Classification Using Kernel Density Estimates : Multi-scale Analysis and Visualization

Anil K. Ghosh[+], Probal Chaudhuri[+] and Debasis Sengupta[*]

[+]Theoretical Statistics and Mathematics Unit, [*]Applied Statistics Unit, Indian Statistical Institute, 203, B. T. Road, Calcutta-700108, India.

email : res9812@isical.ac.in, probal@isical.ac.in, sdebasis@isical.ac.in

Abstract

The use of kernel density estimates in discriminant analysis is quite well known among scientists and engineers interested in statistical pattern recognition. The use of a kernel density estimate involves proper selection of 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, while 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 multi-class 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 multi-scale 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.

Keywords and Phrases : majority voting, misclassification rates, optimal bandwidths, pairwise coupling, posterior probability, P-value type measure, 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 review). The basic problem in classification or discriminant analysis

is to formulate a decision rule $\mathtt{d}(\mathbf{x}) : R^d \to \{1, 2, \ldots, J\}$ 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 $\mathtt{d}_B(\mathbf{x}) = j^*$ 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 they 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 j^{th} population ($j = 1, 2, \ldots, J$), the kernel estimate $\hat{f}_{jh_j}(\mathbf{x})$ of the j^{th} population density is given by $\hat{f}_{jh_j}(\mathbf{x}) = n_j^{-1} h_j^{-d} \sum_{k=1}^{n_j} K\left\{h_j^{-1}(\mathbf{x}_{jk} - \mathbf{x})\right\}$, 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 density estimates are plugged in $\mathtt{d}_B(\mathbf{x})$ to build a nonparametric classification method called as 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). Gaussian kernel $K(t) = (2\pi)^{-1/2} e^{-t^2/2}$ is a popular choice for the kernel function $K(\cdot)$, and we will use it throughout this article.

Clearly, the performance of this nonparametric classifier depends critically on the values of bandwidth parameters. There are many different techniques available in the literature (see e.g., Hall, 1983; Stone, 1984; Silverman, 1986; Hall *et. al.*, 1991; Sheather and Jones, 1991; Scott, 1992; Wand and Jones, 1995; Jones, Marron and Sheather, 1996) for choosing optimal bandwidths from the data. 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 the increasing values of bandwidth, the bias of a kernel density estimate gets increased while its variance gets smaller. A detail discussion of the effect of this bias and variance on misclassification rates is available in 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 for choosing bandwidths in practice due to 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. Two other important points to keep in mind in the case of discriminant analysis using kernel density estimates are the following :

(i) The choice of bandwidths should depend on the specific observation to be classified in addition to depending on the population densities, and 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.

(ii) In a multi-class 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 it is compared with the density estimates for different competing classes 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 multi-scale version of kernel discriminant analysis. Over the last few years, multi-scale 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 multi-scale 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 the 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 is to be done by some judicious combination of all information obtained at different levels of smoothing, and we will discuss some appropriate ways for doing that. In the presence of more than two competing populations, we follow the same procedure taking each pair of classes, and then use the method of majority voting (see e.g., Friedman, 1996) or the method of pairwise coupling (see e.g., Hastie and Tibshirani, 1998) to combine the results of these pairwise comparisons.

2 Multi-scale visualization of discrimination measures

Suppose that $\mathbf{x}_{j1}, \mathbf{x}_{j2}, \ldots, \mathbf{x}_{jn_j}$ are training sample observations from the j^{th} class, where $1 \leq j \leq J$. In order to classify an observation $\mathbf{x}$ into one of the J classes, first we need to obtain the density estimates $\hat{f}_{jh_j}(\mathbf{x})$ at the point $\mathbf{x}$ for all $j = 1, 2, \ldots, J$. In practice, before computing the class density estimate, one 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 making use of a common bandwidth h_j for all co-ordinate variables more justified. The density estimate for the original data vectors can be obtained from that of the standardized data vectors by 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, the posterior probability estimate for the first population is given by

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

We can use a wide range of values for h_1 and h_2 to compute these estimated posteriors, and they can be plotted using grey 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).

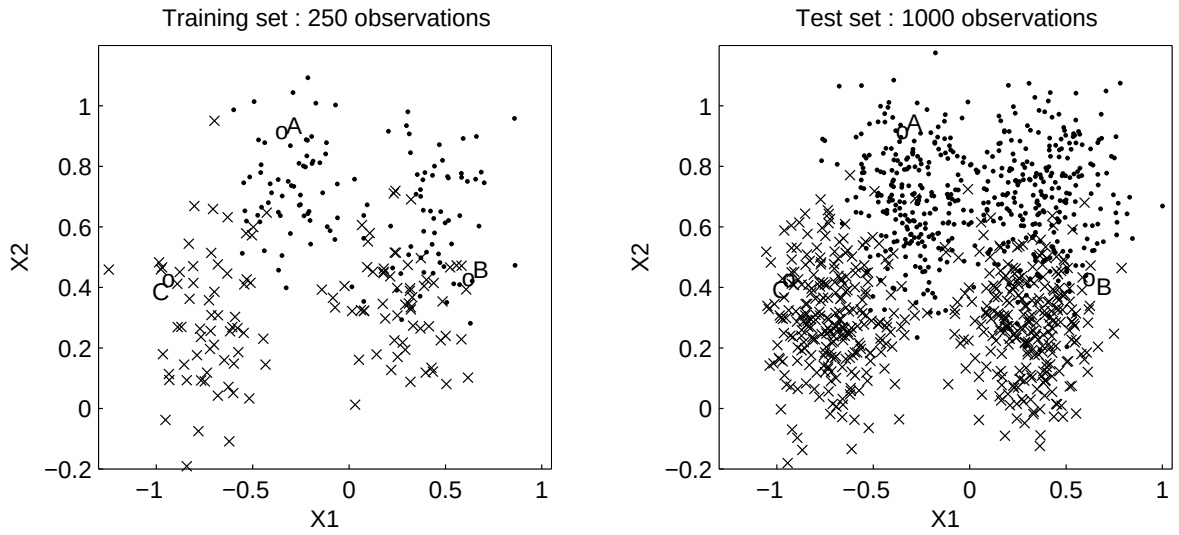

Figure 2.1 : Scatter plots for synthetic data

To demonstrate our methodology, we consider an example data set from Ripley (1994), which is popularly known as the "synthetic data". This data set is related to a two-class problem, where both the classes are equal mixtures of two bivariate normal populations differing only in their location parameters. This data set contains a training sample of size 250 (125 for each class) and a test sample of size 1000 (500 from each class). It is available from http://www.lib.stat.cmu.edu. Scatter plots for the training and the test samples of synthetic data are given in Figure 2.1, where the dots ($\cdot$) and the crosses ($\times$) represent the observations coming form the two classes.

