To appear
Computational Statistics & Data Analysis
(Special Issue of *CSDA* on Data Visualization)

Class cover catch digraphs for latent class discovery in gene expression monitoring by DNA microarrays*

Carey E. Priebe[1*], Jeffrey L. Solka[2], David J. Marchette[2], B. Ted Clark[2]
[1]Department of Mathematical Sciences, Johns Hopkins University, Baltimore, MD 21218-2682, USA
[2]Naval Surface Warfare Center, Code B10, Dahlgren, VA 22448-5000, USA

February 27, 2002

Abstract

The purpose of this article is to introduce a data visualization technique for class cover catch digraphs which allows for the discovery of latent subclasses. We illustrate the technique via a pedagogical example and an application to data sets from artificial nose chemical sensing and gene expression monitoring by DNA microarrays. Of particular interest is the discovery of latent subclasses representing chemical concentration in the artificial nose data and two subtypes of acute lymphoblastic leukemia in the gene expression data and the associated conjectures pertaining to the geometry of these subclasses in their respective high-dimensional observation spaces.

Keywords: Random graphs; Statistical genetics; Exploratory data analysis

*The work of CEP was partially supported by Office of Naval Research Grant N00014-01-1-0011 and DARPA Grant F49620-01-1-0395. The work of JLS, DJM and BTC was partially supported by the Office of Naval Research through the NSWCDD In-house Laboratory Independent Research Program. This work was performed while CEP was ASEE/ONR Sabbatical Leave Fellow 2000–2001 (N00014-97-C-0171 and N00014-97-1-1055) at NSWCDD. The authors thank Prof. Michael Trosset of William and Mary College for Figure 8.
*Corresponding author. Tel.: +1-410-516-7198; fax: +1-410-516-7459.
E-mail address: cep@jhu.edu (C.E. Priebe).

1 Introduction

Techniques for the analysis of high dimensional and/or massive data sets are of critical importance in many areas of science in general, and the analysis of genetic data in particular. We term these techniques *statistical data mining* in the sense of Edward J. Wegman (1999, 2000); to paraphrase: "Data Mining is an extension of exploratory data analysis and has basically the same goals: the discovery of unknown and unanticipated structure in the data. The chief distinction between the two topics resides in the size and dimensionality of the data sets involved. Data mining in general deals with much more massive data sets for which highly interactive analysis is not fully feasible." Thus, the main scientific goal of statistical data mining is the discovery of unknown and unanticipated structure in data, leading to new working hypotheses which can subsequently be tested. In this paper, we describe a set of techniques for the analysis and visualization of high dimensional data for the purposes of discovering patterns in the data. These techniques are applied to a gene expression data set (see Golub et al. 1999), resulting in an interesting and potentially important working hypothesis about the relationships between different types of leukemia.

2 Gene Expression I

The Golub et al. (1999) data set, produced by Affymetrix DNA microarrays, involves two general classes of leukemia, ALL (acute lymphoblastic leukemia) and AML (acute myeloid leukemia). Each observation is a patient, with $n_{ALL} = 47$, $n_{AML} = 25$; $n = n_{ALL} + n_{AML} = 72$. Each observation is a point in 7129-dimensional Euclidean space; there are 6817 unique human genes monitored, augmented with some redundant probes and control elements. (See also Getz, Levine and Domany (2000).)

Note that this is not a "simple" data set. For example, a principal component analysis scree plot (Cattell 1978) suggests that as many as ten or more dimensions are necessary to adequately account for the variability in the data set.

The ALL class has two (latent) subclasses, T-cell and B-cell, with $n_T = 9$, $n_B = 38$; $n_{ALL} = n_T + n_B = 47$. Note, however, that this subclass information is not used in building the model presented below. In fact, we were unaware at the time of the analysis that these subclasses possess the geometry in the high-dimensional "gene expression space" required for discovery. When we investigated the subclasses produced by our methodology, the T-cell/B-cell dichotomy emerged. The result of our procedure is the discovery of T-cell/B-cell as potential latent subclasses, and a potentially scientifically valuable conjecture pertaining to the geometry of these subclasses in gene expression space. This result is described in detail below.

