Classification Using Kernel Density Estimates: Multiscale Analysis and Visualization

Anil K. GHOSH and Probal CHAUDHURI

Theoretical Statistics and Mathematics Unit
Indian Statistical Institute
Kolkata-700108, India
(*anilkghosh@rediffmail.com*; *probal@isical.ac.in*)

Debasis SENGUPTA

Applied Statistics Unit
Indian Statistical Institute
Kolkata-700108, India
(*sdebasis@isical.ac.in*)

The use of kernel density estimates in discriminant analysis is quite well known among scientists and engineers interested in statistical pattern recognition. Using a kernel density estimate involves properly selecting the scale of smoothing, namely the bandwidth parameter. The bandwidth that is optimum for the mean integrated square error of a class density estimator may not always be good for discriminant analysis, where the main emphasis is on the minimization of misclassification rates. On the other hand, cross-validation–based methods for bandwidth selection, which try to minimize estimated misclassification rates, may require huge computation when there are several competing populations. Besides, such methods usually allow only one bandwidth for each population density estimate, whereas in a classification problem, the optimum bandwidth for a class density estimate may vary significantly, depending on its competing class densities and their prior probabilities. Therefore, in a multiclass problem, it would be more meaningful to have different bandwidths for a class density when it is compared with different competing class densities. Moreover, good choice of bandwidths should also depend on the specific observation to be classified. Consequently, instead of concentrating on a single optimum bandwidth for each population density estimate, it is more useful in practice to look at the results for different scales of smoothing for the kernel density estimates. This article presents such a multiscale approach along with a graphical device leading to a more informative discriminant analysis than the usual approach based on a single optimum scale of smoothing for each class density estimate. When there are more than two competing classes, this method splits the problem into a number of two-class problems, which allows the flexibility of using different bandwidths for different pairs of competing classes and at the same time reduces the computational burden that one faces for usual cross-validation–based bandwidth selection in the presence of several competing populations. We present some benchmark examples to illustrate the usefulness of the proposed methodology.

KEY WORDS: Majority voting; Misclassification rates; MISE; Optimal bandwidths; p value-type measure; Pairwise coupling; Posterior probability; Weighted posterior.

1. INTRODUCTION

Classification based on kernel density estimates has been widely discussed in the literature on pattern recognition and statistical learning (see, e.g., Duda, Hart, and Stork 2000; Hastie, Tibshirani, and Friedman 2001 for some recent reviews). The basic problem in classification or discriminant analysis is to formulate a decision rule, $\mathsf{d}(\mathbf{x}) : \mathbb{R}^d \to \{1, 2, \ldots, J\}$, for classifying a d-dimensional observation $\mathbf{x}$ into one of J competing classes. For instance, the optimal Bayes rule assigns an observation to the class $\mathsf{d}_B(\mathbf{x}) = j^*$ such that $j^* = \arg\max_j \pi_j f_j(\mathbf{x})$, where the π_j's are the prior probabilities and the $f_j(\mathbf{x})$'s are the probability density functions of the respective classes ($j = 1, 2, \ldots, J$). These probability density functions are usually unknown in practice and can be estimated from the training sample using some parametric or nonparametric methods. Kernel density estimation (see, e.g., Muller 1984; Silverman 1986; Scott 1992; Wand and Jones 1995) is a well-known method for constructing nonparametric estimates of population densities. If $\mathbf{x}_{j1}, \mathbf{x}_{j2}, \ldots, \mathbf{x}_{jn_j}$ are d-dimensional observations in the training sample from the jth population ($j = 1, 2, \ldots, J$), then the kernel estimate $\hat{f}_{jh_j}(\mathbf{x})$ of the jth population density is given by $\hat{f}_{jh_j}(\mathbf{x}) = n_j^{-1} h_j^{-d} \sum_{k=1}^{n_j} K\{h_j^{-1}(\mathbf{x}_{jk} - \mathbf{x})\}$, where the kernel function $K(\cdot)$ is a d-dimensional density function and $h_j > 0$ is a smoothing parameter commonly known as the bandwidth. These kernel density estimates are plugged into $\mathsf{d}_B(\mathbf{x})$ to form a nonparametric classification method called *kernel discriminant analysis* (see, e.g., Devijver and Kittler 1982; Hand 1982; Coomans and Broeckaert 1986; Hall and Wand 1988; Ripley 1996; Cooley and MacEachern 1998; Duda et al. 2000; Hastie et al. 2001). Using a single bandwidth parameter rather than separate bandwidths for each of the d dimensions requires some preliminary transformation of the data to make the variability approximately equal in each dimension, which can be done by standardizing the dataset using the sample dispersion matrix. The Gaussian kernel $K(\mathbf{t}) = (2\pi)^{-d/2} e^{-\mathbf{t}'\mathbf{t}/2}$ is a popular choice for the kernel function $K(\cdot)$, and we use it throughout this article.

Clearly, the performance of this nonparametric classifier depends critically on the values of bandwidth parameters. Many different techniques for choosing optimal bandwidths from the data are available in the literature (see, e.g., Hall 1983; Stone 1984; Silverman 1986; Hall, Sheather, Jones, and Marron 1991; Sheather and Jones 1991; Scott 1992; Wand and Jones 1995; Jones, Marron, and Sheather 1996). But, instead of minimizing the misclassification rate, most of these bandwidth selection methods target to minimize the mean integrated square error ($MISE = E[\int \{\hat{f}_{jh}(\mathbf{x}) - f_j(\mathbf{x})\}^2 \, d\mathbf{x}]$) of the class density estimate.

As a result, they may lead to rather poor misclassification rates for the resulting classifier. For discriminant analysis using kernel density estimates, Hall and Wand (1988) proposed a bandwidth selection rule by minimizing the MISE of the estimate of difference of class densities. It has been observed by Ghosh and Chaudhuri (2004) that sometimes the bandwidth minimizing misclassification rate might be much larger that the bandwidth minimizing MISE. It is well known that with increasing values of bandwidth, the bias of a kernel density estimate increases while its variance decreases. A detailed discussion of the effect of this bias and variance on misclassification rates was provided by Friedman (1997).

On the other hand, popular V-fold cross-validation (see, e.g., Stone 1977; Ripley 1996) and similar methods for selecting the smoothing parameter in a nonparametric classification problem may not guide one very well in choosing bandwidths in practice, because of the piecewise-constant nature of estimated misclassification probability functions with infinitely many minima. Further, all such cross-validation–based techniques require a huge computation when there are several competing classes. Three other important points to keep in mind in the case of discriminant analysis using kernel density estimates are the following:

1. The choice of bandwidths should depend on the specific observation to be classified as well as on the population densities.
2. Given a specific observation to be classified, one needs to assess the strength of the evidence in favor of one population or the other for varying choices of bandwidths for density estimates corresponding to different competing populations.
3. In a multiclass discrimination problem, instead of using a single bandwidth for each population density estimate, it is more meaningful to use different bandwidths for a class density estimate when comparing it with different competing class density estimates for classifying a specific observation.

In this article, for each population we consider a family of density estimates $\{\hat{f}_{jh_j} : h_j \in H_j\}$ over a wide range of bandwidths to carry out a multiscale version of kernel discriminant analysis. Over the last few years, multiscale methodology has emerged as a powerful exploratory and visualization tool for statistical data analysis. Minnotte and Scott (1993) and Minnotte, Marchette, and Wegman (1998) used multiscale techniques for finding modes in univariate and bivariate density estimation problems. Chaudhuri and Marron (1999, 2000) and Godtliebsen, Marron, and Chaudhuri (2002, 2004) used similar methods to find significant features in regression and density estimates. Simultaneous consideration of different levels of smoothing is expected to yield more useful information for classification than that obtained in an approach based on a single optimum bandwidth for each class density estimate. The results of multiscale analysis are presented using two-dimensional plots, which are specific to an observation to be classified, and there one can visually compare the strength of the evidence in favor of different competing classes over wide ranges of smoothing parameters. Statistical uncertainties at various locations in the plots are also quantified on the basis of appropriately estimated misclassification probabilities, and they too are displayed using some two-dimensional plots to facilitate the decision about classification. Of course, the final classification of an observation must be done by some judicious combination of all information obtained at different levels of smoothing, and we discuss some appropriate ways to do this. In the presence of more than two competing populations, we follow the same procedure, taking each pair of classes and then using the method of majority voting (see, e.g., Friedman 1996) or pairwise coupling (see, e.g., Hastie and Tibshirani 1998) to combine the results of these pairwise comparisons.

2. MULTISCALE VISUALIZATION OF DISCRIMINATION MEASURES

Suppose that $\mathbf{x}_{j1}, \mathbf{x}_{j2}, \ldots, \mathbf{x}_{jn_j}$ are training sample observations from the jth class, where $1 \le j \le J$. To classify an observation $\mathbf{x}$ into one of the J classes, we first need to obtain the density estimates $\hat{f}_{jh_j}(\mathbf{x})$ at the point $\mathbf{x}$ for all $j = 1, 2, \ldots, J$. As we pointed out in the previous section, before computing the class density estimate, we can standardize the data points in a class using an estimate of the class dispersion matrix to make the data more spherical in nature and thereby further justify the use of a common bandwidth h_j for all coordinate variables. The density estimate for the original data vectors can be obtained from that of the standardized data vectors using the simple transformation formula for a probability density function when the random vectors undergo a linear transformation. For a given pair of competing classes, say, class 1 and class 2, and a fixed pair of bandwidths h_1 and h_2 for the two class density estimates, there is an ordering between the functions $\pi_1 \hat{f}_{1h_1}(\mathbf{x})$ and $\pi_2 \hat{f}_{2h_2}(\mathbf{x})$ that determines which one of the two classes is more favorable. We now consider some measures for the strength of this evidence in favor of one class or the other.

2.1 Posterior Probability

In a two-class problem, for a given observation $\mathbf{x}$ and a given pair of bandwidths h_1 and h_2, a posterior probability estimate for the first population is given by

$$\mathcal{P}_{h_1,h_2}(1|\mathbf{x}) = \frac{\pi_1 \hat{f}_{1h_1}(\mathbf{x})}{\pi_1 \hat{f}_{1h_1}(\mathbf{x}) + \pi_2 \hat{f}_{2h_2}(\mathbf{x})}.$$

We can use a wide range of values for h_1 and h_2 to compute these estimated posteriors, and plot these using gray scale in a two-dimensional diagram, where 0 corresponds to black (i.e., the lowest possible posterior for class 1) and 1 corresponds to white (i.e., the highest possible posterior for class 1).

