Statistica Sinica **14**(2004), 457-483

OPTIMAL SMOOTHING IN KERNEL DISCRIMINANT ANALYSIS

Anil K. Ghosh and Probal Chaudhuri

Indian Statistical Institute, Calcutta

Abstract: One well-known use of kernel density estimates is in nonparametric discriminant analysis, and its popularity is evident in its implementation in some commonly used statistical softwares (e.g., SAS). In this paper, we make a critical investigation into the influence of the value of the bandwidth on the behavior of the average misclassification probability of a classifier that is based on kernel density estimates. In the course of this investigation, we have observed some counter-intuitive results. For instance, the use of bandwidths that minimize mean integrated square errors of kernel estimates of population densities may lead to rather poor average misclassification rates. Further, the best choice of smoothing parameters in classification problems not only depends on the underlying true densities and sample sizes but also on prior probabilities. In particular, if the prior probabilities are all equal, the behavior of the average misclassification probability turns out to be quite interesting when both the sample sizes and the bandwidths are large. Our theoretical analysis provides some new insights into the problem of smoothing in nonparametric discriminant analysis. We also observe that popular cross-validation techniques (e.g., leave-one-out or V-fold) may not be very effective for selecting the bandwidth in practice. As a by-product of our investigation, we present a method for choosing appropriate values of the bandwidths when kernel density estimates are fitted to the training sample in a classification problem. The performance of the proposed method has been demonstrated using some simulation experiments as well as analysis of benchmark data sets, and its asymptotic properties have been studied under some regularity conditions.

Key words and phrases: Average misclassification probability, bandwidth selection, Bayes' risk, cross-validation techniques, location-shift models, scale space, spherical symmetry.

1. Introduction

In a discriminant analysis problem, one uses a decision rule $d(\mathbf{x}) : R^d \to \{1, \ldots, J\}$ for classifying a d-dimensional observation $\mathbf{x}$ into one of the J classes. The optimal Bayes rule (see e.g., Rao (1973) and Anderson (1984)) assigns an observation to the class with the largest posterior probability. It can be described as

$$d(\mathbf{x}) = \arg\max_j p(j \mid \mathbf{x}) = \arg\max_j \pi_j f_j(\mathbf{x}),$$

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 ($j = 1, \ldots, J$). Throughout this paper, we evaluate the performance of a discrimination rule by its average misclassification probability

$$\Delta = \sum_{j=1}^{J} \pi_j P\{d(\mathbf{x}) \neq j \mid \mathbf{x} \in j\text{th population}\}.$$

Density functions $f_j(\mathbf{x})$'s are usually unknown in practice, and can be estimated from the "training data set" either parametrically or nonparametrically. In parametric approaches (see e.g., Rao (1973), Mardia, Kent and Bibby (1979), Anderson (1984), James (1985), Fukunaga (1990), McLachlan (1992) and Ripley (1996)), the underlying population distributions are assumed to be known except for some unknown parameters (e.g., mean vector, dispersion matrix). Consequently, the performance of a parametric discrimination rule largely depends on the validity of those parametric models. Nonparametric classification techniques (see e.g., Breiman, Friedman, Olshen and Stone (1984), Loh and Vanichsetakul (1988), Dasarathy (1991), Hastie, Tibshirani and Buja (1994), Hastie and Tibshirani (1996), Bose (1996), Kooperberg, Bose and Stone (1997), Loh and Shih (1997) and Kim and Loh (2001)), however, are more flexible in nature and free from such parametric model assumptions. Kernel density estimation (see e.g., Muller (1984), Silverman (1986), Scott (1992) and Wand and Jones (1995)) is a well-known method for constructing nonparametric estimates of population densities. The use of kernel density estimates in discriminant analysis is quite popular in the existing literature (see e.g., Hand (1982), Coomans and Broeckaert (1986), Silverman (1986), Hall and Wand (1988), Scott (1992), Ripley (1996), Duda, Hart and Stork (2000), Hastie, Tibshirani and Friedman (2001) and Bensmail and Bozdogan (2002)) and in many standard softwares (e.g., SAS).

