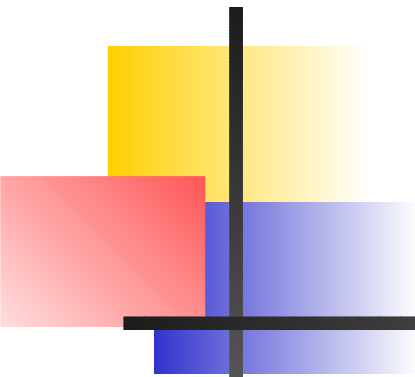


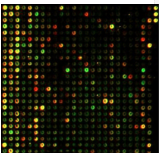


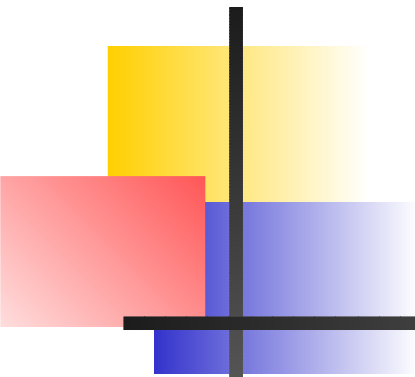
Graph Theoretic Latent Class Discovery

Jeff Solka

jsolka@gmu.edu

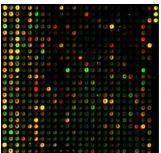
NSWCDD/GMU

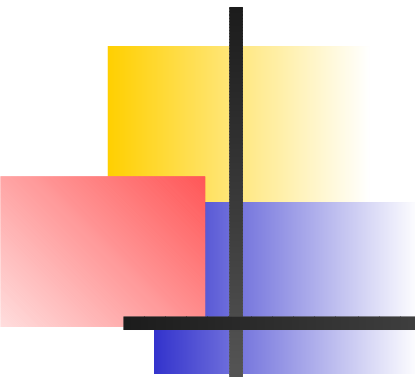




Agenda

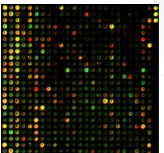
- What is latent class discovery?
- What are some approaches to the latent class discovery process?
- The class cover catch digraph classifier.
- Latent class discovery results on a gene expression data set.
- Wrap-up and conclusions.

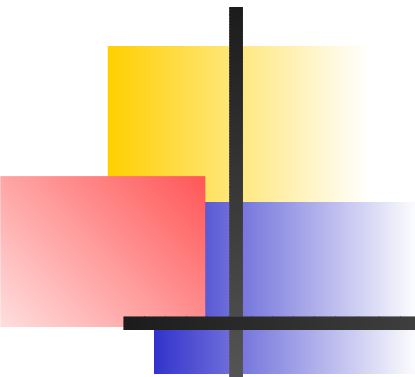




Acknowledgments

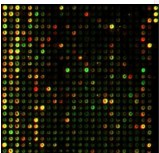
- John Grefenstette
- Office of Naval Research through their ILIR Program for funding this effort



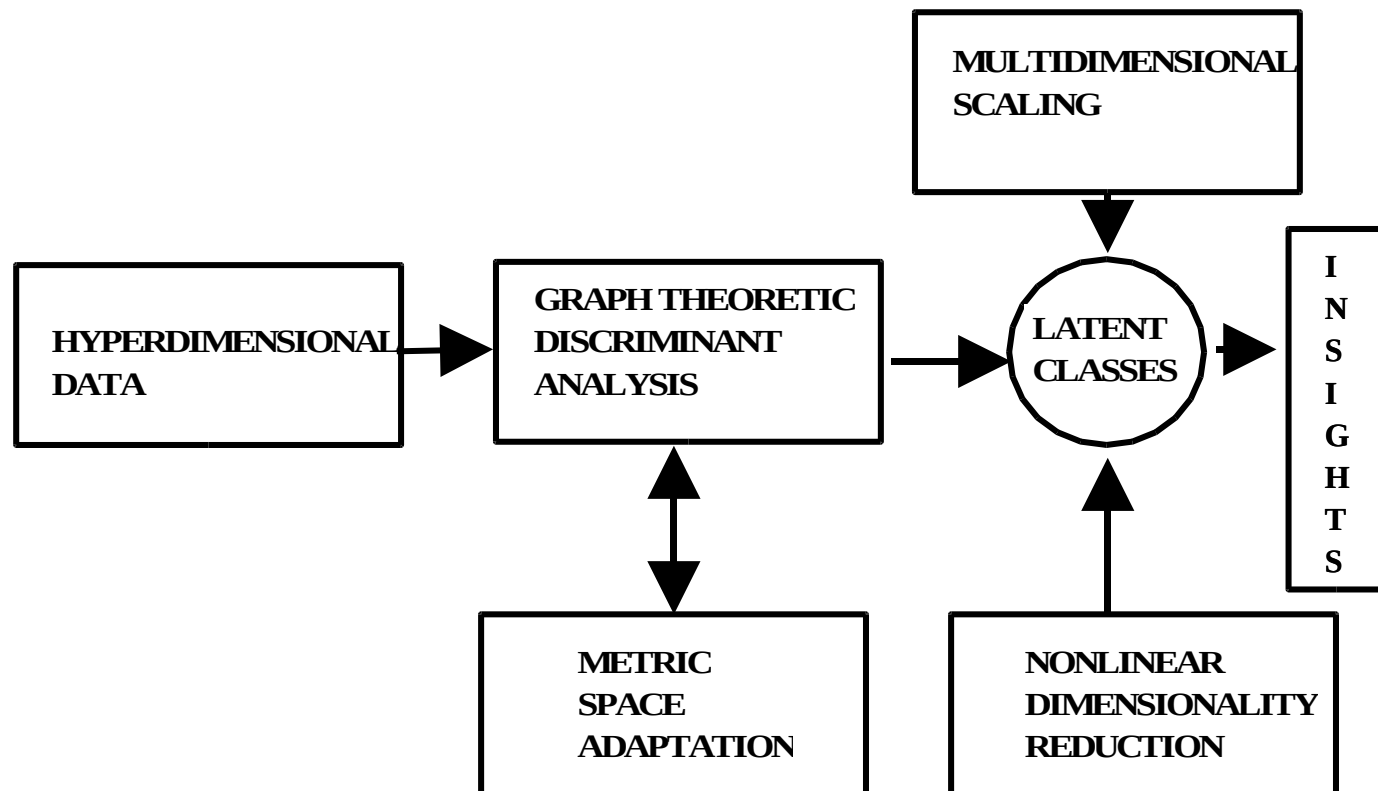


What is Latent Class Discovery?