To demonstrate our methodology, we consider an example dataset from Ripley (1994) popularly known as the "synthetic data." This dataset is related to a two-class problem, where both classes are equal mixtures of two bivariate normal populations differing only in their location parameters. This dataset contains a training sample of size 250 (125 from each class) and a test sample of size 1,000 (500 from each class); it is available at *http://www.lib.stat.cmu.edu*. Scatterplots for the training and test samples of synthetic data are given in Figure 1, where the dots ($\cdot$) and the crosses ($\times$) represent the observations from the two classes.

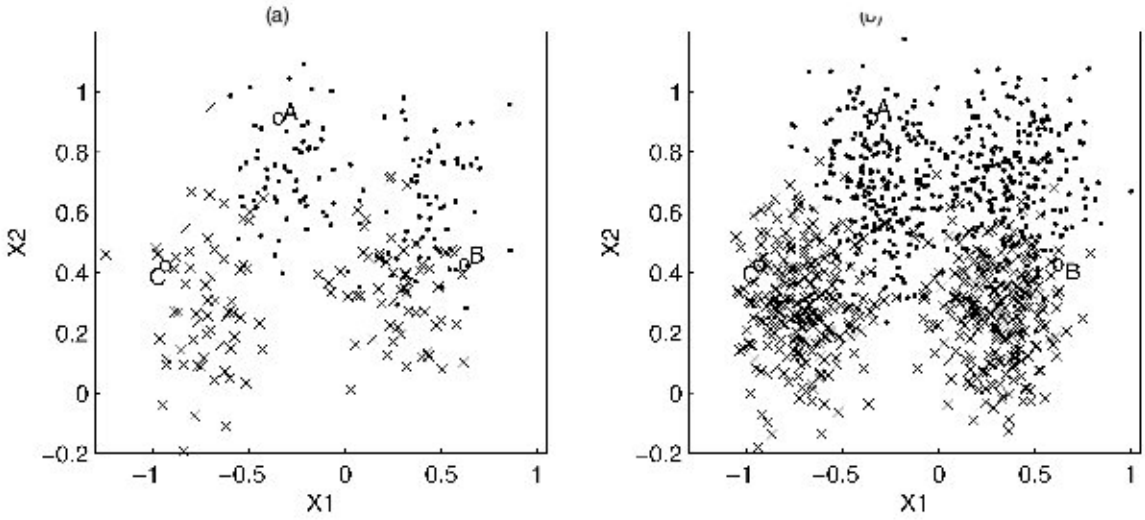

Figure 1. Scatterplots for Synthetic Data. (a) Training set: 250 observations. (b) Test set: 1,000 observations.

We have chosen three observations (indicated by ○ in Fig. 1) from the test set and labeled them as A, B, and C. These three points are purposely chosen from three different parts of the data. Observation A lies well within the cluster of observations from population 1, whereas C clearly belongs to population 2. The observation B is taken near the class boundary, where both populations have more or less equal strength. We performed usual linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) on this data to classify the entire test set observations using the training sample. Some observations were misclassified by both methods; B is one of these. Although B is originally from population 1, both LDA and QDA gave decisions in favor of population 2. Of course, both A and C were correctly classified by both the linear and the quadratic classifiers. We used a wide range of bandwidths for both populations to evaluate the posterior probabilities $\mathcal{P}_{h_1,h_2}(1|\mathbf{x})$ for different levels of smoothing. Because there are equal numbers of observations in these two classes, prior probabilities for our analysis are taken to be equal.

The top row of Figure 2 gives the gray-scale representation of posterior probabilities for the three cases, where the natural logarithms of the bandwidths of the first and second populations are plotted along the horizontal and the vertical axes. Here, white (high posterior) indicates the regions in favor of the first population, whereas black (low posterior) points toward the other population. The intensity of color varies with the magnitude of the posterior probabilities, which helps us determine the regions for strong evidence in favor of one of the two populations. As expected, we observe a dominance of light-colored regions in the case of observation A and dominance of dark regions in the case of observation C. However, for observation B, which lies near the class boundary, the evidence is not so clear in favor of either of the two populations.

2.2 A *p* Value-Type Discrimination Measure

In two-class kernel discriminant analysis, we classify an observation $\mathbf{x}$ into population 1 if $\pi_1\hat{f}_{1h_1}(\mathbf{x}) > \pi_2\hat{f}_{2h_2}(\mathbf{x})$. For a given observation $\mathbf{x}$, consider the probability

$$P_{h_1,h_2}(\mathbf{x}) = P\{\pi_1\hat{f}_{1h}(\mathbf{x}) > \pi_2\hat{f}_{2h}(\mathbf{x})|\mathbf{x}\}.$$

Clearly, high values of this probability give a decision in favor of population 1; low values, in favor of population 2. For fixed h_1 and h_2, because the density estimates are averages of iid random variables and density estimates for different populations are based on independent sets of observations, we can conveniently use normal approximation to evaluate the foregoing probability with a great degree of accuracy for even moderately large training sample sizes. Note that for a fixed value of h_j, this asymptotic normality follows from the standard central limit theorem for an iid sequence of random variables. One can also let $h_j \to 0$ as $n_j \to \infty$, but in that case one requires the condition $n_j h_j^d \to \infty$ as $n_j \to \infty$ for asymptotic normality of kernel density estimates (see, e.g., Lindebarg's condition for the central limit theorem for triangular arrays in Hall and Heyde 1980). Using such a normal approximation with estimated means and variances, we get

$$P_{h_1,h_2}(\mathbf{x}) \simeq \Phi\left(\frac{\{\pi_1 E[\hat{f}_{1h_1}(\mathbf{x})|\mathbf{x}] - \pi_2 E[\hat{f}_{2h_2}(\mathbf{x})|\mathbf{x}]\}}{\sqrt{\pi_1^2 \operatorname{var}[\hat{f}_{1h_1}(\mathbf{x})|\mathbf{x}] + \pi_2^2 \operatorname{var}[\hat{f}_{2h_2}(\mathbf{x})|\mathbf{x}]}}\right)$$

$$\simeq \Phi\left(\frac{\{\pi_1\hat{f}_{1h_1}(\mathbf{x}) - \pi_2\hat{f}_{2h_2}(\mathbf{x})\}}{\sqrt{\pi_1^2 s_{1h_1}^2(\mathbf{x})/n_1 + \pi_2^2 s_{2h_2}^2(\mathbf{x})/n_2}}\right),$$

where Φ is the standard normal distribution function, n_1 and n_2 are the training sample sizes for the two classes, and $s_{jh_j}^2(\mathbf{x})/n_j$ is the variance of $\hat{f}_{jh_j}(\mathbf{x})$ $(j = 1, 2)$, which can be estimated from the training sample using the sample variance of $h_j^{-d}K\{h_j^{-1} \times (\mathbf{x}_{j1} - \mathbf{x})\}, \ldots, h_j^{-d}K\{h_j^{-1}(\mathbf{x}_{jn_j} - \mathbf{x})\}$.

An alternative interesting interpretation of the foregoing normal approximation of $P_{h_1,h_2}(\mathbf{x})$ can be given as follows. For a given observation $\mathbf{x}$ and a pair of bandwidths h_1 and h_2, let us imagine a pair of hypotheses, $H_0 : \pi_1 E\{\hat{f}_{1h}(\mathbf{x})\} \geq \pi_2 E\{\hat{f}_{2h}(\mathbf{x})\}$ and $H_A : \pi_1 E\{\hat{f}_{1h}(\mathbf{x})\} < \pi_2 E\{\hat{f}_{2h}(\mathbf{x})\}$. If the training sample is used to test these hypotheses using kernel density estimates, which can be viewed as statistics like sample means used in two-sample problems, then the foregoing normal approximation can be taken as the one-sided p value associated with that testing problem. This is why we have chosen to call it a p value-type measure of the strength of discrimination.

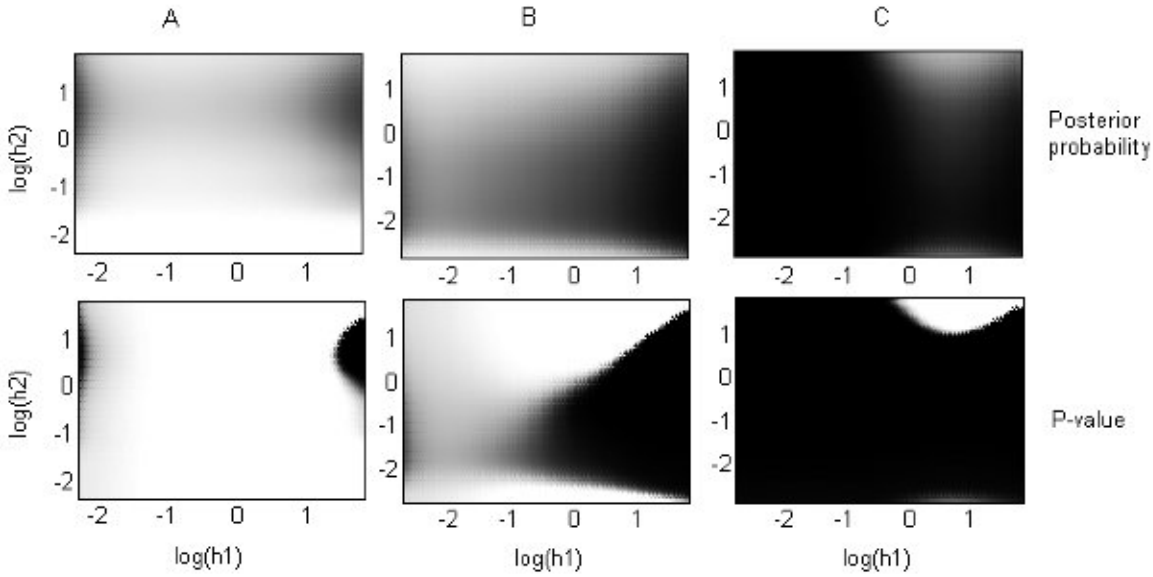

Figure 2. Multiscale Analysis of Synthetic Data.

The bottom row in Figure 2 shows these p values in two-dimensional plots for the observations A, B, and C using gray scales for various choices of h_1 and h_2. As before, the white region corresponding to high values of $P_{h_1,h_2}(\mathbf{x})$ favors population 1, whereas the black region corresponding to low values of $P_{h_1,h_2}(\mathbf{x})$ favors population 2. Once again, the plots give some idea for classifying observations A and C, but not for B. For observation B, the nearly equal spread of white and black regions gives an indication of nearly equal strength of evidence for each of the two populations depending on different choices of bandwidths.