3 Methodology

Our methodology for latent class discovery, which involves building a (random) graph model for a (supervised) two-class classification problem and the subsequent (unsupervised) investigation of this model for latent subclasses, is described in this section. At nearly every stage there are generalizations which can (and often should) be employed; we present here a simplified version. Additional details can be found in Marchette and Priebe, 2002; Priebe et al., 2002.

We are given two disjoint sets of d-dimensional observations, $\mathcal{X} = \{x_1, \cdots, x_n\} \subset \Re^d$ and $\mathcal{Y} = \{y_1, \cdots, y_m\} \subset \Re^d$. We begin by choosing $\mathcal{X}$ as the "target class"; our procedure is asymmetric in target class. (Development of a methodology for *classification*, as opposed to the *latent class discovery* described herein, requires symmetrization by considering each class as the target class in turn; see Priebe et al., 2001.)

Following Priebe, DeVinney and Marchette (2001) and DeVinney and Priebe (2001), the class cover catch digraph (*cccd*) $D = (V, A)$ for $\mathcal{X}$ against $\mathcal{Y}$ is defined as follows. Let $V = \mathcal{X}$ (the set of target class observations). For each $v \in V$, let $B_v := B(v, \min_{y \in \mathcal{Y}} \rho(v, y)) := \{z \in \Re^d : \rho(v, z) < \min_{y \in \mathcal{Y}} \rho(v, y)\}$ for some distance or pseudo-distance function $\rho : \Re^d \times \Re^d \to \Re_+ := [0, \infty)$; we will use the L_2 (Euclidean) distance. That is, for each target class observation v, B_v is the (open) ball around v of maximal radius such that the ball contains no non-target class observations. Then the arc (directed edge) $vw \in A \iff w \in B_v$.

A *dominating set* S for D is a set $S \subset V$ such that, for all $w \in V$, either $w \in S$ or $vw \in A$ for some $v \in S$. The invariant $\gamma(D)$ is defined as the cardinality of the smallest dominating set(s) of D; $1 \leq \gamma(D) \leq \text{cardinality}(V) = n$. A minimum dominating set for D is defined as a dominating set with

cardinality $\gamma(D)$. Finding a minimum dominating set in a general digraph is NP-Hard; an (approximate) minimum dominating set $\widehat{S}$ can be obtained in polynomial time using a well-known greedy algorithm (see Priebe, DeVinney and Marchette (2001) and references therein). Our estimate for the domination number of the digraph D is $\widehat{\gamma} = \text{cardinality}(\widehat{S})$.

For each $v \in V$ there is an associated radius; $r_v := \min_{y \in \mathcal{Y}} \rho(v, y)$. We employ agglomerative clustering on the radii $\{r_v : v \in \widehat{S}\}$, yielding a dendogram, or cluster tree (Everitt 1980; Hartigan 1975). The leaves of this dendogram correspond to the $\widehat{\gamma}$ elements of $\widehat{S}$.

The dendogram provides a sequence of "cluster maps" $m_k : R^d \to R_+^k$ for each $k = 1, \cdots, \widehat{\gamma}$. The cluster map with a given range-space dimensionality k is based on a disjoint partition of $\widehat{S}$ and can be conceptualized by visually "cutting" the dendogram horizontally at a level which yields k branches, or clusters, $\widehat{S}_1, \cdots, \widehat{S}_k$. The k^{th} cluster map is then defined as $m_k(x) = [\rho(x, \widehat{S}_1), \cdots, \rho(x, \widehat{S}_k)]'$, where the distance $\rho(x, S)$ from a point x to a set S is defined as the minimum over $s \in S$ of the distances $\rho(x, s)$.

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

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

The empirical risk $\widehat{L}_{\widehat{\gamma}} = 0$ by construction, whereas $\widehat{L}_k$ may be nonzero for $k < \widehat{\gamma}$. The goal is to use the empirical risk as a function of k to determine a reasonable cluster map dimensionality; this "model complexity selection" is a notoriously difficult task, but is necessary nonetheless.

