ADVANCES IN APPLIED MATHEMATICS **19,** 319–331 (1997)
ARTICLE NO. AM970532

Randomized Nonlinear Projections Uncover High-Dimensional Structure

Lenore J. Cowen and Carey E. Priebe

Department of Mathematical Sciences,
Johns Hopkins University,
Baltimore, Maryland 21218

Received December 2, 1996; accepted December 30, 1996

We consider the problem of investigating the "structure" of a set of points in high-dimensional space (n points in d-dimensional Euclidean space) when $n \ll d$. The analysis of such data sets is a notoriously difficult problem in both combinatorial optimization and statistics due to an exponential explosion in d. A randomized nonlinear projection method is presented that maps these observations to a low-dimensional space, while approximately preserving salient features of the original data. Classical statistical analyses can then be applied, and results from the multiple lower-dimensional projected spaces are combined to yield information about the high-dimensional structure. We apply our dimension reduction techniques to a pattern recognition problem involving PET scan brain volumes.

THE PROBLEM. Let $x_1, \ldots, x_n$ be a collection of n observations in $\Re^d$. The goal is to cluster the observations into g groups. We are concerned with this problem in the unsupervised case, where nothing is known a priori about the classification of any of the data, as well as in the supervised case, where some of the x_i have associated with them a class label c_i. The extreme case of this is called the classification problem, where all observations x_i have associated with them a class label c_i, and the goal is to use this training data to develop a discriminant rule which then can be used to classify unlabeled observations. These problems have been studied extensively in the probability, statistics, engineering, and application-specific literature. See, for instance, [1, 2] and the references contained therein.

Approximate distance. Linial, London, and Rabinovich [3] investigated algorithmic constructions for embedding points from a high-dimensional Euclidean space into a lower dimensional space while approximately

0196-8858/97 $25.00

preserving all the pairwise distances between points. Their work extended results of [4–7] from the realm of functional analysis. Bourgain [4] showed that:

THEOREM. *Let X be a collection of n points in $\Re^d$, with distances computed under the L_2 norm. Then there is a function ψ which maps X to $\psi(X)$, a set of points in r-dimensional Euclidean space, such that $\forall x_i, x_j \in X$,*

$$\|x_i - x_j\| \geq \|\psi(x_i) - \psi(x_j)\| \geq \frac{1}{C \log n}\|x_i - x_j\|.$$

Furthermore, such a ψ can be found in randomized polynomial time.

The work of [3] includes explicit constructions that are highly combinatorial in nature, where the construction of ψ combines small random subsets of the sample points themselves. Our point of departure from [3] is based on the observation that for pattern recognition purposes it is necessary only to preserve class or cluster separability, rather than all interpoint distances. If the intra-cluster distances collapse to near zero, this will only magnify the ease of recognizing clustering behavior in the lower-dimensional space. This idea is also the basis of a deterministic method based on optimization techniques that appears in [8].

OVERVIEW OF THE METHOD

The method we employ has three stages: (1) find a useful set of projections to low-dimensional space, (2) explore the classification or clustering properties of the data in the projections, and (3) use the information obtained to determine structure in the high-dimensional space. Here we focus mainly on (1), since this has been the bottleneck in successfully attacking pattern recognition problems for high-dimensional data. The reasoning behind using low-dimensional projections is that there are many classical pattern recognition methods that work well in low dimensions. Therefore, while the statistical method needs to be tailored to the problem at hand, stage (2) can be accomplished by standard statistical techniques. How to combine and use the information from stage (2) will also depend on the problem at hand; in this paper, we will investigate stage (3) in specific cases only.

The ADC maps. We now define the family of maps that result in the nonlinear projections which are considered. We will call these the ADC maps, for *approximate distance clustering*. The one-dimensional ADC map

is defined in terms of a subset D of the original data. Each subset D specifies a one-dimensional ADC map; later we will choose D of a small fixed size randomly from the data.