One noteworthy feature of the plots in the two rows in Figure 2 is that the plots corresponding to p values at the bottom are much sharper than those corresponding to posterior probabilities at the top. Thus the plots in the bottom row provide easier visualization of the strength of evidence in favor of one of the two populations for different choices of bandwidths. The following theorem explains the reason for such a difference in sharpness for the two sets of plots.

Theorem 1. Suppose that for a given observation $\mathbf{x}$, $E[K^2\{h_j^{-1}(\mathbf{x}-\mathbf{x}_{j1})\}|\mathbf{x}] < \infty$ for $j = 1, 2$. Further, assume that $n_j/N \to \lambda_j$ $(j = 1, 2)$ as $N = n_1 + n_2 \to \infty$ $(0 < \lambda_1, \lambda_2 < 1)$. Then,

(a) $|\mathcal{P}_{h_1,h_2}(1|\mathbf{x}) - \frac{\pi_1 \mathcal{S}_{1h_1}(\mathbf{x})}{\pi_1 \mathcal{S}_{1h_1}(\mathbf{x}) + \pi_2 \mathcal{S}_{2h_2}(\mathbf{x})}| = O_P(N^{-1/2})$, where $\mathcal{S}_{jh_j}(\mathbf{x}) = E\{\hat{f}_{jh_j}(\mathbf{x})\}$ for $j = 1, 2$; and

(b) $|P_{h_1,h_2}(\mathbf{x}) - I\{\pi_1 \mathcal{S}_{1h_1}(\mathbf{x}) > \pi_2 \mathcal{S}_{2h_2}(\mathbf{x})\}| = O_P(N^{-1/2} \times e^{-CN})$ for some $C > 0$.

The main implication of this theorem is as follows. For any given $\mathbf{x}$ and a given pair of bandwidths (h_1, h_2), the estimated posterior probability $\mathcal{P}_{h_1,h_2}(1|\mathbf{x})$ converges to $\pi_1 \mathcal{S}_{1h_1}(\mathbf{x})/[\pi_1 \mathcal{S}_{1h_1}(\mathbf{x}) + \pi_2 \mathcal{S}_{2h_2}(\mathbf{x})]$ at a rate $O(N^{-1/2})$, but, depending on $\mathcal{S}_{1h_1}(\mathbf{x})$, $\mathcal{S}_{2h_2}(\mathbf{x})$, and the prior probabilities, the p value $P_{h_1,h_2}(\mathbf{x})$ converges to either 0 or to 1, also at an exponential rate. For instance, if $\pi_1 \mathcal{S}_{1h_1}(\mathbf{x}) < \pi_2 \mathcal{S}_{2h_2}(\mathbf{x})$, then $\mathcal{P}_{h_1,h_2}(1|\mathbf{x})$ has a $\sqrt{N}$ rate of convergence to a value $<.5$, but the corresponding p value-type measure converges to 0 at a much faster exponential rate. Therefore, for any given (h_1, h_2), as the training sample size grows, after some stage $P_{h_1,h_2}(\mathbf{x})$ will always give stronger evidence than $\mathcal{P}_{h_1,h_2}(1|\mathbf{x})$ for or against population 1.

In practice, the choice of bandwidth ranges in Figure 2 is an important issue. It can be shown (see Thm. 2 and its proof) that under fairly general conditions on population densities and with the use of Gaussian kernel, as the bandwidths tend to infinity, the posterior estimates derived from kernel density estimates tend to .5 near the line $h_2/h_1 = (\pi_2/\pi_1)^{1/d}$ [i.e., $\log(h_2) - \log(h_1) = \{\log(\pi_2) - \log(\pi_1)\}/d$ in the logarithmic scale] in the plots. On one of the two sides of this line, with increasing bandwidth, the posterior estimate for one population tends to be larger, and on the other side of the line, the posterior estimate for the other population tends to be larger. For $\pi_1 = \pi_2 = .5$, as in the case of Figure 2, this line is the diagonal line. Because this is true irrespective of the training sample and the specific observation to be classified, the plot will not carry any useful evidence for classification purposes in the region corresponding to very large values of the bandwidths. In the case of p value plots, which are sharper than the posterior plots with each pixel more white or more black than in the case of posterior plots, one would expect to see mostly black on one side of this line and mostly white on the other side of it for very large bandwidths. Of course, the computational cost will increase rapidly with the increasing range of bandwidths. Keeping all of these issues in mind, here we have adopted a rule of using an upper limit for the bandwidths that is about as large as the maximum pairwise distance of standardized data points in a population in the training set. For Figure 2, both upper limits for h [$\log(h)$ resp.] turn out to be about 6 (1.8 resp.). The lower bound of the bandwidths must be specified as well. If we use very small bandwidths, then there may not be sufficient observations in the effective regions around the data points. So including those small bandwidths may increase the computational cost while giving unreliable and possibly misleading information for classification. One can take a conservative approach by setting the minimum c[illegible]wise distances as this lower

limit. But in some cases this turns out to be 0. We have chosen one-third of the first percentile of the pairwise distances of standardized data points in a population as the lower limit of the bandwidth for that class. Using the factor one-third is motivated by the fact that here the kernel is a Gaussian kernel. However, if this first percentile is smaller than the distance between the specific data point and its nearest neighbor, then using the foregoing lower limit makes no sense. In such cases, one-third of the distance between that data point and its nearest neighbor is taken as the lower limit; see the discussion at the end of Section 3.3.

When using plots like those in Figure 2, one must keep in mind that the evidence at a point (h_1, h_2) in favor of or against a class needs to be properly supplemented by the reliability of the evidence as measured by the misclassification rate at that point (see Sec. 3). Hence, although the plots in Figure 2 are definitely useful as the first step for forming a visual evidence for the multiscale classification results, one cannot just use the visible sizes of white and black regions in the plots for making the final classification. Instead, one must carefully weigh the evidence at each point using appropriate weight functions, as described in the following section.

3. AGGREGATION OF CLASSIFICATION RESULTS

To arrive at the final classification for an observation, one must aggregate the results obtained at different levels of smoothing. A natural way to combine these results is to form some appropriate weighted average of the posterior probabilities computed for different choices of (h_1, h_2). Bagging (see, e.g., Breiman 1996), boosting (see, e.g., Schapire, Fruend, Bartlett, and Lee 1998; Friedman, Hastie, and Tibshirani 2000) and arcing classifier (see, e.g., Breiman 1998) are some of the well-known aggregation methods that adopt a similar procedure for combining the results of several classification techniques. They assign different weights to different classifiers based on their corresponding misclassification probabilities, and those weights are then used to build up the aggregated classification rule.

3.1 Misclassification Rates

For any fixed choice of (h_1, h_2), the average misclassification probability of a kernel classifier for a two-class problem is given by

$$\Delta(h_1, h_2) = \pi_1 \int_{\mathbf{x} \in \mathcal{R}^c_{h_1,h_2}} f_1(\mathbf{x})\, d\mathbf{x} + \pi_2 \int_{\mathbf{x} \in \mathcal{R}_{h_1,h_2}} f_2(\mathbf{x})\, d\mathbf{x},$$

where $\mathcal{R}_{h_1,h_2}$ is the set of all $\mathbf{x}$'s classified into class 1 by the classifier and $\mathcal{R}^c_{h_1,h_2}$ is the complementary set. Usual cross-validation techniques (see, e.g., Stone 1977) estimate this misclassification rate, $\Delta(h_1, h_2)$, by some kind of empirical proportion of misclassified cases, and as a result, they lead to estimates that are usually piecewise constant even when the true $\Delta(h_1, h_2)$ is a smooth function. This problem was discussed in detail by Ghosh and Chaudhuri (2004). For varying choices of bandwidths, these authors proposed a smooth and more accurate estimate of the misclassification probability for [illegible]ased on kernel density estimates. Their estimates use normal approximation to the distribution of a kernel density estimate, which is an average of iid random variables. In this article we use the method of Ghosh and Chaudhuri (2004) to estimate $\Delta(h_1, h_2)$, and consider the plot of the corresponding probability of correct classification $[= 1 - \widehat{\Delta}(h_1, h_2)]$ in a two-dimensional figure using their gray-scale values. Figure 3 shows such a plot for the "synthetic data" discussed in the preceding section. Here white represents high probability of correct classification, and black represents the opposite. The bandwidth pair that minimizes the MISE of the density estimates (marked by $\circ$), and the bandwidth pair that minimizes $\widehat{\Delta}(h_1, h_2)$ (marked by $*$) are also indicated in the figure. One of the striking features of the plot is the existence of a wide range of bandwidth pairs with very low misclassification rates, that have performance comparable to the bandwidth pair marked by $*$. Note that as $\log(h_1)$ and $\log(h_2)$ approach the upper limit, the plot begins to get darker, indicating less credibility of classification beyond this limit (see discussion at the end of Sec. 2). This plot gives a useful visualization of statistical uncertainties in classification obtained by using the kernel density estimates for different levels of smoothing, and also demonstrates the importance of looking at a wide range of bandwidth pairs instead of some single optimal pair.

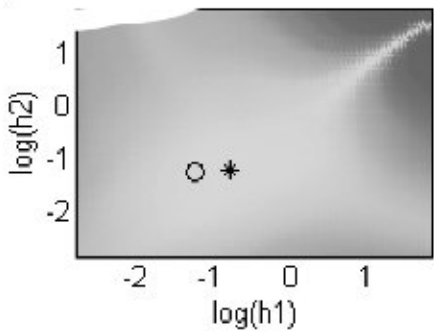

Figure 3. Plot for Probability of Correct Classification (synthetic data). The optimal MISE bandwidth pair ($\circ$) and the optimal bandwidth pair ($*$) for misclassification rate are marked.