We proceed by defining the "scale dimension" $\widehat{d}^\star$ to be the cluster map dimension that minimizes a dimensionality-penalized empirical risk; $\widehat{d}^\star_\delta := \min\{\arg\min_k \widehat{L}_k + \delta \cdot k\}$ for some penalty coefficient $\delta \in [0, 1]$. (Some will prefer a logarithmic penalty $\delta \cdot \log(k)$ or Bayesian model selection; alterations such as these can of course be accommodated.) Again, by construction we have $\widehat{d}^\star_\delta = \min\{k : \widehat{L}_k = 0\}$ for $\delta = 0$ and $\widehat{d}^\star_\delta = 1$ for $\delta = 1$. The choice of δ determines the sharpness required to define the "elbow" in the curve of empirical risk versus cluster map dimension. Thus, the scale dimension is, loosely, the x-coordinate of the elbow in the curve. (Note that $\widehat{d}^\star_\delta$ is an estimate of $d^\star_\delta$, the cluster map dimension which minimizes the penalized probability of misclassification.)

The result of this methodology is $m_{\widehat{d}^\star_\delta}$, the cluster map of interest for exploratory data analysis. It is this map that is investigated in the examples below. For instance, we may consider the assignment of each observation x_i in the target class to that cluster (or those clusters) for which $x_i \in \cup_{v \in \widehat{S}_k} B(v, \min_{w \in \widehat{S}_k} r_w)$. Our "latent class discovery" will derive from information revealed via this assignment; there may ultimately be a (known or unknown) latent variable or geometric structure that is responsible for the particular set of target observations that reside in a particular cluster. Another way to think about this is that we are investigating the structure of the target observations based on their distance to the non-target observations. Figure 1 presents an example that illustrates the process; the target regions labeled C_1 are closer to the non-target region (Class 0) than is the C_2 target region. In this case we would expect to discover latent classes. (Note, however, that the leftmost C_1 subclass may *not* be discovered, as the observations therein may fall in a ball centered at a C_2 observation – a large radius ball.)

4 2-Dimensional Simulation

Let us consider, for the purpose of illustration, a simple 2-dimensional simulation example. For this case the domain space class-conditional scatter plot and the algorithmically produced dominating set $\widehat{S}$ (with $\widehat{\gamma} = 6$) and the associated radii (one large and a collection of five smaller) for the target class observations are presented in Figure 2 (a), the dendogram for the complete linkage clustering of these six radii is presented in Figure 2 (b), and the 2-dimensional range space class-conditional scatter plot (the result of the application of the cluster map m_2 to the observations of Figure 2 (a)) is presented in Figure 2 (c). Figure 3 shows that the scale dimension $\widehat{d}^\star = 2$. (Precisely, $\widehat{d}^\star_\delta = 2$ for $\delta \in [0.03, 0.92)$.)

Note that the type of hierarchical clustering employed (e.g., complete linkage versus single linkage) will affect the process. Since different dendograms (and different clusters) can be produced by using different linkage criteria, the cluster maps and hence the choice of $\widehat{d}^\star_\delta$ are dependent on the criterion employed.

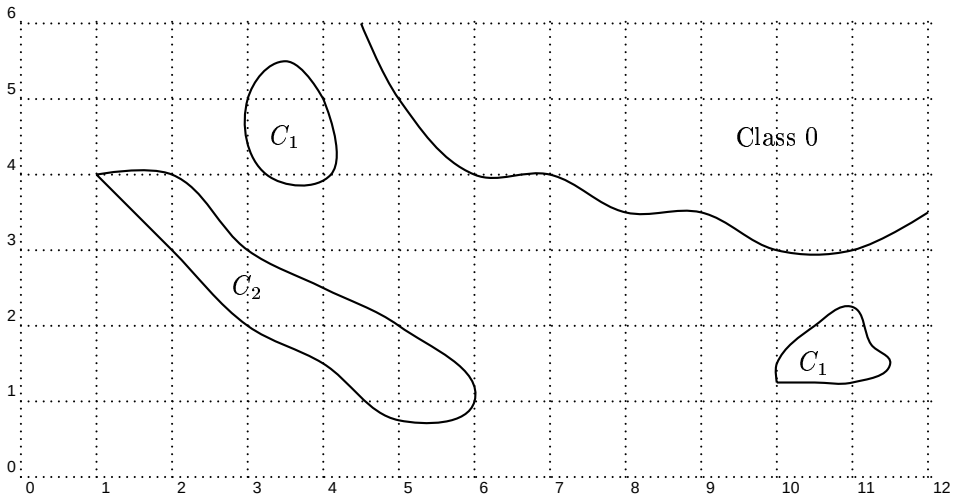

Figure 1: Our "latent class discovery" is similar to a clustering of target class observations based on their distance to non-target class (Class 0). C_1 and C_2 are latent subclasses of the target class.

