Trevor Hastie is an associate professor in the Statistics and Biostatistics departments at Stanford University. He received a B.Sc from Rhodes University, South Africa, a M.Sc. from the University of Cape Town, South Africa, and a Ph.D in statistics from Stanford University in 1984. His Ph.D thesis is titled *Principal Curves and Surfaces*, which is now an active area of research, and bears a close relationship to self organizing maps. He spent 9 years in the Statistics and Data Analysis research department at AT&T Bell laboratories, Murray Hill, New Jersey. His research has focused on applied modelling problems, especially nonparametric regression and classification. He coauthored *Generalized Additive Models, Chapman and Hall, 1990* with R. Tibshirani, and coedited *Statistical Models in S, Chapman and Hall, 1991* with J. Chambers. He is a member of the International Statistics Institute, and a fellow of the Royal Statistical Society.

Robert Tibshirani is a professor in the Statistics and Biostatistics departments at University of Toronto. He received a B.Sc. degree from the University of Waterloo, and a Ph.D in statistics from Stanford University in 1984. He has many research articles on nonparametric regression and classification. He co-authored *Generalized Additive Models, Chapman and Hall, 1990* with T. Hastie, and coauthored *An Introduction to the Bootstrap, Chapman and Hall, 1993* with B. Efron. He has been an active researcher on bootstrap technology for the past 11 years. His 1984 Ph.D thesis spawned the currently lively research area known as Local Likelihood. He is a recent recipient of a Guggenheim Foundation fellowship. He is a fellow of the American Statistical association and the Institute of Mathematical Statistics.

reduction.

[10] proposed a technique close to ours for the two class problem. In our terminology they used our metric with $\mathbf{W} = \mathbf{I}$ and $\epsilon = 0$, with $\mathbf{B}$ determined locally in a neighborhood of size K_M. In effect this extends the neighborhood infinitely in the null space of the local between class directions, but they restrict this neighborhood to the original K_M observations. This amounts to projecting the local data onto the line joining the two local centroids. In our experiments this approach tended to perform on average 10% worse than our metric, and we did not pursue it further. [11] extended this to $J > 2$ classes, but here their approach differs even more from ours. They computed a weighted average of the J local centroids from the overall average, and project the data onto it, a one-dimensional projection. Even with $\epsilon = 0$ we project the data onto the subspace containing the local centroids, and deform the metric appropriately in that subspace. [12] recognized a shortfall of the Short and Fukanaga approach, since the averaging can cause cancellation, and proposed other metrics to avoid this. Although their metrics differ from ours, the Chi-squared motivation for our metric (3) was inspired by the metrics developed in their paper. We have not tested out their proposals, but they report results of experiments with far more modest improvements over standard nearest neighbors than we achieved.

[6] proposes a number of techniques for flexible metric nearest neighbor classification. These techniques use a recursive partitioning style strategy to adaptively shrink and shape rectangular neighborhoods around the test point. Friedman also uses derived variables in the process, including discriminant variates. With the latter variables, his procedures have some similarity to the discriminant adaptive nearest neighbor approach.

Other recent work that is somewhat related to this is that of [13]. He estimates the covariance matrix in a variable kernel classifier using a neural network approach.

There are a number of ways in which this work might be generalized. In some discrimination problems, it is natural to use specialized distance measures that capture invariances in the feature space. For example [14],[15] use a transformation-invariant metric to measure distance between digitized images of handwritten numerals in a nearest neighbor rule. The invariances include local transformations of images such as rotation, shear and stroke-thickness. An invariant distance measure might be used in a linear discriminant analysis and hence in the DANN procedure.

Near neighbor techniques are used in the regression setting as well. Local polynomial regression [16] is currently very popular, where, for example, locally weighted linear surfaces are fit in modest sized neighborhoods. Analogs of K-NN classification for small K are used less frequently. In this case the response variable is quantitative rather than a class label.

[17] invented a technique called *sliced inverse regression*, a dimension reduction tool for situations where the regression function changes in a lower-dimensional space. They show that under symmetry conditions of the marginal distribution of X, the inverse regression curve $E(X|Y)$ is concentrated in the same lower-dimensional subspace. They estimate the curve by slicing Y into intervals, and computing conditional means of X in each interval, followed by a principal component analysis. There are obvious similarities with our `DANN` procedure, and the following generalizations of `DANN` are suggested for regression:

- locally we use the $\mathbf{B}$ matrix of the sliced means to form our `DANN` metric, and then perform local regression in the deformed neighborhoods.
- The local $\mathbf{B}(i)$ matrices can be pooled as in `subDANN` to extract global subspaces for regression. This has an apparent advantage over the approach of [17]: we only require symmetry locally, a condition that is locally encouraged by the convolution of the data with a spherical kernel.

Acknowledgments

The authors thank Jerome Friedman for sharing his work on related problems, and for many enjoyable conversations.

References

[1] T.M. Cover, "Rates of convergence for nearest neighbor procedures", in *Proc. Hawaii Inter. Conf. on Systems Sciences.* 1968, pp. 413–415, Western Periodicals, Honolulu.

[2] Richard O. Duda and Peter E. Hart, *Pattern classification and scene analysis*, Wiley, New York, 1973.

[3] Geoffrey J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York, 1992.

[4] T.M. Cover and P. Hart, "Nearest neighbor pattern classification", *Proc. IEEE Trans. Inform. Theory*, pp. 21–27, 1967.

[5] T. Hastie, A. Buja, and R. Tibshirani, "Penalized discriminant analysis", To appear, Annals of Statistics, 1994.

[6] J. Friedman, "Flexible metric nearest neighbour classification", Tech. Rep., Stanford University, November 1994.

[7] B.D. Ripley, "Neural networks and related methods for classification", *J. Royal Statist. Soc. (Series B) (with discussion)*, 1994.

[8] L. Breiman, J.H. Friedman, R. Olshen, and C.J. Stone, *Classification and Regression Trees*, Wadsworth, 1984.

[9] D. Michie, D. Spigelhalter, and C. Taylor, Eds., *Machine Learning, Neural and Statistical Classification*, Ellis Horwood series in Artificial Intelligence. Ellis Horwood, 1994.

[10] R. Short and K. Fukanaga, "A new nearest neighbor distance measure", in *Proc. 5th IEEE Int. Conf. on Pattern Recognition*, 1980, pp. 81–86.

[11] R. Short and K. Fukanaga, "The optimal distance measure for nearest neighbor classification", *IEEE transactions of Information Theory*, vol. IT-27, pp. 622–627, 1981.

[12] J.P. Myles and D. J. Hand, "The multi-class metric problem in nearest neighbour discrimination rules", *Pattern Recognition*, vol. 23, pp. 1291–1297, 1990.