We have chosen three observations (indicated by '$\circ$' in Figure 2.1) from the test set and labeled them as 'A', 'B' and 'C'. These three points are purposively 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 the populations have more or less equal strength. We performed usual linear (LDA) and quadratic discriminant analysis (QDA) on this data to classify the entire test set observations using the training sample. There are some observations that got misclassified by both the methods, and 'B' is one of them.

Though it is originally from population-1, both LDA and QDA gave decisions in favor of population-2. Of course, both of 'A' and 'C' were correctly classified by the linear and the quadratic classifiers. We used a wide range of bandwidths for each of the two populations to evaluate the posterior probabilities $\mathcal{P}_{h_1,h_2}(1 \mid \mathbf{x})$ for different levels of smoothing. As there are equal numbers of observations in these two classes, prior probabilities for our analysis are taken to be equal.

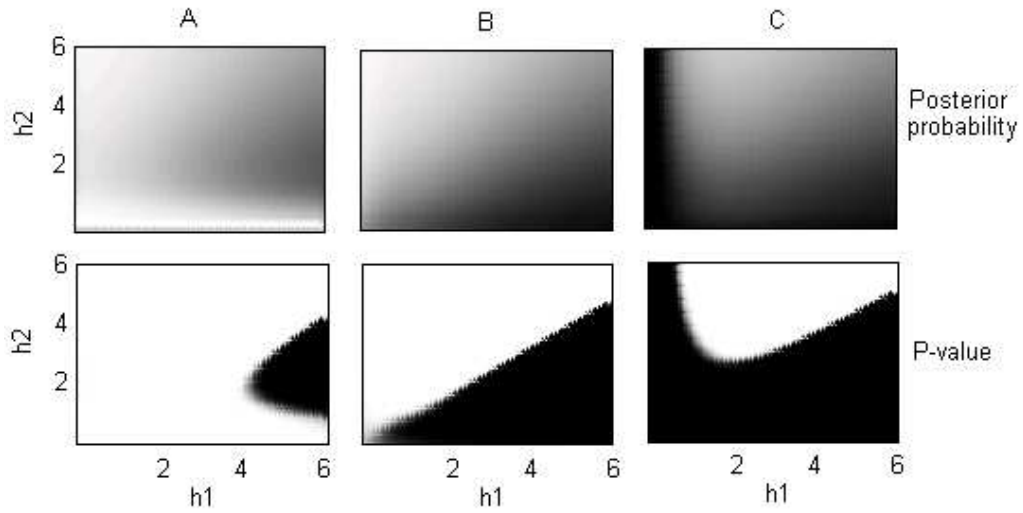

Figure 2.2 : Multi-scale analysis of synthetic data

The top row of Figure 2.2 gives the grey scale representation of posterior probabilities for the three cases, where the bandwidths of the first and the second populations are plotted along the horizontal and the vertical axes respectively. Here white color (high posterior) indicates the regions in favor of the first population whereas black color (low posterior) points towards the other. Intensity of the color varies with the magnitudes of the posterior probabilities, and this helps us to find out the regions for strong evidence in favor of one of the two populations. As it is expected, we observe a dominance of light colored region in the case of observation 'A' and that of the dark region 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 any 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

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

Clearly, high and low values of this probability give the decisions in favor of the first and the second population, respectively. For fixed h_1 and h_2, since the density estimates are averages of i.i.d. random variables, and density estimates for different populations are based on independent sets of observations, we can conveniently use normal approximation to evaluate the above probability with a great degree of accuracy even for 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 i.i.d. sequence of random

variables. However, 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 central limit theorem for triangular arrays in Hall and Heyde, 1980). Using such normal approximation with estimated means and variances we get

$$\begin{aligned} P_{h_1,h_2}(\mathbf{x}) &\simeq \Phi\left(\left\{\pi_1 E[\hat{f}_{1h_1}(\mathbf{x}) \mid \mathbf{x}] - \pi_2 E[\hat{f}_{2h_2}(\mathbf{x}) \mid \mathbf{x}]\right\} / \sqrt{\pi_1^2 Var[\hat{f}_{1h_1}(\mathbf{x}) \mid \mathbf{x}] + \pi_2^2 Var[\hat{f}_{2h_2}(\mathbf{x}) \mid \mathbf{x}]}\right) \\ &\simeq \Phi\left(\left\{\pi_1 \hat{f}_{1h_1}(\mathbf{x}) - \pi_2 \hat{f}_{2h_2}(\mathbf{x})\right\} / \sqrt{\pi_1^2 s_{1h_1}^2(\mathbf{x})/n_1 + \pi_2^2 s_{2h_2}^2(\mathbf{x})/n_2}\right), \end{aligned}$$

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 estimated variance of $\hat{f}_{jh_j}(\mathbf{x})$ $(j = 1, 2)$ obtained from the training sample using the sample variance of $h_j^{-d} K\{h_j^{-1}(\mathbf{x}_{j1} - \mathbf{x})\}, \ldots, h_j^{-d} K\{h_j^{-1}(\mathbf{x}_{jn_j} - \mathbf{x})\}$.

An alternative interesting interpretation of the above 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 above 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.

In the bottom row in Figure 2.2, we have plotted these P-values in two-dimensional plots for the observations 'A' , 'B' and 'C' using grey scales for various choices of h_1 and h_2. Like before, the white region corresponding to high values of $P_{h_1,h_2}(\mathbf{x})$ favors the first population while the black region corresponding to low values of $P_{h_1,h_2}(\mathbf{x})$ favors the second. Once again, the plots give some idea for classification of 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 the bandwidths.

One note-worthy feature of the plots in the two rows in Figure 2.2 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 second row provide an 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 difference in the sharpness for the two sets of plots.

