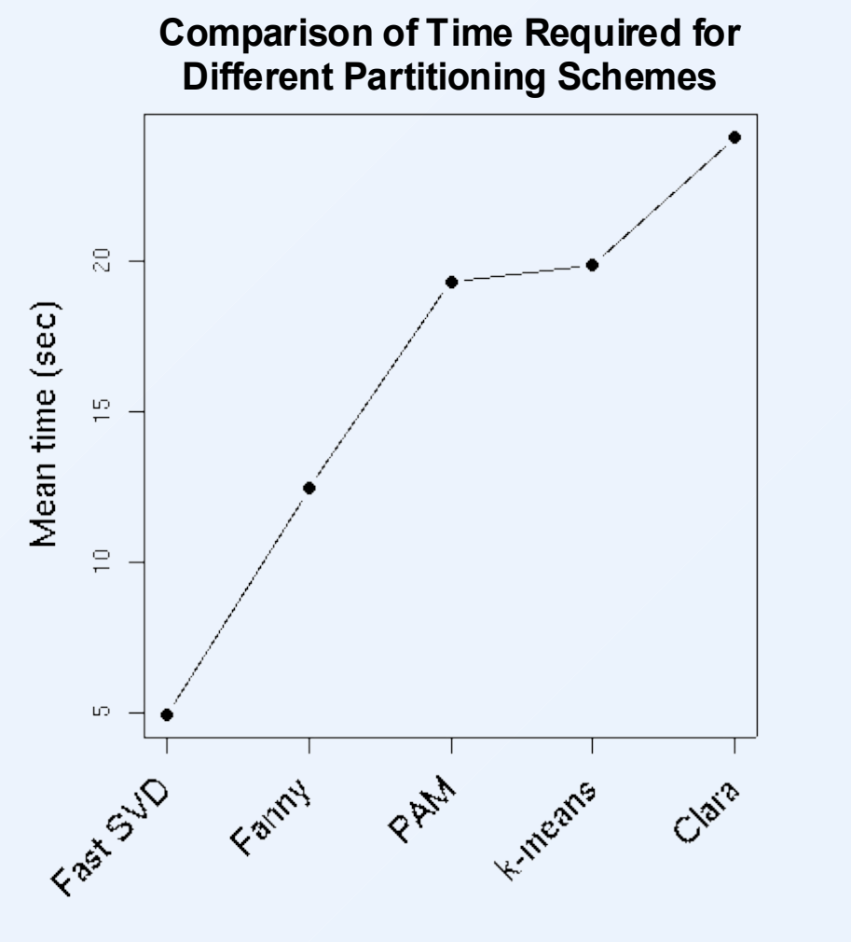
- Benchmarked on a 4337 x 4337 matrix
- Each case was run 10 times
- No significant error until less than 10% of the rows are sampled
- ***Simple SVD = 368s***