[13] D. G. Lowe, "Similarity metric learning for a variable kernel classifier", Tech. Rep., Dept. of Comp Sci, Univ. of British Columbia, 1993.

[14] P. Y. Simard, Y. LeCun, and J. Denker, "Efficient pattern recognition using a new transformation distance", in *Advances in Neural Information Processing Systems*, San Mateo, CA, 1993, pp. 50–58, Morgan Kaufman.

[15] T. Hastie, P. Simard, and Eduard Sackinger, "Learning prototype models for tangent distance", Tech. Rep., AT& Bell Labs, 1993.

[16] W Cleveland, "Robust locally-weighted regression and smoothing scatterplots", *Journal of the American Statistical Society*, vol. 74, pp. 829–836, 1979.

[17] N. Duan and K-C Li, "Slicing regression: a link-free regression method", *Annals of Statistics*, pp. 505–530, 1991.

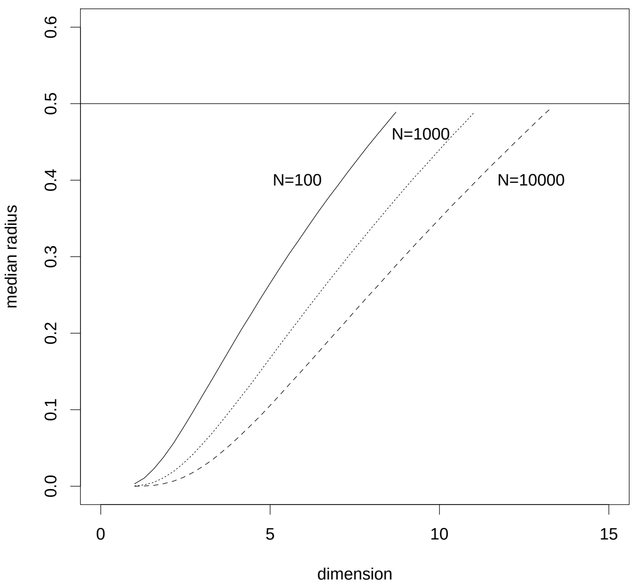

Fig. 9. Median radius of a one-nearest neighborhood as a function of dimension and size of training sample.

Here $v_d = 2\pi^{d/2}/(d \cdot \Gamma(d/2))$ and $v_d r^d$ is the volume of the sphere of radius r in d dimensions. From this we can compute the median of R:

$$\text{med}(R) = v_d^{-1/d}(1 - .5^{1/N})^{1/d}. \qquad (10)$$

Figure 9 shows the median radius as a function of the dimension and size of the training sample. A horizontal line is drawn at the maximum radius 0.5, We see that for $N = 100$ the median radius reaches the maximum by dimension 9, and is already as large as 0.25 by dimension 5. For $N = 10000$ the situation is only a little better.

Now suppose we have a two-class problem with $Pr(Y = 1|\mathbf{x}) = p(\mathbf{x})$. We form a nearest neighborhood at a point $\mathbf{x}_0$ and find the nearest neighbor $\mathbf{X}$ having class $Y(\mathbf{X})$ (the upper case letter denotes a random variable). Our estimate of $p(\mathbf{x}_0)$ is simply $Y(\mathbf{X})$. The bias and variance of $Y(\mathbf{X})$ are

$$\begin{aligned} \text{Bias} &= \text{E}p(\mathbf{X}) - p(\mathbf{x}_0) \\ \text{Var} &= \text{var}[\text{E}(Y(\mathbf{X})|\mathbf{X})] + \text{E}[\text{var}(Y(\mathbf{X})|\mathbf{X})] \\ &= \text{E}p(\mathbf{X})[1 - \text{E}p(\mathbf{X})]. \end{aligned} \qquad (11)$$

The expectations in the final expression for bias and variance are with respect to the distribution of the nearest neighbor $\mathbf{X}$. Now as dimension increases, we have seen above that the distance between nearest neighbor $\mathbf{X}$ and $\mathbf{x}_0$ increases. Therefore if $p(\mathbf{x})$ changes appreciably over the space, the difference between $p(\mathbf{x})$ and $p(\mathbf{x}_0)$ will tend to increase and hence the bias will increase. But (11) also tells us that the variance will often increase as well. Suppose we have equal numbers in each class. Note that $p(1-p)$ takes is maximum at $p = 1/2$. Now if $\mathbf{x}_0$ is in a reasonably pure region (that is $p(\mathbf{x}_0)$ near 0 or 1), $p(\mathbf{x}_0)(1 - p(\mathbf{x}_0))$ will be small. But as the nearest neighbor $\mathbf{X}$ moves farther away from $\mathbf{x}_0$, $\text{E}(p(\mathbf{X}))$ will tend towards 0.5 and hence the variance will increase.

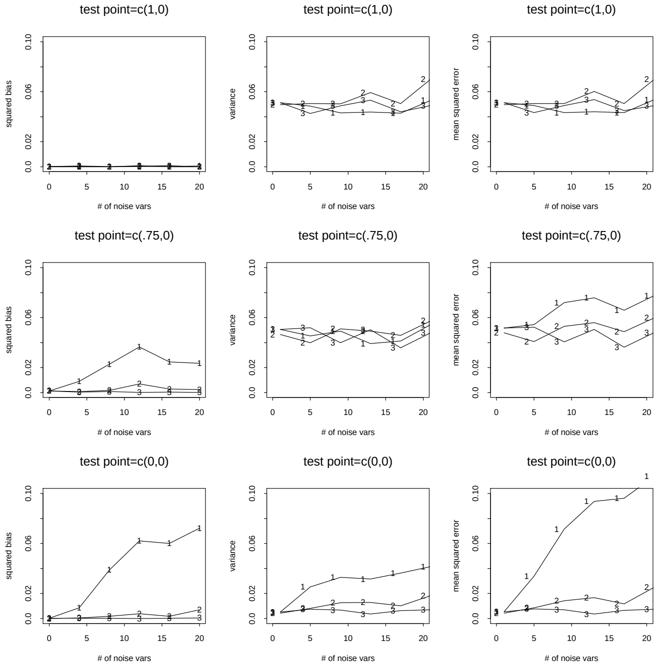

Fig. 10. Bias, variance and mean squared error of 5-NN (1), DANN (2) and 5-NN restricted to the first predictor (3). There are two classes with centers (0, 0) and (2, 0). In the top, middle and bottom rows the prediction is done at (1, 0), (.75, 0) and (0, 0) respectively. The simulation standard errors of the bias and variance estimates are roughly .015 and .003 respectively.

For the two-dimensional normal example with noise, we carried out some simulations to estimate the bias and variance of the 5-NN, DANN, and 5-NN restricted to the first predictor. Figure 10 shows the bias and variance of class probability estimates from each method, as the number of noise variables increases. The estimates were obtained from 100 training samples of size 200.