Theorem 2.1 : *Suppose that, for a given observation* $\mathbf{x}$, $E\left[K^2\left\{h_j^{-1}(\mathbf{x} - \mathbf{x}_{j1})\right\} \middle| \mathbf{x}\right] < \infty$ *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) $\mid \mathcal{P}_{h_1,h_2}(1 \mid \mathbf{x}) - \frac{\pi_1 \mathcal{S}_{1h_1}(\mathbf{x})}{\pi_1 \mathcal{S}_{1h_1}(\mathbf{x}) + \pi_2 \mathcal{S}_{2h_2}(\mathbf{x})} \mid = O_P(N^{-1/2})$, *where* $\mathcal{S}_{jh_j}(\mathbf{x}) = E\{\hat{f}_{jh_j}(\mathbf{x})\}$ *for* $j = 1, 2$, *and*

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

The main implication of the above theorem is as follows. For any given $\mathbf{x}$ and a given pair of bandwidths (h_1, h_2), the estimated posterior probability $\mathcal{P}_{h_1,h_2}(1 \mid \mathbf{x})$ 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 $\mathcal{S}_{1h_1}(\mathbf{x})$, $\mathcal{S}_{2h2}(\mathbf{x})$ and the prior probabilities, the P-value $P_{h_1,h_2}(\mathbf{x})$ either converges to 0 or to 1 and that too at an exponential rate. For instance, if $\pi_1 \mathcal{S}_{1h_1}(\mathbf{x}) < \pi_2 \mathcal{S}_{2h_2}(\mathbf{x})$, $\mathcal{P}_{h_1,h_2}(1 \mid \mathbf{x})$ has a $\sqrt{N}$ rate of convergence to a value less than 0.5 but the

corresponding P-value type measure converges to zero 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 a stronger evidence than $\mathcal{P}_{h_1,h_2}(1 \mid \mathbf{x})$ for or against population-1.

In practice, the choice of bandwidth ranges in Figure 2.2 is an importance issue. It can be shown that (see Theorem 4.1 and its proof) 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 0.5 near the line $h_2 = (\pi_2/\pi_1)^{1/d} h_1$ 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 $\pi_1 = \pi_2 = 0.5$, as it is in the case of Figure 2.2, this line is the diagonal line. Since 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 purpose 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 these issues in mind, here we have adopted a rule of using an upper limit $(h^{(1)}, h^{(2)})$ 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 2.2 this upper limit works out to be about 6). Also, it is possible to look at the plots for various choices of $(h^{(1)}, h^{(2)})$ to form an idea about the classification results and the strength of the evidence.

When using the plots like those in Figure 2.2, one has to keep in mind that the evidence at a pixel (i.e. at a bandwidth pair (h_1, h_2)) in favor or against a class needs to be properly supplemented by the reliability of the evidence as measured by the misclassification rate at that pixel (see Section 3). Hence, while the plots in Figure 2.2 are definitely useful as the first step for forming a visual evidence for the multi-scale classification results, one cannot just use the visible sizes of white and black regions in the plots for making the final classification. Instead, one needs to carefully weigh the evidence at each pixel 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 needs to aggregate the results obtained at different levels of smoothing. A natural way of combining 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 *et. al.*, 1998; Friedman, Hastie and Tibshirani, 2000) and arcing classifier (see e.g., Breiman, 1998) are some of the well known aggregation methods which adopt 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

$$\Delta(h_1, h_2) = \pi_1 \int\limits_{\mathbf{x} \in \mathcal{R}^c_{h1,h_2}} f_1(\mathbf{x}) d\mathbf{x} \; + \; \pi_2 \int\limits_{\mathbf{x} \in \mathcal{R}_{h_1,h_2}} f_2(\mathbf{x}) d\mathbf{x} \; ,$$

where $\mathcal{R}_{h_1,h_2}$ is the set of all $\mathbf{x}$ that are 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 in Ghosh and Chaudhuri (2004). For varying choices of bandwidths, they proposed a smooth and more accurate estimate of the misclassification probability for classifiers, which are based on kernel density estimates. Their estimates use normal approximation to the distribution of a kernel density estimate, which is an average of i.i.d. random variables. In this article, we used their method to estimate $\Delta(h_1, h_2)$, and the corresponding probability of correct classification $\{= 1 - \widehat{\Delta}(h_1, h_2)\}$ has been plotted in a two-dimensional figure using their grey scale values. Figure 3.1 shows such a plot for the "synthetic data" discussed in the preceding section. Here white color represents high probability of correct classification and black color represents the opposite. 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, and these bandwidth pairs have comparable performance to the bandwidth pair marked by '$*$'. This plot gives a useful visualization of statistical uncertainties in classification obtained by using the kernel density estimates for different levels of smoothing, and it also demonstrates the importance of looking at a wide range of bandwidth pairs instead of some single optimal pair.

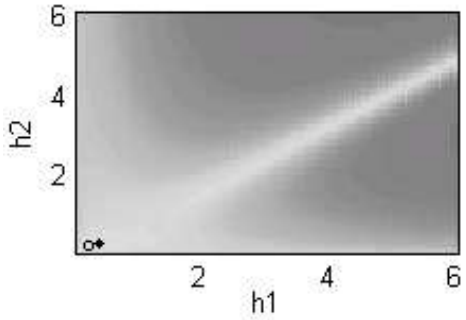

Figure 3.1 : Plot for probability of correct classification (synthetic data). The optimal $MISE$ bandwidth pair ('$\circ$') and the optimal bandwidth pair ('$*$') for misclassification rate are marked in the figure. Note a wide range bandwidth pairs with very low misclassification rates having comparable performance to the bandwidth pair ('$*$') with minimal misclassification rate.

3.2 Weight function derived from misclassification rates

In Figure 3.1, the white strip near the diagonal clearly suggests that this region is most reliable for classification, and the weight function $w(h_1, h_2)$ should take higher values near this line. 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 *et. al.*, 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 can be found in Opitz and Maclin (1999). These 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, in our method, we do not require any resampling techniques for generating the classifiers. Use of different value 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 which have reasonably good misclassification rates. But for some values of (h_1, h_2), the kernel classifier may lead to very poor classification. One has to appropriately weigh down these classification rules. The log function used in boosting decreases with misclassification probability at a very slow rate. Instead, if one chooses a Gaussian-type function, which decreases at a faster rate, the poor classifiers would be weighted down 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 since the performance of the classifier then turns out to be poorer than that of a trivial classifier, which classifies all observations to the class having the larger prior. Keeping these in view, in all our numerical work, we have used a Gaussian-type weight function

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

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 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 Gaussian-type weight function above 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 much sensitive to the weighting procedure as long as any reasonable weight function (which decreases appropriately with misclassification rates) is used.