8

Comparison of Timings

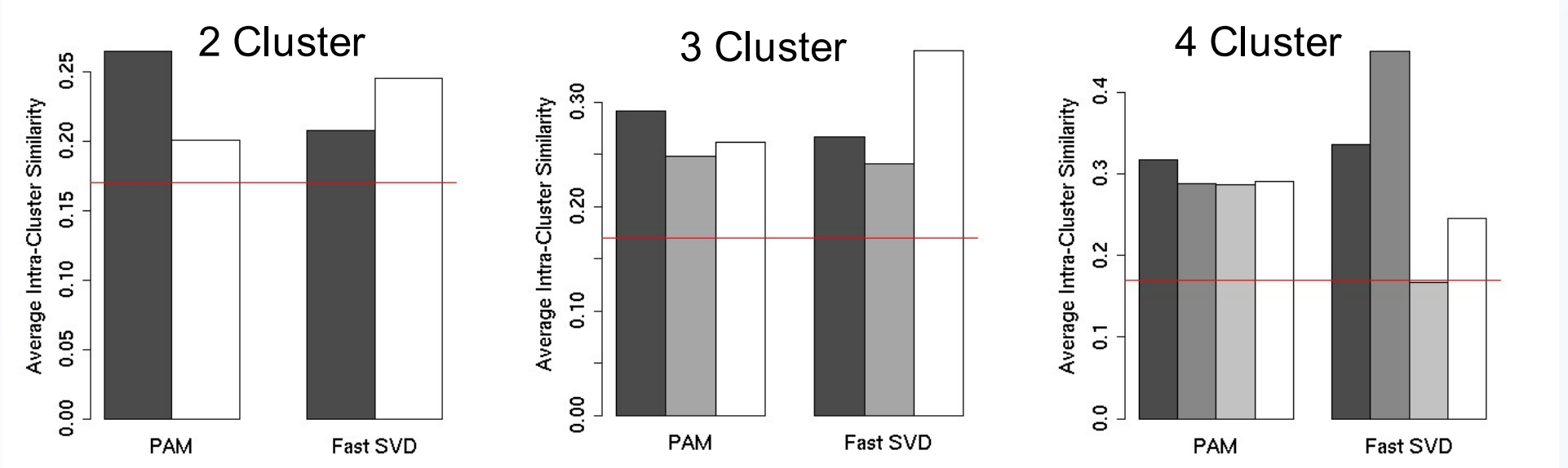
- PAM is a more robust version of *k*-means
- Clara is an extension of PAM but is more suitable for large datasets as it uses a sampling process
- However, the fast SVD is significantly faster than all other partitioning methods

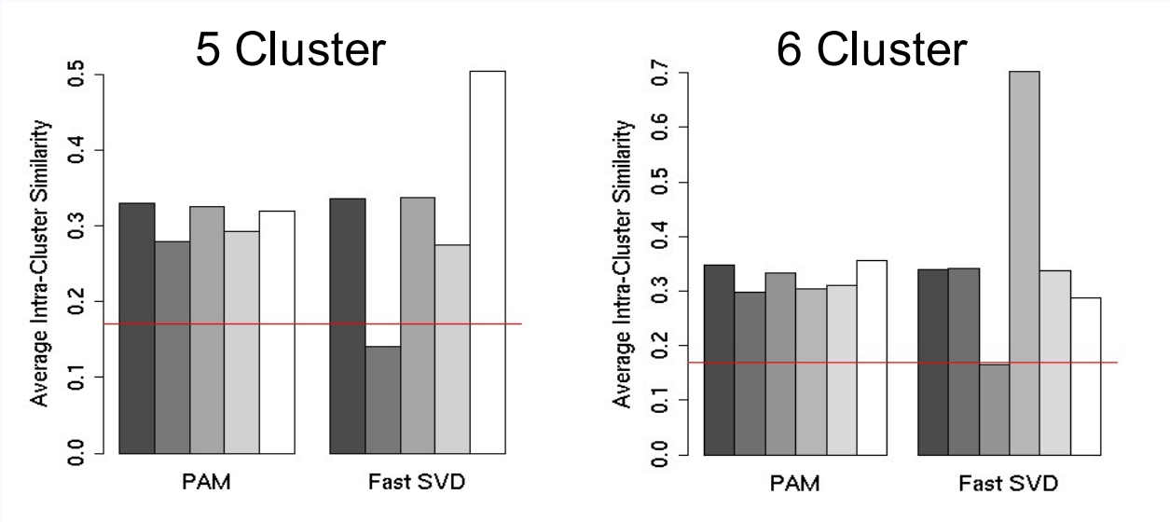


For all cases, *k* = 2 and the times reported are the average of 10 runs. The AMES dataset (4337 compounds) was used