Recall that the class centers are at $(0, 0, 0, \ldots 0)$ and $(2, 0, 0, \ldots 0)$. In the top, middle and bottom rows of the figure, the classification was done at the points $(1, 0, \ldots 0), (.5, 0, 0, \ldots 0)$ and $(0, 0, \ldots 0)$ respectively, for which the true probability of class 1 are 0.5, 0.729 and 0.981.

The variance of the three methods is similar, since they all use a fixed number of neighbors (5). The bias of 5-NN increases with increasing number of noise variables, but DANN retains a low bias, similar to that of reduced 5-NN by concentrating on the directions orthogonal to maximum centroid separation.

VII. Discussion

We have developed an adaptive form of nearest neighbor classification that can offer substantial improvements over standard nearest neighbors method in some problems. We have also proposed a method that uses local discrimination information to estimate a subspace for global dimension

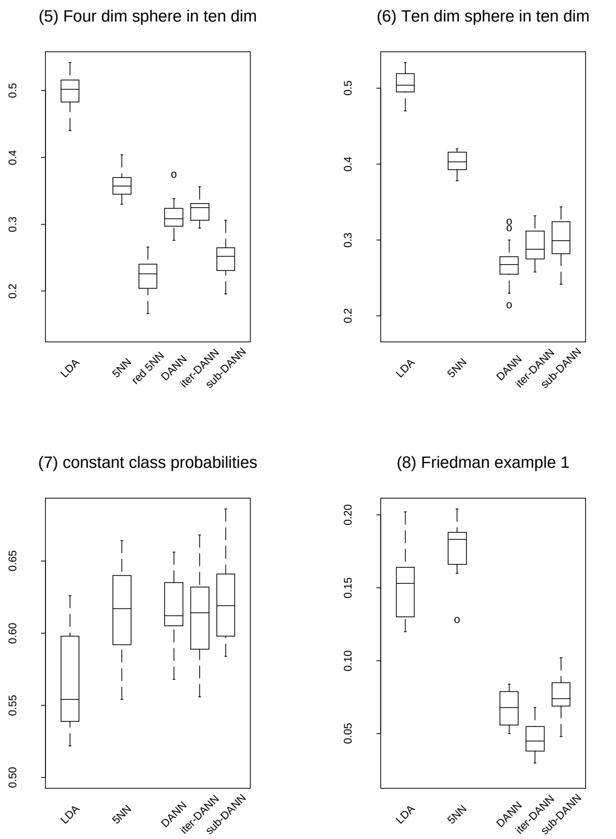

Fig. 6. Boxplots of error rates over 20 simulations, second four simulated examples.

are taken from the data archive of the STATLOG [9]. [1]. The goal is to classify each pixel into one of 7 land types: *red soil, cotton, vegetation stubble, mixture, grey soil, damp grey soil, very damp grey soil.* We extract for each pixel its 8-neighbors, giving us $(8+1) \times 4 = 36$ features (the pixel intensities) per pixel to be classified. The data come scrambled, with 4435 training pixels and 2000 test pixels, each with their 36 features and the known classification. Included in figure 8 is the true classification, as well as that produced by linear discriminant analysis. The right panel compares **DANN** to all the procedures used in STATLOG, and we see the results are favorable.

VI. Some bias-variance calculations

In this section we examine the bias and variance of the nearest neighbor and discriminant adaptive nearest neighbor rules.

First we derive the distribution of the radius of the nearest neighborhood. Suppose we have N data points uniformly distributed in the unit cube $[-.5,.5]^d$. Consider a spherical (one)-nearest neighborhood centered at the origin. Let R be the radius of the neighborhood. Then

$$Prob(R \geq r) = (1 - v_d r^d)^N. \tag{9}$$

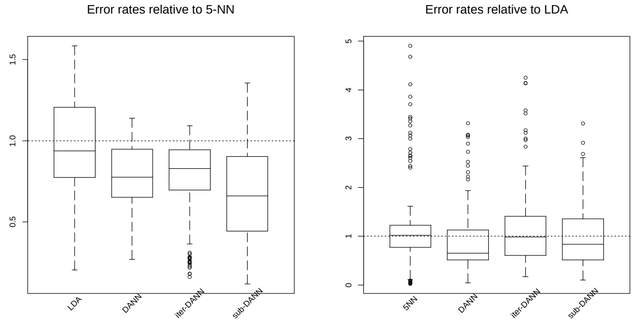

Fig. 7. Relative error rates of the methods across the 8 simulated problems. In the top panel the error rate has been divided by the error rate 5-NN, on a simulation by simulation basis. In the bottom panel we have divided by the error rate of LDA.

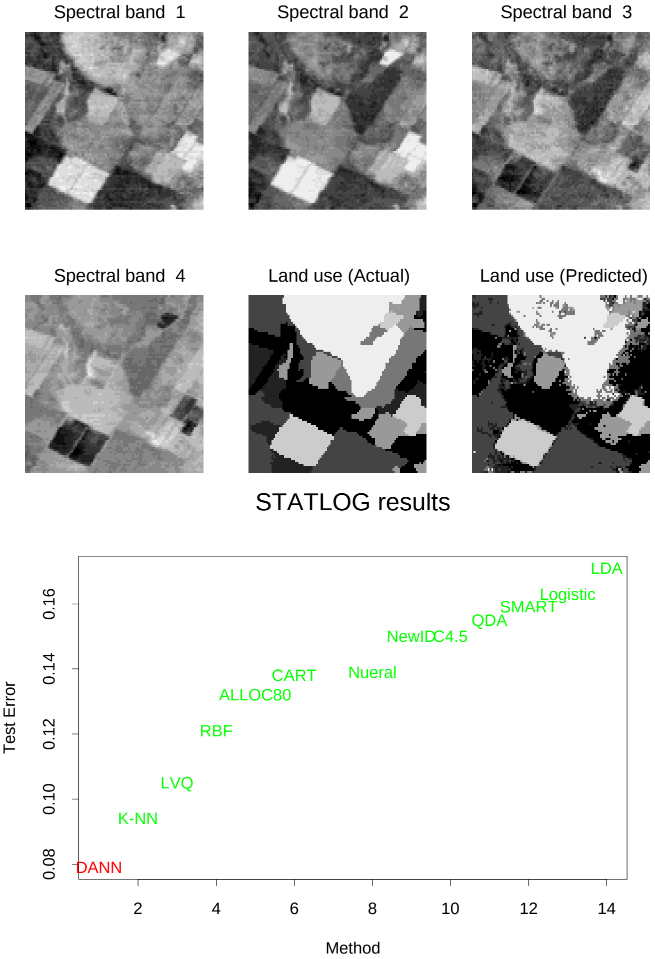

Fig. 8. The first four images are the satellite images in the four spectral bands. The fifth image represents the known classification, and the final image is the classification map produced by linear discriminant analysis. The right panel shows the misclassification results of a variety of classification procedures on the satellite image test data (taken from [9].) **DANN** is the overall winner.