If $\mathbf{X}_{j1}, \ldots, \mathbf{X}_{jn_j}$ are d-dimensional observations in the training sample from the jth population, the kernel estimate of the density $f_j(\mathbf{x})$ is given by

$$\hat{f}_{jh}(\mathbf{x}) = {n_j}^{-1} h^{-d} \sum_{k=1}^{n_j} K\left\{h^{-1}(\mathbf{X}_{jk} - \mathbf{x})\right\} ,$$

where the kernel function $K(.)$ is a density function on the d-dimensional space, and $h > 0$ is a smoothing parameter popularly known as the bandwidth. A classification rule based on these kernel density estimates can be described as

$$d^*(\mathbf{x}) = \arg\max_j \pi_j \hat{f}_{jh}(\mathbf{x}).$$

For usual density estimation problems, the optimal bandwidth is generally taken to be the one that minimizes the mean integrated square error ($MISE = E[\int \{\hat{f}_{jh}(\mathbf{x}) - f_j(\mathbf{x})\}^2 \, d\mathbf{x}]$, see e.g., Silverman (1986) and Scott (1992)). As the

performance of a nonparametric classifier depends on the corresponding class density estimates, the choice of the smoothing parameter does have an important role in classification problems also. A question that naturally arises at this point is: how good is the average misclassification rate when the bandwidth that minimizes $MISE$ for the density estimation problem is used for classification?

In an attempt to investigate this question, we begin by considering a very simple two class problem with equal priors, where the classes are multivariate normal with the same dispersion matrix $\mathbf{\Sigma} = \mathbf{I}$ but different mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. In this setting, the bandwidth that minimizes the $MISE$ is the same for both classes if one has equal numbers of data points from the two classes in the training sample. Further, if we use normal kernel, it is possible to compute the bandwidth that minimizes $MISE$ analytically for normally distributed data. For such a problem, the average misclassification probability (Δ) can also be evaluated and plotted as a function of the bandwidth (h). Since the kernel density estimate is an average of i.i.d. random variables, one can conveniently use a normal approximation for its distribution. The mean and the variance of this normal approximation have nice analytic expressions when both the distribution of the data and the kernel are normal. We have tried to evaluate $\Delta(h)$ for a given value of h by two different procedures, one by using the normal approximation (described above) and the other by a large scale Monte-Carlo simulation. There was no visible difference in the plotted values of $\Delta(h)$ for these two different approaches – it seems that our sample size ($n_1 = n_2 = 50$) was good enough for a very high degree of accuracy in the normal approximation for the distribution of kernel density estimates.

In Figure 1.1, $\Delta(h)$ has been plotted for varying choices of h and for different dimensions ($d = 1, 2, 4, 6$), where we have chosen $\boldsymbol{\mu}_1 = (0, \ldots, 0)$ and $\boldsymbol{\mu}_2 = (2, 0, \ldots, 0)$, and the sample sizes are taken as 50 for both classes. This figure clearly shows striking difference between the optimal bandwidth for *the usual density estimation problem* and that for *the classification problem*. For different dimensions, optimal bandwidths for the classification problems (i.e., the bandwidths leading to the lowest misclassification probabilities) are marked by '$*$' in the figure, and the bandwidths that minimize $MISE$ are marked by '$\circ$'. This difference between the two bandwidths becomes larger as the dimension d increases. For dimension $d = 6$, the best bandwidth for the classification problem reduces the average misclassification rate by almost 32% when compared to the error rate corresponding to the optimal bandwidth that minimizes the $MISE$ in the density estimation problem.