3.3 Super-imposition of weight function over discrimination measures

Super-imposition 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 strength of different classes. Figure 2.2 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 fail to differentiate between the easier and the harder cases. Super-imposed versions of discrimination measures become helpful in such situations.

Let us consider an example with two multivariate normal populations differing only in their location parameters. Suppose that the populations have the mean vectors $\boldsymbol{\mu}_1 = (2, 0, \ldots, 0)$, $\boldsymbol{\mu}_2 = (0, 0, \ldots, 0)$ and the common dispersion matrix $\mathbf{I}_6$. We also consider the prior probabilities ($\pi_1 = \pi_2 = 0.5$) for the two classes to be equal and generate equal number 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, respectively, while $x_1 = 1$ represents a point on the class boundary. Therefore, one expects 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 3.2, where the maximum of the pairwise distances between the standardized data points is chosen as the upper limit of bandwidths. From this figure it is quite transparent that both posterior probabilities and P-values (first two rows of Figure 3.2) fail to reflect the difference in the strength of classification in these three cases. In these plots, though the white region extends as we move on from $x_1 = 0$ to $x_1 = 2$, still in all the cases we have almost equal split in favor of the classes indicated by white and black regions.

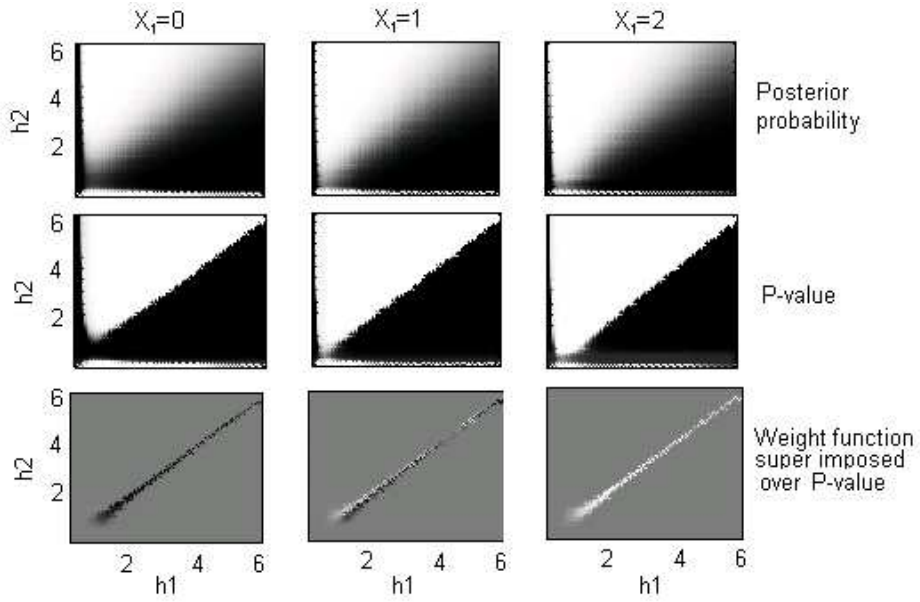

Figure 3.2 : Multi-scale analysis of simulated data

However, the difference in the classification result becomes evident if we look at the P-values super-

imposed over the weight function (last row of Figure 3.2). The weighted P-value that has been plotted against h_1 and h_2 is

$$P^S_{h_1,h_2}(\mathbf{x}) = 0.5 + \{P_{h_1,h_2}(\mathbf{x}) - 0.5\} w^*(h_1, h_2),$$

where w^* is the re-scaled version of the weight function which has the 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 0.5, which is indicated by the grey regions in the plots. However, in more reliable regions (pairs having high weights), we get stronger evidence as $P_{h_1,h_2}(\mathbf{x})$ moves away (in either direction) from 0.5. When $x_1 = 0$ (or $x_1 = 2$), we observe a black (or white) shade in this region, which gives a clear idea about the direction and the 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 shades 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 super-imposed 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, one may also notice a white or a black streak near both the axes. This is because, for the given sample sizes, use of such small bandwidth makes one density estimate very close to zero, and therefore the competing class density estimate turns out to be the winner. However, these streaks appear in a region of the plot where we have high misclassification probability. Consequently, the weight function becomes zero in this regions, and the plots of $P^S_{h_1,h_2}$ do not have such odd looking streaks.

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 average of posterior probabilities. The weighted P-value $P^S_{h_1,h_2}$ defined in Section 3.2 can be used for this purpose, as it makes sense to rely more on those bandwidth pairs, which lead to stronger and reliable evidence for one of the two classes. We choose the adjusted weight function

$$w_{\mathbf{X}}(h_1, h_2) = w^*(h_1, h_2)\, |P_{h_1,h_2}(\mathbf{x}) - 0.5| = \left| P^S_{h_1,h_2}(\mathbf{x}) - 0.5 \right|,$$

and use it to aggregate the posterior probabilities obtained by different classifiers. These adjusted weights not only depend 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 has to fix the range of bandwidths as well. Our empirical experience suggests that in a two class problem, after a certain level, if we keep on increasing the upper limit of bandwidths, the difference between the weighted averages of posteriors gets reduced but the classification result remains same in almost all cases. After standardizing a data set, one can compute all pairwise distances between the standardized observations and determine the α-th quantile $(0 < \alpha < 1)$of these distances. One can use this quantile as the upper limit of bandwidth for some large values of α like $\alpha = 0.9$ or 0.95. To reduce the computational cost, instead of calculating

all pairwise distances, one can also approximate this upper limit using the some high percentile of appropriate chi-square distribution if the data distribution is approximately normal. However, this choice of upper limit is subjective, and one may use some smaller or larger values for upper limits as well. Use of bandwidths larger than this proposed upper limit usually increases the computational complexity but it does not improve the performance of multi-scale method in terms of error rates. We will investigate this in detail in Section 5.2, where some benchmark data sets will be used for classification. In all our numerical work in this article, we will use $\alpha = 0.95$ and the corresponding upper limits will be denoted by $\lambda_\alpha = \lambda_{0.95} = (h^{(1)}_{0.95}, h^{(2)}_{0.95})$. Note that for visualization purpose in Sections 2.1, 2.2, 3.1 and 3.3, we used a simpler and more conservative rule setting the upper limit for bandwidths as the maximum of the pairwise distances between the standardized data points.