[1] The authors thank C. Taylor and D. Spiegelhalter for making these images and data available

9. *Sonar data.* This example has 60 predictors, two classes ("mines" and "rocks") and 104 observations in both the training and test data set. It was obtained from the benchmark collection maintained by Scott Fahlman at Carnegie Mellon University, and was contributed by Terry Sejnowski.
10. *Vowel data.* This example is a popular benchmark for neural network algorithms, and consists of training and test data with 10 predictors and 11 classes. It was also obtained from the benchmark collection maintained by Scott Fahlman at Carnegie Mellon University.
11. *Glass data.* This data was taken from [7]. The goal is to classify 4 types of forensic glass from 10 chemical attributes. There are 89 training cases and 96 test cases.
12. *Heart data.* These data were analyzed in [8]. We received it from Leo Breiman, who credits Elizabeth Gilpin and and Richard Olshen with helping him to obtain it. After removing two variables with many missing values, the dataset consists of 19 measurements on 779 patients who had recently suffered a heart attack. The objective was to predict survival to 30 days. There were 77 deaths in the dataset. We randomly chose a training sample of size 679 and a test sample of size 100.

C. Discussion of results

The results for the simulated examples are summarized in Figures 5 and 6. Each experiment was repeated 20 times, and the boxplots are a convenient summary of these 20 results for each configuration. The results for the real data examples are given in Table I.

DANN seems to do as well as 5-NN across the board, and offers significant improvements in problems with noise variables (2, 4, 6, and 8). DANN does not do as well as reduced nearest neighbors in problems 2 or 5: this is not surprising since in effect we are giving the nearest neighbor rule the information that DANN is trying to infer from the training data. A nearest neighbor method with variable selection might do well in these problems: however this procedure can be foiled by by rotating the relevant subspace away from the coordinate directions. In Friedman's example 1, the error rates of DANN are roughly the same as the rates for the "machete" and "scythe" reported in [6].

On the average there seems to be no advantage in carrying out more than one iteration of the DANN procedure. The subspace DANN procedure is the overall winner, producing big gains in problems admitting global dimension reduction.

In the sonar example, DANN outperforms 5-NN, although it should be noted that 1-NN gives only 3.8% errors. The performance of DANN on the vowel data is particularly encouraging: the error rate of 38.3% is the lowest that we know of for any procedure. In the glass and heart data, there is little to choose among any of the methods.

The top panel of Figure 7 shows error rates relative to 5-NN, accumulated across the 8×20 simulated problems. The bottom panel shows the rates relative to LDA.

We see that DANN is 20-30% better than 5-NN on the average, and is at most 20% worse. DANN is also better than LDA on the average but can be three times worse (in problem 2).

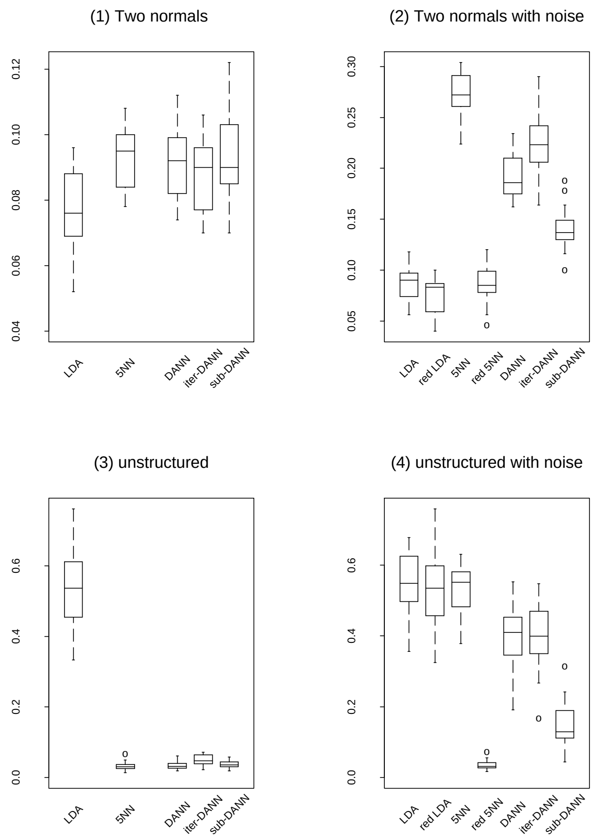

Fig. 5. Boxplots of error rates over 20 simulations, first four simulated examples.

TABLE I
Results for real data examples

Problem	LDA	5-NN	DANN	DANN	iter-DANN
			$\epsilon = 1$	best ϵ	$\epsilon = 1$
sonar	24.0	18.2	4.8	4.8	4.8
vowel	55.6	50.0	40.3	38.3	42.4
glass	40.6	40.0	40.4	36.5	42.7
heart	9.0	9.0	11.0	10.0	10.0

V. Image Classification Example

Here we consider an image classification problem. The data consist of 4 LANDSAT images in different spectral bands of a small area of the earths surface, and the goal is to classify into soil and vegetation types. Figure 8 shows the four spectral bands, two in the visible spectrum (red and green) and two in the infra red spectrum. These data

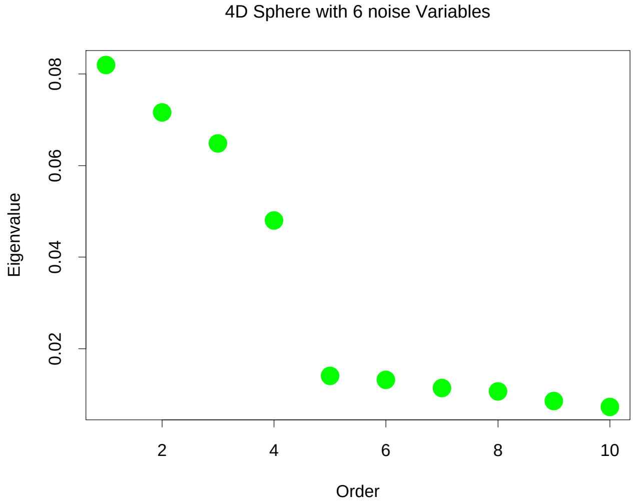

Fig. 4. The eigenvalues of the average between matrix for the 4D sphere + 6 noise variable problem. Using these first four dimensions followed by our DANN nearest neighbor routine, we get better performance than 5NN in the real 4D subspace.