DEFINITION. Let $X = \{x_i\}$ be a collection of data in $\Re^d$. Let $D \subset X$. The associated one-dimensional ADC map is defined as the function that, $\forall x_i \in X$, maps x_i to the scalar quantity $\min_{d \in D} \|x_i - d\|$. An r-dimensional ADC map is now defined in terms of subsets $D_1, \ldots, D_r$, where each $x_i \in X$ is mapped to the point in $\Re^r$ whose jth coordinate is its value under the one-dimensional ADC map with associated subset D_j.

Notice that specifying stage (1) now means simply choosing r and determining how to choose sets of subsets $D_1, \ldots, D_r$. In fact, we will be choosing the D (of a particular size k, dependent on characteristics the data, specified later) *in a randomized fashion* and, as has been the case in many algorithmic problems studied of late (cf. [9]), it seems that the randomness itself is what helps us to get over the main computational bottleneck.

Before we can begin a theoretical examination of how good the ADC projections are for finding candidate clusterings, however, we must be able to define what a good clustering is. This has not been, in general, an easy notion [10]. However, it is fairly easy to write down a restrictive definition of when data clusters into two clusters in one dimension as follows.

DEFINITION. Let $x_1, \ldots, x_n \in \Re$, so that $x_{(1)}, \ldots, x_{(n)}$ are the ordered points s.t. $x_{(i)} \leq x_{(j)}$. Let $y_i = x_{(i+1)} - x_{(i)}$ denote the spacings between points. Let $y_j = y_{(n-1)}$ be the largest gap. If either $x_{(j)} - x_{(1)} < y_{(n-1)}$ or $x_{(n)} - x_{(j+1)} < y_{(n-1)}$ then $x_1, \ldots, x_n$ are said to cluster perfectly in 1-dimension. The clusters are $x_{(1)} \cdots x_{(j)}$ and $x_{(j+1)} \cdots x_{(n)}$.

Based on this and for expository purposes, we first present a randomized algorithm based on ADC projections into just one dimension. In this simple case we can describe completely our algorithm's computational complexity for finding clusters for a data set whose underlying distribution falls into two clusters. We apply the method to a simulation and to a set of PET data. The extension to multiple dimensions and multiple clusters is then discussed in the following section.

MEASURES OF CLUSTERING

In what follows, let C be a cluster, with $C = \{x_1, \ldots, x_n\}$. Let $E = \{x_{n+1}, \ldots, x_{n+m}\}$, where $C \cup E$ is a set of $n + m$ points embedded in a d-dimensional metric space with norm $\|\cdot\|$ (For the purposes of this paper, we will always consider the L_2 norm). Here $d \gg n$.

DEFINITION. We say that C is k-clusterable if there exists a subset D of C such that $|D| = k$ and

$$2 \max_{c \in C} \min_{d \in D} \|c - d\| < \min_{e \in E} \min_{d \in D} \|e - d\|.$$

The set D is called a *witness* that C is k-clusterable.

For the purposes of this paper, k is constant, independent of n. More generally, k could be allowed to be a function of n, so that the size of D required will grow as n grows, for example, allowing $\log n$-clusterable sets.

LEMMA. *Let C and E be as above. If C is k-clusterable, then the mapping which sends each point $u \in C \cup E$ to $\min_{d \in D} \|u - d\|$, where D is a clusterability witness for C, clusters C perfectly in one-dimension.*

Proof. Let D be the clusterability witness above and f the associated mapping. First notice that if $x \in C$ and $y \in E$; then $\min_{d \in D} \|x - d\| < \min_{d \in D} \|y - d\|$ by the property of D. Thus f maps all the points of E to the right of all the points in C, on the real line. Let $y_{(j)}$ be the gap between the rightmost point in C and the leftmost point in E; i.e., $x_{(j)} \in C$ and $x_{(j+1)} \in E$. Then $x_{(j)} - x_{(1)} = x_{(j)}$, since f maps points in C which are also in D to 0. Thus we need to show that $x_{(j)} < x_{(j+1)} - x_{(j)}$, or $2x_{(j)} < x_{(j+1)}$. But $2x_{(j)} = 2 \max_{c \in C} \min_{d \in D} \|c - d\| < \min_{e \in E} \min_{d \in D} \|e - d\| = x_{(j+1)}$. ∎