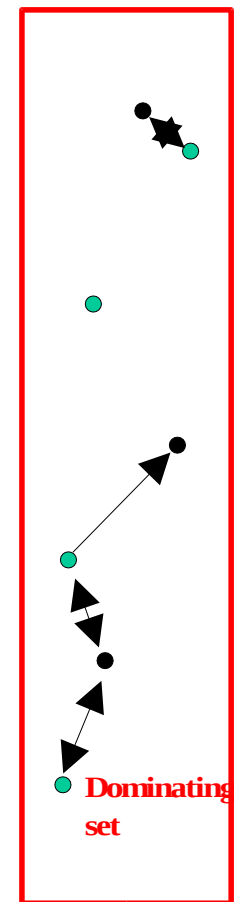
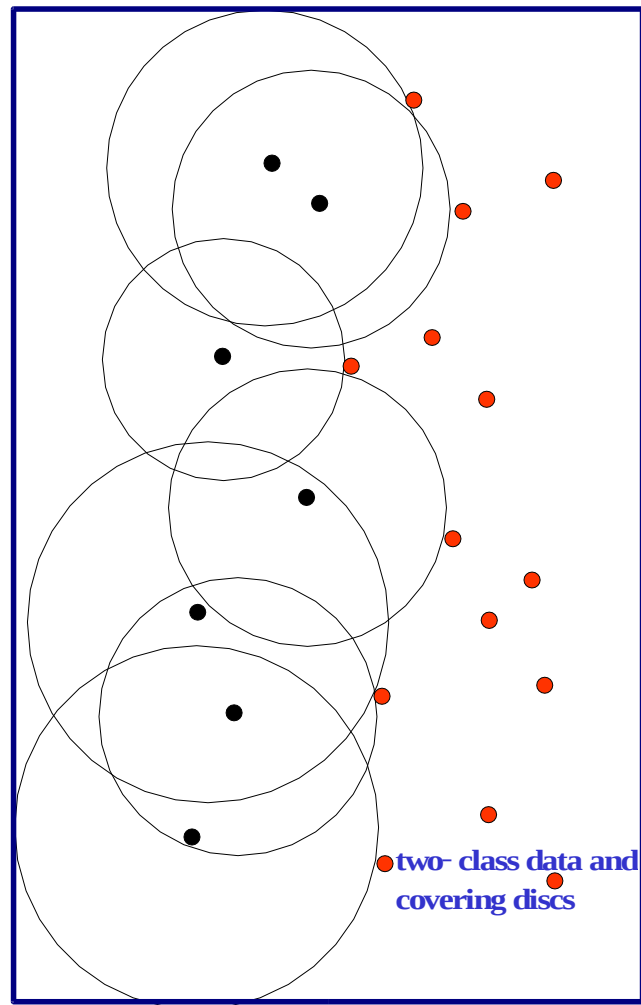
- A latent class is a class of observations that reside undiscovered within a known class of observations.
- Develop a general methodology for the discernment of latent class structure during discriminant analysis.
 - Moderately large hyperdimensional data sets.
 - During training or testing.
- Explore applications of developed methodologies to the analysis of data sets in the areas of hyperdimensional image analysis, artificial olfactory systems, computer security data, gene expression data, and text data mining.