IV. Examples

A. The methods

In the following examples we compare several classification approaches:

- LDA—linear discriminant analysis
- reduced LDA—linear discriminant restricted to the relevant subspace, where appropriate. In example 2 below, for instance, the relevant subspace is defined by the first two predictors. Reduced LDA is included for comparison purposes, as one would not normally know the relevant subspace in a problem.
- 5-NN: 5 nearest neighbor classification
- reduced 5-NN: 5 nearest neighbor classification, restricted to the relevant subspace as above. Again this subspace would not normally be known.
- DANN— Discriminant adaptive nearest neighbor, one iteration.
- iter-DANN— Discriminant adaptive nearest neighbor, five iterations.
- sub-DANN— Discriminant adaptive nearest neighbor, with automatic subspace reduction. This is described in section III.

For all methods, the predictors were first standardized so as to have zero mean and unit variance over the training set, and the test set predictors was standardized by the corresponding training mean and variance. The training and test set sizes were 200 and 500, unless indicated otherwise.

B. The problems

1. *2 Dimensional Gaussian.* Two Gaussian classes in two dimensions (X_1, X_2) separated by 2 units in X_1. The predictors have variance $(1, 2)$ and correlation 0.75.
2. *2 Dimensional Gaussian with 14 Noise.* As in (1), augmented with 14 predictors having independent standard Gaussian distributions.
3. *Unstructured.* In this example we simulated data with extremely disconnected class structure. There are 4 classes each with 3 spherical bivariate normal subclasses, having standard deviation 0.25. The means of the 12 subclasses were chosen at random (without replacement) from the integers $[1, 2, \cdots 5] \times [1, 2, \cdots 5]$. Each training sample had 20 observations per subclass, for a total of 240 observations.
4. *Unstructured with 8 Noise.* As in 3 above, but augmented with 8 predictors having independent standard Gaussian distributions.
5. *4 Dimensional Spheres with 6 Noise.* In this example there are 10 predictors and 2 classes. The last 6 predictors are noise variables, with standard Gaussian distributions, independent of each other and the class membership. The first four predictors in class 1 are independent standard normal, conditioned on the radius being greater than 3, while the first four predictors in class 2 are independent standard normal without the restriction. The first class almost completely surrounds the second class in the four dimensional subspace of the first four predictors. This example was designed to see if DANN could improve upon nearest neighbors in the presence of noise variables.
6. *10 Dimensional Spheres.* As in the previous example there are 10 predictors and 2 classes. Now all 10 predictors in class 1 are independent standard normal, conditioned on the radius being greater than 22.4 and less than 40, while the predictors in class 2 are independent standard normal without the restriction. In this example there are no pure noise variables, the kind that a nearest neighbor subset selection rule might be able weed out. At any given point in the feature space, the class discrimination occurs along only one direction. However this direction changes as we move across the feature space and all variables are important somewhere in the space. The first class almost completely surrounds the second class in the full ten-dimensional space.
7. *Constant Class Probabilities.* This is 4 class problem, with class probabilities (.1,.2,.2,.5) independent of $\mathbf{x}$. The $\mathbf{x}$ vectors were independent standard Gaussian in 6 dimensions. The training sample size was 100. The idea here was to investigate the cost of using an adaptive method like DANN in a scenario where adaptation is not needed.
8. *Friedman's Example 1*: This example is taken from [6]. There are two classes in 10 dimensions, 200 training observations, 500 test observations. The predictors in class 1 are independent standard normal; those in class 2 are independent normal with mean $\sqrt{j}/2$ and variance $1/j$, for $j = 1, 2, \ldots 10$. All predictors are important here, but the ones with higher index j are more so.

dominate. This is the value we used in our examples below.

III. Dimension Reduction using Local Discriminant Information

So far our technique has been entirely "memory based", in that we locally adapt a neighborhood about a query point at the time of classification. Here we describe a method for performing a global dimension reduction, by pooling the local dimension information over all points in the training set. In a nutshell we consider subspaces corresponding to eigenvectors of the average local between sum-of-squares matrices. One would first project the data onto the chosen subspace, where classification would then take place (using $K-NN$, DANN or any other classifier). The novelty here is in choosing the subspace using pooled local information.

Consider first how linear discriminant analysis (LDA) works. After sphering the data, it concentrates in the space spanned by the class means $\bar{\mathbf{x}}_j$ or a reduced rank space that lies close to these means. If $\bar{\mathbf{x}}$ denote the overall mean, this subspace is exactly the principal component hyperplane for the data points $\bar{\mathbf{x}}_j - \bar{\mathbf{x}}$, weighted by the class proportions.

Our idea to compute the deviations $\bar{\mathbf{x}}_j - \bar{\mathbf{x}}$ locally in a neighborhood around each of the N training points, and then do an overall principal components analysis for the $N \times J$ deviations. Here are the details. Let $\mathbf{x}_j(i)$ be the mean of class j vectors in a neighborhood of the ith training point, and $\bar{\mathbf{x}}(i)$ be the overall mean. All means are weighted by the local class membership proportions $\pi_j(i),\ j = 1, \ldots, J$. Let $\tilde{\mathbf{x}}_j(i) = \mathbf{x}_j(i) - \bar{\mathbf{x}}(i)$, the local centroid deviations. We seek a subspace that gets close in average weighted squared distance to all $N \times J$ of these. Denoting by $\mathbf{U}$ $(p \times J)$ an orthonormal basis for the $k < p$ dimensional subspace,we minimize the criterion

$$RSS(\mathbf{U}) = \sum_{i=1}^{N}\sum_{j=1}^{J} \pi_j(i)\tilde{\mathbf{x}}_j^T(i)(\mathbf{I} - \mathbf{U}\mathbf{U}^T)\tilde{\mathbf{x}}_j(i),$$

or the total weighted residual sum of squares. It is not hard to show that minimizing $RSS(\mathbf{U})$ amounts to maximizing

$$\mathrm{tr}\ \mathbf{U}^T \left(\sum_{i=1}^{N} \mathbf{B}(i)\right) \mathbf{U}$$

where the $\mathbf{B}(i)$ are the local between sum-of-squares matrices. This latter problem is solved by finding the largest eigenvectors of the average between sum-of-squares matrix $\sum_{i=1}^{N} \mathbf{B}(i)/N$.

Figure 3 shows a simple illustrative example. The two classes are Gaussian with substantial within class covariance between the two predictors X_1 and X_2. In the left panel, the solid line is the Gaussian decision boundary that optimally separates the classes. The orthogonal vector labeled S is a one-dimensional subspace onto which we can project the data and perform classification. Using the knowledge that the data are Gaussian, it is the leading discriminant direction. The broken lines are the boundaries and equivalent subspace produced by our procedure. In the right panel, each line segment represents the local between information centered at that point. Our procedure uses a principal components analysis of these $N \times J$ line segments to produce the broken line subspace in the left panel.