We introduce a weaker notion called k-separability as well, by dropping the 2 in the definition above. In this case, instead of perfect clusters in one-dimension, the weaker property holds that all the points of C appear to the left of all the points of E in the image of f. This formulation is generalized to higher dimensions and multiple clusters below.

DEFINITION. Let C and E be as above. Fix $\epsilon > 0$. We say that C is strongly (k, ϵ)-clusterable, if for a random subset D of C of size k,

$$\text{Prob}\left[2 \max_{c \in C} \min_{d \in D} \|c - d\| < \min_{e \in E} \min_{d \in D} \|e - d\|\right] \geq \epsilon.$$

If there exists any $\epsilon > 0$ such that C is strongly (k, ϵ)-clusterable, we say C is strongly k-clusterable.

DEFINITION. Fix $\epsilon > 0$. Let $C \overset{\text{i.i.d.}}{\sim} F_C$ with $|C| = n$, and $E \overset{\text{i.i.d.}}{\sim} F_E$, with $|E| = m$. We say C is strongly (k, ϵ)-clusterable in distribution if for a random subset D of size k drawn uniformly from C, $\text{Prob}[2 \max_{c \in C} \min_{d \in D} \|c - d\| < \min_{e \in E} \min_{d \in D} \|e - d\|] \geq \epsilon$.

The same notions for this and the previous definitions are made for k-separable clusters by dropping the factor of 2.

THE ALGORITHM

Consider the following procedure. Suppose $x_1, \ldots, x_n$ are strongly (k, ϵ) clusterable into clusters C and $\overline{C}$. We choose D of size k[1] at random from among the data points (if we have partial classification information available, choosing D either entirely or partially from data points known to lie within C speeds up the algorithm). We form the associated one-dimensional ADC map and check whether the data clusters perfectly in one dimension. (In the case of partial classification information, we also require that all points known to be within C lie to the left of the biggest gap. In the case of total classification information, we remove the requirement of perfect clustering and simply require that the points in C lie to the left of the points in $\overline{C}$.) If this stage (2) analysis is successful in uncovering two resultant clusters, our goal has been met. If not, we select D at random again. How many times do we need to do this before we find a "good" ADC map? The following theorem provides an answer to this, the fundamental question of stage (1).

THEOREM. *Let δ be a lower bound on the fraction of the data points that lie within C. Then if the data is completely unclassified, the algorithm runs in time $O((c(\alpha)\delta^{-k}\epsilon^{-1})(nk + n \log n))$ and recovers C with probability $> 1 - 2^{-\alpha}$, for any fixed $\alpha > 0$. In the case that the algorithm can sample from points know to be in C, the algorithm runs in time $O((c(\alpha)\epsilon^{-1})(nk + n \log n))$ and recovers C with probability $> 1 - 2^{-\alpha}$, for any fixed $\alpha > 0$.*

Proof. We prove the second statement first. If $x_1, \ldots, x_n$ are strongly (k, ϵ) clusterable into clusters C and $\overline{C}$, checking $2/\epsilon$ samples R_i each chosen from points already classified as belonging to C gives probability $< (1 - \epsilon)^{2/\epsilon} < \frac{1}{2}$ that no sample is a witness, and this can be increased to $1 - 2^{-\alpha}$, for any fixed α by simply multiplying the number of samples chosen by a constant. For the first statement, the probability that a sample of k points lies entirely within C is at least δ^k. Thus if we choose $O(\delta^{-k})$ samples, we expect to have chosen a sample entirely from within C.