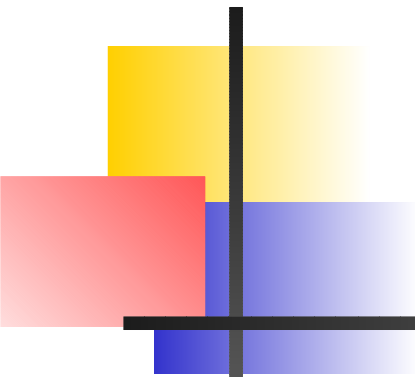


Flow Chart

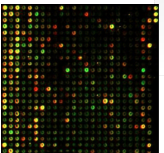


Dominating Set

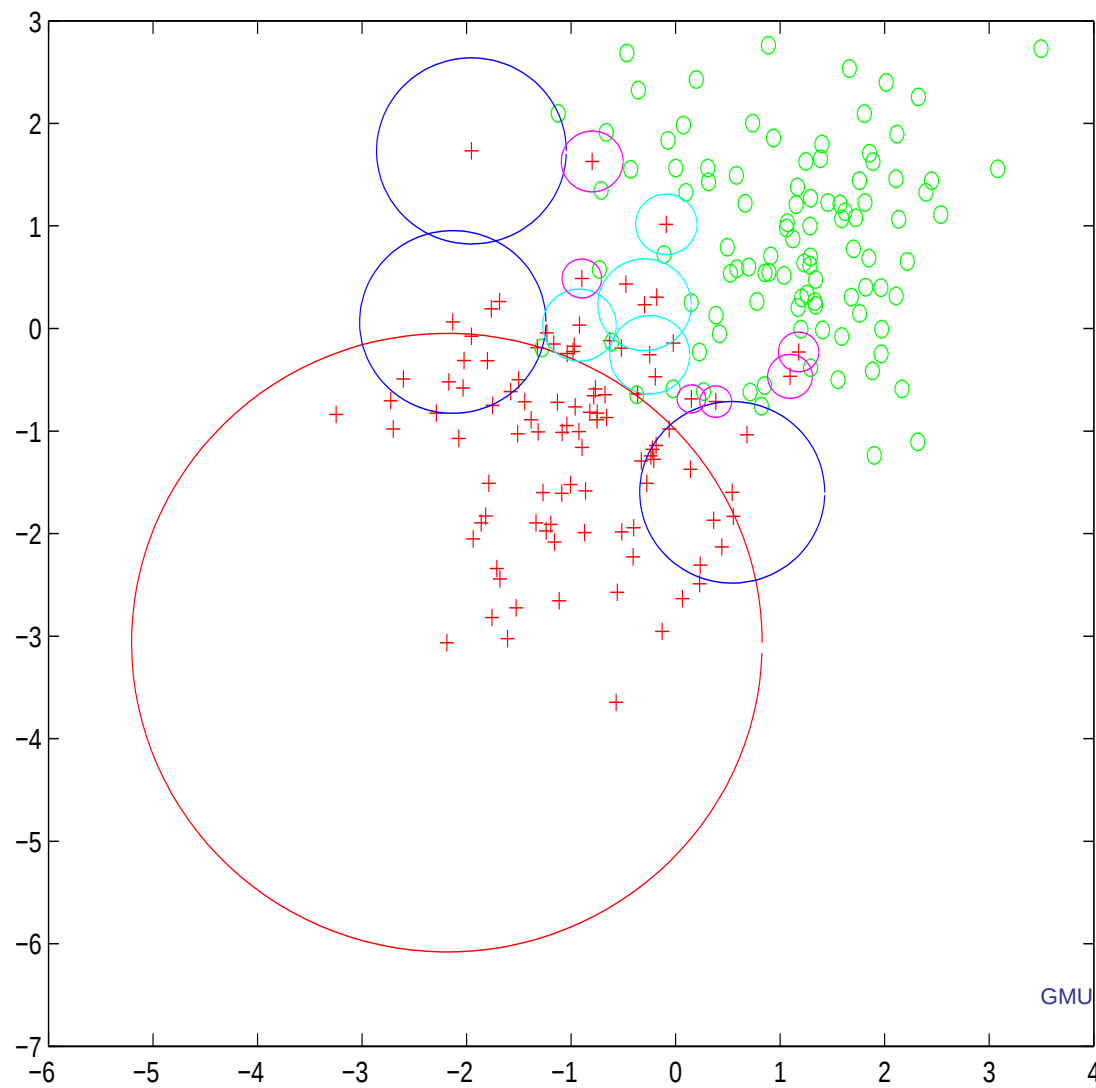




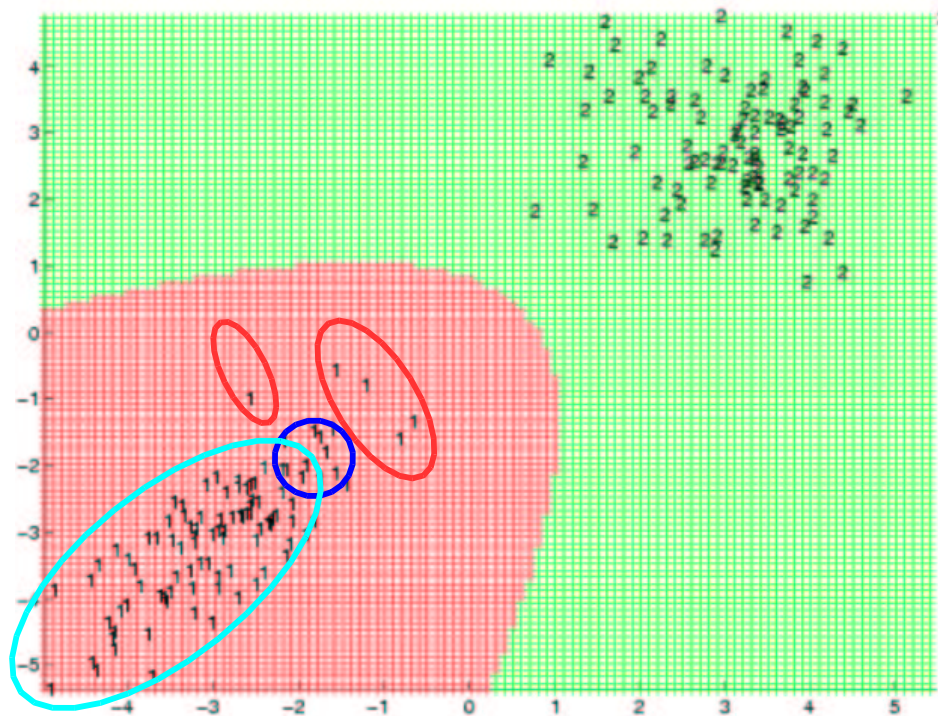
A Brief Movie



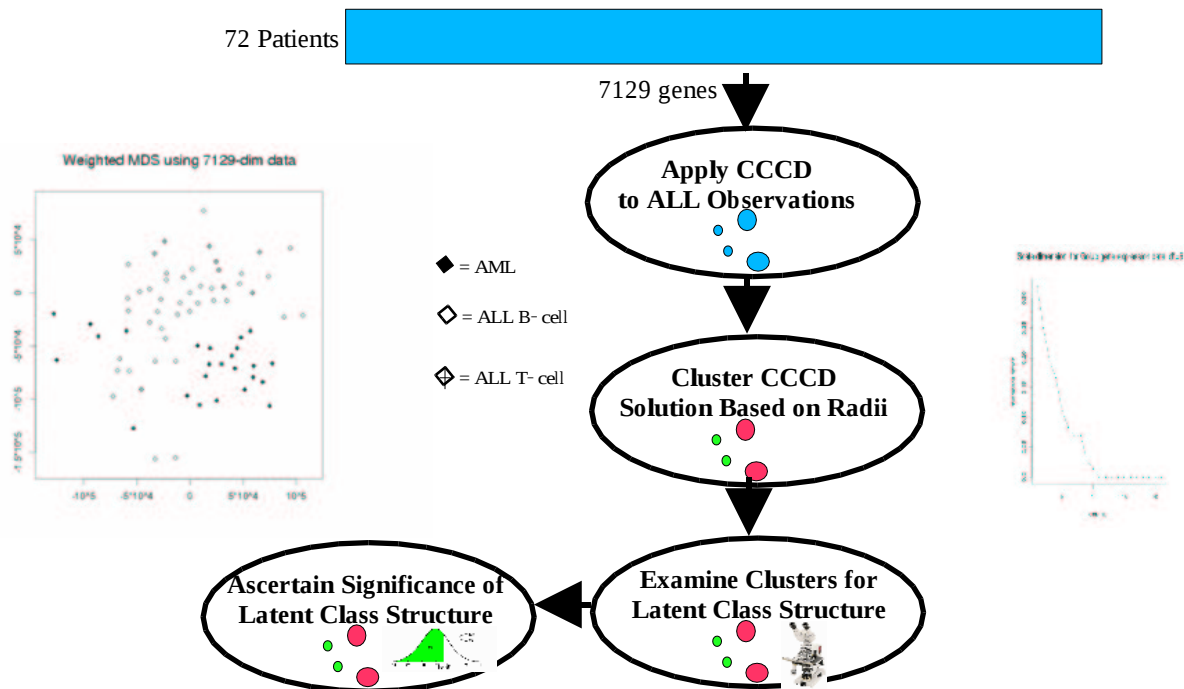
CCCD-Based Latent Class Discovery

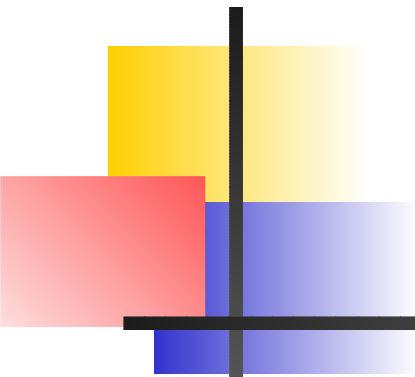


Quadratic Classifier-Based Latent Class Discovery



ALL/AML Leukemia Gene Expression Analysis

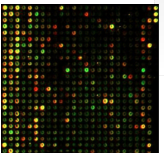


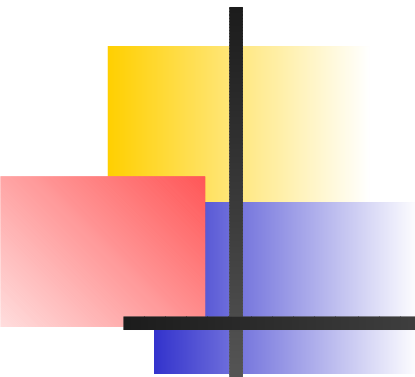


Resubstitution Error Rate Estimate

For each $k = 1, \dots, \hat{\gamma}$ an empirical risk (resubstitution error rate estimate) $\hat{L}_k$ is calculated as