What is even more interesting and counter-intuitive in Figure 1.1 is the behavior of $\Delta(h)$ for large values of h. It is well-known that for the density estimation problem, the $MISE$ turns out to be large for very small values of the bandwidth (due to large variance) as well as for very large values of the bandwidth

(due to large bias) (see e.g., Silverman (1986), Scott (1992) and Wand and Jones (1995) for detailed discussion). However, in all the cases in Figure 1.1, $\Delta(h)$ becomes almost flat after reaching its minimum value. Unlike what happens in the case of usual density estimation, large bandwidths do not seem to be a bad choice for the classification problems considered here. By changing $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$, we get different figures for $\Delta(h)$ but the basic pattern remains the same.

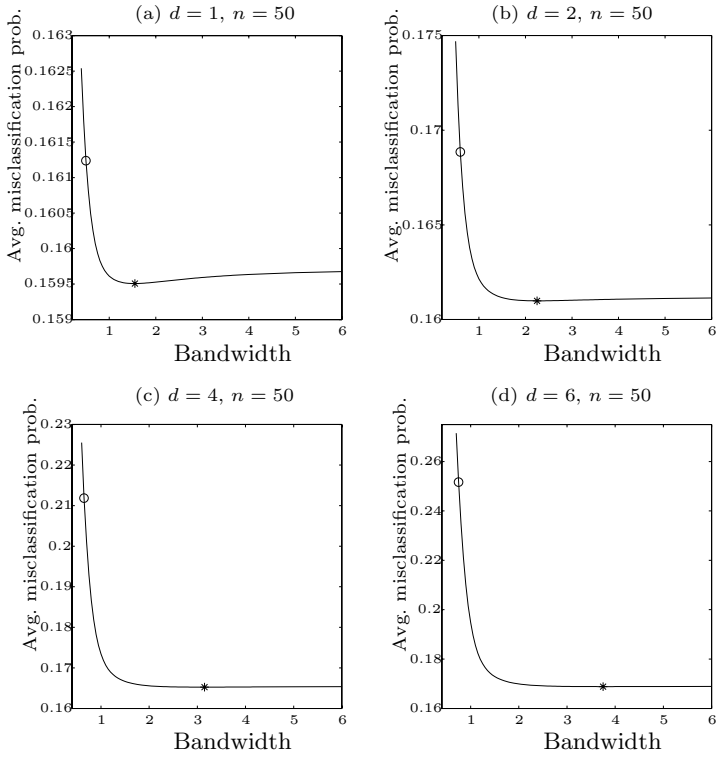

Figure 1.1. True Δ-function and optimal bandwidths (equal prior cases).

There are other popular methods for choosing the bandwidth in a classification problem based on cross-validation techniques (see e.g., Breiman, Friedman, Olshen and Stone (1984), Ripley (1996) and Duda, Hart and Stork (2000)). For instance, V-fold cross-validation divides the whole training sample into V parts of size as nearly equal as possible. Usually stratified random sampling is used to form the folds, where observations belonging to different classes are used as different strata. Then, taking one fold at a time as a test sample, one uses different bandwidths to classify its members based on a training sample formed by all the observations belonging to the other $V - 1$ folds. This procedure is repeated over the V folds, and the overall proportion of misclassification is used to estimate $\Delta(h)$. The bandwidth h, for which estimated value of $\Delta(h)$ is minimum, is

considered as the optimal bandwidth. When we have a total of N observations, leave-one-out or N-fold cross-validation (see e.g., Mosteller and Wallace (1963), Hills (1966) and Lachenbruch and Mickey (1968)) can be viewed as a special case of this procedure, where each fold consists of a single observation.

As the observed proportions of misclassifications are used to estimate $\Delta(h)$, the estimates are like step functions instead of being smooth curves even when the true $\Delta(h)$ is a nice smooth function. Consequently, instead of a single unique minimum, this procedure frequently leads to an interval or a union of some intervals as the possible choices for smoothing parameter from which it is difficult to choose a single optimum value.