For each sample R_i, taking the distances of n points to the k points in the sample takes $O(kn)$ time. The cost of extracting the maximum gap is linear and dominated by the $O(n \log n)$ cost of sorting the $x_{(j)}$. ∎

Notice that in the absence of any classification data there is an extra factor of δ^k, which is exponential in k. Thus only for small constant k, and large constant $\delta \leq \frac{1}{2}$, will this result in a feasible bound on the number of samples required before the clustering is found. However, if there is even

[1] In practice, k is not known; the algorithm is run by hypothesizing $k_{\max}$ and k is first set to 1 and increased incrementally.

partial training data (i.e., samples known to come within C), for reasonable ϵ the procedure will be nearly linear in k, and in practice we have indeed found that an interesting structure is found, based on a small number of ADC projections. In practice, for clusterable data, setting a tolerance on running time and examining the ''best'' few projections produces dramatic results. For separable data one may expect only that the best projections order the data correctly; recognition of these projections requires additional information. In reality, of course, even separability is a strong assumption. In the absence of fully classified training data one must ''look for clumps.'' If an independent statistical test is on hand to check the quality of each candidate projection, however, the algorithm above becomes a classification algorithm. Given a fully classified training sample, each projection is evaluated for its utility in classification, using, for example, a leave-one-cut cross-validation procedure. The few best projections can then be used to classify unlabeled observations.

EXAMPLES

The power of k. Performance of ADC in practice is fundamentally dependent on the choice of k. For any fixed k, we could estimate ϵ, i.e. what percentage of ADC maps give perfect (100% correct) clustering. Once this percentage is nonnegligible, choosing samples uniformly at random will *decrease* the percentage as k increases, since clustering structure is typically observed when all k samples are drawn from the same cluster, which decreases in probability exponentially in k. *If we were able* to restrict our consideration to those ADC maps where samples were all chosen from within the same cluster, the quality of the clusters found and the percentage of the ADC maps that give perfect clustering would increase since within-cluster points are more likely to have a small distance to one of the within-cluster points in the sample. In the general case, we of course are not able to do this, since we do not know which points lie in each cluster. Note, however, that sometimes all k samples will lie in the same cluster by chance. Thus when the percentage of ADC maps which give 100% correct clustering (or above any fixed percentage p correct clustering) is negligible or 0, increasing k can increase the percentage of good ADC maps.

We remark that when the percentage of good ADC maps is below half, the information we wish to retrieve lies only in a percentage of ''best'' ADC maps, and stage (3) comes into play. In the one-dimensional ADC map, we have the ''perfect clustering test'' described above and we simply discard the bad projections in favor of the good ones. More complicated implementations of stage (3) can recover clustering structure that is more

delicate. Also, since ADC is a randomized method, when deciding whether to reject the hypothesis of ''no clustering,'' it is important to estimate and correct for the percentage of correctly classified observations one would expect to see by chance.

The following simulation example demonstrates the power of k. Consider $n = 100$ observations in $d = 10$-dimensional Euclidean space comprising two classes with $n_i = 50$ observations per class, $i = 1, 2$. The class distributions N_i differ only in two of the 10 dimensions: for eight of the dimensions we have $N_i = \text{MultivartiateNormal}\ (\mathbf{0}, \mathbf{1})$. For the last two dimensions, uncorrelated from the first 8, we take each cluster to be a mixture of multivariate normals. Letting I^2 be the two-dimensional identity matrix, dimensions 9 and 10 are distributed

$$1/2\left(N\left(\begin{bmatrix} 2.5 \\ 3.5a \end{bmatrix}, I^2\right)\right) + 1/2\left(N\left(\begin{bmatrix} -2.5 \\ 3.5a \end{bmatrix}, I^2\right)\right)$$

for each cluster, where $a = 1$ for cluster 1 and -1 for cluster 2. These data are then rotated via a random orthogonal matrix (so that the clusters cannot be found by one of the standard basis projections). One thousand random one-dimensional ADC projections were calculated. The largest gap between two projected points (not including the points in the random sample) was calculated, and the points were classified into two clusters based on whether they were to the left or the right of the gap. The results from this simulation (Table I) indicate that, when $k = 2$ there is a higher percentage of projections that yield an 80% or higher correct classification rate. As we demand a greater than 80% correct classification, the percentage gap between $k = 2$ and $k = 1$ widens: and only 0.5% of the $k = 1$ maps give a 100% classification, whereas 1.5% of the $k = 2$ maps do. The table shows the percentage of ADC projections that correctly cluster the given percentage of the data.