We conclude this section by considering once again the "synthetic data" for the purpose of illustration. In case of observations 'A' and 'C' (see Section 2.1), when we use $\lambda_{0.95} = (3.263, 3.258)$ as the upper limit of bandwidths, the weighted average of the posteriors (with $\tau = 3$) in favor of the first population came out to be 0.873 and 0.189 respectively, which give a clear indication about the classes to which they belong. Inclusion of large bandwidths in aggregation reduces the difference between the weighted posteriors but generally it does not change the classification result. For instance, if one uses $(10, 10)$ as the upper limit of the bandwidths, $\mathcal{P}_{h_1,h_2}(1 \mid \mathbf{x})$ for 'A' and 'C' tuned out to be 0.778 and 0.273, respectively. However, in the case of observation 'B', for both choices of range $\mathcal{P}_{h_1,h_2}(1 \mid \mathbf{x})$ was found to be very close to 0.5 (0.482 for λ_α and 0.489 for (10,10)), as one would expect in view of the fact that this observation lies near the class boundary where both the classes have almost equal strength. One should 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 find out 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 multi-class problems into a number of binary classification problems taking a pair of classes at a time and proceed in the same way as before. The results of all 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 which has the maximum number of votes. However, this voting method sometimes may lead to a region of indecision, where 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 MISE bandwidths

As we have mentioned before, the bandwidths that minimizes $MISE$ of the density estimates, sometimes lead to poor performance in discriminant analysis. For example, consider the classification problem with six dimensional simulated data set as discussed in Section 3.3. Since the population distributions are themselves spherical, without any standardization one can use a single common bandwidth in all the directions. Moreover, because of the same 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.

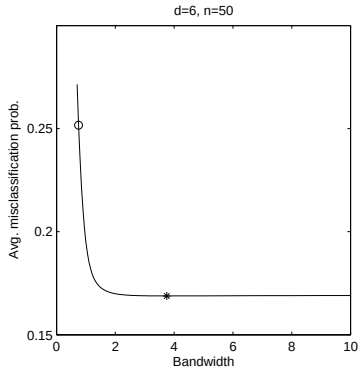

Figure 4.1 : Average misclassification probability and optimal bandwidth

In Figure 4.1 (taken from Ghosh and Chaudhuri, 2004), the true average misclassification probability has been plotted 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 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 number of observations from these two classes. We generated a test set of size 1000 (500 from each class) and classified them using 100 training set observations (50 from each class). The bandwidth pair (h_1, h_2) that minimize the MISE of the density estimates led to a misclassification rate of 22.3%, while the kernel classifier with the best bandwidth for classification (which can also be viewed as a weighted averaging method with $\tau = 0$) could reduce it to 18.6%. In this example, the optimal Bayes classifier wrongly classified 16.2% of the test set observations. Like what we observed in Figure 4.1, here also $\mathbf{h}_* = (4.45, 4.30)$ was found to be much larger than $\mathbf{h}_\circ = (0.75, 0.75)$. In density estimation problems, use of large bandwidths generally leads to large bias and hence large $MISE$ for the density estimates. Therefore, in density estimation, with the increasing sample size, one usually shrinks the bandwidth to zero in order to get good performance. But that is not the case for kernel discriminant analysis. Here, depending on competing population densities, use of 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 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, 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 4.1 : *Suppose that f_1 and f_2 are such that $\int \|\mathbf{x}\|^6 f_j(\mathbf{x})d\mathbf{x} < \infty$ 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, 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

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

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

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

If f_j's are spherically symmetric and they satisfy a location shift model, and if the kernel function K is also spherical (note that $\nabla^2 K(\mathbf{0})$ is negative definite), 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 j-th population ($j = 1, 2$). It is to be noted that the linear classifier described above is the optimal Bayes classifier under this set up. Therefore, in this particular case, use of large bandwidth leads to misclassification probability close to the optimal Bayes risk.

However, the use of $\mathbf{h}_*$ does not necessarily lead to better estimates for the posterior probabilities. In Figure 4.2, the estimated posterior probabilities for the simulated data set are plotted against the true posteriors of different observations. When $\mathbf{h}_\circ$ is used for classification, the posteriors get more scattered (right column of Figure 4.2) but this choice of bandwidth leads to very little bias for the posterior probability estimates. On the other hand, for $\mathbf{h}_*$ (left column of Figure 4.2) 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. The use of large bandwidths reduce the variance of the kernel density estimate at the cost of increased bias in order to preserve the ordering of the true posteriors, and this is precisely the fact that reflected in Figure 4.2. A detailed discussion on the effect of such bias and variance on misclassification error rates is available in Friedman (1997). While $\mathbf{h}_\circ$ leads to a mean square error of 0.046 for posterior estimates, h_* increases it to 0.112. In this case, the method based on weighted averaging of posterior with $\tau = 3$ amounts to a compromise

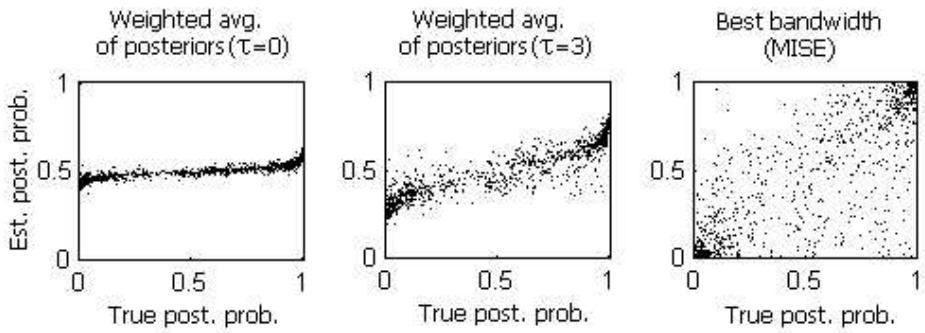

Figure 4.2 : Estimated posterior probabilities for simulated data set

between the preceding two (middle row). It improves the mean square error (0.060) of posterior estimates significantly without sacrificing much accuracy in terms of misclassification rates (19.0%).

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 in Reaven and Miller (1979). This data set 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 data set, if we use bandwidths that minimize estimated $MISE$ of population density estimates, we get leave-one-out cross-validated misclassification rate of 12.41%. This error rate is higher than that obtained for even simple linear and quadratic discriminant analysis, which showed leave-one-out misclassification rates of 11.03% and 9.66%, respectively. However, for our multi-scale analysis followed by the weighted averaging of posteriors led to a leave-one-out cross-validated error rate of 5.52% and 6.21% for $\tau = 0$ and $\tau = 3$, respectively. 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 data set, majority voting did not lead to any tied case either for $\tau = 0$ or for $\tau = 3$.