In Figure 1.2, we demonstrate the limitations of cross-validation based techniques. Here, we consider the same problem as in Figure 1.1 and generate samples from the same normal populations. The true and estimated (by leave-one-out and 10-fold cross-validations) average misclassification probabilities are plotted simultaneously in Figure 1.2. The estimated curves not only behave like step functions but also miss the proper locations of optimum bandwidths by wide margins in some cases.

2. Behavior of $\Delta(h)$ as h varies

We know that for very large bandwidths, the $MISE$ of a kernel density estimate becomes large due to large bias, and we have observed in the examples in the preceding section that $\Delta(h)$ reaches a minimum and then remains nearly flat for a wide range of large values of h. We first try to explain such an apparently anomalous behavior of $\Delta(h)$ in those examples. Throughout this section, we assume that we have n observations in the training sample from each population, and a common bandwidth h is used for different population density estimates (which is justified in cases like location-shift population models).

For varying choices of the smoothing parameter h, following the ideas and the terminology in Chaudhuri and Marron (1999, 2000), $E\{\hat{f}_h(\mathbf{x})\}$ and $\hat{f}_h(\mathbf{x})$ can be viewed as *the theoretical* and *the empirical scale space functions* respectively. Theoretical scale space functions $E\{\hat{f}_{jh}(\mathbf{x})\}$ are the convolutions of the true densities $f_j(\mathbf{x})$ with a kernel K with bandwidth h. We know that with growing sample size, the variance of a kernel density estimate (which is an average of a set of i.i.d. random variables) gets smaller and, as a consequence, for any fixed bandwidth h the distribution of $\hat{f}_h(\mathbf{x})$ tends to be almost degenerate at $E\{\hat{f}_h(\mathbf{x})\}$ when the sample size is large.