TABLE 1
The Simulation Example: If We Wish a High Percentage of the Points to Be Correctly Classified, Then It Is Better to Select $k = 2$

	Percentage of correct clustering								
	55	65	75	80	85	90	95	99	100
% ADC correct: $k = 1$	23.5	9.6	7.9	6.3	5.3	4.5	3.5	1.2	.5
% ADC correct: $k = 2$	18.5	9.1	7.2	6.3	6.1	6.0	5.4	3.2	1.8

Note. The crossover occurs at precisely the percentage of ADC maps correctly classifying at least 80% of the points correctly.

PET EXAMPLE. We now present an example from positron emission tomography (PET). Twenty six subjects (14 schizophrenic and 12 normal) are scanned in each of three conditions: rest (R), sensory control (SC), and tone recognition (TR). These three conditions give rise to two contrasts of interest for each subject: SC-R and TR-SC, yielding a data set of $n = 26$ PET scan volume pairs. Each $65 \times 87 \times 26$ voxel scan has been normalized and aligned using the SPM statistical image analysis software package [11]. A voxel's value represents a measure of the change in the amount of blood flow to that area of the brain and, thus, regional neural activity [12]. The resulting scan volume pairs are not readily clusterable into ''schizophrenics'' and ''normals'' visually, due to a large interclass variance.

This volumetric image data is naively represented as having a dimension for each voxel. Thus we consider clustering $n = 26$ observations in $d = 2(65 \times 87 \times 26) = 294{,}060$ dimensions. In the PET community it is common practice, once SPM registration has been performed, to proceed by analyzing each voxel separately, i.e. disregarding spatial context [11]. Thus PET is a particularly good example for expository purposes; we can demonstrate the ease of applying our methods to small sample sizes in extremely high-dimensional settings and obtain significant results without needing to incorporate complicating spatial preanalysis. Figure 1 shows one midbrain transverse slice for one subject.

There are 26 one-dimensional ADC projections for this problem that use one point per set ($k = 1$) and none provide any useful clustering—not surprising, perhaps, as the high-dimensional structure is no doubt complicated. Figure 2 shows the results of the one-dimensional ADC map for stage (1) of this problem using five observations per set. Both qualitatively and quantitatively (using the leftmost mode of the probability density estimator) it is indicated that normals seem to be more tightly clustered than schizophrenics. That is, for normals the majority of the probability mass tends to be at the smaller distances, whereas for the schizophrenics there is more at the larger distances.

There is also behavioral evidence based on performance versus reaction time which proposes to explain the additional structure seen in Fig. 2. Five of the schizophrenics behave similarly to the normals—i.e., some schizophrenics perform this task normally. These may be indicated in Fig. 2 by the schizophrenic subjects with significant probability mass at the small distances and could indicate that the behavioral result has a physiological manifestation.

Discriminant analysis (leave-one-out cross-validation) based on ADC projections yields correct classification of 25 of the 26 observations into the classes normal and schizophrenic. This is a pleasant surprise since the volumetric images themselves, even for experienced PET scan analysts, do not readily yield the conclusion that such a classification rule exists.