3.2 Weight Function Derived From Misclassification Rates

Light colors in various regions in Figure 3 clearly suggest that these regions are most reliable for classification, and that the weight function $w(h_1, h_2)$ should take higher values there. We define $\widehat{\Delta}_\circ = \min_{h_1,h_2} \widehat{\Delta}(h_1, h_2)$ and consider $w(h_1, h_2)$ to be a decreasing function of $\widehat{\Delta}(h_1, h_2)$ or, equivalently, of $\widehat{\Delta}(h_1, h_2) - \widehat{\Delta}_\circ$. Boosting (see, e.g., Friedman, Hastie, and Tibshirani 2000) uses the same idea for aggregation where $w = \log\{(1 - \Delta)/\Delta\}$ is taken as the weight function. Clearly, this weight function takes higher values for those classifiers that lead to lower misclassification rates, and it decreases gradually as the misclassification rate increases. Bagging (see, e.g., Breiman 1996) of course uses equal weights for all classifiers. A comparative empirical study of bagging (see, e.g., Breiman 1996), boosting, and other ensemble methods has been given by Opitz and Maclin (1999). Bagging and boosting methods use bootstrap (or weighted bootstrap) technique to generate different samples from the training data, and, based on

these different samples, different classifiers are developed. The results of these classification rules are aggregated using the weight functions. However, our method does not require any resampling technique for generating the classifiers; using different values of (h_1, h_2) leads to different classification rules, which are aggregated using some weight function. Bagging or boosting generally aggregates those base classifiers that have reasonably good misclassification rates. But for some values of (h_1, h_2), the kernel classifier may lead to very poor classification. One must appropriately downweight these classification rules. The log function used in boosting decreases with misclassification probability at a very slow rate. But if one chooses a Gaussian-type function, which decreases at a faster rate, then the poor classifiers would be downweighted appropriately. Further, $w(h_1, h_2)$ should vanish whenever the corresponding $\widehat{\Delta}(h_1, h_2)$ exceeds any of the two prior probabilities, because the performance of the classifier then turns out to be poorer than that of a trivial classifier, which classifies all observations into the class having the larger prior. Keeping these in view, in all our numerical work, we have used a Gaussian-type weight function,

$$w(h_1,h_2) = \begin{cases} \exp\left\{-\frac{1}{2}\frac{(\widehat{\Delta}(h_1,h_2)-\widehat{\Delta}_\circ)^2}{\widehat{\Delta}_\circ(1-\widehat{\Delta}_\circ)/N}\right\} \\ \quad \text{if } \dfrac{\widehat{\Delta}(h_1,h_2)-\widehat{\Delta}_\circ}{[\widehat{\Delta}_\circ(1-\widehat{\Delta}_\circ)/N]^{1/2}} \le \tau \text{ and} \\ \quad \widehat{\Delta}(h_1,h_2) < \min\{\pi_1,\pi_2\} \\ 0 \quad \text{otherwise.} \end{cases}$$

Here, for $N = n_1 + n_2$, $\widehat{\Delta}_\circ$ and $\widehat{\Delta}_\circ(1-\widehat{\Delta}_\circ)/N$ can be viewed as estimates for the mean and the variance of the empirical misclassification rate of the best classifier based on kernel density estimates when such a classifier is used to classify N independent observations. The constant τ determines the maximum amount of deviation from the minimal estimated misclassification rate in a standardized scale beyond which the weighting scheme ignores the bandwidth pair (h_1, h_2) by putting zero weight on them. Clearly, $\tau = 0$ corresponds to the situation of putting all of the weights only on the bandwidth pairs (h_1, h_2) for which $\widehat{\Delta}(h_1, h_2) = \widehat{\Delta}_\circ$. Note also that the choice of the foregoing Gaussian-type weight function implies that for practical purposes, there is no need to consider a value of τ larger than 3. This choice of the weight function is somewhat subjective, and one may use other suitable functions for the same purpose. However, it is our empirical experience that the final result is not very sensitive to the weighting procedure as long as any reasonable weight function (which decreases appropriately with misclassification rates) is used.

3.3 Superimposition of Weight Function Over Discrimination Measures

Superimposition of this weight function over the plots of discrimination measures provides a useful visual device for classification problems. In Section 2.1 we demonstrated the use of posterior probabilities and p values for visual comparison between the strengths of different classes. Figure 2 gave some rough idea about the final classification for observations A and C, and it could identify the borderline case (observation B) as well. But in higher dimension, the plot of these discrimination measures often fails to differentiate between the easier and the harder cases. Superimposed versions of discrimination measures become helpful in such situations.

Let us consider an example with two six-variate normal populations differing only in their location parameters. Suppose that the populations have mean vectors $\boldsymbol{\mu}_1 = (2, 0, \ldots, 0)$ and $\boldsymbol{\mu}_2 = (0, 0, \ldots, 0)$ and common dispersion matrix $\mathbf{I}_6$. We also consider the prior probabilities for the two classes to be equal $(\pi_1 = \pi_2 = .5)$ and generate equal numbers of observations $(n_1 = n_2 = n = 50)$ from these two classes to construct the training set. Next, consider an observation $\mathbf{x} = (x_1, 0, 0, 0, 0, 0)$. Clearly, $x_1 = 0$ and $x_1 = 2$ give the centers for population 2 and population 1, whereas $x_1 = 1$ represents a point on the class boundary. Therefore, one would expect to have three different behavior of the classification methodology at these three points. The plots of the discrimination measures for these three cases are given in Figure 4, where the upper and the lower limits of bandwidths are chosen using the same rule used in Figure 2. From this figure, it is clearly evident that both posterior probabilities and p values (top and middle rows of Fig. 4) fail to reflect the differences in strength of classification in these three cases. In these plots, although the white region extends as we move on from $x_1 = 0$ to $x_1 = 2$, still in all of the cases we have an almost equal split in favor of the classes indicated by white and black regions.

However, the difference in the classification result becomes evident if we look at the p values superimposed over the weight function (bottom row of Fig. 4). The weighted p value that has been plotted against h_1 and h_2 is

$$P^S_{h_1,h_2}(\mathbf{x}) = .5 + \{P_{h_1,h_2}(\mathbf{x}) - .5\}w^*(h_1, h_2),$$

where w^* is the rescaled version of the weight function that has minimum value 0 and maximum value 1. From the definition, it is quite clear that when the pair (h_1, h_2) has low weight, $P^S_{h_1,h_2}(\mathbf{x})$ is expected to be very close to .5, which is indicated by the gray regions in the plots. However, in more reliable regions (i.e., pairs with high weights), we get stronger evidence as $P_{h_1,h_2}(\mathbf{x})$ moves away (in either direction) from .5. When $x_1 = 0$ (or $x_1 = 2$), we observe a black (or white) color in this region, which gives a clear idea of the direction and strength of the decision. Evidence for classification is very strong in these cases. For $x_1 = 1$, we observe some white as well as some black color of almost equal intensity. Clearly, the evidence is poor in this case, and the plot gives a clear indication of a borderline case. Instead of p values, one may also consider the superimposed version of posterior probabilities for visualization, but we use the p values because of its sharpness.

In the plots of posterior probabilities and p values, sometimes (specially when the data point is in a sparse region) one may notice a white or a black streak near both axes (see Fig. 2). This is because, for the given sample sizes, using such a small bandwidth makes one density estimate very close to 0, and thus the competing class density estimate turns out to be the winner. However, these streaks appear in a region of the plot where we have a high misclassification probability. Consequently, the weight function becomes 0 in th[illegible]ion which makes the plots

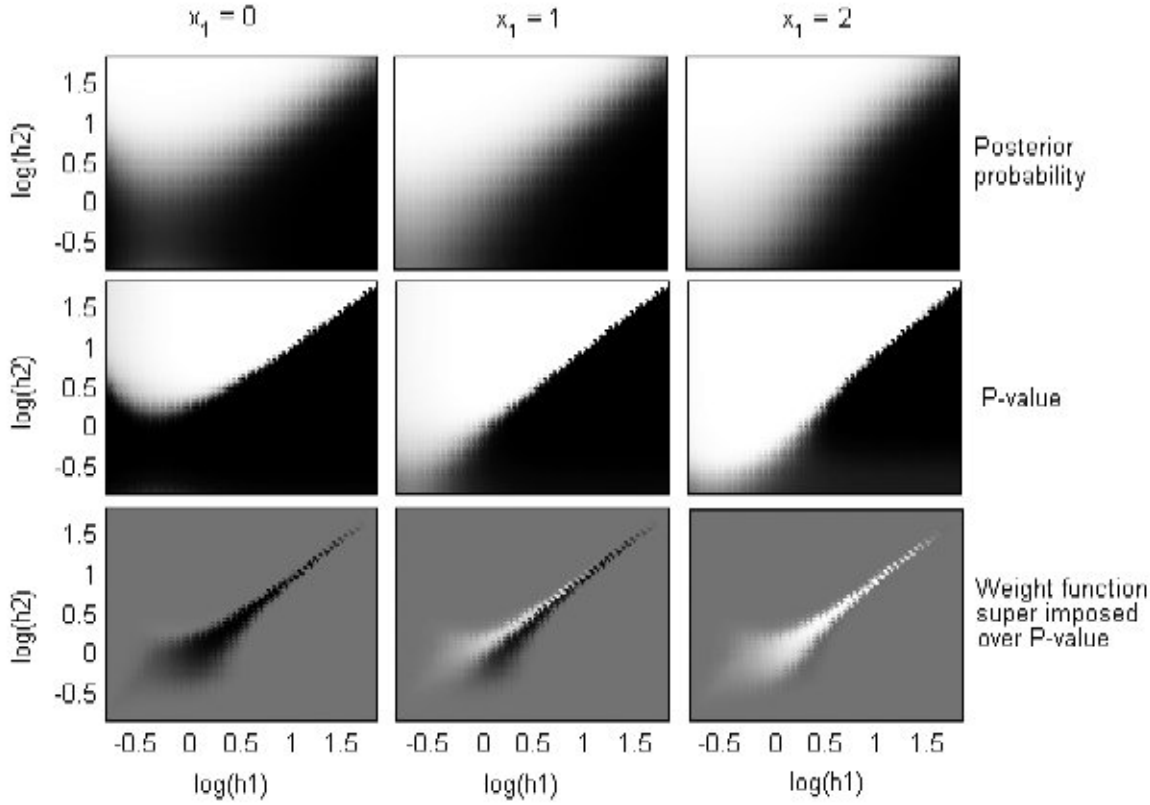

Figure 4. Multiscale Analysis of Simulated Data.

of $P^S_{h_1,h_2}$ free from such odd-looking streaks. The plots of $P^S_{h_1,h_2}$ also give us an idea of the effective range of bandwidths that one should look at for classifying a specific observation. Reconstructing the plots of discrimination measures using that range may help get rid of odd-looking streaks and provide better visualization.

3.4 Aggregation by Weighted Averaging

As it has been mentioned earlier, a natural way to combine the results of different classifiers is to use appropriate weighted averages of posterior probabilities. The weighted p value, $P^S_{h_1,h_2}$, defined in Section 3.3 can be used for this purpose, because it makes sense to rely more on those bandwidth pairs, which lead to stronger and more reliable evidence for one of the two classes. We choose the adjusted weight function

$$w_{\mathbf{x}}(h_1,h_2) = w^*(h_1,h_2)\left|P_{h_1,h_2}(\mathbf{x}) - .5\right| = \left|P^S_{h_1,h_2}(\mathbf{x}) - .5\right|$$

and use it to aggregate the posterior probabilities obtained by different classifiers. The resulting weighted posteriors, $\mathcal{P}^*(j|\mathbf{x})$ $(j = 1, 2)$, can be expressed as

$$\mathcal{P}^*(j|\mathbf{x}) = \sum_{h_1,h_2} w_{\mathbf{x}}(h_1,h_2)\mathcal{P}_{h_1,h_2}(j|\mathbf{x}) \Big/ \sum_{h_1,h_2} w_{\mathbf{x}}(h_1,h_2).$$

Note that the adjusted weights $w_{\mathbf{x}}(h_1,h_2)$ depend not only on the estimated overall misclassification probabilities, but also on the particular observation to be classified. This data-dependent adjustment of weight function provides more flexibility to the classification methodology.