In this pedagogical example, there are clearly two latent subclasses, and these subclasses are associated with the two clusters of radii. (We would not expect to encounter such simplicity in the analysis of real world data sets.)

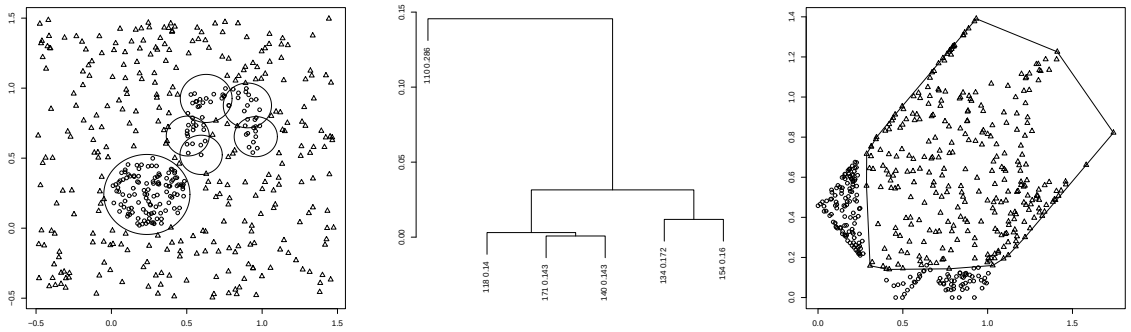

Figure 2: Depiction of the simulation example. (a) The domain space class-conditional scatter plot, with dominating set $\widehat{S}$ ($\widehat{\gamma} = 6$) for the target class observations (represented by "o"s). The two axes represent two canonical dimensions. (b) The dendogram for the six radii. (c) The class-conditional scatter plot resulting from cluster map m_2 (with the convex hull of the projected non-target class observations). The two axes represent $m_2(\cdot) = [\rho(\cdot, \widehat{S}_1), \rho(\cdot, \widehat{S}_2)]'$.

5 Artificial Nose Example

Before returning to the gene expression data we present results for an artificial nose chemical sensing data set. The results obtained for these data are qualitatively similar to those obtained below for the gene expression data. Furthermore, the structure of these artificial nose data, and the subsequently discovered latent subclasses, are perhaps better understood.

The data are taken from a fiber optical sensor constructed at Tufts University; see Priebe (2001) for details. Each observation is a multivariate time series – 19 fiber responses at each of two wavelengths, sampled at 60 equally spaced time steps, for a total "dimensionality" of $d = 2280$. The data set is designed

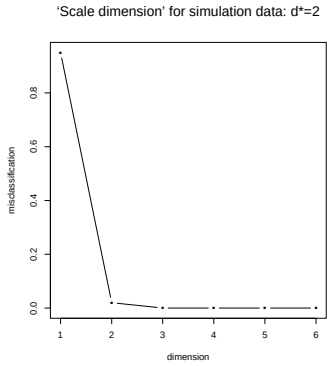

Figure 3: For the simulation example the scale dimension $\widehat{d}^{\star} = 2$.

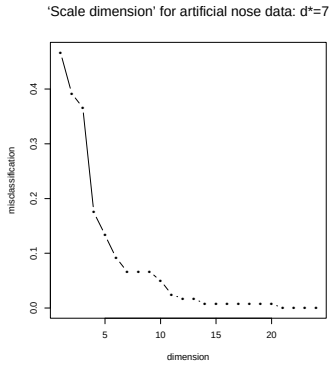

Figure 4: The empirical risk (y-axis) against the cluster map dimension for the artificial nose data set. This plot suggests $\widehat{d}^{\star} \approx 7$.

for the investigation of the detection of trichloroethylene (TCE), a carcinogenic industrial solvent. For this example we consider a subset of the Tufts data set consisting of those observations containing chloroform as a confounder. This yields $n = 80$ target class observations (observations with TCE in a mixture of chloroform and air) and $m = 40$ non-target class observations (observations with just chloroform and air, and no TCE).