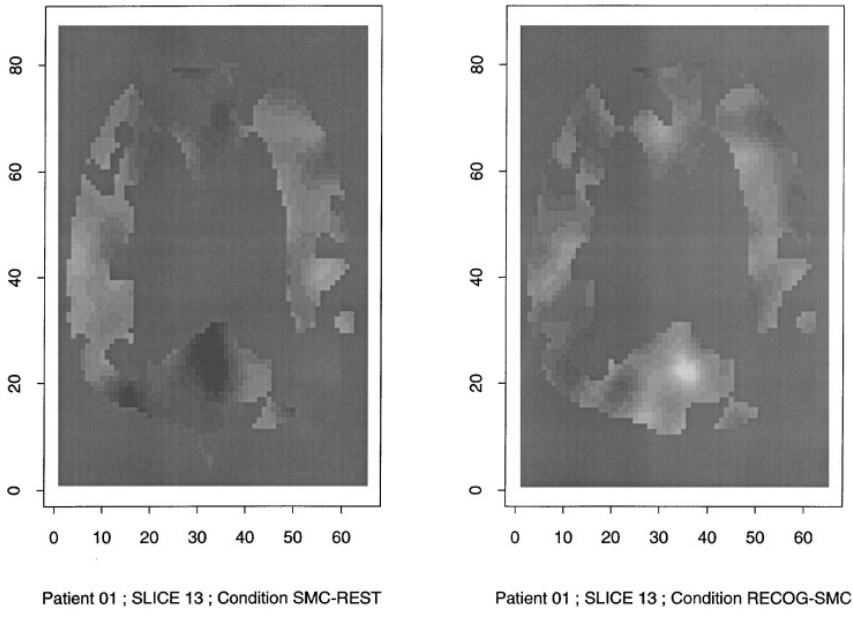

FIG. 1. This shows one transverse slice from one of the 26 scan pairs.

This naive voxel-as-dimension approach is obviously not a solution to the general image clustering application, as ignoring spatial information may be foolhardy for inherently contextual data. However, image analysis is a notoriously difficult application, and the ultimate solution to any image clustering problem will involve a compendium of techniques. Furthermore, in PET—a subtractive application in which registration is possible and for which the features of interest are normalizable, anatomically based, equi-located changes in blood flow—the noncontextual approach described above has potential application.

MULTIPLE CLUSTERS

In some sense, using the one-dimensional method to find two clusters is the general case from an algorithmic point of view, since, given a set of data that clusters into g groups, we can separate cluster C_i from the rest of the data, remove C_i, and iterate. However, from an approximate distance versus dimension reduction point of view, it is interesting to ask for projections that can represent multiple clusters simultaneously. Here we show how the required dimensionality will, in general, increase as a function of the number of clusters.