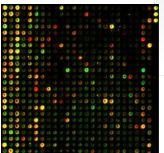
$$\begin{aligned} \hat{L}_k &:= (1/(n+m)) \left(\sum_{i=1}^n I\{x_i \notin \cup_{j=1, \dots, k} \cup_{v \in \hat{S}_j} B(v, \min_{w \in \hat{S}_j} r_w)\} \right. \\ &\quad \left. + \sum_{i=1}^m I\{y_i \in \cup_{j=1, \dots, k} \cup_{v \in \hat{S}_j} B(v, \min_{w \in \hat{S}_j} r_w)\} \right) \end{aligned}$$





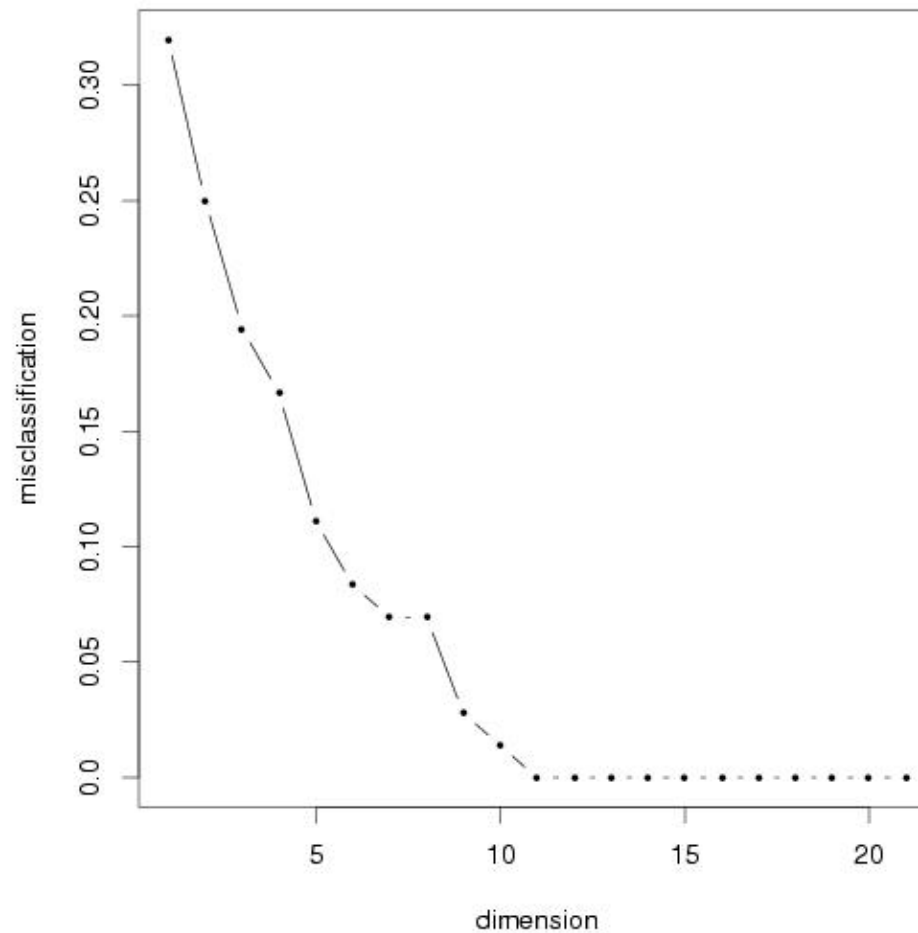
Classification Dimension

We proceed by defining the “scale dimension” $\hat{d}^\star$ to be the cluster map dimension that minimizes a dimensionality-penalized empirical risk; $\hat{d}_\delta^\star := \min\{\arg \min_k \hat{L}_k + \delta \cdot k\}$ for some penalty coefficient $\delta \in [0, 1]$.