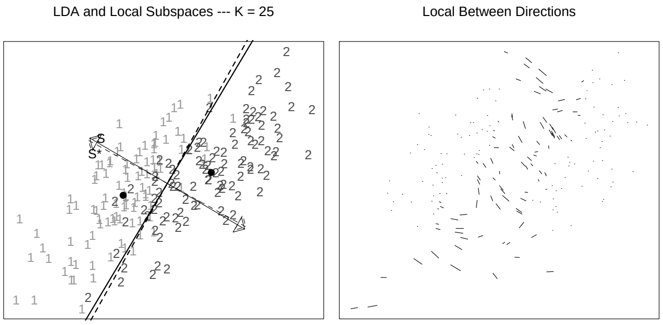

Fig. 3. [Left Panel] Two dimensional gaussian data with two classes and correlation 0.65. The solid lines are the LDA decision boundary and its equivalent subspace for classification. The dashed lines were produced by the local procedure described in this section. [Right panel] Each line segment represents the local between information centered at that point.

To allow combination of the local between information in a meaningful way, notice that we have not sphered the data locally before computing the mean deviations. A justification for this is that any local spherical window containing two classes, say, will have approximately a linear decision boundary orthogonal to the vector joining the two means.

Figure 4 shows the eigenvalues of the average between matrix for an instance of a two-class, 4 dimensional sphere model with 6 noise dimensions. The decision boundary is a 4 dimensional sphere, although locally linear (full details of this example are given in the next section). For this demonstration we randomly rotated the 10 dimensional data, so that the dimensions to be trimmed are not coordinate directions. The eigenvalues show a distinct change after 4 (the correct dimension), and using our DANN classifier in these four dimensions actually beats ordinary 5NN in the *known* four dimensional sphere subspace in many simulation realizations (because DANN does additional local neighborhood adjustments.)

It is desirable to automate the dimension reduction operation. Since our local information is based on spherical neighborhoods (potentially in high dimensions), we find an iterative approach most successful. We apply this procedure in the full space, and use cross-validated DANN to find the best nested subspace (with a built in bias towards larger subspaces). We then successively repeat these operations in the new subspaces, until no further reduction is deemed suitable by CV. Using DANN in this final subspace is what we have labelled **sub-DANN** in the boxplots of figures 5 and 6.

$\mathbf{W}$ and $\sum_j p(j|\mathbf{x}_0)(\mu_j - \bar{\mu})(\mu_j - \bar{\mu})^T$ by $\mathbf{B}$ gives the first term in the metric (2).

By allowing prior uncertainty for the class means μ_j, that is, assuming $\mu_j \sim N(\nu_j, \epsilon\mathbf{I})$ in the sphered space, we obtain the second term in the metric (2).

A. Details of the implementation

Define a weight function at $\mathbf{x}_0$ by

$$k(\mathbf{x}, \mathbf{x}_0; \Sigma, h) = \phi_h(||\Sigma_0^{1/2}(\mathbf{x} - \mathbf{x}_0)||). \tag{6}$$

Here Σ_0 is an initial non-negative metric (often I), and ϕ_h is a symmetric real-valued function depending on a parameter h. We use a *tri-cube* function defined over a K-nearest neighborhood $N_K(\mathbf{x}_0)$ of $\mathbf{x}_0$. Formally, we define $d_i = ||\Sigma^{1/2}(\mathbf{x}_i - \mathbf{x}_0)||$, $h = \max_{i \in N_K(\mathbf{x}_0)} d_i$ and define

$$k(\mathbf{x}_i, \mathbf{x}_0; \Sigma, h) = [1 - (d_i/h)^3]^3 I(|d_i| < h). \tag{7}$$

Let $\mathbf{B}(\mathbf{x}_0; \Sigma_0, h)$ and $\mathbf{W}(\mathbf{x}_0; \Sigma_0, h)$ be the weighted between and within class sum-of-squares matrices, where the weights assigned to the *ith* observation are given by $w_i = k(\mathbf{x}_i, \mathbf{x}_0; \Sigma_0, h)$. That is,

$$\begin{aligned} \mathbf{B}(\mathbf{x}_0; \Sigma_0, h) &= \sum_{j=1}^{J} \hat{\pi}_j(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^T \\ \hat{\pi}_j &= \frac{\sum_{y_i=j} w_i}{\sum_{i=1}^{N} w_i} \\ \mathbf{W}(\mathbf{x}_0; \Sigma_0, h) &= \sum_{j=1}^{J} \sum_{y_i=j} w_i(\mathbf{x}_i - \bar{\mathbf{x}}_j)(\mathbf{x}_i - \bar{\mathbf{x}}_j)^T / \sum_{i=1}^{N} w_i \end{aligned} \tag{8}$$

where $\bar{\mathbf{x}}_j$ is the weighted mean of the N_j observations in the jth group. Finally, we let $\mathbf{B}(\mathbf{x}_0; \Sigma_0, h)$ and $\mathbf{W}(\mathbf{x}_0; \Sigma_0, h)$ determine the metric Σ in (2).

Notice that equations (8) and (2) produce a mapping $\Sigma_0 \rightarrow \Sigma$, say $\Sigma = g(\Sigma_0)$. An approach we explore is to start with $\Sigma_0 = I$ (the identity matrix) and iterate this procedure. The result is a metric Σ for use in a nearest neighbor classification rule at $\mathbf{x}_0$. In our examples we explore the effect of this iterative procedure.

B. Some remarks about the DANN metric

It is natural to ask whether the mapping $g(\cdot)$ has a fixed point, and if it does, whether an iteration of the form $\Sigma \leftarrow g(\Sigma)$ converges to it. These questions seem difficult to answer in general. To get some insight, it is helpful to consider an equivalent form of the iteration. At each step we take a spherical neighborhood around the test point, estimate the metric Σ, and then transform the predictors via $\mathbf{x}^{new} = \Sigma^{1/2}\mathbf{x}^{old}$. At completion we use a spherical nearest neighbor rule in the final transformed space. It is easy to show that this procedure is equivalent to the one given above. If the metrics estimated in j iterations are $\Sigma_1, \Sigma_2, \ldots \Sigma_j$, then the effective metric for the original coordinates is $\Sigma_j^{1/2}\Sigma_{j-1}^{1/2} \cdots \Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2} \cdots \Sigma_{j-1}^{1/2}\Sigma_j^{1/2}$.

Expressed in this way, the fixed points of the iteration satisfy $\mathbf{W}^{-1}\mathbf{B}\mathbf{W}^{-1} + \epsilon\mathbf{W}^{-1} = c\mathbf{I}$. In particular a fixed point occurs when $\mathbf{B}$ is zero and $\mathbf{W}$ is proportional to the identity matrix, in the space of the transformed coordinates.

In practice we find it more effective to estimate only the diagonal elements of $\mathbf{W}$, and assume that the off diagonal elements are zero. This is especially true if the dimension of the predictor space is large, as there will be insufficient data locally to estimate the $O(p^2)$ elements of $\mathbf{W}$. With the diagonal approximation, the two forms of the algorithm are not equivalent: we use the version that transforms the space at each step since a diagonal approximation makes most sense in the transformed coordinates.