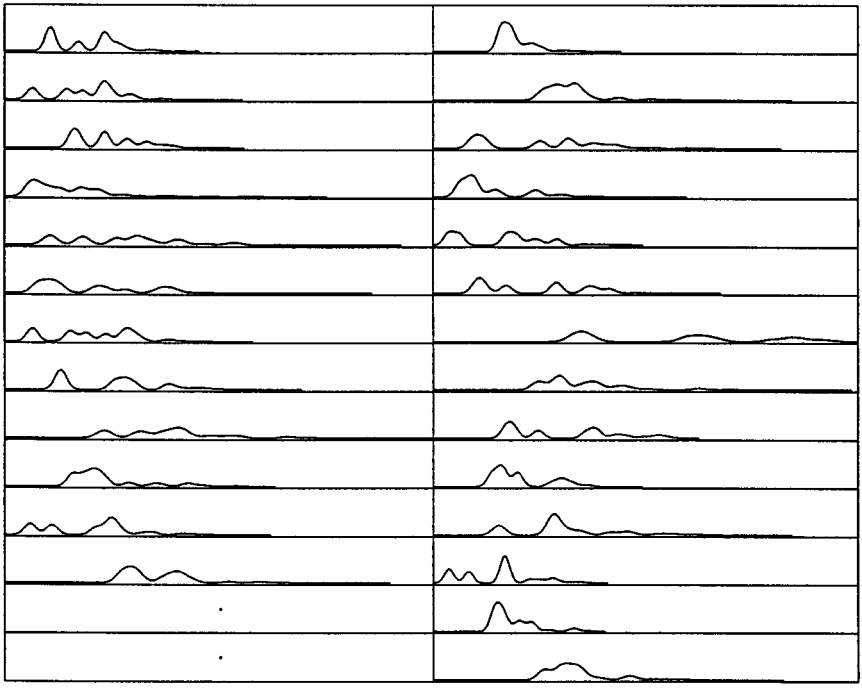

FIG. 2. This presents kernel density estimates of the distances for 2500 random ADC projections using a random sample of $k = 5$ observations per set. The subjects on the left are normals, and those on the right are schizophrenics.

Suppose d-dimensional data clusters into g clusters, $C_1, \ldots, C_g$, with $g > 2$. As before, we wish to represent this data in fewer than d dimensions, in such a way that the separation between clusters is not lost. In this section, it will be cleaner to generalize the one-dimensional k-separable, rather than k-clusterable definitions.

DEFINITION. Clusters $C_1, \ldots, C_g$ in $\Re^s$ are *linearly separable* if there exists $g - 1$ $(s - 1)$-dimensional hyperplanes that partition $\Re^s$ such that each C is contained in its own region.

Let C be a collection of points in $\Re^d$. Given a collection of s subsets $D_1, \ldots, D_s$, $D \subset C$, we define the associated s-dimensional ADC map M: $\Re^d \to \Re^s$ to be the map which, for each $c \in C$, $M(c) = (m_1, \ldots, m_s)$, where $m_i = \min_{d \in D_i} \|d - c\|$. In the theorem below, k-separable refers to the two clusters C_j and all the rest of the data, $C \setminus C_j$.

THEOREM. *Suppose* $\mathbf{C} = C_1, \ldots, C_g$ *lie in* $\Re^d$ *so that for each* C_j, C_j *is* k*-separable from* $\mathbf{C} \setminus C_j$. *Then* (1) *there exists an ADC map into* $s = g - 1$ *dimensions so that the images of* $C_1, \ldots, C_g$ *are linearly separable*; (2) *there exists a collection* $\mathbf{C}$ *as above so that no ADC map with one point per set into fewer than* $g - 1$ *dimensions results in linear separable clusters in the image*; (3) *there exists a collection* $\mathbf{C}$ *as above so that no ADC map into fewer than* $\log_2 g$ *dimensions results in linear separable clusters in the image*.

Proof. (Sketch). (1) For $i = 1 \cdots g - 1$, let D_i separate C_i from $\mathbf{C} \setminus C_i$ and let M be the ADC map associated with $D_1 \cdots D_{g-1}$. Let $\epsilon_i = \max_{c \in \mathbf{C}} \min_{d \in D_i} \|c - d\|$. Define h_i to be the $(g - 1)$-dimensional hyperplane which contains each of the axes except i and goes through the ith axis at the coordinate ϵ_i. Clearly h_i separates C_i from the rest of $\mathbf{C}$. (2) Let $\mathbf{C}$ be as follows: C_i consists of identical points, all of whom have 0s, except a 1 in the ith coordinate. Then each of the C_i is 1-separable from $\mathbf{C} \setminus C_i$. Choosing a single point in C_i, all the points in C_i map to 0, whereas all the non-C_i points map to 1. Any point $c \in C_i$ will map to a point with coordinates 0 or 1 under any ADC map: 0 in a coordinate where a point from C_i is chosen; 1 otherwise. Thus if only one point is chosen per set, we need a point from each set and $g - 1$ dimensions are required. (3) Choose the same collection $\mathbf{C}$ as in part (2). Suppose we have an ADC map that clusters $\mathbf{C}$ in $s < \log_2 g$ dimensions. We can represent each cluster C_i by a binary string of length s, where we put a 1 in position i if some element of C_i was chosen in D_i. If $s < \log_2 g$, some C_i and some C_j will be represented by the same binary string (by the pigeonhole principle) and, hence, their elements will be mapped to the same points in $\Re^d$. ∎

One easy specialization of the theorem above is the observation that the one-dimensional method cannot in general represent three distinct clusters. For example, if three clusters form points of an equilateral triangle in 2-space, any point in one cluster will be equidistant from points in both of the other clusters, no matter how well-separated the three clusters are in 2-space. However, the one-dimensional method can sometimes recover three clusters, for example, when they are at different distances from each other. This representation of more clusters than the theory might indicate can occur frequently in practice.