In practice, for finding weighted posterior probabilities, one must fix the range of bandwidths as well. Our empirical experience suggests that in a two-class problem, if we keep on increasing the range of bandwidths, then, after a certain level, the classification results based on weighted averaging of posteriors remain unaltered in almost all cases. After standardizing a dataset, one can compute all pairwise distances between the standardized observations in a class and determine the αth quantile, λ_α $(0 < \alpha < 1)$, of these distances. One can use this quantile as the upper limit of the bandwidth for some large values of α, like $\alpha = .9$ or .95. In all of our numerical work for aggregation purposes, we use $\alpha = .95$ and denote the corresponding upper limits by $\lambda_{.95}$. Similarly, for setting the lower limits of bandwidths for aggregation purpose, we use $\lambda_{.05}/3$, where the factor $1/3$ is motivated by our use of Gaussian kernel. Note that for visualization purposes in Sections 2.1, 2.2, 3.1, and 3.3, we used a more conservative rule for setting the upper and lower limits for the bandwidths. Depending on the length of the interval, we use 60–100 equidistant values of h_j $(j = 1, 2)$ on its range, and then combine the results for different pairs of bandwidths to arrive at the final aggregated decision.

We conclude this section by considering once again the "synthetic data" for the purposes of illustration. When we used $\lambda_{.95} = (3.263, 3.258)$ as the upper limit of the bandwidths, for observations A and C (see Sec. 2.1), the weighted average of the posteriors (with $\tau = 3$) in favor of the first population turned out to be .873 and .189, which give a clear indication of the classes to which they belonged. Including large bandwidths in aggregation reduces the difference between the weighted posteriors, but generally does not change the classification result. For instance, if we used (10, 10) as the upper limit of the bandwidths,

then $\mathcal{P}_{h_1,h_2}(1|\mathbf{x})$ for A and C turned out to be .778 and .273. However, in the case of observation B, for both choices of range $\mathcal{P}_{h_1,h_2}(1|\mathbf{x})$ was found to be very close to .5 [.482 for $\lambda_{.95}$ and .489 for (10, 10)], as would be expected in view of the fact that this observation lies near the class boundary where both classes have almost equal strength. Note that these posterior estimates may not always be very accurate, and one may get better estimates using other classification methods. For instance, in the case of the synthetic data, where it is known that both the populations are equal mixtures of normal populations, one should expect to get better posterior estimates using mixture discriminant analysis (see, e.g., Hastie and Tibshirani 1996).

3.5 Classification Among More Than Two Populations

In the presence of more than two competing populations, it becomes computationally difficult to determine the optimum bandwidths by minimizing the estimate of overall average misclassification probability $\Delta(h_1, h_2, \ldots, h_J)$. In these situations we can decompose the multiclass problem into a number of binary classification problems, taking a pair of classes at a time and proceeding in the same way as before. The results of all of these pairwise classifications are combined together to come up with the final decision rule. The method of majority voting (see, e.g., Friedman 1996) is the simplest procedure for combining these results. In a J-class problem, after $\binom{J}{2}$ pairwise comparisons, this method classifies an observation to the class that has the maximum number of votes. But this voting method may sometimes lead to a region of indecision, in which more than one class can have the maximum number of votes. One can avoid this problem using alternative techniques like the method of pairwise coupling (see, e.g., Hastie and Tibshirani 1998), which combines the estimated posteriors for different pairwise classifications to determine the final posteriors for different competing classes.

4. EFFECT OF BANDWIDTHS ON MISCLASSIFICATION RATES: INADEQUACY OF MINIMUM MEAN INTEGRATED SQUARED ERROR BANDWIDTHS

As we mentioned earlier, the bandwidths that minimize MISE of the density estimates sometimes lead to poor performance in discriminant analysis. For example, consider the classification problem with two six-dimensional normal distributions as discussed in Section 3.3. Because the population distributions are themselves spherical, without any standardization one can use a single common bandwidth in all directions. Moreover, because of the similar dispersion structure of these two populations, it is quite reasonable to use the same bandwidth h for both of them. Therefore, in this case the average misclassification probability can be viewed as a function of a single bandwidth parameter, h.

In Figure 5 (taken from Ghosh and Chaudhuri 2004) we plot the true average misclassification probability for varying choices of h. This figure clearly shows the striking difference between the optimal bandwidth for usual density estimation (marked by $\circ$) and that for the classification problem (marked by $*$). The best possible bandwidth for the classification problem (h_*) leads to a significantly lower misclassification error

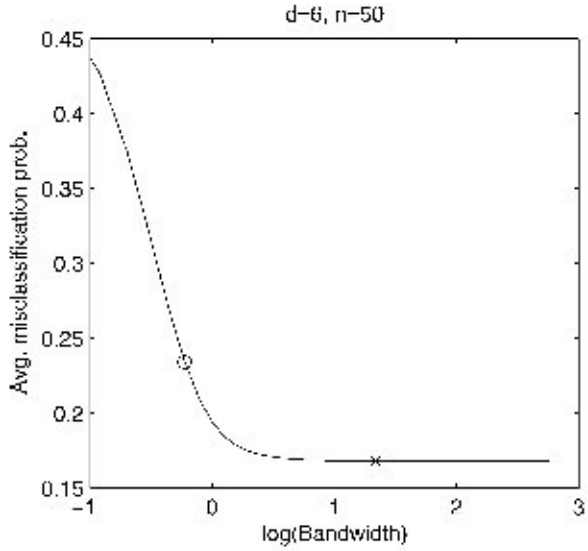

Figure 5. Average Misclassification Probability and Optimal Bandwidth.

rate than that obtained by using the bandwidth ($h_\circ$) that minimizes the MISE of the density estimates.

We also carried out a simulation study taking equal numbers of observations from these two classes. We generated a test set of size 1,000 (500 from each class) and classified them using 100 training set observations (50 from each class). The bandwidth pair that minimizes the estimated MISE of the density estimates ($\mathbf{h}_\circ$) led to a misclassification rate of 22.3%. In contrast, the kernel classifier with the bandwidth estimated by minimizing the leave-one-out misclassification rate ($\mathbf{h}_*$) (which relates to the weighted averaging method with $\tau = 0$) could reduce the misclassification rate to 18.6%. In this example, the optimal Bayes classifier based on true densities wrongly classified 16.2% of the test set observations. Similar to what we observed in Figure 5, in our simulated dataset $\mathbf{h}_* = (4.45, 4.30)$ was found to be much larger than $\mathbf{h}_\circ = (.75, .75)$. In density estimation problems, using large bandwidths generally leads to large bias and hence large MISE for the density estimates. Therefore, in density estimation, with increasing sample size, one usually shrinks the bandwidth to 0 to get good performance. But this is not necessarily the case for kernel discriminant analysis. Here, depending on competing population densities, using large bandwidths may also lead to lower misclassification rates in some special situations (see, e.g., Hand 1982; Scott 1992; Ghosh and Chaudhuri 2004). As observed by Scott (1992), a kernel discriminant function based on the Gaussian kernel tends to behave like the standard linear discriminant function as the bandwidth parameters tend to infinity. If the competing populations are location shifts of a spherically symmetric distribution, then this linear classifier coincides with the optimal Bayes classifier. The following theorem on misclassification rates provides some useful insights into the asymptotic behavior of misclassification rates as the bandwidth parameters tend to infinity.

Theorem 2. Suppose that f_1 and f_2 are such that $\int \|\mathbf{x}\|^6 \times f_j(\mathbf{x})\, d\mathbf{x} < \infty$ for $j = 1, 2$, and the kernel K is a d-dimensional density function with a mode at $\mathbf{0}$ and bounded third derivatives. Define a constant $C_\pi = \pi_2/\pi_1$ and assume that h_1 and h_2 vary

in such a way that $h_2/h_1 = C_h$, a constant. Now as $h_1 \to \infty$, $\Delta(h_1, h_2)$ has the following asymptotic behavior:

(a) When $C_\pi > C_h^d$, as $n_1, n_2 \to \infty$, $\Delta(h_1, h_2) \to \pi_1$.
(b) When $C_\pi < C_h^d$, as $n_1, n_2 \to \infty$, $\Delta(h_1, h_2) \to \pi_2$.
(c) When $C_\pi = C_h^d$, as $n_1, n_2 \to \infty$, $\Delta(h_1, h_2)$ tends to the misclassification probability of a quadratic classification rule given by

$$\mathsf{d}_Q(\mathbf{x}) = \begin{cases} 1 & \text{if } C_h^2 E_{f_1}\{(\mathbf{x}-\mathbf{X})'\nabla^2 K(\mathbf{0})(\mathbf{x}-\mathbf{X})\} \\ & \quad > E_{f_2}\{(\mathbf{x}-\mathbf{X})'\nabla^2 K(\mathbf{0})(\mathbf{x}-\mathbf{X})\} \\ 2 & \text{otherwise.} \end{cases}$$

When $C_\pi = C_h = 1$, the foregoing quadratic classifier actually turns out to be a linear classifier,

$$\mathsf{d}_L(\mathbf{x}) = \arg\min_j \left[\mathbf{x}'\nabla^2 K(\mathbf{0}) E_{f_j}(\mathbf{X}) - \frac{1}{2} E_{f_j}\{\mathbf{X}'\nabla^2 K(\mathbf{0})\mathbf{X}\} \right].$$

If the f_j's are spherically symmetric and satisfy a location shift model, and if the kernel function K is also spherical [note that $\nabla^2 K(\mathbf{0})$ is negative definite], then this linear classifier can be expressed in a further simplified form,

$$d_l(\mathbf{x}) = \arg\max_j \left\{ \mathbf{x}'\boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_j'\boldsymbol{\mu}_j \right\},$$

where $\boldsymbol{\mu}_j$ is the location parameter for the jth population ($j = 1, 2$). Note that the linear classifier described above is the optimal Bayes classifier under this setup. Therefore, in this particular case, using large bandwidth leads to a misclassification probability close to the optimal Bayes risk.