If the predictors are spatially or temporally related, we might use a penalized estimate of $\mathbf{W}$ that downweights components of the covariance that correspond to spatially noisy signals [5]. A related approach is to pre-filter the predictors using a smooth basis, and then operate in the reduced domain.

In the final neighborhood we perform K nearest neighbor classification. An alternative approach would be to use discriminant analysis to perform the classification, using the locally determined parameters. We are currently investigating this approach.

C. Choice of tuning parameters

The DANN procedure has a number of adjustable tuning parameters:

K_M : the number of nearest neighbors in the neighborhood $N_{K_M}(\mathbf{x}_0)$ for estimation of the metric;

K : the number of neighbors in the final nearest neighbor rule;

ϵ : the "softening" parameter in the metric.

While K is common to all near-neighbor rules, our procedure has introduced two new parameters. We do not have any theory on which to base their choice, but have experimented with different ranges of values. Test sets or cross validation could be used to estimate a optimal values for these parameters. In the examples in the next section we instead use fixed choices, based on previous experimentation.

The value of K_M must be reasonably large since the initial neighborhood is used to estimate a covariance: we use $K_M = \max(N/5, 50)$. To ensure consistency one should take K_M to be a vanishing fraction of N, and should also use larger values for higher dimensional problems. Often a smaller number of neighbors is preferable for the final classification rule to avoid bias: we used $K = 5$, and compared it to standard 5 nearest neighbors. Since our metric adapts to avoid bias, it is conceivable that in cases where a small K is needed for pure nearest neighbors, we might be able to support a larger K and reduce the variance.

Note that the metric (2) is invariant under nonsingular transformations of the predictors, and hence it is not unreasonable to consider fixed values of ϵ. In many of our experiments we tried values of ϵ in the set $\{0, 0.01, 0.1, .2, .5, 1, 2, 5\}$; in all cases any value greater than 0 was better than 0, and there was little to distinguish amongst them, with a value of $\epsilon = 1$ appearing to

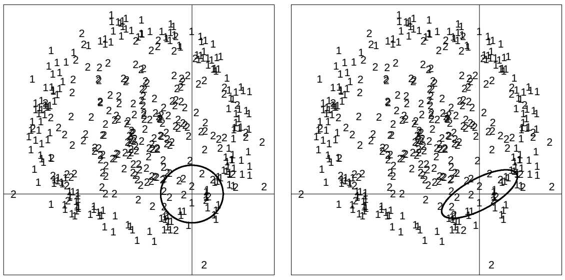

Fig. 2. The left panel shows the spherical neighborhood containing 25 points. The right panel shows the ellipsoidal neighborhood found by the DANN procedure, also containing 25 points. The latter is elongated along the true decision boundary, and flattened orthogonal to it.

these lines in the literature. A summary of previous work is provided in section VII.

II. Discriminant adaptive nearest neighbors

Our proposal is motivated as follows. Consider first a standard linear discriminant (LDA) classification procedure with K classes. Let $\mathbf{B}$ and $\mathbf{W}$ denote the $p \times p$ between and within sum-of-squares matrices. In LDA the data are first sphered with respect to $\mathbf{W}$, then the target point is classified to the class of the closest centroid (with a correction for the class prior membership probabilities). Since only relative distances are relevant, any distances in the complement of the subspace spanned by the sphered centroids can be ignored. This complement corresponds to the null space of $\mathbf{B}$.

We propose to estimate $\mathbf{B}$ and $\mathbf{W}$ locally, and use them to form a local metric that approximately behaves like the LDA metric. One such candidate is

$$\begin{aligned}\Sigma &= \mathbf{W}^{-1}\mathbf{B}\mathbf{W}^{-1}\\ &= \mathbf{W}^{-1/2}(\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2})\mathbf{W}^{-1/2}\\ &= \mathbf{W}^{-1/2}\mathbf{B}^{*}\mathbf{W}^{-1/2}. \qquad (1)\end{aligned}$$

where $\mathbf{B}^*$ is the between sum-of-squares in the sphered space. Consider the action of Σ as the matrix in a quadratic metric for computing (squared) distances $(\mathbf{x}-\mathbf{x}_0)^T\Sigma(\mathbf{x}-\mathbf{x}_0)$:

- it first spheres the space using $\mathbf{W}$;
- components of distance in the null space of $\mathbf{B}^*$ are ignored;
- other components are weighted according to the eigenvalues of $\mathbf{B}^*$ when there are more than 2 classes — directions in which the centroids are more spread out are weighted more than those in which they are close.

Thus this metric would result in neighborhoods similar to the narrow strip in figure 1: infinitely long in the null space of $\mathbf{B}$, and then deformed appropriately in the centroid subspace according to how they are placed. It is dangerous to allow neighborhoods to extend infinitely in any direction, so we need to limit this stretching. Our proposal is

$$\begin{aligned}\Sigma &= \mathbf{W}^{-1/2}[\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}+\epsilon\mathbf{I}]\mathbf{W}^{-1/2}\\ &= \mathbf{W}^{-1/2}[\mathbf{B}^{*}+\epsilon\mathbf{I}]\mathbf{W}^{-1/2} \qquad (2)\end{aligned}$$

where ϵ is some small tuning parameter to be determined, and $\mathbf{I}$ is the p-dimensional identity matrix. The metric shrinks the neighborhood in directions in which the local class centroids differ, with the intention of ending up with a neighborhood in which the class centroids coincide (and hence nearest neighbor classification is appropriate). With this goal in mind one can think of iterating the procedure, and thus successively shrinking in directions in which the class centroids do not coincide.

Here is a summary of the proposal.

```
Discriminant Adaptive Nearest Neighbor Classifier
  0. Initialize the metric $\Sigma = \mathbf{I}$, the identity matrix.
  1. Spread out a nearest neighborhood of $K_M$ points around the test point $\mathbf{x}_0$, in the metric $\Sigma$.
  2. Calculate the weighted within and between sum-of-squares matrices $\mathbf{W}$ and $\mathbf{B}$ using the points in the neighborhood (see formula (8 below).
  3. Define a new metric $\Sigma = \mathbf{W}^{-1/2}[\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}+\epsilon\mathbf{I}]\mathbf{W}^{-1/2}$.
  4. Iterate steps 1, 2, and 3.
  5. At completion, use the metric $\Sigma$ for $K$-nearest neighbor classification at the test point $\mathbf{x}_0$.
```

The metric (2) can be given a more formal justification. Suppose we are classifying at a test point $\mathbf{x}_0$ and find a single nearest neighbor $\mathbf{X}$ according to a metric $d(\mathbf{X},\mathbf{x}_0)$. Let $p(j|\mathbf{x})$ be the true probability of class j at point $\mathbf{x}$.