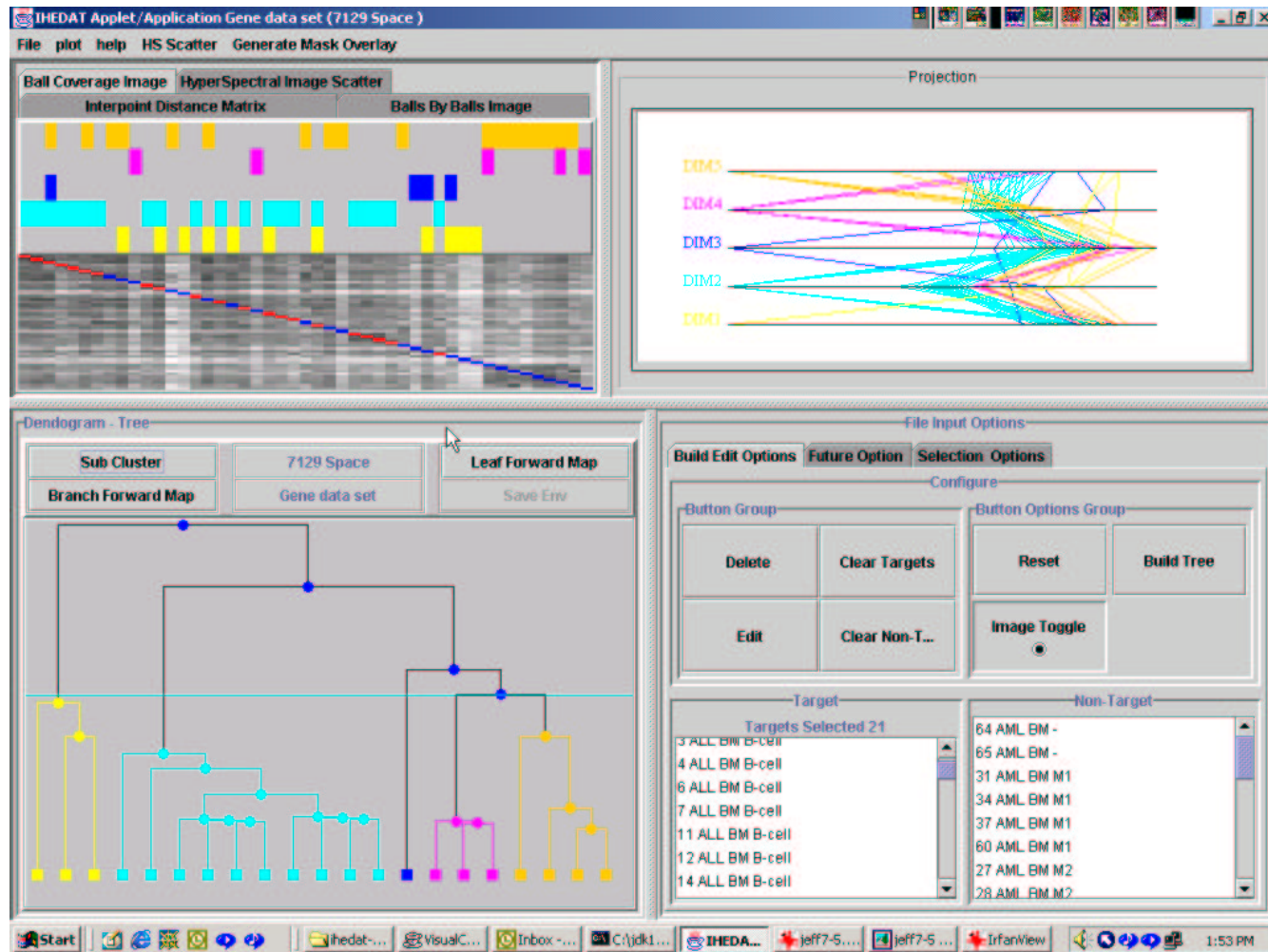


ALL/AML Classification Dimension Plot

'Scale dimension' for Golub gene expression data: $d^*=5$

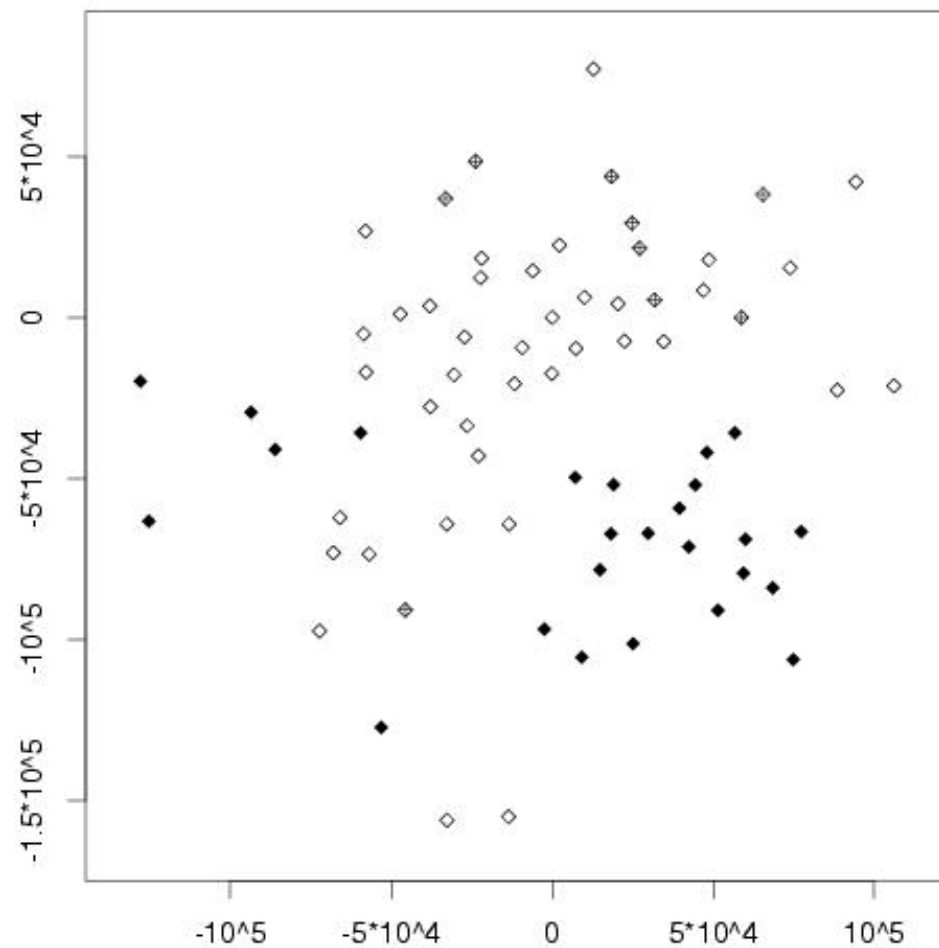


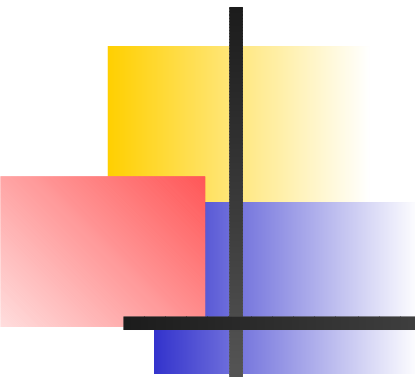
Gene Latent Class Discovery



ALL/AML MDS Plot

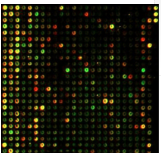
Weighted MDS using 7129-dim data

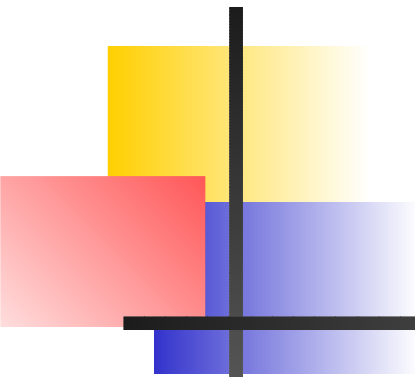




How Robust is the Methodology?