9

Aqueous Solubility - Class Similarity



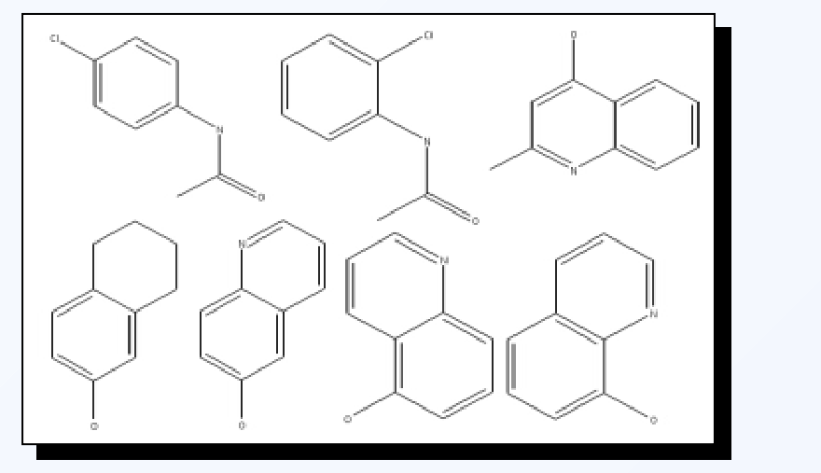


- In most cases, the average intra-class similarity is higher than the overall value
- SVD clustering appears to favor specific classes
- Results are competitive with PAM