We consider the *Chi-squared* distance

$$r(\mathbf{X},\mathbf{x}_0) = \sum_{j=1}^{J}\frac{[p(j|\mathbf{X})-p(j|\mathbf{x}_0)]^2}{p(j|\mathbf{x}_0)} \qquad (3)$$

which measures the distance (appropriately weighted) between the true and estimated posteriors. Small $r(\mathbf{X},\mathbf{x}_0)$ implies that the misclassification error rate will be close to the asymptotic error rate for 1NN, which is achieved when $\mathbf{X} = \mathbf{x}_0$ or more generally when $p(j|\mathbf{X}) = p(j|\mathbf{x}_0)$. We show that the first term in our metric (2) approximates $r(\mathbf{X},\mathbf{x}_0)$.

Assuming that in the neighborhood $\mathbf{x}|j$ has a Gaussian distribution with mean μ_j and covariance Σ, we obtain by a simple first-order Taylor approximation

$$p(j|\mathbf{X}) \approx p(j|\mathbf{x}_0) - p(j|\mathbf{x}_0)(\mu_j-\bar{\mu})^T\Sigma^{-1}(\mathbf{X}-\mathbf{x}_0) (4)$$

where $\bar{\mu} = \sum_j p(j|\mathbf{x}_0)\mu_j$. Plugging this into (3) we get

$$r(\mathbf{X},\mathbf{x}_0) \approx \sum_{j=1}^{J} p(j|\mathbf{x}_0)\left[(\mu_j-\bar{\mu})^T\Sigma^{-1}(\mathbf{X}-\mathbf{x}_0)\right]^2. \qquad (5)$$

Thus the approximately best distance metric is $\Sigma^{-1}\sum_j p(j|\mathbf{x}_0)(\mu_j-\bar{\mu})(\mu_j-\bar{\mu})^T\Sigma^{-1}$. Estimating Σ by

Discriminant Adaptive Nearest Neighbor Classification

Trevor Hastie, Robert Tibshirani

Abstract— **Nearest neighbor classification expects the class conditional probabilities to be locally constant, and suffers from bias in high dimensions. We propose a locally adaptive form of nearest neighbor classification to try to ameliorate this curse of dimensionality. We use a local linear discriminant analysis to estimate an effective metric for computing neighborhoods. We determine the local decision boundaries from centroid information, and then shrink neighborhoods in directions orthogonal to these local decision boundaries, and elongate them parallel to the boundaries. Thereafter, any neighborhood-based classifier can be employed, using the modified neighborhoods. The posterior probabilities tend to be more homogeneous in the modified neighborhoods. We also propose a method for global dimension reduction, that combines local dimension information. In a number of examples, the methods demonstrate the potential for substantial improvements over nearest neighbor classification.**

Keywords— **classification, nearest neighbors, linear discriminant analysis**

I. Introduction

We consider a discrimination problem with J classes and N training observations. The training observations consist of predictor measurements $\mathbf{x} = (x_1, x_2, \ldots x_p)$ on p predictors and the known class memberships. Our goal is to predict the class membership of an observation with predictor vector $\mathbf{x}_0$

Nearest neighbor classification is a simple and appealing approach to this problem. We find the set of K nearest neighbors in the training set to $\mathbf{x}_0$ and then classify $\mathbf{x}_0$ as the most frequent class among the K neighbors. Nearest neighbors is an extremely flexible classification scheme, and does not involve any pre-processing (fitting) of the training data. This can offer both space and speed advantages in very large problems: see [1],[2], or [3] for background material on nearest neighborhood classification.

[4] show that the one nearest neighbor rule has asymptotic error rate at most twice the Bayes rate. However in finite samples the curse of dimensionality can severely hurt the nearest neighbor rule. The relative radius of the nearest-neighbor sphere grows like $r^{1/p}$ where p is the dimension and r the radius for $p = 1$, resulting in severe bias at the target point $\mathbf{x}$ (see section VI). Figure 1 illustrates the situation for a simple example, where the data in each class are uniformly distributed in a half-square, and perfectly separated by a vertical line. . Our illustration here is based on a 1-NN rule, but the same phenomenon occurs for k-NN rules as well. Nearest neighbor techniques are based on the assumption that locally the class posterior probabilities are approximately constant (in figure 1, the true posterior probabilities are constant along any vertical line.)

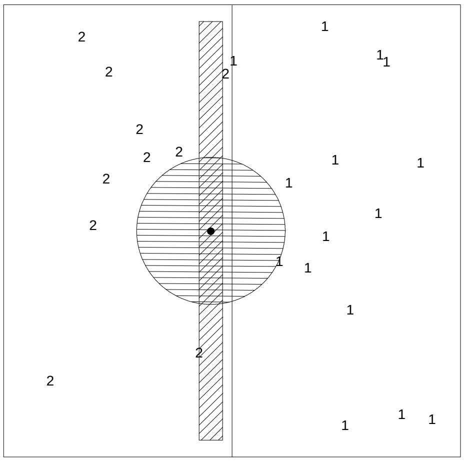

Fig. 1. The points are uniform in the square, with the vertical line separating class 1 and 2. The vertical strip denotes the NN region using only the horizontal coordinate to find the nearest neighbor for the target point (solid dot). The sphere shows the NN region using both coordinates, and we see in this case it has extended into the class 1 region (and found the wrong class in this instance).

Using only the horizontal coordinate in figure 1, we create narrow vertical-strip neighborhoods, for which this assumption is approximately true. The same size neighborhood using both coordinates is too wide in the horizontal direction—the direction in which the posterior probabilities change.

Figure 2 shows an example of our discriminant adaptive nearest neighbor (DANN) metric. There are two classes in two dimensions, one of which almost completely surrounds the other. The left panel shows a nearest neighborhood of size 25 at the target point (shown as the origin), which is chosen to be near the class boundary. The right panel shows the same size neighborhood using a DANN metric. Notice how the modified neighborhood extends further in the direction parallel to the decision boundary. As we will see in our simulation studies, this new neighborhood can often provide improvement in classification performance.

While the idea of local adaptation of the nearest neighbor metric may seem obvious, we could find few proposals along

Trevor Hastie is Associate Professor in the Departments of Statistics and Biostatistics, Stanford University, California 94305; trevor@playfair.stanford.edu. His work was partially supported by NSF grant DMS-9504495. Robert Tibshirani is Professor in the Department of Preventive Medicine and Biostatistics, and Department of Statistics, University of Toronto; tibs@playfair.stanford.edu; this work was completed while he was on sabbatical leave at Stanford University in 1994-95, and was supported by a grant from the Guggenheim foundation and the Natural Sciences and Engineering Research Council of Canada.