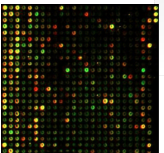
- One other “success” story using artificial nose data.
- What if we had used another dominating set in our analysis?
- Is the discovered latent class structure independent of the dominating set used?



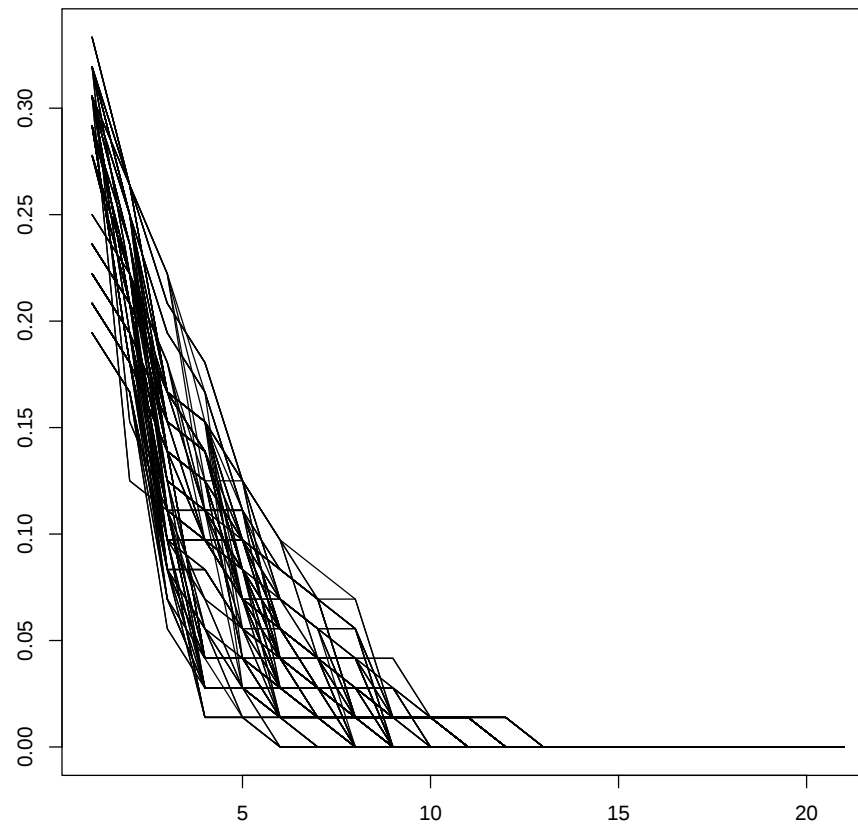


An Exhaustive Enumeration of All Possible Dominating Sets for the Gene Data

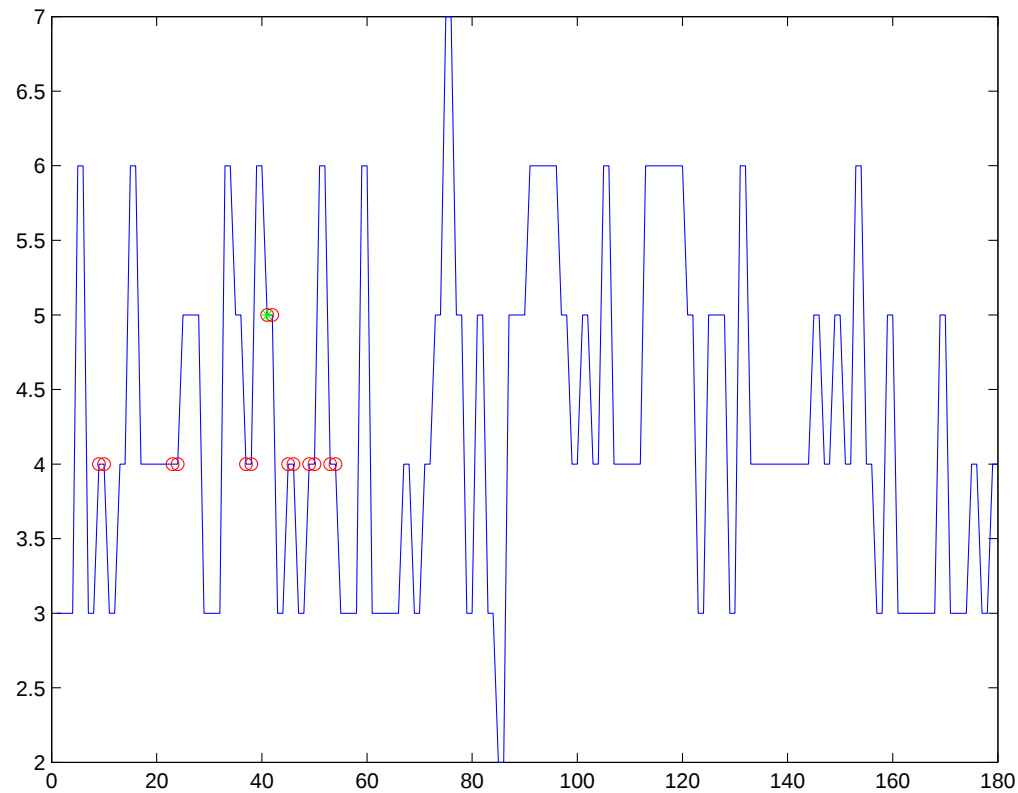
- 180 21 node solutions
- 16 of the nodes remain fixed across the solutions
- 14 greedy solutions