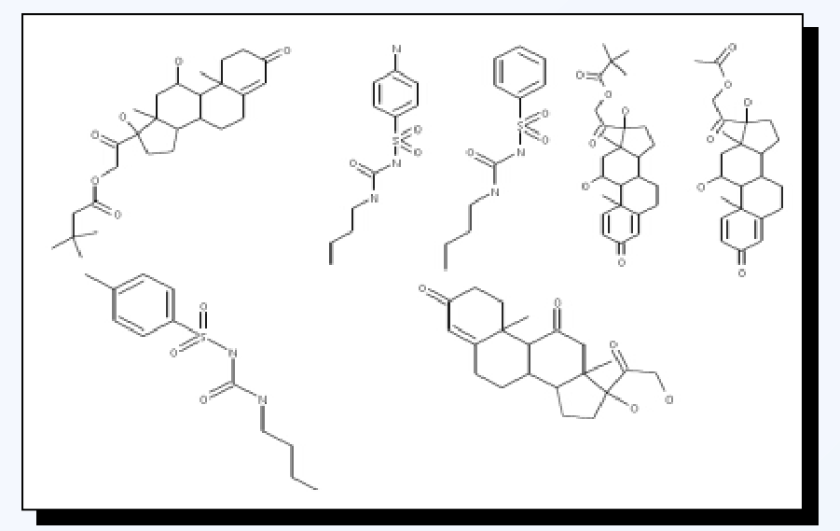
10

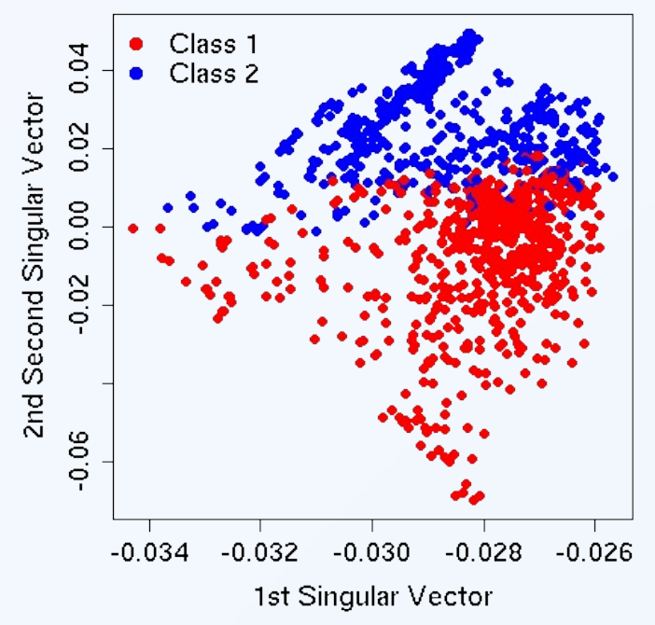
Aqueous Solubility - Class Members

Class 1



Class 2



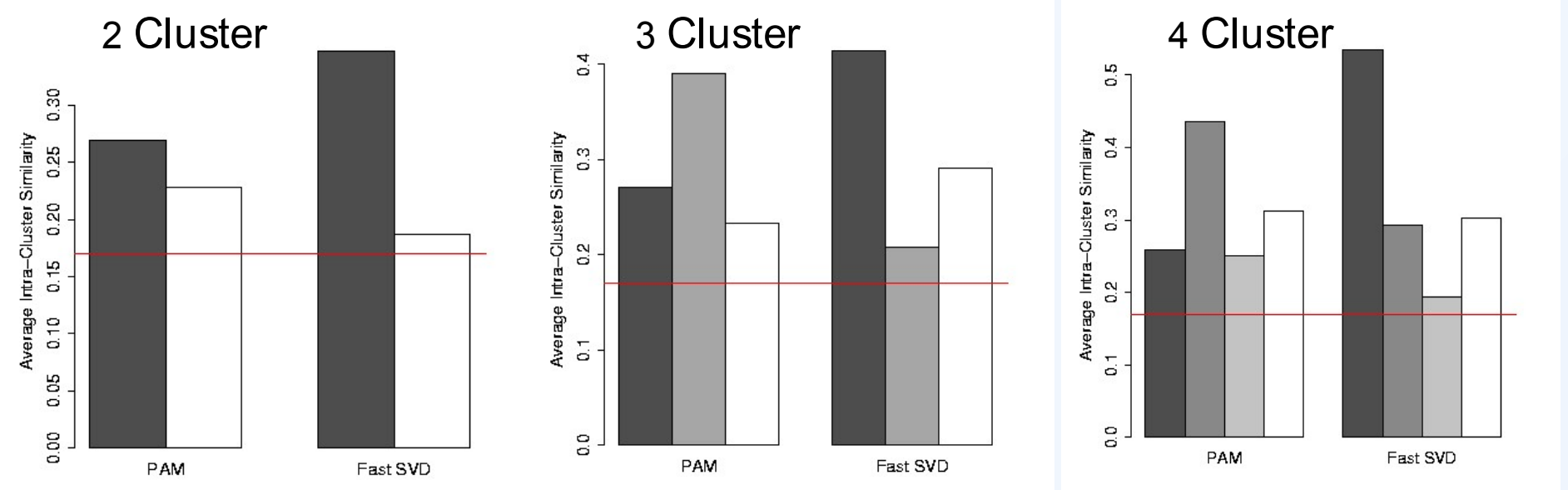


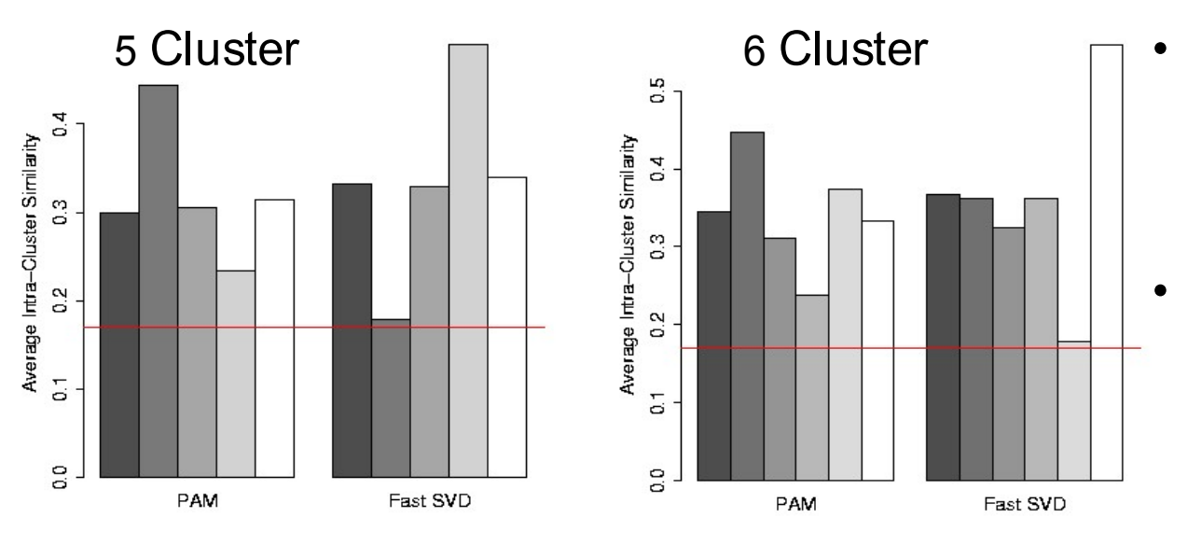
- Plot of the first two singular vectors obtained from a 2 class clustering
- The class structure is not very clear. This is not surprising since the histogram of the pair-wise similarities not significantly multi-modal