DISCUSSION

We have introduced a new method for finding clusters in high-dimensional space based on the preservation of approximate distances between clusters. At the heart of the method is a randomized algorithm: in some sense, we can say that our definition of when a clustering structure can be

recovered is a randomized one. The randomness can allow us to automatically find dense cluster regions and thus pull out cluster structure. J. Michael Steele has asked us: is there a way to define clustering along these lines, but in such a way that the criteria is *norm independent* or at least is invariant over a wide class of distance metrics?

Other open questions include: for what data is it possible to give a dimension reduction map that represents g clusters in less than $f(g)$ dimensions? What alternative distance metrics should be considered?

For the purpose of this paper, we examine data that clustered; in many practical examples, the raw data would not cluster because of misclassified sample points, or outliers. There has been a large literature in the optimization and machine learning communities about optimizing misclassification rates. The definition of k-clusterable can be generalized to deal with misclassification, so that error rates can be better studied, and this is an important area of future research.

Finally, as far as the PET data sets themselves are concerned, we have recently found that restricting the region of interest via preprocessing a la ''statistical parametric mapping'' [11] or spatial smoothing [13] is an effective way to combine the ADC procedure with information about spatial dependencies.

ACKNOWLEDGMENTS

The authors are very grateful to Nati Linial and his student Eran London for bringing their work on the geometry of graphs to our attention. Thanks to Henry Holcomb for the PET data and expertise and to Dave Marchette and Clyde Schoolfield for their assistance with the simulations. Thanks to Bill Bogstad and Margaret Zhao. LJC is supported in part by an ONR Young Investigator Grant N00014-96-1-0829. CEP is supported in part by ONR Young Investigator Grant N00014-95-1-0777 and ONR Grant N00004-96-1-0313.

REFERENCES

1. R. Duda and P. Hart, ''Pattern Classification and Scene Analysis,'' Wiley, New York, 1973.
2. L. Devroye, L. Gyorfi, and G. Lugosi, ''A Probabilistic Theory of Pattern Recognition,'' Springer-Verlag, New York, 1996.
3. N. Linial, E. London, and Y. Rabinovich, *Combinatorica* **15** (1995), 215–245.
4. J. Bourgain, *Israel J. Math.* **52** (1985), 46–52.
5. W. Johnson, and J. Lindenstrauss, *Contemporary Mathematics* **26** (1984), 189–206.
6. W. Johnson, J. Lindenstrauss, and G. Schechtman, ''Geometric Aspects of Functional Analysis,'' pp. 177–184, Springer-Verlag, Berlin/New York, 1987.
7. J. Matousek, *Comment. Math. Univ. Carolinae.* **33** (1992), 51–55.

8. K. Fukunaga, ''Introduction to Statistical Pattern Recognition,'' Academic Press, New York, 1972.
9. R. Motwani, and P. Raghavan, ''Randomized Algorithms,'' Cambridge Univ. Press, 1995
10. J. Hartigan (1975) ''Clustering Algorithms,'' (Wiley, New York).
11. K. Friston, A. Holmes, K. Worsley, J. Poline, C. Frith, and R. Frackowiak, *Human Brain Mapping* **2** (1995), 189–210.
12. Z.-H. Cho, J. Jones, and M. Singh, *Foundations of Medical Imaging*. (1993). (Wiley, New York).
13. C. Priebe, H. Holcomb, and D. Marchette, *in* ''Proc., 1996 Joint Statistical Meetings, 1996.''