The methodology described above yields $\widehat{\gamma} = 24$. Figure 4 depicts the plot of empirical risk versus cluster map dimension; $\widehat{d}^{\star} = 7$ for this case.

The exploratory analysis of the data was performed using the Interactive Hyperdimensional Exploratory Data Analysis Tool (IHEDAT) (see Solka et al., 2001). IHEDAT is a java-based system developed to study the interaction between the *cccd* classification methodology and various high dimensional data sets. Figure 5 depicts the results of an exploratory analysis of this data set using $\widehat{d}^{\star} = 7$. The lower left panel depicts the agglomerative clustering dendogram for the radii associated with the approximate minimal dominating set $\widehat{S}$ for the class cover catch digraph D of the target class observations with respect to the non-target class observations. The upper right panel is a parallel coordinates plot (Wegman, 1990) depiction of the associated cluster map from the original 2280-dimensional space to the associated 7-dimensional "scale dimension space." The upper left panel depicts the clustering (into $\widehat{d}^{\star} = 7$ clusters) associated with the seven clusters of dominating set elements. It is in this panel that the latent class discovery emerges. In the "data image" (see Minnotte and West, 1998) depicted here, the target class observations are in order of decreasing *concentration* of TCE in the observation. (The data set contains TCE at different concentrations.) Each observation is

represented by a column in the data image. Above the data image we color-code for the cluster(s) into which a particular observation falls. We see, thanks to this ordering, that the seven clusters are highly correlated with concentration. In general, these clusters can be thought of as a regression on concentration; the latent classes discovered in this manner are associated with the various concentrations. In particular, the magenta and blue clusters contain, almost exclusively, the 50 lowest concentration observations. Further investigation indicates that these lowest concentration clusters (magenta and blue) are associated with dominating set elements with small radii, and the five clusters which contain (again, almost exclusively) the 30 highest concentration observations are associated with dominating set elements with larger radii. We conclude this example with the claim that the latent class discovery depicted for this nose data set is in keeping with our (limited) understanding of the (high-dimensional) geometry of the problem: the radii are determined by the distance of the dominating set elements to the non-target class, and *low concentration target class observations should be closer to the non-target class, and should thus be associated with the smallest radii.*