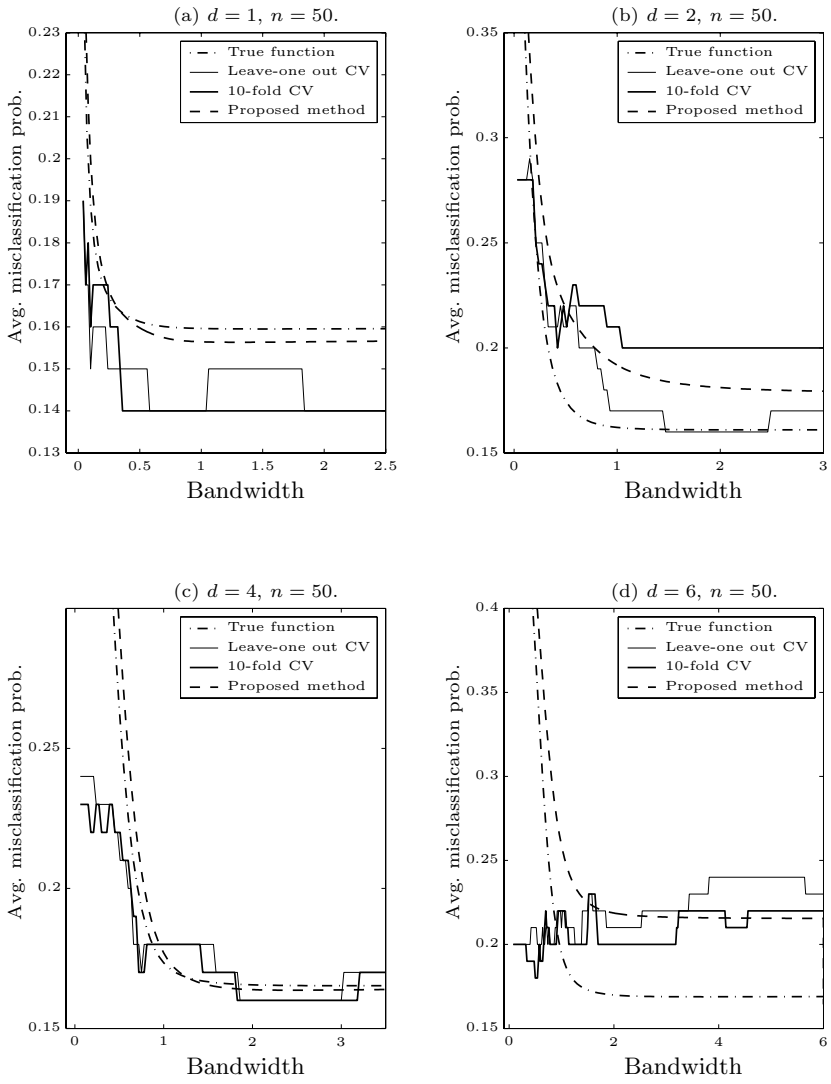

Figure 1.2. Average misclassification probabilities (equal prior cases).

When the prior probabilities for different populations are all equal, for a fixed value of h, as the sample size n tends to infinity the kernel density estimate based classifier tends to classify an observation into the class which has the largest value for the theoretical scale space function. When f and K both happen to be spherically symmetric and strictly decreasing functions of the distance from their centers of symmetry, the same holds for the convolution and, in that case for all values of h, theoretical scale space functions preserve the ordering among the original densities when they satisfy a location-shift model.

Theorem 2.1. *Suppose that $f_1, \ldots, f_J$ and K are all spherically symmetric densities and the f_j's satisfy the location-shift model i.e., $f_j(\mathbf{x}) = g(\mathbf{x} - \boldsymbol{\mu}_j)$ for some common density g with zero mean and location parameter $\boldsymbol{\mu}_j$. Assume also that the f_j's and K are strictly decreasing functions of the distance from their centers of symmetry. Then, for any positive h as $n \to \infty$, the average misclassification probability of the kernel density estimate based classifier tends to the optimal Bayes risk provided the prior probabilities are equal.*

This theorem explains the reason behind the counter-intuitive behavior of $\Delta(h)$ observed in Figure 1.1. The next theorem throws some light on the behavior of kernel density estimate based classifiers for large sample sizes and large bandwidths when the population densities do not necessarily satisfy symmetry condition.

Theorem 2.2. *Suppose that $f_j(\mathbf{x})$'s are density functions satisfying $\int \|\mathbf{x}\|^6 f_j(\mathbf{x})d\mathbf{x} < \infty$ for all $j = 1, \ldots, J$, and the kernel K is a density with a mode at $\mathbf{0}$ and bounded third derivatives. Then, if the priors are equal, as n, $h \to \infty$ the average misclassification probability of the kernel density estimate based classifier tends to that of a linear classifier given by*

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

Note that when the kernel K is spherically symmetric and a strictly decreasing function of the norm of its argument, the limiting linear classifier obtained in the preceding theorem for a large bandwidth and a large sample size is nearly equivalent to the classifier that classifies an observation $\mathbf{x}$ into the class j_0 that maximizes $\mathbf{x}' E_{f_j}(\mathbf{X}) - (1/2) E_{f_j}(\mathbf{X}\mathbf{X}')$ or minimizes $E_{f_j}(\|\mathbf{x} - \mathbf{X}\|^2)$ for $1 \le j \le J$.

Interestingly, the behavior of the average misclassification probability turns out to be quite different when the prior probabilities are different for different populations. As an example, we consider the same distributions as discussed in Figures 1.1 and 1.2 but now we set the priors at 0.6 and 0.4, respectively, for the two populations. The results obtained are summarized in Figure 2.1. Once again, in some of the cases, the bandwidth that minimizes $\Delta(h)$ (marked by '$*$') and the bandwidth minimizing the $MISE$ for the kernel density estimate (marked by '$\circ$') turn out to be quite different. However, more importantly, $\Delta(h)$ now shows a completely different behavior as h varies. After reaching its minimum value, $\Delta(h)$ increases significantly before becoming flat. Large bandwidths do not seem to be a good choice for the classification problem here. The following theorem describes the behavior of the kernel density estimate based classifiers for large training sample sizes and large bandwidths when the priors are not necessarily equal.