- Some of the most similar members from the two classes obtained from the fast SVD clustering using MACSS fingerprints

11

AMES Dataset – Class Similarity

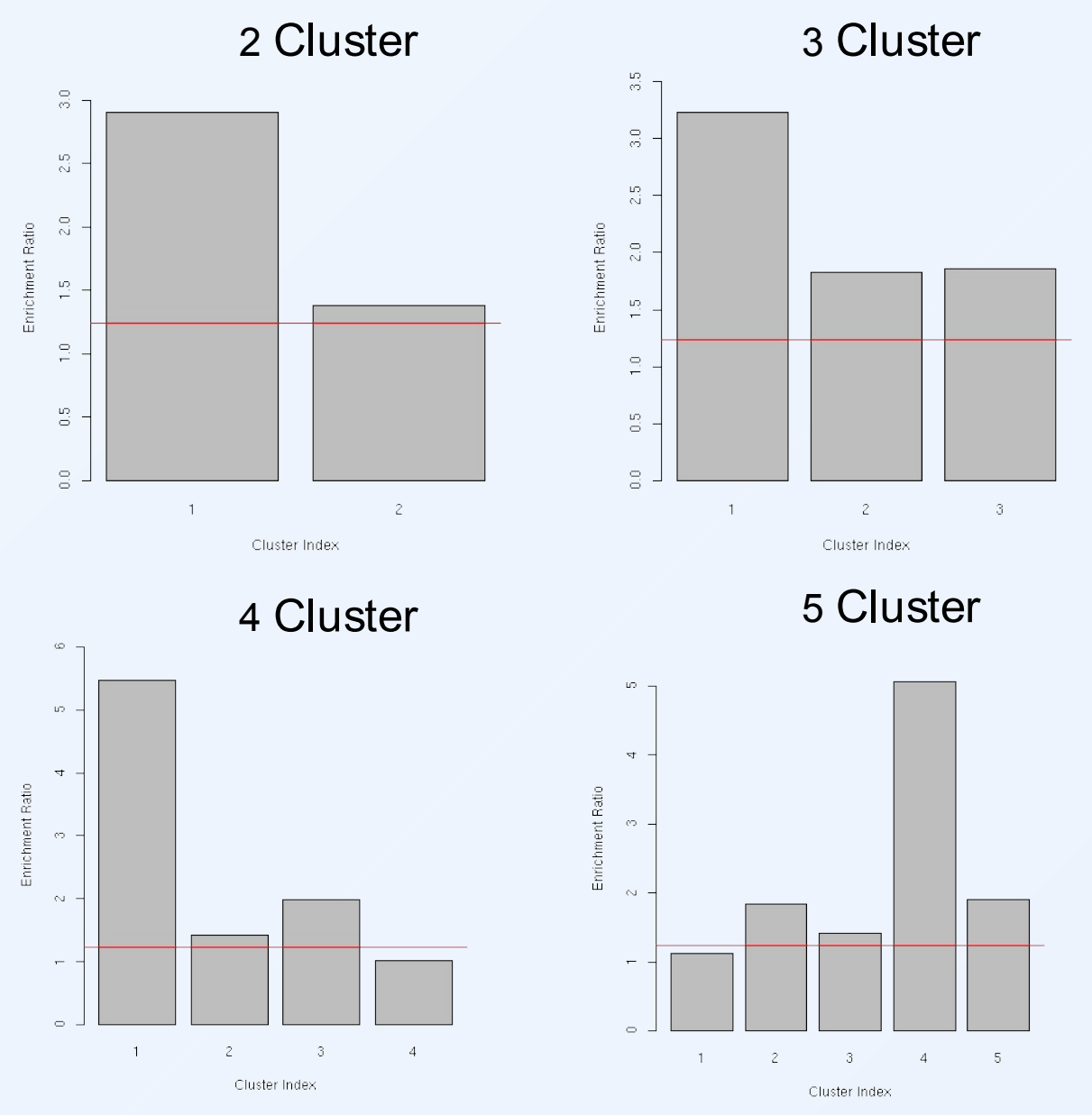




- As before, spectral clustering leads to general improvements, over PAM
- However, certain clusters are degraded by this method

12

AMES Dataset - Class Enrichment

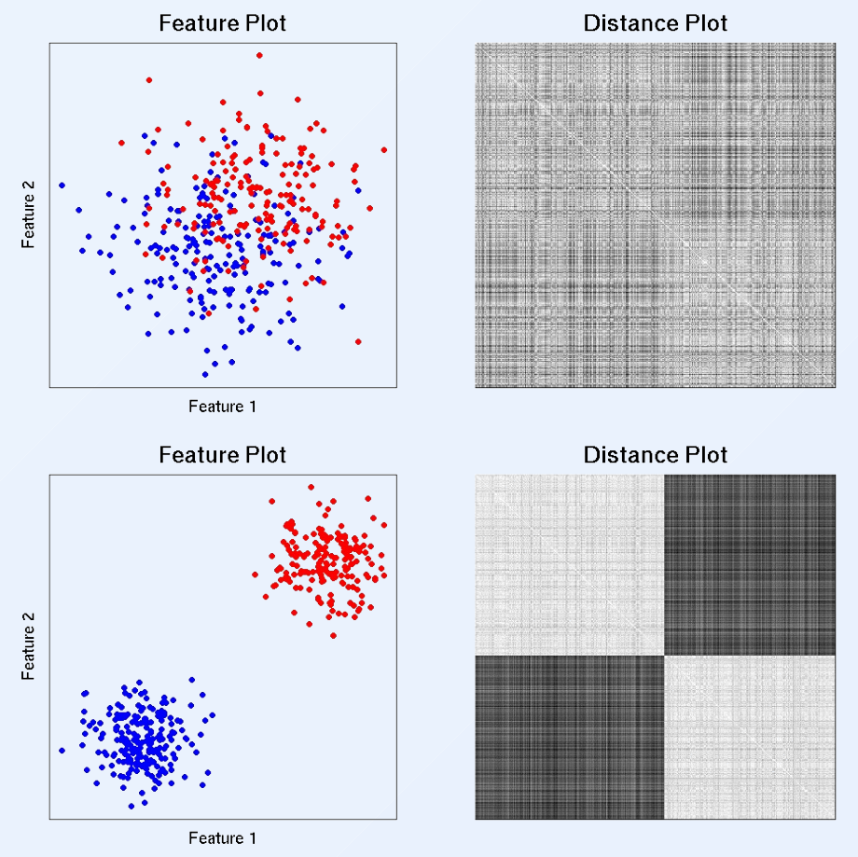


- Class enrichment is defined by the ratio between the sizes of the larger class to the smaller class
- For the overall dataset mutagen : non-mutagen = 1.24
- The spectral clustering algorithm enriches specific clusters, rather than each cluster
- The class enrichment does not always correspond to average cluster similarity