5 Case studies using benchmark data sets

In this section, we report our findings based on some benchmark data sets 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 combined the results by using majority voting (Friedman, 1996) as well as by pairwise coupling (Hastie and Tibshirani, 1998). Misclassification error rates for usual linear and quadratic discriminant analysis (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 those tied cases are considered as "misclassification". Therefore, the reported results on voting are actually the proportion of misclassifications in the worst possible cases. The data

sets we consider here have been analyzed before in the literature, where nonparametric methods like classification trees (see e.g., Breiman *et. al.*, 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) was 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 priors. Apart from the vowel recognition data, all the data sets that are considered in this section are available at http://www.lib.stat.cmu.edu.

• **Synthetic data :** Description of this data set has already been given in Section 2.1. Ripley (1994) used this data to compare the performance of different classification algorithms. The class distributions were chosen to have a Bayes risk of 8.0%. In this data set, LDA and QDA could achieve test set error rates of 10.8% and 10.2%, respectively. Classification tree (CART) (see Breiman *et. al.*, 1984) also misclassified more than 10% observations (see Table 5.1). Performance of other nonparametric method were fairly similar. Weighted averaging of the posterior achieved the best error rate when $\tau = 0$ is used.

• **Vowel recognition data :** This data 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 data set a challenging one for any classification method. A scatter plot of this data set is given in Figure 5.1 where the numbers represent the labels of the different classes ('0' represents the 10-th class).

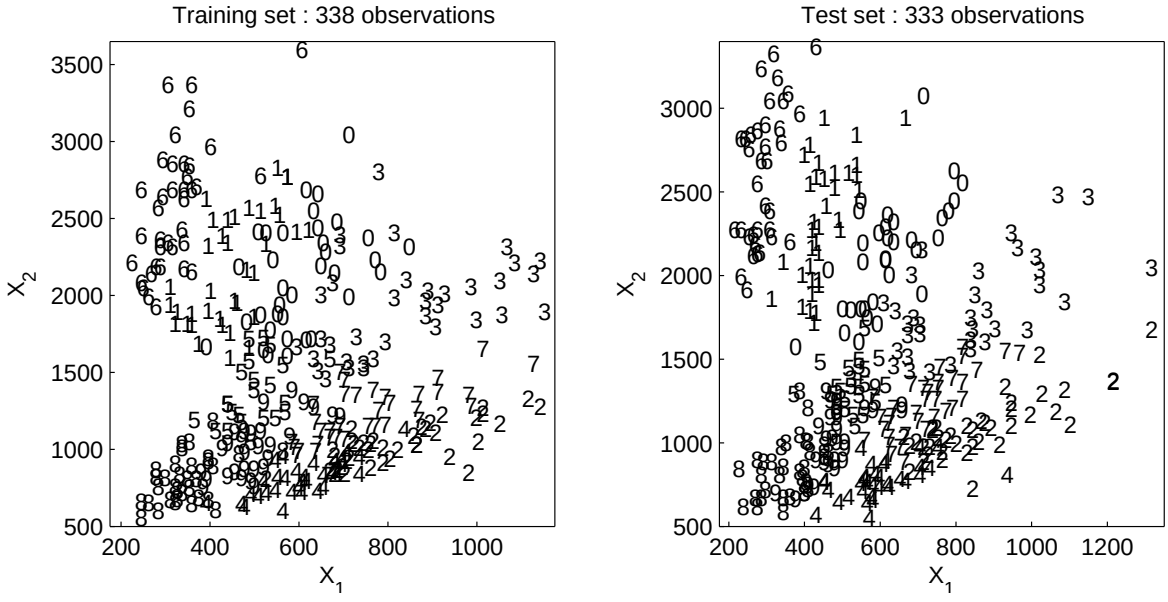

Figure 5.1 : Scatter plots for vowel recognition data

This data set has been extensively analyzed by many authors (see e.g., Lee and Lippman, 1989;

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

When pairwise coupling was used for final classification instead of majority voting, we obtained an error rate of 24.6% for weighted averaging of the posteriors with $\tau = 0$. We suspect that since the optimal bandwidth minimizing the misclassification rate do not always lead to good estimates for posterior probabilities as we have seen before, the performance of pairwise coupling method turns out to be so bad due to the presence of a large number of overlapping populations. 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 $\tau = 3$ followed by an application of pairwise coupling method.

• **Sonar data :** This data set was used by Gorman and Sejnowski (1988). It 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 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 0.0 to 1.0, each of which represents the energy within a particular frequency band, integrated over a certain period of time. To reduce co-ordinate-wise dependence, the data were averaged in a band of three making the number of measurement variables 20. The data set 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 data set are available in Ripley (1994) and Coolie and MacEachern (1998). QDA in this data set performed quite well as compared to other classification methods like LDA, FDA-MARS, CART and neural nets (see Table 5.1). Kernel method with $\tau = 3$ led to even better performance.

Table 5.1 : Percentage of misclassifications for different classification methods

Data sets	LDA	QDA	FDA-MARS		CART	Neural	Kernel	Kernel (wt. avg.)	
			deg.1	deg.2		networks$^+$	(MISE)	$\tau = 0$	$\tau = 3$
Synthetic	10.8	10.2	9.3	9.6	10.1	9.4	9.3	9.0	9.1
Vowel$^\circ$	25.2	19.8	20.7	19.8	23.7	18.6	18.9	19.2	17.7
Sonar	20.2	15.4	22.1	19.2	20.2	19.2	17.3	17.3	13.5

$\circ$ Majority voting is used for final classification.

5.1 Forensic glass data : a challenging problem for kernel discriminant analysis

This data set contains information on refractive index and eight other different 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 data set but most of them are window float (70) and window non-float (76) glasses. The rest of the classes, namely vehicle glass (17), containers (13), tableware (9) and vehicle headlamp (29), contain much smaller number of observations, and this makes it a difficult high dimensional classification problem.