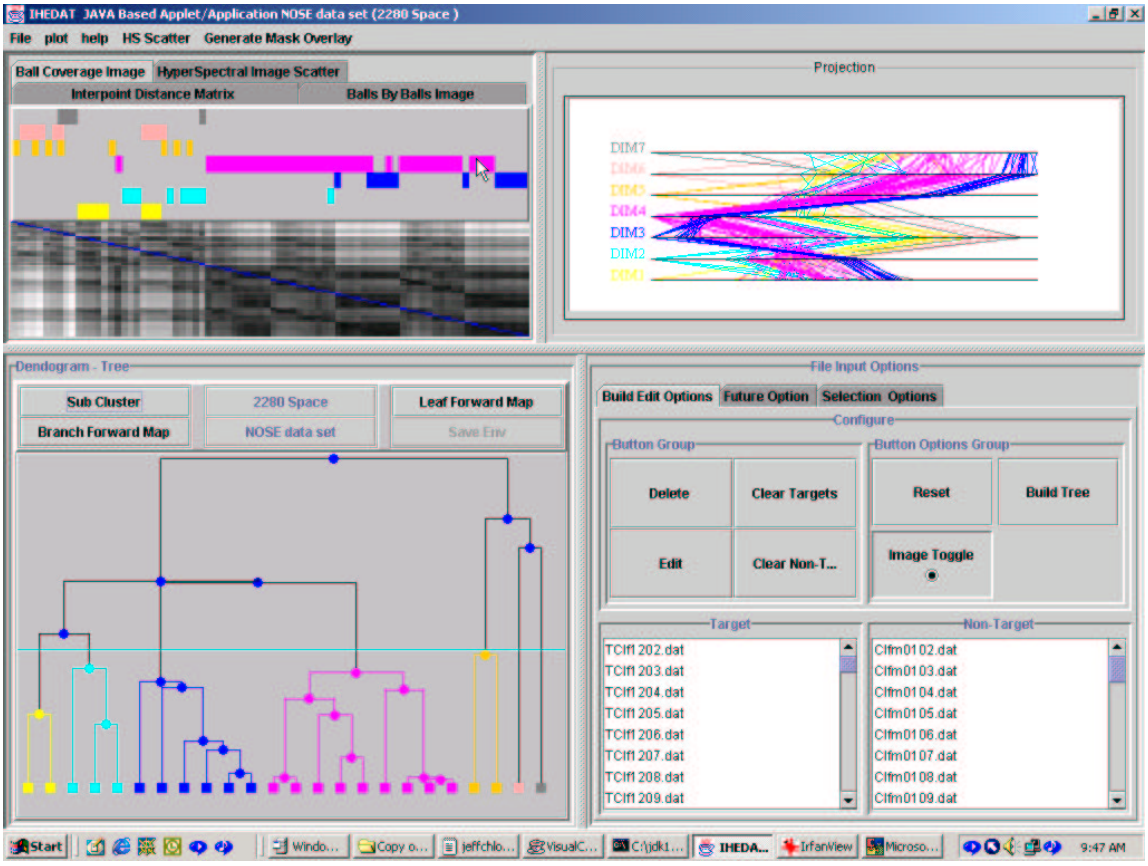

Figure 5: This graphic displays the results of using class cover catch digraphs for latent class discovery on a data set of artificial olfactory observations (the Tufts artificial nose data set). The upper left panel depicts the clustering (into $\widehat{d}^* = 7$ clusters) associated with the seven clusters of dominating set elements. It is in this panel that the latent class discovery emerges. See text for details.

6 Gene Expression II: Exploratory Data Analysis

We return now to the gene expression data. Recall that our procedure is asymmetric in target class; for this example we choose ALL to be the target class.

Figure 6 shows that $\widehat{d^\star} = 5$ for the gene expression data set; Figure 7 depicts the results of our exploratory data analysis at $\widehat{d^\star} = 5$. For this display the data image and cluster coverage, depicted in the upper left panel, is ordered so that the nine T-cell observations are the rightmost in the display. Investigation uncovers that the clusters with the smallest radii (the third, fourth, and fifth rows, referenced from the top, of the five rows in the cluster coverage display) are made up entirely of B-cell observations (although it is not the case that all B-cell observations fall into these clusters). Furthermore, the topmost (orange) cluster contains eight of the nine T-cell observations (as well as some B-cell observations) and is associated with the largest radii. This yields the (possibly scientifically valuable?) conjecture, analogous to that obtained in the artificial nose investigation, that the B-cell subclass of the ALL leukemia is "closer" to the AML leukemia in "gene expression space" than is the T-cell subclass; see also Figure 8. This conjecture is obtained via exploratory data analysis with no prior indication that the B-cell/T-cell subclasses even existed.

(Historical note: the ordering used for illustration in Figure 7 was employed post-discovery. In practice, we observed that the (arbitrarily ordered) cluster coverage display indicated (nearly) disjoint clusters and delved into the descriptors associated with the data in an effort to find a descriptor that correlated with the clusters. The T-cell/B-cell subclass was discovered thusly.)

It is hoped that results such as this one – novel working hypotheses – can be used to drive future scientific investigations.

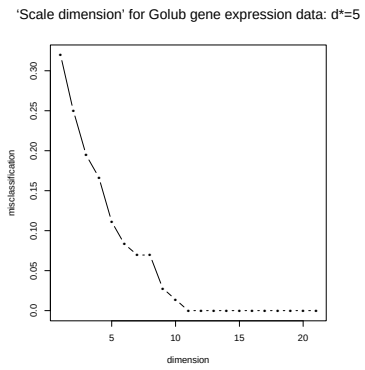

Figure 6: The empirical risk (y-axis) against the cluster map dimension for the gene expression data set. For this data set, $\widehat{d^\star} \approx 5$.

7 Gene Expression III: Inferential Statistical Analysis

Given our working hypothesis, a logical next step is to attempt to build support for (or evidence against) this hypothesis via inferential statistical analysis. We present here a simple but useful step in this direction.

Let η_B (respectively, η_T) represent the median (location) of the distribution of distances from B-cell ALL leukemia (respectively, T-cell ALL leukemia) to AML leukemia. The Wilcoxon rank sum test (equivalent to the Mann-Whitney test; see, e.g., Conover 1980 or Lehmann 1975) for the null hypothesis H_0: $\eta_B - \eta_T = 0$ versus the general alternative H_A: $\eta_B - \eta_T \neq 0$ yields the (two-sided) p-value = 0.0051.