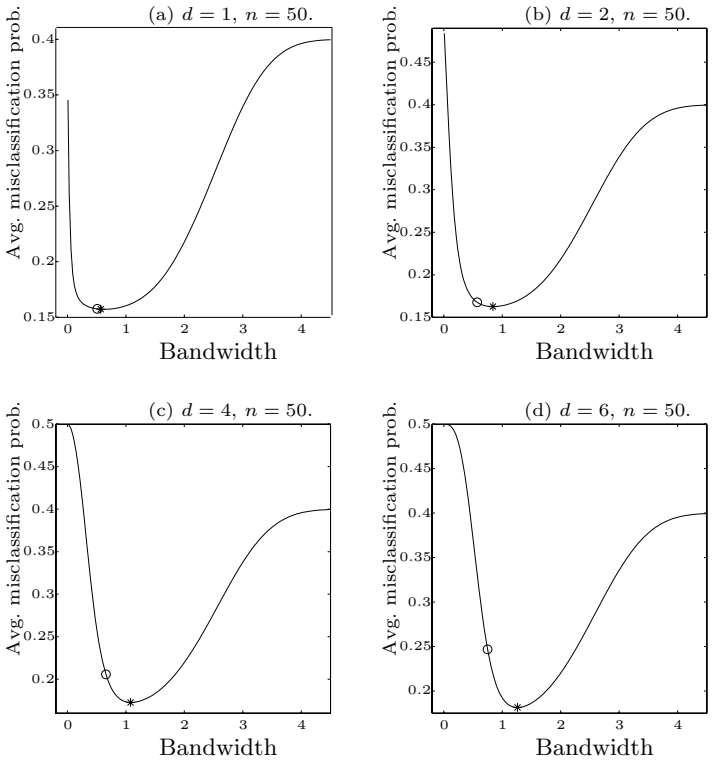

Figure 2.1. True Δ-function and optimal bandwidths (unequal prior cases).

Theorem 2.3. *Suppose that the density functions $f_1, \ldots, f_J$ and the kernel K satisfy the conditions of Theorem 2.2. Assume further that the densities f_j's satisfy the location-shift model in the sense that for all $j = 1, \ldots, J$, $f_j(\mathbf{x}) = g(\mathbf{x} - \boldsymbol{\mu}_j)$ for a common density g with zero mean and location parameter $\boldsymbol{\mu}_j$. Then, as n, $h \to \infty$ the average misclassification probability of kernel density estimate based classifier behaves in the following way.*
(a) *If $\pi_1 = \cdots = \pi_J$, the average misclassification rate of the classifier tends to that of a linear classifier given by*

$$d_l(\mathbf{x}) = \arg\min_j \left[\mathbf{x}' \nabla^2 K(\mathbf{0}) \boldsymbol{\mu}_j - \{ \boldsymbol{\mu}_j{}' \nabla^2 K(\mathbf{0}) \boldsymbol{\mu}_j \}/2 \right].$$

(b) *If there exists a j_0 such that $\pi_{j_0} > \pi_j$ for all $j \neq j_0$, the average misclassification rate of the classifier tends to that of the trivial classifier which classifies all observations to the population j_0. (This also holds for finite n and h tending to infinity.)*
(c) *If there exist m maxima among the prior probabilities, $\pi_{j_1} = \cdots = \pi_{j_m} > \pi_j$ for all $j \notin \{j_1, \ldots, j_m\}$, the average misclassification probability of the classifier*

tends to that of a linear classifier for an m-class problem, where the classes are those which have the maximum prior probability.