13

Block Structure & Spectral Clustering

- It has been shown that good clusterings correspond to block diagonal distance (affinity) matrices
- The aim is to enhance the block diagonal character of a distance matrix
- Analysis of the block structure can be also be used for hierarchical clustering
- The decomposition of the affinity matrix can be clustered using the *k*-lines algorithm



Visualization of clusters generated from a normal distribution. In the upper panel, there is little cluster separation and the block structure is weak.In the lower panel the distinct clusters are clearly represented in the block structure

Scott, G.L.; Longuet-Higgins, H.C., *British Machine Vision Conf.*, 1990, 103-108
Ng, A., et al., in *Adv. in Neural Inf. Proc.*, 2002, **14**, MIT Press

14

Asymmetric Spectral Clustering

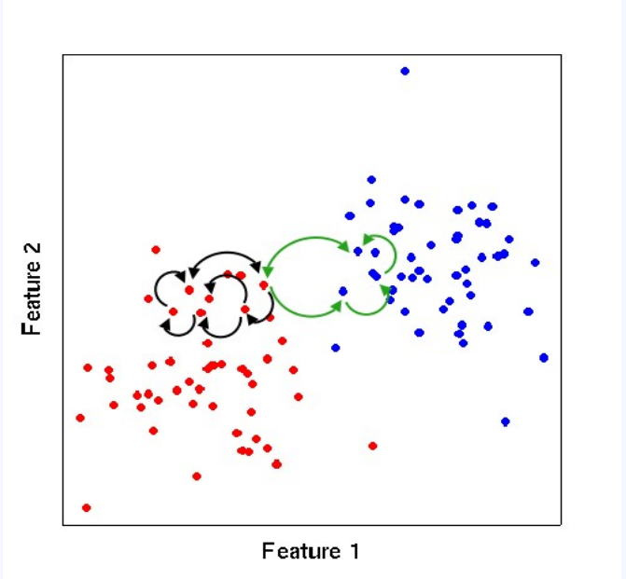
- A Gaussian kernel can be used to create an asymmetric affinity matrix:

$$A(i,j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)$$

- The bandwidth is different for each observation
- Individual bandwidths can be determined by

$$\sum_{j=1}^{n} \exp\left(\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right) = \tau$$

- Here, τ is termed the neighborhood size
- The affinity matrix is then used to calculate a *conductance matrix*



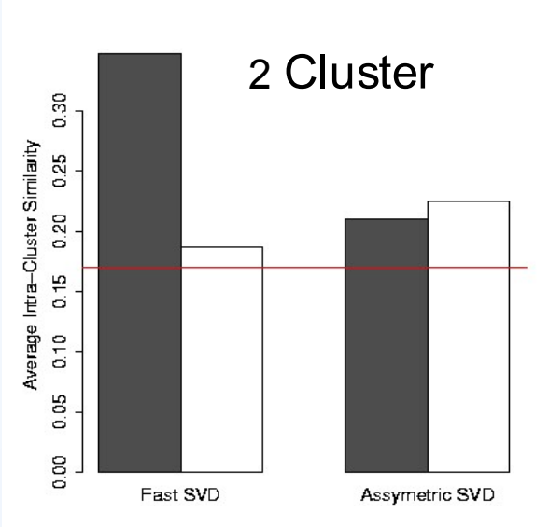
- A cluster can be considered as a continous concentration of points
- Cluster membership can be quantified by the *conductance* – analogous to conductance in electrical circuits
- The conductance matrix is then decomposed and clustered