Thus an inferential statistical analysis, undertaken in response to the latent subclass discovery conjecture formed based on exploratory statistical analysis, yields strongly significant evidence that B-cell ALL is closer

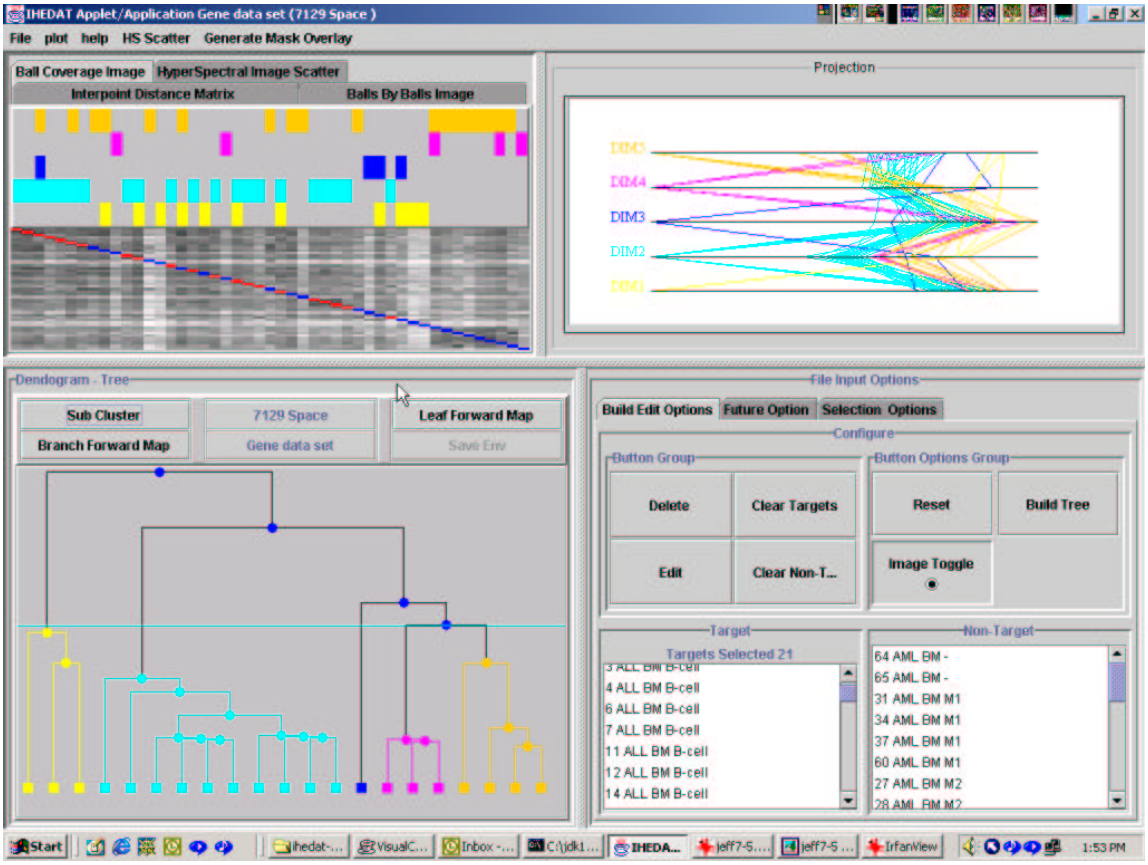

Figure 7: This graphic displays the intriguing results of using class cover catch digraphs for latent class discovery on the gene expression data set. The upper left panel depicts the clustering (into $\widehat{d}^* = 5$ clusters) associated with the five clusters of dominating set elements. It is in this panel that the latent class discovery emerges. See text for details.

to AML than is T-cell ALL. (This could be dominated by just a few genes, or just a few AML observations; the investigation of these issues is the subject of ongoing effort.)

Alas, this inference is biased, as we chose the (potential) subclasses after (exploratory) data analysis; valid inferential statistics requires an independent test set. "Nevertheless," as noted by Bickel and Doksum in their discussion of data-based model selection (2001, p. 8), "we can draw guidelines from our numbers and cautiously proceed."