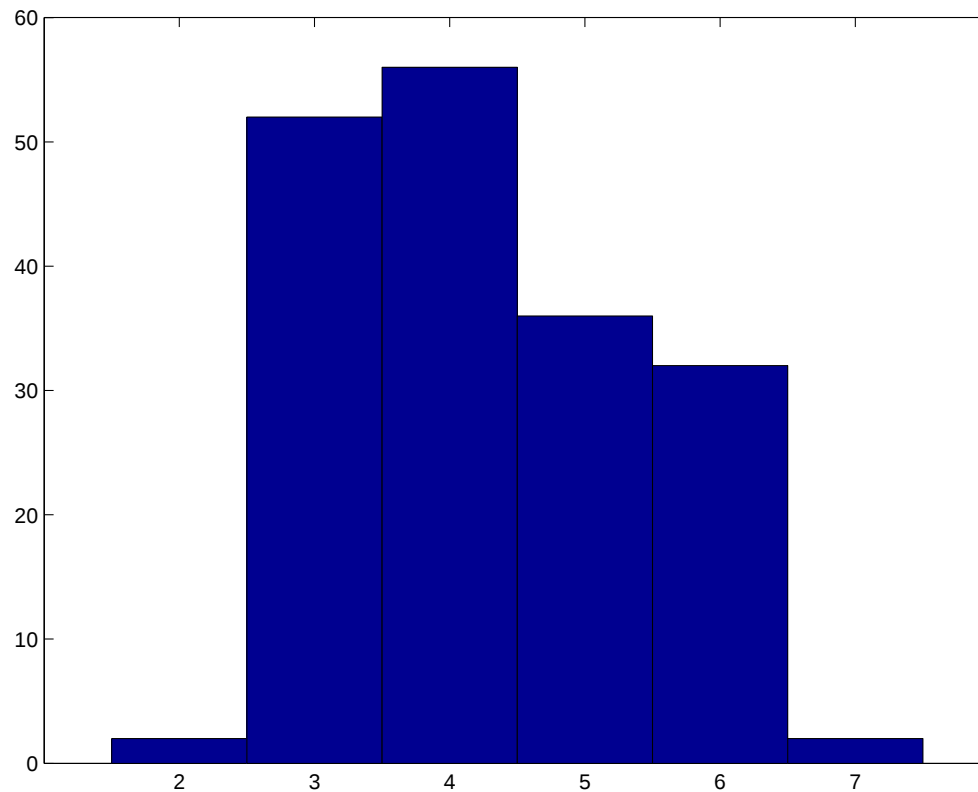
Classification Space Curves for the 180 Solutions



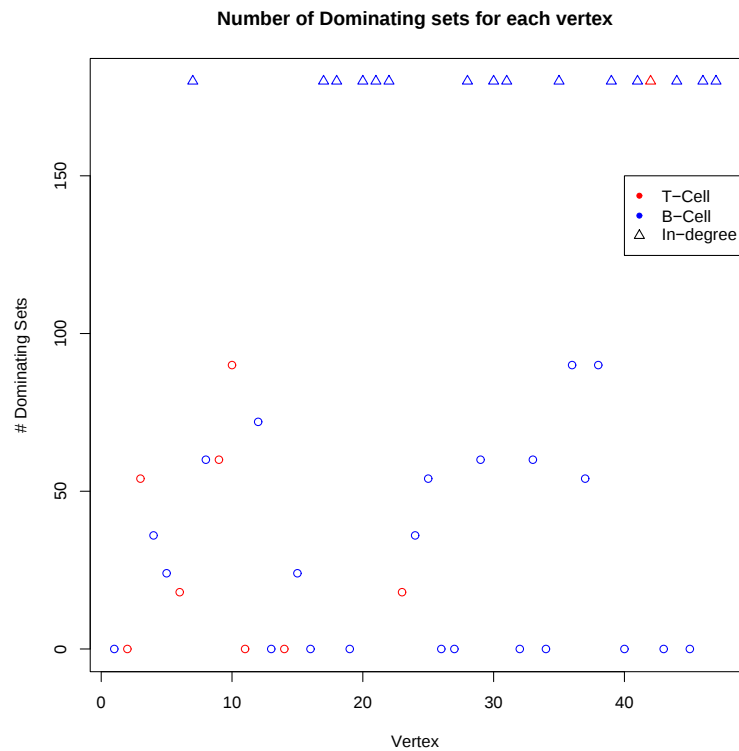
Classification Dimension for the 180 Solutions (red o Greedy Solutions, Green * Previous Solution)



Classification Dimension for the 180 Solutions



Number of Dominating Sets for Each Vertex



Digraph Analysis

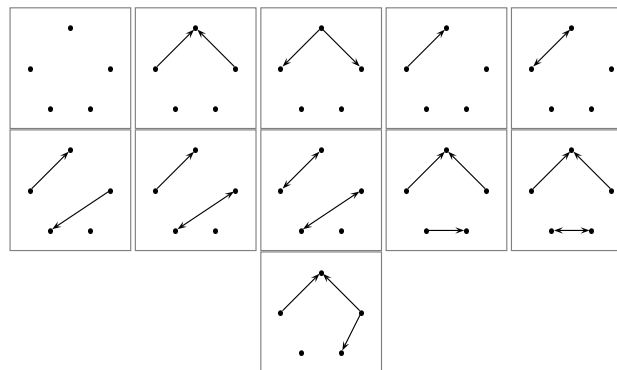


Figure 1: Unique induced subgraphs on the 5 changing vertices in the 180 dominating sets.

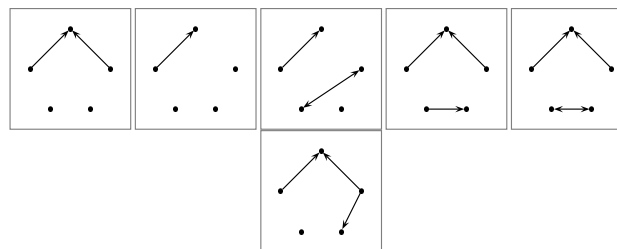
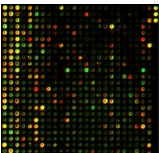
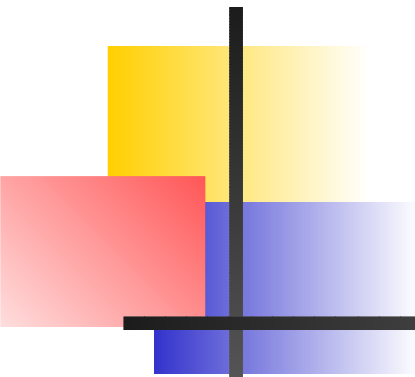


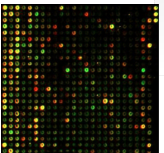
Figure 2: Unique induced subgraphs on the 5 changing vertices in the 14 dominating sets that could result from a greedy algorithm.