But using $\mathbf{h}_*$ does not necessarily lead to better estimates for the posterior probabilities. Figure 6 plots the estimated posterior probabilities for the simulated dataset against the true posteriors of different observations. When $\mathbf{h}_\circ$ is used for classification, the posteriors become more scattered [Fig. 6(c)], but this choice of bandwidth leads to very little bias for the posterior probability estimates. In contrast, for $\mathbf{h}_*$ [Fig. 6(a)], the scatter shrinks to the horizontal line at the center, indicating a reduction in variance of the estimates, but the bias of the posterior probability estimates increases considerably. Using large bandwidths reduces the variance of the kernel density estimate at the cost of increased bias to preserve the ordering of the true posteriors, as reflected in Figure 6. (A detailed discussion on the effect of such bias and variance on misclassification error rates was given in Friedman 1997.) Whereas $\mathbf{h}_\circ$ leads to a mean squared error of .046 for posterior estimates, h_* increases it to .112. In this case the method based on weighted averaging of posterior with $\tau = 3$ amounts to a compromise between the preceding two [Fig. 6(b)]. It improves the mean squared error (.060) of the posterior estimates significantly without sacrificing much accuracy in terms of misclassification rates (19.1%).

We have observed the inadequacy of $\mathbf{h}_\circ$ as bandwidth for kernel discriminant analysis in some real data as well. As an example, consider the diabetes data reported by Reaven and Miller (1979). This dataset consists of five measurement variables (fasting plasma glucose level, steady-state plasma glucose level, glucose area, insulin area, and relative weight) and three classes of individuals ("overt diabetic," "chemical diabetic," and "normal"). There are 145 individuals with 33, 36, and 76 in the three classes according to some clinical classification. For this dataset, if we use bandwidths that minimize the estimated MISE of population density estimates, we get a leave-one-out cross-validated misclassification rate of 12.41%. This error rate is higher than that obtained for simple LDA and QDA, which had leave-one-out misclassification rates of 11.03% and 9.66%. However, our multiscale analysis followed by the weighted averaging of posteriors led to leave-one-out cross-validated error rates of 5.52% for $\tau = 0$ and 6.21% for $\tau = 3$. Note that this is a three-class problem, and we have used the method of "majority voting" to combine the results of pairwise comparisons to arrive at the final classification. Fortunately, in this dataset, majority voting did not lead to any tied case for either $\tau = 0$ or $\tau = 3$.

5. CASE STUDIES USING BENCHMARK DATASETS

In this section we report our findings based on some benchmark datasets that illustrate the utility of the proposed method. Results of the kernel discriminant analysis based on bandwidths that minimize MISE and that based on the weighted averaging of posteriors (both with $\tau = 0$ and $\tau = 3$) are presented to compare their performance. For classification problems with more than two populations, we adopt the pairwise classification method and combine the results using majority voting (Friedman 1996), as well as pairwise coupling (Hastie and Tibshirani 1998). Misclassification error rates for usual

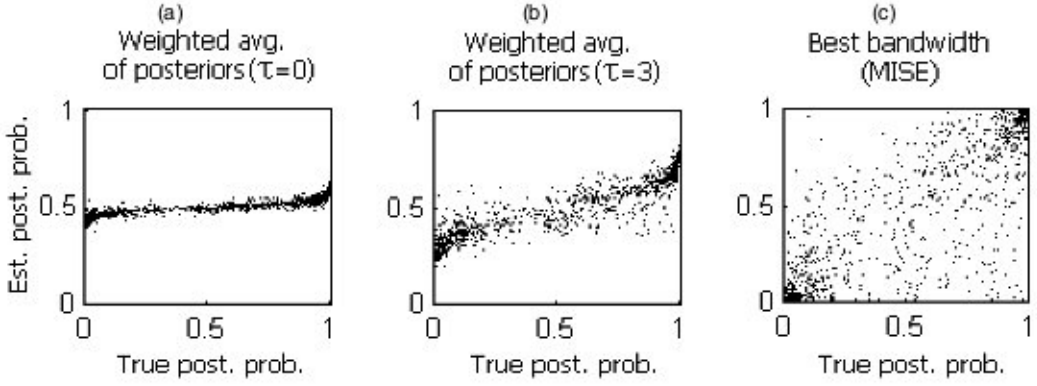

Figure 6. Estimated Posterior Probabilities for the Simulated Dataset.

LDA and QDA are also given to facilitate the comparison. As we have discussed earlier, in a few cases the voting method may end up with a tied situation. Here, all of those tied cases are considered "misclassification." Therefore, the reported results on voting are actually the proportion of misclassifications in the worst possible cases. The datasets that we consider here have been analyzed before in the literature, where nonparametric methods like classification trees (see, e.g., Breiman, Friedman, Olshen, and Stone 1984; Loh and Vanichsetakul 1988; Kim and Loh 2001), neural nets (see, e.g., Cheng and Titterington 1994; Ripley 1994, 1996) and flexible discriminant analysis (FDA) (see Hastie, Tibshirani, and Buja 1994) based on multivariate adaptive regression splines (MARS) (see Friedman 1991) were used to classify the observations. We have quoted those results directly from the available literature. Throughout these experiments, sample proportions for different classes are used as their prior probabilities. Apart from the vowel recognition data, all of the datasets considered in this section are available at *http://www.lib.stat.cmu.edu*.

Synthetic Data. Description of this dataset has already been given in Section 2.1. Ripley (1994) used these data to compare the performance of different classification algorithms. The class distributions were chosen to have a Bayes risk of 8.0%. In this dataset, LDA and QDA could achieve test set error rates of 10.8% and 10.2%. Classification tree (CART) also misclassified more than 10% observations (see Table 1). The performances of other nonparametric methods were fairly similar. Weighted averaging of the posterior achieved the best error rate when $\tau = 0$ was used.

Vowel Recognition Data. This dataset was created by Peterson and Barney (1952) by a spectrographic analysis of vowels in words formed by an "h" followed by a vowel and then followed by a "d." There were 67 persons who spoke different words, and the two lowest-resonant frequencies of a speaker's vocal track were noted for 10 different vowels. The observations were then randomly divided into a training set consisting of 338 observations and a test set consisting of 333 observations. Here the classes have significant overlaps between them, which makes the dataset a challenging one for any classification method. A scatterplot of this dataset is given in Figure 7, where the numbers represent the labels of the different classes (0 represents the 10th class).

This dataset has been extensively analyzed by many authors (see, e.g., Lee and Lippman 1989; Bose 1996; Cooley and MacEachern 1998). Bose reported a test set error rate of 18.6% for neural network methods when 20 hidden nodes were used, the lowest error rate reported for such methods. Error rates for LDA and CART were much higher than those for the other classifiers. For this dataset, the best test set misclassification rate reported by earlier authors is 17.4%, which was achieved by the k nearest-neighbor algorithm (see Lee and Lippman 1989). In this dataset, the method based on weighted averaging of posteriors with $\tau = 3$ together with the majority voting rule led to an error rate of 17.7% and had a clear edge over most of the other classifiers.

When pairwise coupling instead of majority voting was used for final classification, we obtained an error rate of 24.6% for weighted averaging of the posteriors with $\tau = 0$. We suspect that the performance of the pairwise coupling method in the presence of a large number of overlapping populations turns out to be bad because the optimal bandwidth minimizing the misclassification rate does not always lead to good estimates of posterior probabilities—as we have seen before. The posterior estimates may become better when $\tau = 3$ is used instead of $\tau = 0$. Perhaps this is the reason for improved performance of the classifier leading to an error rate of 21.3% when we used weighted averaging of posteriors with $\tau = 3$ together with the pairwise coupling method.

Sonar Data. This dataset, used by Gorman and Sejnowski (1988), contains 111 patterns obtained by bouncing sonar signals off a metal cylinder and 97 patterns obtained from rocks at various angles and under various conditions. The transmitted sonar signal is a frequency-modulated chirp, rising in

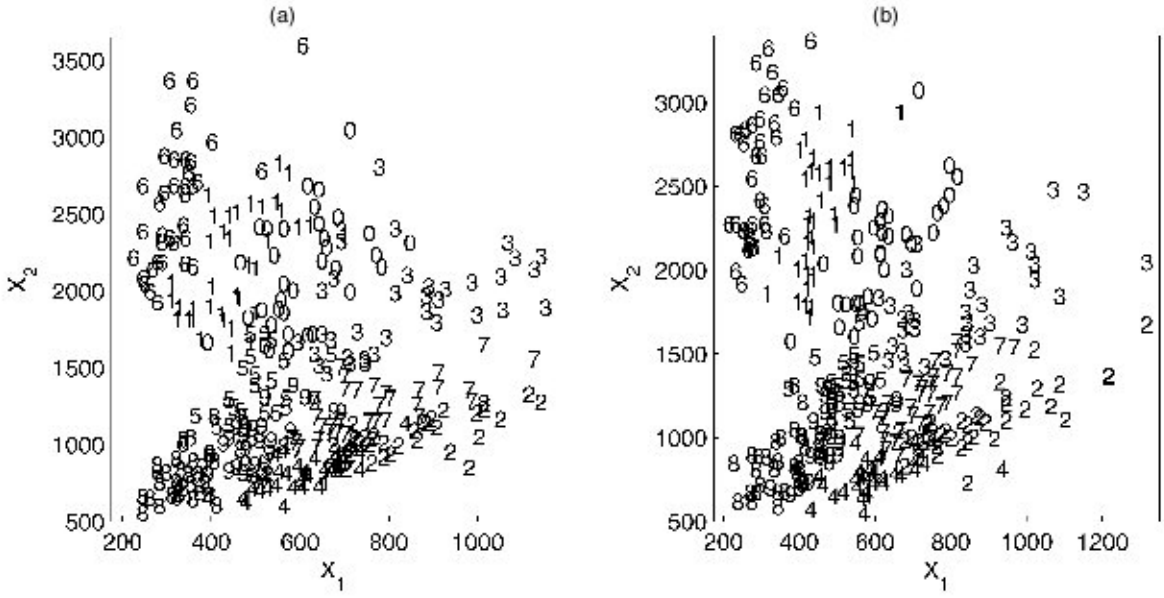

Figure 7. Scatterplots for Vowel Recognition Data. (a) Training set: 338 observations. (b) Test set: 333 observations

Datasets	LDA	QDA	FDA–MARS Degree 1	FDA–MARS Degree 2	CART	Neural networks	Kernel (MISE)	Kernel (weighted average) $\tau = 0$	Kernel (weighted average) $\tau = 3$
Synthetic	10.8	10.2	9.3	9.6	10.1	9.4	9.3	9.0	9.1
Vowel°	25.2	19.8	20.7	19.8	23.7	18.6	18.9	18.9	17.7
Sonar	20.2	15.4	22.1	19.2	20.2	19.2	17.3	16.3	13.5

°Majority voting is used for final classification.

frequency. Signals were obtained from a variety of different aspect angles, spanning 90 degrees for the cylinder and 180 degrees for the rock. Each observation is a set of 60 numbers in the range 0 to 1.0, each of which represents the energy within a particular frequency band, integrated over a certain time period. To reduce coordinatewise dependence, the data were averaged in bands of 3, making the number of measurement variables 20. The dataset was split into training and test sets each of size 104 using a cluster analysis method to ensure even matching.