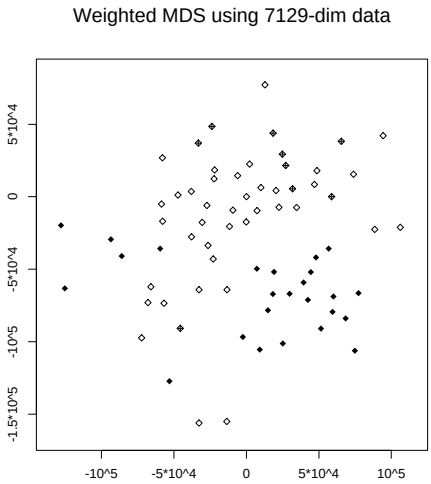

Figure 8: Multidimensional scaling map of the gene expression data. Observe that, in general, T-cell ALL observations (cross-hatched diamonds) are further from AML observations (filled diamonds) than are B-cell ALL observations (open diamonds).

8 Conclusions

We have presented a new methodology for latent class discovery based on the visualization of class cover catch digraphs, and we have applied the methodology to data sets from artificial nose chemical sensing and gene expression monitoring. In both cases, interesting latent class structure emerged. The discovery of latent subclasses leads to associated conjectures pertaining to the geometry of these subclasses in their respective high-dimensional observation spaces. In particular, the latent classes discovered in the investigation of the artificial nose data set are in keeping with our belief, based on our understanding of the high-dimensional geometry of the problem, that the low concentration target class observations should be associated with the smallest radii. For the gene expression data set, the conjecture is that B-cell ALL is "closer" to AML than is T-cell ALL in "gene expression space". The ultimate utility of our methodology in the discovery of new class distinctions for any given application will require subsequent scientific investigation of the subclass conjectures.

References

Bickel, P.J., Doksum, K.A., 2001. Mathematical Statistics: Basic Ideas and Selected Topics, Vol I, 2nd ed. Prentice Hall.

Cattell, R.B., 1978. The Scientific Use of Factor Analysis. Plenum, New York.

Conover, W.J., 1980. Practical Nonparametric Statistics, 2nd ed. Wiley, New York.

DeVinney, J.G., Priebe, C.E., 2001. Class cover catch digraphs. Johns Hopkins University Department of Mathematical Sciences Technical Report #633.

Everitt, B., 1980. Cluster Analysis, 2nd ed. Halsted, New York.

Getz, G., Levine, E., Domany, E., 2000. Coupled two-way clustering analysis of gene microarray data. *PNAS*, 97(22), 12079-12084.

Golub, T.R., Slonim, D.K., et al., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531-537.

Hartigan, J.A., 1975. Clustering Algorithms. Wiley. New York.

Lehmann, E.L., 1975. Nonparametrics: Statistical Methods Based on Ranks. Holden and Day, San Francisco.

Marchette, D.J., Priebe, C.E., 2002. Characterizing the scale dimension of a high-dimensional classification problem. *Pattern Recognition*, to appear.

Minnotte, M., West, W., 1998. The data image: a tool for exploring high dimensional data sets. *1998 Proceedings of the ASA Section on Statistical Graphics*, 25-33.

Priebe, C.E., 2001. Olfactory classification via interpoint distance analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4), 404–413.

Priebe, C.E., DeVinney, J.G., Marchette, D.J., 2001. On the distribution of the domination number of random class cover catch digraphs. *Statistics and Probability Letters*, 55(3), 239-246.

Priebe, C.E., Marchette, D.J., DeVinney, J.G., Socolinsky, D., 2002. Classification using class cover catch digraphs. Johns Hopkins University Department of Mathematical Sciences Technical Report #628.

Solka, J.L., Clark, B.T., Priebe, C.E., 2002. A visualization framework for the analysis of hyperdimensional data. *International Journal of Image and Graphics*, 2(1), 1-17.

Wegman, E.J., 1990. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85, 664-675.

Wegman, E.J., 1999. Visions: The evolution of statistics. Research in Official Statistics, 2(1), 7-19.

Wegman, E.J., 2000. Visions: New techniques and technologies in statistics. *Computational Statistics*, 15, 133-144.