Ripley (1996) analyzed this data set extensively 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 re-scaled. This re-scaled nearest neighbor algorithm had an error rate of 23.6%. Misclassification error rate for usual nearest neighbor method was found to be 26.6% for $k = 1$. Neural networks with 4 to 8 hidden nodes were reported to have error rates between 24.8% and 29.9%. LDA in this data set led to a cross validated error rate of 37.9% which was even worse for quadratic discriminant analysis which had 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 the error rates 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% respectively) than the other nonparametric classifiers.

As four out of the nine measurement variables (oxides of Mg, K, Ba and Fe) have a significant number of zeros among their observed values, we decided to carry out our analysis with the remaining five variables. However, even after using this subset of measurement variables, we could achieve a competitive performance for classifiers based on kernel density estimates. When the bandwidths, which minimize MISE of the density estimates, are used for classification, it led to a fairly good performance. The leave-one-out estimate for the misclassification error was found to be 31.3%. Using the method of weighted averaging of posterior, we obtained even better performance. The error rates for $\tau = 0$ and $\tau = 3$ were found to be 29.9% and 27.6%, respectively, when majority voting is used. When pairwise coupling method was applied to this data set after weighted averaging of posteriors, the aforesaid error rates increased to 30.4% and 36.4%, respectively.

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 purpose, and this has been followed throughout this section. For a given data set, the 0.95-th quantile of the pairwise distances (denoted by $\lambda_{0.95}$) is taken as the upper limit. Of course this choice is subjective and one may use smaller or larger values of upper limits as well. However, in Table 5.2 below, the misclassification rates for different choices of ranges of bandwidths are found to be almost equal for all the benchmark data sets analyzed here.

Table 5.2 : Percentage of misclassifications for different choice of ranges for bandwidths

Upper limit	$\lambda_{0.95}/3$		$\lambda_{0.95}/2$		$\lambda_{0.95}$		$2\lambda_{0.95}$		$3\lambda_{0.95}$	
of bandwidths	$\tau=0$	$\tau=3$	$\tau=0$	$\tau=3$	$\tau=0$	$\tau=3$	$\tau=0$	$\tau=3$	$\tau=0$	$\tau=3$
Synthetic	9.0	9.0	9.0	8.9	9.0	9.1	9.0	9.1	9.0	9.3
Sonar	16.3	13.5	16.3	12.5	17.3	13.5	16.3	13.5	17.3	13.5
Vowel$^{\circ}$	18.6	19.2	18.6	18.9	19.2	17.7	19.8	17.4	19.8	17.7
Glass$^{\circ,+}$	28.5	27.6	30.8	29.0	29.9	27.6	28.5	29.9	29.9	29.0

$\circ$ In multi-class problems, majority voting is used for final classification.

$+$ Numbers represent leave-one-out error rates

Acknowledgement

We are thankful to the associate editor and the two referees for their careful reading of an earlier version of the paper. Their constructive criticisms and suggestions led to substantial improvement of the paper.

Appendix

Proof of Theorem 2.1 : (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, as T_j is an average of i.i.d. random variables, from Central limit theorem, it follows that under the assumed moment condition, for large sample sizes, T_j tends to be normally distributed with mean $\tau_j = \pi_j \mathcal{S}_{jh_j}(\mathbf{x})$ and variance $\upsilon_j = \pi_j^2 s^2_{jhj}(\mathbf{x})/n_j$.

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 they are independent. Moreover, the function ψ is continuously differentiable in T_1 and T_2. Therefore, the usual asymptotic Taylor expansion leads to :

$$\frac{\{\psi(T_1,T_2)-\psi(\tau_1,\tau_2)\}}{\mathcal{V}} \xrightarrow{L} Normal\ (0,1), \quad \text{where } \mathcal{V} = \left\{ \sum_{j=1}^{2} \upsilon_j \ (\partial\psi/\partial T_j)^2_{T_1=\tau_1, T_2=\tau_2} \right\}^{1/2}.$$

Since $n_j/N \to \lambda_j > 0$ 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 $\tau_1 > \tau_2$ i.e. $I\{\pi_1 \mathcal{S}_{1h_1}(\mathbf{x}) > \pi_2 \mathcal{S}_{2h_2}(\mathbf{x})\} = 1$. Now, for some fixed h_1, h_2 and $\mathbf{x}$, from part (a) of this theorem, it follows that,

$$\frac{1}{\sqrt{\upsilon_1+\upsilon_2}} \left[(T_1 - T_2) - (\tau_1 - \tau_2)\right] \xrightarrow{L} Normal(0,1) \ \text{ as } N \to \infty.$$

Now define $Z_{h_1,h_2}(\mathbf{x}) = \frac{1}{\sqrt{\upsilon_1+\upsilon_2}}(T_1 - T_2) = \sqrt{N}(T_1 - T_2)/V$, where $V = \{\pi_1^2 s^2_{1h_1}(\mathbf{x})/\lambda_1 + \pi_2^2 s^2_{2h_2}(\mathbf{x})/\lambda_2\}^{1/2}$. 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 $\frac{1}{x}\phi(x) < 1 - \Phi(x) < \left(\frac{1}{x} - \frac{1}{x^3}\right)\phi(x)$, (where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the pdf and the cdf of a standard normal distribution), we get $1 - P_{h_1,h_2}(\mathbf{x}) = 1 - \Phi\left(Z_{h_1,h_2}(\mathbf{x})\right) = O_p(N^{-1/2} e^{-CN})$.

Proof of Theorem 4.1 : First note that

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

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

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

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

$$K\{(\mathbf{x}-\mathbf{X})/h_j\} = K(\mathbf{0}) + (1/2h_j^2)\{(\mathbf{x}-\mathbf{X})'\nabla^2 K(\mathbf{0})(\mathbf{x}-\mathbf{X})\} + (1/6h_j^3)\sum_{i,k,l} Y_{i,k,l}, \quad (\text{since } \nabla K(\mathbf{0}) = 0)$$

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,

$$\begin{aligned} E_{f_j}\{\hat{f}_{jh_j}(\mathbf{x})\} &= h_j^{-d}\left[K(\mathbf{0}) + (1/2h_j^2)E_{f_j}\{(\mathbf{x}-\mathbf{X})'\nabla^2 K(\mathbf{0})(\mathbf{x}-\mathbf{X})\} + O(h_j^{-3})\right] \text{ and} \\ Var_{f_j}\{\hat{f}_{jh_j}(\mathbf{x})\} &= (4n_j h_j^{2d+4})^{-1}\left[Var_{f_j}\{(\mathbf{x}-\mathbf{X})'\nabla^2 K(\mathbf{0})(\mathbf{x}-\mathbf{X})\} + O(h_j^{-1})\right] \end{aligned}$$

using the fact 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 zero, 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

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

(a) When $C_\pi < {C_h}^d$, for large h_1 and $h_2 = C_h h_1$, the above inequality holds whatever be 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 above 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.