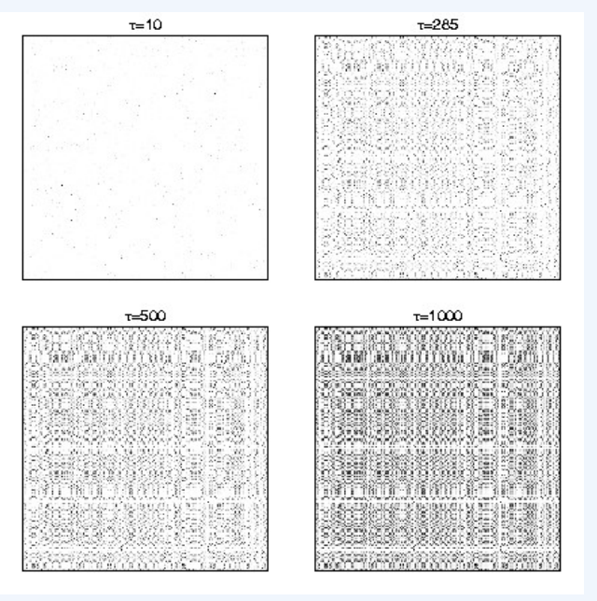
Fischer, I., Poland, J., *Proc. 14th Annual Machine Conf. Of Belgium and the Netherlands*, 2005, 21-28
Ben-Hur, A. et al., *J. Mach. Learn.*, 2001, **2**, 125-137

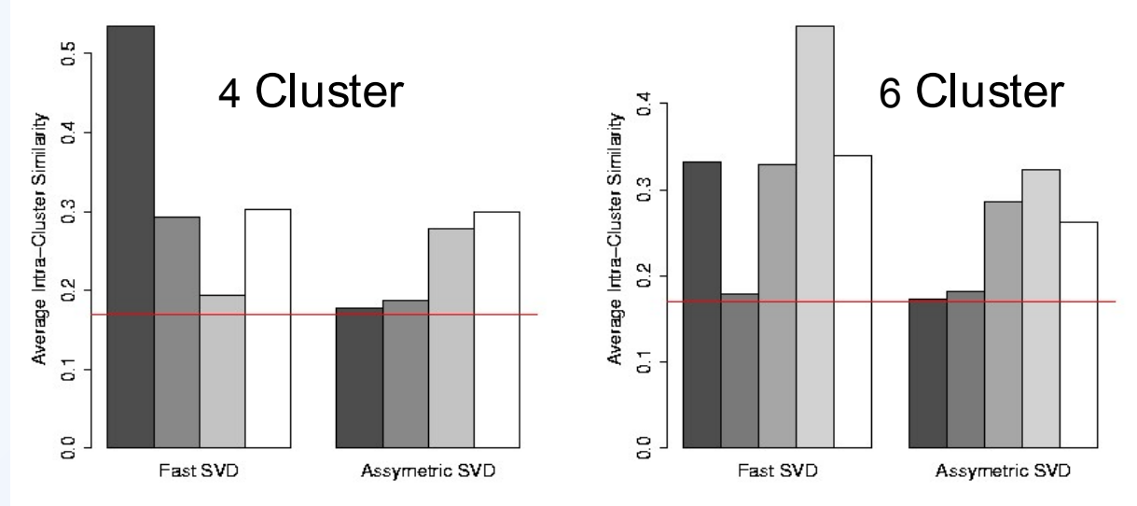
15

Asymmetric Spectral Clustering

- •For the AMES dataset the asymmetric approach does not work very well
- •One explanation is that the band amplification was not successful
- •In addition the dataset is not distinctly clustered overall







Plots of the asymmetric distance matrix for varying values of τ. It is clear that the block structure for this dataset is quite indistinct.

In the above plots, white indicates affinities equal to the minimum and black indicates affinities equal to the maximum

16

Summary

- Fast SVD based clustering gives nearly identical results compared to slow SVD based clustering at significantly higher speed
- The average intra-cluster similarities are comparable to PAM and *k*-means
- SVD based clustering appears to emphasize specific clusters over others
- The algorithm appears to handle correlated and information-poor descriptors well

17

Further Work

- Use feature selection when considering real-valued descriptors
- Combine the asymmetric affinity matrix with the fast SVD algorithm
- Investigate methods to avoid the distance matrix calculation

18

Acknowledgements

- Dr. Petros Drineas for clarifications about the fast SVD algorithm
- Dr. Igor Fischer for providing code to test the asymmetric affinity approach
- Chemical Computing Group for providing MOE