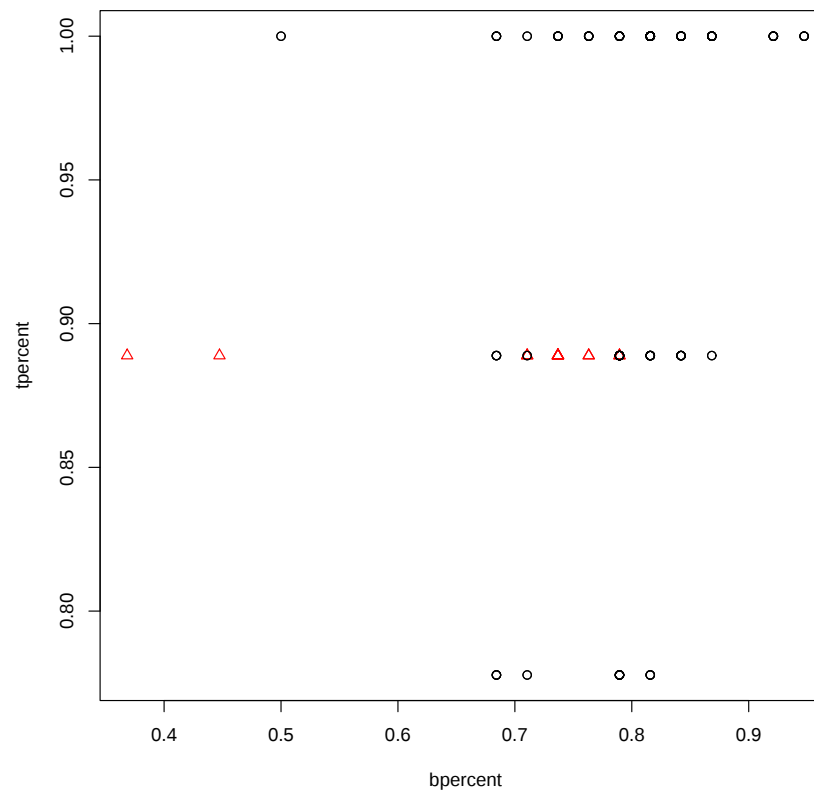


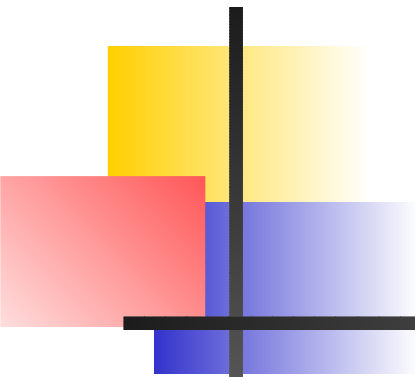
Latent Class Discovery Figures of Merit

- How can we be assured that all of the greedy dominating set solutions discover the same latent classes?
- Previous greedy solution had 3 clusters that are pure B and 1 cluster that contained 8/9 of the T observations
- Percentage of B points that are in pure B clusters and the highest percentage of T points in any one cluster



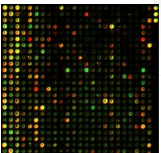
Purity (Latent Class Discovery) for the Golub Gene Data , Red Triangles are the Greedy Solutions

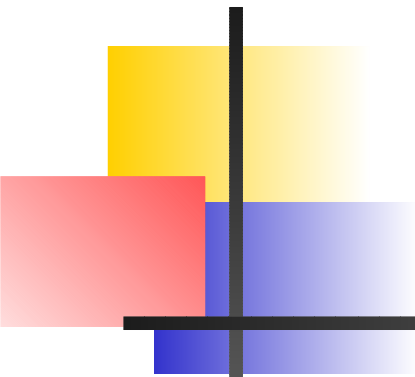




Remaining Questions

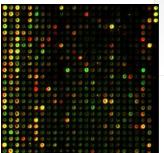
- Demonstrated similar latent class discovery among all of the greedy dominating set solutions
- Many of the 7129 variates (genes) are superfluous to the discriminant analysis problem
- Work is ongoing to examine the discovered latent classes based on subsets of the genes
- Various figures of merit have been used to choose the subsets of the genes

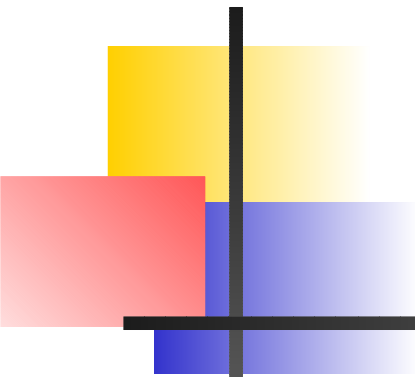




Conclusions

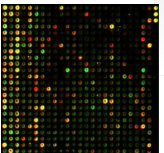
- Developed a new concept for latent class discovery during discriminant analysis
- Illustrated one graph theoretic methodology for the discovery of the latent classes
- Illustrated this methodology with a gene expression data set.
- Presented some preliminary results examining the robustness of the discovery process to the cccd process

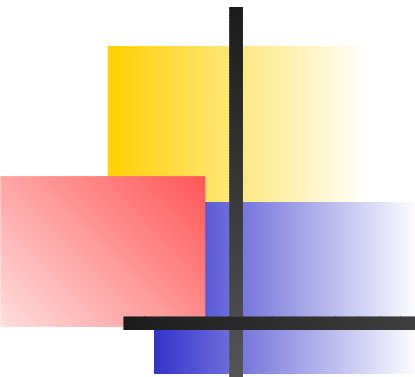




Readings

- C. E. Priebe, J. L. Solka, D. J. Marchette, and B. T. Clark, 2003, “Class Cover Catch Digraphs for Latent Class Discovery in Gene Expression Monitoring by DNA Microarrays,” *Computational Statistics and Data Analysis on Statistical*, Vol. 43, pp. 621–632.
- J. L. Solka, C. E. Priebe, and B. T. Clark, 2002, “A Visualization Framework for the Analysis of Hyperdimensional Data,” *International Journal of Image and Graphics Special Issue on Graphical Methods in Data Mining*, pp. 145-161.
- Marchette, D.J., Priebe, C.E., 2002, “Characterizing the scale dimension of a high-dimensional classification problem,” *Pattern Recognition*, Vol. 36, pp. 45–60.





Questions?