Results for different classification methods on this dataset were given by Ripley (1994) and Cooley and MacEachern (1998). QDA performed quite well in this dataset compared with other classification methods like LDA, FDA–MARS, CART, and neural nets (see Table 1). The kernel method with $\tau = 3$ led to even better performance.

5.1 Forensic Glass Data: A Challenging Problem for Kernel Discriminant Analysis

This dataset contains information on refractive index and eight other components (weight percentage of oxides of Na, Mg, Al, Si, K, Ca, Ba, and Fe) for each of the six different types of glasses. There are 214 observations in the dataset, but most of them are window float (70) and window nonfloat (76) glass. The rest of the classes, namely vehicle glass (17), containers (13), tableware (9), and vehicle headlamp (29), contain much smaller number of observations, making this a difficult high-dimensional classification problem.

Ripley (1996) extensively analyzed this dataset and reported cross-validated error rates for different classifiers. The best result was reported for the k nearest-neighbor (see, e.g., Cover and Hart 1968; Duda et al. 2000) method with $k = 1$ when the measurement variables were suitably rescaled. This rescaled nearest-neighbor algorithm had an error rate of 23.6%. The misclassification error rate for the usual nearest-neighbor method was found to be 26.6% for $k = 1$. Neural networks with four to eight hidden nodes had been reported to have error rates between 24.8% and 29.9%. LDA in this dataset led to a cross-validated error rate of 37.9%; QDA was even worse, with an error rate of 40.2%. CART had error rates ranging from 31% to 42% for different types of pruning. FDA–MARS (with degree 1) could achieve an error rate of 32.2%, which was reduced to 29% when interactions were taken into consideration. Logistic discriminant analysis (see, e.g., Ripley 1996; Hastie et al. 2001) and projection pursuit (see, e.g., Huber 1985) had higher error rates (36% and 35.5%) than the other nonparametric classifiers.

Because four of the nine measurement variables (oxides of Mg, K, Ba, and Fe) had a significant number of 0's among their observed values, we decided to carry out our analysis with the remaining five variables. But even after using this subset of measurement variables, we could achieve a competitive performance for classifiers based on kernel density estimates. Using the bandwidths, which minimize the MISE of the density estimates, for classification led to fairly good performance. The leave-one-out estimate for the misclassification error was found to be 31.3%. We obtained even better performance using the method of weighted averaging of posterior. The error rates for $\tau = 0$ and $\tau = 3$ were found to be 29.9% and 28.5%, when majority voting was used. When pairwise coupling method was applied to this dataset after weighted averaging of posteriors, the aforementioned error rates increased to 31.3% and 36.4%.

5.2 Effect of Bandwidth Ranges on Misclassification Rates

In Section 3.4 we proposed a working rule for choosing the bandwidth ranges for aggregation purposes. For a given dataset, $\lambda_{.05}/3$ and $\lambda_{.95}$ are taken as the lower and upper limits of the bandwidths. Our empirical experience suggests that for moderately large sample sizes, if we use bandwidths smaller than $\lambda_{.05}/3$, then the classification result generally remains the same. For almost all observations, either because of the high variance of the density estimates or because of the poor misclassification rates of the classifier, the adjusted weight function $w_{\mathbf{x}}(h_1, h_2)$ becomes virtually 0 in this region. As a consequence, these regions usually have no effect on the weighted posterior probabilities of different classes.

In contrast, as we discussed in Section 3.4, increasing the upper limit of the bandwidths generally decreases the difference between the weighted posteriors without disturbing their orderings. Therefore, the misclassification rate of the weighted aggregation method remains more or less unaffected. As shown in Table 2, the misclassification rates for different choices of the upper limit of the bandwidths are almost equal for all of the benchmark datasets analyzed here.

Table 2. Percentage of Misclassifications for Different Choices of Upper Limits for Bandwidths

Upper limit		$\lambda_{.95}/3$	$\lambda_{.95}/2$	$\lambda_{.95}$	$2\lambda_{.95}$	$3\lambda_{.95}$
Synthetic	$\tau = 0$	9.0	9.0	9.0	9.0	9.0
	$\tau = 3$	9.1	9.1	9.1	9.1	9.3
Sonar	$\tau = 0$	15.4	16.3	16.3	16.3	15.4
	$\tau = 3$	12.5	13.5	13.5	13.5	14.4
Vowel°	$\tau = 0$	18.6	18.0	18.9	18.6	19.2
	$\tau = 3$	19.2	18.9	17.7	17.4	17.7
Glass°,+	$\tau = 0$	28.0	30.8	29.9	29.4	30.4
	$\tau = 3$	28.5	29.9	28.5	30.4	29.4

°Majority voting is used for final classification.
+Numbers represent leave-one-out error rates.

ACKNOWLEDGMENTS

The authors thank an associate editor and two referees for their careful reading of earlier versions of the manuscript. Their constructive criticisms and valuable suggestions led to a substantial improvement of the article.

APPENDIX: PROOFS

Proof of Theorem 1

(a) To make the expressions notationally simpler, let us define $T_j = \pi_j \hat{f}_{jh_j}(\mathbf{x})$ for $j = 1, 2$. Now, because T_j is an average of iid random variables, from the central limit theorem, it follows that under the assumed moment condition, for large sample sizes, T_j tends to be normally distributed with mean $\tau_j = \pi_j S_{jh_j}(\mathbf{x})$ and variance $\upsilon_j = \pi_j^2 s_{jh_j}^2(\mathbf{x})/n_j$.

Now define a function $\psi(T_1, T_2) = T_1/(T_1 + T_2)$. Here T_1 and T_2 are both positive-valued random variables and are independent. Moreover, the function ψ is continuously differentiable in T_1 and T_2. Therefore, the usual asymptotic Taylor expansion leads to

$$\frac{\{\psi(T_1, T_2) - \psi(\tau_1, \tau_2)\}}{\mathcal{V}} \xrightarrow{L} \text{Normal}(0, 1), \quad \text{where}$$

$$\mathcal{V} = \left\{ \sum_{j=1}^{2} \upsilon_j \left(\frac{\partial \psi}{\partial T_j} \right)^2_{T_1=\tau_1, T_2=\tau_2} \right\}^{1/2}.$$

Because $n_j/N \to \lambda_j > 0$ as $N \to \infty$ $(j = 1, 2)$, we have $|\psi(T_1, T_2) - \psi(\tau_1, \tau_2)| = O_P(N^{-1/2})$.

(b) Without loss of generality, let us assume that $\tau_1 > \tau_2$, that is, $I\{\pi_1 S_{1h_1}(\mathbf{x}) > \pi_2 S_{2h_2}(\mathbf{x})\} = 1$. Now, for some fixed h_1, h_2, and $\mathbf{x}$, from part (a), it follows that

$$\frac{1}{\sqrt{v_1 + v_2}}[(T_1 - T_2) - (\tau_1 - \tau_2)] \xrightarrow{L} \text{Normal}(0, 1) \quad \text{as } N \to \infty.$$

Now define $Z_{h_1,h_2}(\mathbf{x}) = \frac{1}{\sqrt{v_1+v_2}}(T_1 - T_2) = \sqrt{N}(T_1 - T_2)/V$, where $V = \{\pi_1^2 s_{1h_1}^2(\mathbf{x})/\lambda_1 + \pi_2^2 s_{2h_2}^2(\mathbf{x})/\lambda_2\}^{1/2}$. Therefore, $Z_{h_1,h_2}(\mathbf{x}) = O_p(N^{1/2})$ and $\frac{1}{\sqrt{N}} Z_{h_1,h_2}(\mathbf{x}) \xrightarrow{P} (\tau_1 - \tau_2)/V = C$ (say). For $x > 0$, using the fact that $\frac{1}{x}\phi(x) < 1 - \Phi(x) < (\frac{1}{x} - \frac{1}{x^3})\phi(x)$ [where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the pdf and the cdf of a standard normal distribution], we get $1 - P_{h_1,h_2}(\mathbf{x}) = 1 - \Phi(Z_{h_1,h_2}(\mathbf{x})) = O_p(N^{-1/2} e^{-CN})$.

Proof of Theorem 2

First, note that

$$\Delta(h_1, h_2) = \pi_1 E_{f_1}\left\{ I\left(\pi_1 \hat{f}_{1h_1} < \pi_2 \hat{f}_{2h_2}\right)\right\} + \pi_2 E_{f_2}\left\{ I\left(\pi_1 \hat{f}_{1h_1} > \pi_2 \hat{f}_{2h_2}\right)\right\}.$$

From the definition of $\hat{f}_{jh_j}(\mathbf{x})$ $(j = 1, 2)$, it is easy to see that

$$E_{f_j}\left\{\hat{f}_{h_j}(\mathbf{x})\right\} = h_j^{-d} E_{f_j}\left[K\left\{\frac{\mathbf{x} - \mathbf{X}}{h_j}\right\}\right] \quad \text{and}$$

$$\text{var}_{f_j}\left\{\hat{f}_{h_j}(\mathbf{x})\right\} = n_j^{-1} h_j^{-2d} \text{var}_{f_j}\left[K\left\{\frac{\mathbf{x} - \mathbf{X}}{h_j}\right\}\right].$$

Using a Taylor expansion about $\mathbf{0}$, $K\{(\mathbf{x} - \mathbf{X})/h_j\}$ can be expressed as

$$K\left\{\frac{\mathbf{x} - \mathbf{X}}{h_j}\right\} = K(\mathbf{0}) + \frac{1}{2h_j^2}\{(\mathbf{x} - \mathbf{X})' \nabla^2 K(\mathbf{0})(\mathbf{x} - \mathbf{X})\} + \frac{1}{6h_j^3} \sum_{i,k,l} Y_{i,k,l} \quad [\text{because } \nabla K(\mathbf{0}) = 0],$$

where $Y_{i,k,l} = (x_i - X_i)(x_k - X_k)(x_l - X_l) \frac{\partial^3 K(\mathbf{t})}{\partial t_i \partial t_k \partial t_l}|_{\mathbf{t}=\boldsymbol{\xi}}$ for some intermediate vector $\boldsymbol{\xi}$ between $\mathbf{0}$ and $(\mathbf{x} - \mathbf{X})/h_j$. Therefore,

$$E_{f_j}\left\{\hat{f}_{jh_j}(\mathbf{x})\right\} = h_j^{-d}\left[K(\mathbf{0}) + \frac{1}{2h_j^2} E_{f_j}\{(\mathbf{x} - \mathbf{X})' \nabla^2 K(\mathbf{0})(\mathbf{x} - \mathbf{X})\} + O(h_j^{-3})\right]$$

and

$$\text{var}_{f_j}\left\{\hat{f}_{jh_j}(\mathbf{x})\right\} = (4n_j h_j^{2d+4})^{-1}\left[\text{var}_{f_j}\{(\mathbf{x} - \mathbf{X})' \nabla^2 K(\mathbf{0})(\mathbf{x} - \mathbf{X})\} + O(h_j^{-1})\right],$$

using the facts that K has bounded third derivatives and $\int \|\mathbf{x}\|^6 f_j(\mathbf{x})\, d\mathbf{x} < \infty$.