References

[1] Bose, S. (1996) Classification using splines. *Computational Statistics and Data Analysis*, **22**, 505-525.

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

[3] Breiman, L. (1996) Bagging predictors. *Machine Learning*, **24**, 123-140.

[4] Breiman, L. (1998) Arcing classifiers (with discussion) *Ann. Statist.*, **26**, 801-849.

[5] Chaudhuri, P. and Marron, J. S. (1999) SiZer for exploration of structures in curves. *J. Amer. Statist. Assoc.*, **94**, 807-823.

[6] Chaudhuri, P. and Marron, J. S. (2000) Scale space view of curve estimation. *Ann. Statist.*, **28**, 408-428.

[7] Cheng, B. and Titterington, D. M. (1994) Neural networks : a review from a statistical perspective (with discussion). *Statistical Science*, **9**, 2-54.

[8] Cooley, C.A. and S.N. MacEachern (1998) Classification via kernel product estimators. *Biometrika*, **85**, 823-833.

[9] Coomans, D. and Broeckaert, I. (1986) *Potential Pattern Recognition in Chemical and Medical Decision Making*. Research Studies Press, Letchworth.

[10] Cover, T. M. and Hart, P. E. (1968) Nearest neighbor pattern classification, *IEEE Trans. Info. Theory*, **13**, 21-27.

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

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

[13] Friedman, J. H. (1991) Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, **19**, 1–141.

[14] Friedman, J. H. (1996) Another approach to polychotomous classification. *Tech. Rep., Dept. of Stat., Stanford University*.

[15] Friedman, J. H. (1997) On bias, variance, 0-1 loss, and the curse of dimensionality. *Data Mining and Knowledge Discovery*, **1**, 55-77.

[16] Friedman, J. H., Hastie, T. and Tibshirani, R. (2000) Additive logistic regression : a statistical view of boosting (with discussion). *Ann. Statist.*, **28**, 337-407.

[17] Ghosh, A. K. and Chaudhuri, P. (2004) Optimal smoothing in kernel discriminant analysis. *Statistica Sinica*, **14**, 457-483.

[18] Godtliebsen, F., Marron, J. S. and Chaudhuri, P. (2002) Significance in scale space for bivariate density estimation. *J. Comput. Graph. Statist.*, **11**, 1-22.

[19] Godtliebsen, F., Marron, J. S. and Chaudhuri, P. (2004) Statistical significance of features in digital images. To appear in *Image Vision and Computing*.

[20] 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.

[21] Hall, P. and Heyde, C.C. (1980) *Martingale limit theory and its application*. Academic Press, New York.

[22] Hall, P. (1983) Large sample optimality of least squares cross-validations in density estimation. *Ann. Statist.*, **11**, 1156-1174.

[23] Hall, P. and Wand, M. P. (1988) On nonparametric discrimination using density differences. *Biometrika*, **75**, 541-547.

[24] 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.

[25] Hand, D. J. (1982) *Kernel Discriminant Analysis*. Wiley, Chichester.

[26] Hastie, T., Tibshirani, R. and Buja, A. (1994) Flexible discriminant analysis. *J. Amer. Statist. Assoc.*, **89**, 1255-1270.

[27] Hastie, T. and Tibshirani, R. (1996) Discriminant analysis using Gaussian mixtures. *J. Royal Statist. Soc., Series B*, **58**, 155-176.

[28] Hastie, T. and Tibshirani, R. (1998) Classification by pairwise coupling. *Ann. Statist.*, **26**, 451-471.

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

[30] Huber, P. J. (1985) Projection pursuit (with discussion). *Ann. Statist.*, **13**, 435-475.

[31] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996) A brief summary of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.*, **91**, 401-407.

[32] Kim, H. and Loh, W.-Y. (2001) Classification trees with unbiased multiway splits. *J. Amer. Statist. Assoc.*, **96**, 589-604.

[33] Lee, Y. and Lippman, R. P. (1989) Practical characteristics of neural network and conventional pattern classifiers on artificial and speech problems. *Advances in Neaural Information Processing Systems* (Ed. : D. S. Touretzky), San Mateo, California : Morgan Kaufmann, pp. 168-177.

[34] Loh, W.-Y. and Vanichsetakul, N. (1988) Tree-structured classification via generalized discriminant analysis (with discussion). *J. Amer. Statist. Assoc.*, **83**, 715-728.

[35] Minnotte, M. C. and Scott, D. (1993) The mode tree : a tool for visualization of nonparametric density estimates. *J. Comput. Graph. Statist.*, **2**, 51-68.

[36] Minnotte, M. C., Marchette, D. J. and Wegman E. J. (1998) The bumpy road to the mode forest. *J. Comput. Graph. Statist.*, **7**, 239-251.

[37] Muller, H. G. (1984) Smooth optimum kernel estimators of densities, regression curves and modes. *Ann. Statist.*, **12**, 766-774.

[38] Opitz, D. and Maclin, R. (1999) Popular ensemble methods : an empirical study. *J. Art. Intell. Research*, **11**, 169-198.

[39] Peterson, G. E. and Barney, H. L. (1952) Control methods used in a study of vowels. *J. Acoust. Soc. Amer.*, **24**, 175-185.

[40] 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.

[41] Ripley, B. D. (1994) Neural networks and related methods for classification (with discussion). *J. Royal Statist. Soc., Series B*, **56**, 409-456.

[42] Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

[43] Schapire, R. E., Fruend, Y., Bartlett, P. and Lee, W. (1998) Boosting the margin : a new explanation for the effectiveness of voting methods. *Ann. Statist.*, **26**, 1651-1686.

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

[45] Sheather, S. J. and Jones, M. C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *J. Royal Statist. Soc., Series B*, **53**, 683-690.

[46] Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

[47] Stone, C. J. (1984) An asymptotically optimal window selection rule in kernel density estimates. *Ann. Statist.*, **12**, 1285-1297.

[48] Stone, M. (1977) Cross validation : a review. *Mathematische Operationsforschung und Statistik, Series Statistics*, **9**, 127-139.

[49] Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*. Chapman and Hall, London.