As the variance of a kernel density estimates asymptotically converges to 0, for any given observation $\mathbf{x}$ and a given pair of bandwidths (h_1, h_2), the corresponding classifier classifies $\mathbf{x}$ to class 1 if and only if

$$\begin{aligned}
&\pi_1 E_{f_1}\left\{\hat{f}_{1h_1}(\mathbf{x})\right\} > \pi_2 E_{f_2}\left\{\hat{f}_{2h_2}(\mathbf{x})\right\} \\
\Leftrightarrow \quad &\pi_1 h_1^{-d}\left[K(\mathbf{0}) + \frac{1}{2h_1^2} E_{f_1}\{(\mathbf{x} - \mathbf{X})' \nabla^2 K(\mathbf{0})(\mathbf{x} - \mathbf{X})\} + O(h_1^{-3})\right] \\
&\quad > \pi_2 h_2^{-d}\left[K(\mathbf{0}) + \frac{1}{2h_2^2} E_{f_2}\{(\mathbf{x} - \mathbf{X})' \nabla^2 K(\mathbf{0})(\mathbf{x} - \mathbf{X})\} + O(h_2^{-3})\right] \\
\Leftrightarrow \quad &C_\pi C_h^{-d}\left[K(\mathbf{0}) + \frac{1}{2h_1^2} E_{f_1}\{(\mathbf{x} - \mathbf{X})' \nabla^2 K(\mathbf{0})(\mathbf{x} - \mathbf{X})\} + O(h_1^{-3})\right] \\
&\quad > \left[K(\mathbf{0}) + \frac{1}{2h_2^2} E_{f_2}\{(\mathbf{x} - \mathbf{X})' \nabla^2 K(\mathbf{0})(\mathbf{x} - \mathbf{X})\} + O(h_2^{-3})\right].
\end{aligned}$$

(a) When $C_\pi < C_h{}^d$, for large h_1 and $h_2 = C_h h_1$, the foregoing inequality holds whatever the observation $\mathbf{x}$. Consequently, the resulting classifier asymptotically classifies all observations to class 1.

(b) Similarly, when $C_\pi > C_h{}^d$, for every $\mathbf{x}$, the resulting classifier asymptotically always classifies it to class 2.

(c) When $C_\pi = C_h{}^d$, for large values of h_1 and h_2, it is easy to check that the foregoing inequality holds if and only if $C_h{}^2 E_{f_1}\{(\mathbf{x}-\mathbf{X})'\nabla^2 K(\mathbf{0})(\mathbf{x}-\mathbf{X})\} > E_{f_2}\{(\mathbf{x}-\mathbf{X})'\nabla^2 K(\mathbf{0})(\mathbf{x}-\mathbf{X})\}$. This completes the proof.

[Received July 2003. Revised March 2005.]

REFERENCES

Bose, S. (1996), "Classification Using Splines," *Computational Statistics and Data Analysis*, 22, 505–525.

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123–140.

——— (1998), "Arcing Classifiers" (with discussion), *The Annals of Statistics*, 26, 801–849.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Monterrey, CA: Wadsworth & Brooks.

Chaudhuri, P., and Marron, J. S. (1999), "SiZer for Exploration of Structures in Curves," *Journal of the American Statistical Association*, 94, 807–823.

——— (2000), "Scale Space View of Curve Estimation," *The Annals of Statistics*, 28, 408–428.

Cheng, B., and Titterington, D. M. (1994), "Neural Networks: A Review From a Statistical Perspective" (with discussion), *Statistical Science*, 9, 2–54.

Cooley, C. A., and MacEachern, S. N. (1998), "Classification via Kernel Product Estimators," *Biometrika*, 85, 823–833.

Coomans, D., and Broeckaert, I. (1986), *Potential Pattern Recognition in Chemical and Medical Decision Making*, Letchworth, U.K.: Research Studies Press.

Cover, T. M., and Hart, P. E. (1968), "Nearest-Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, 13, 21–27.

Devijer, P. A., and Kittler, J. (1982), *Pattern Recognition: A Statistical Approach*, London: Prentice-Hall.

Duda, R., Hart, P., and Stork, D. G. (2000), *Pattern Classification*, New York: Wiley.

Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines" (with discussion), *The Annals of Statistics*, 19, 1–141.

——— (1996), "Another Approach to Polychotomous Classification," technical report, Stanford University, Dept. of Statistics.

——— (1997), "On Bias, Variance, 0–1 Loss, and the Curse of Dimensionality," *Data Mining and Knowledge Discovery*, 1, 55–77.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2000), "Additive Logistic Regression: A Statistical View of Boosting" (with discussion), *The Annals of Statistics*, 28, 337–407.

Ghosh, A. K., and Chaudhuri, P. (2004), "Optimal Smoothing in Kernel Discriminant Analysis," *Statistica Sinica*, 14, 457–483.

Godtliebsen, F., Marron, J. S., and Chaudhuri, P. (2002), "Significance in Scale Space for Bivariate Density Estimation," *Journal of Computational and Graphical Statistics*, 11, 1–22.

——— (2004), "Statistical Significance of Features in Digital Images," *Image Vision and Computing*, 22, 1093–1104.

Gorman, R. P., and Sejnowski, T. J. (1988), "Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets," *Neural Networks*, 1, 75–89.

Hall, P. (1983), "Large-Sample Optimality of Least Squares Cross-Validations in Density Estimation," *The Annals of Statistics*, 11, 1156–1174.

Hall, P., and Heyde, C. C. (1980), *Martingale Limit Theory and Its Application*, New York: Academic Press.

Hall, P., Sheather, S. J., Jones, M. C., and Marron, J. S. (1991), "On Optimal Data-Based Bandwidth Selection in Kernel Density Estimation," *Biometrika*, 78, 263–270.

Hall, P., and Wand, M. P. (1988), "On Nonparametric Discrimination Using Density Differences," *Biometrika*, 75, 541–547.

Hand, D. J. (1982), *Kernel Discriminant Analysis*, Chichester, U.K.: Wiley.

Hastie, T., and Tibshirani, R. (1996), "Discriminant Analysis Using Gaussian Mixtures," *Journal of the Royal Statistical Society*, Ser. B, 58, 155–176.

——— (1998), "Classification by Pairwise Coupling," *The Annals of Statistics*, 26, 451–471.

Hastie, T., Tibshirani, R., and Buja, A. (1994), "Flexible Discriminant Analysis," *Journal of the American Statistical Association*, 89, 1255–1270.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, New York: Springer-Verlag.

Huber, P. J. (1985), "Projection Pursuit" (with discussion), *The Annals of Statistics*, 13, 435–475.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996), "A Brief Summary of Bandwidth Selection for Density Estimation," *Journal of the American Statistical Association*, 91, 401–407.

Kim, H., and Loh, W.-Y. (2001), "Classification Trees With Unbiased Multiway Splits," *Journal of the American Statistical Association*, 96, 589–604.

Lee, Y., and Lippman, R. P. (1989), "Practical Characteristics of Neural Network and Conventional Pattern Classifiers on Artificial and Speech Problems," in *Advances in Neaural Information Processing Systems*, ed. D. S. Touretzky, San Mateo, CA: Morgan Kaufmann, pp. 168–177.

Loh, W.-Y., and Vanichsetakul, N. (1988), "Tree-Structured Classification via Generalized Discriminant Analysis" (with discussion), *Journal of the American Statistical Association*, 83, 715–728.

Minnotte, M. C., Marchette, D. J., and Wegman, E. J. (1998), "The Bumpy Road to the Mode Forest," *Journal of Computational Graphical Statistics*, 7, 239–251.

Minnotte, M. C., and Scott, D. (1993), "The Mode Tree: A Tool for Visualization of Nonparametric Density Estimates," *Journal of Computational and Graphical Statistics*, 2, 51–68.

Muller, H. G. (1984), "Smooth Optimum Kernel Estimators of Densities, Regression Curves and Modes," *The Annals of Statistics*, 12, 766–774.

Opitz, D., and Maclin, R. (1999), "Popular Ensemble Methods: An Empirical Study," *Journal of Artificial Intelligence Research*, 11, 169–198.

Peterson, G. E., and Barney, H. L. (1952), "Control Methods Used in a Study of Vowels," *Journal of the Acoustical Society of America*, 24, 175–185.

Reaven, G. M., and Miller, R. G. (1979), "An Attempt to Define the Nature of Chemical Diabetes Using a Multidimensional Analysis," *Diabetologia*, 16, 17–24.

Ripley, B. D. (1994), "Neural Networks and Related Methods for Classification" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 56, 409–456.

——— (1996), *Pattern Recognition and Neural Networks*, Cambridge, U.K.: Cambridge University Press.

Schapire, R. E., Fruend, Y., Bartlett, P., and Lee, W. (1998), "Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods," *The Annals of Statistics*, 26, 1651–1686.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice and Visualization*, New York: Wiley.

Sheather, S. J., and Jones, M. C. (1991), "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society*, Ser. B, 53, 683–690.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman & Hall.

Stone, C. J. (1984), "An Asymptotically Optimal Window Selection Rule in Kernel Density Estimates," *The Annals of Statistics*, 12, 1285–1297.

Stone, M. (1977), "Cross-Validation: A Review," *Mathematische Operationsforschung und Statistik, Series Statistics*, 9, 127–139.

Wand, M. P., and Jones, M. C. (1995), *Kernel Smoothing*, London: Chapman & Hall.