Thus, when the prior probabilities for different populations are not equal, one needs to make a careful selection of the bandwidth in order to ensure good performance of kernel density estimate based classifiers.

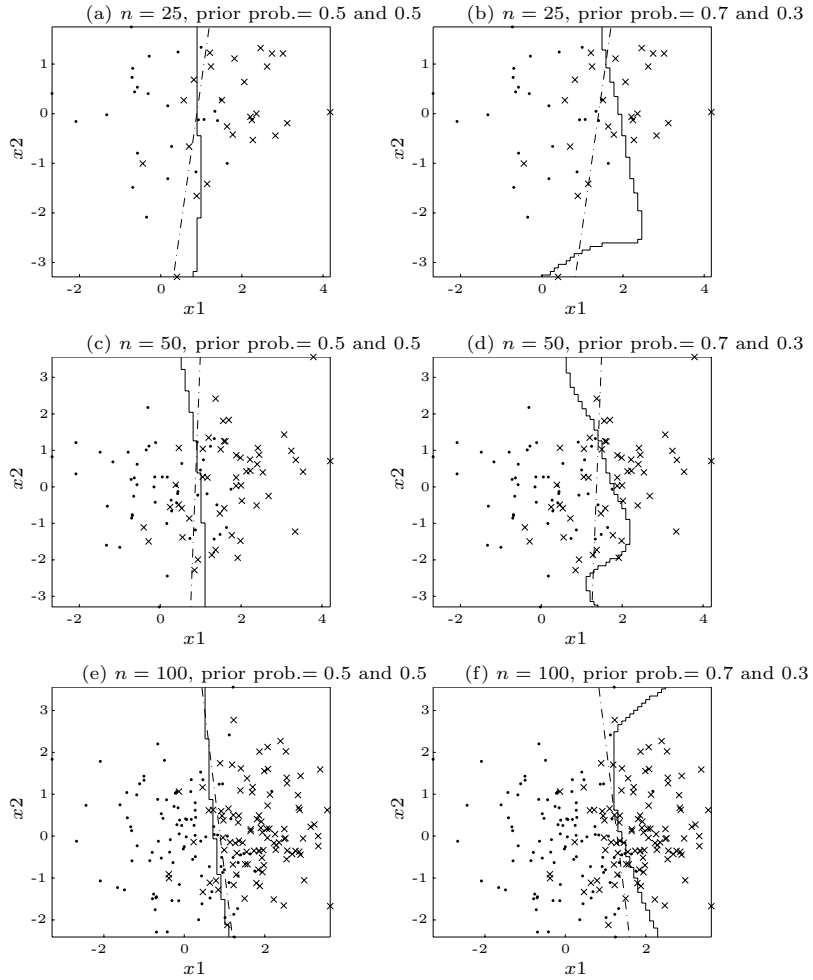

Figure 2.2. Classification boundaries for kernel based and linear classifiers.

Figure 2.2 presents the class boundaries for a two class problem involving spherically symmetric bivariate normal populations with unit variance and location parameters $(0, 0)$ and $(2, 0)$ for the two classes. The dot-dash line gives the class boundary for the usual linear discriminant analysis (LDA) and the continuous solid line gives the boundary for kernel density estimate based classifier where

the optimum bandwidth for classification (h_*) is used to estimate the densities of the two populations. As the population distributions are spherically symmetric and satisfy the location-shift model, LDA performs ideally but, in all these examples, the kernel method also had decent performance. When the priors for the two populations are different (0.7 and 0.3 respectively), the class boundaries for the two methods seem to be quite different but, in equal prior cases, they tend to be the same separating line with increasing sample size. We also know that under this spherically symmetric set up, the misclassification rate for any fixed bandwidth kernel classifier asymptotically converges to the optimal Bayes risk (which is same as the error rate for the standard linear classifier in this case) when the prior probabilities are equal. In Figure 2.3, we have plotted the class

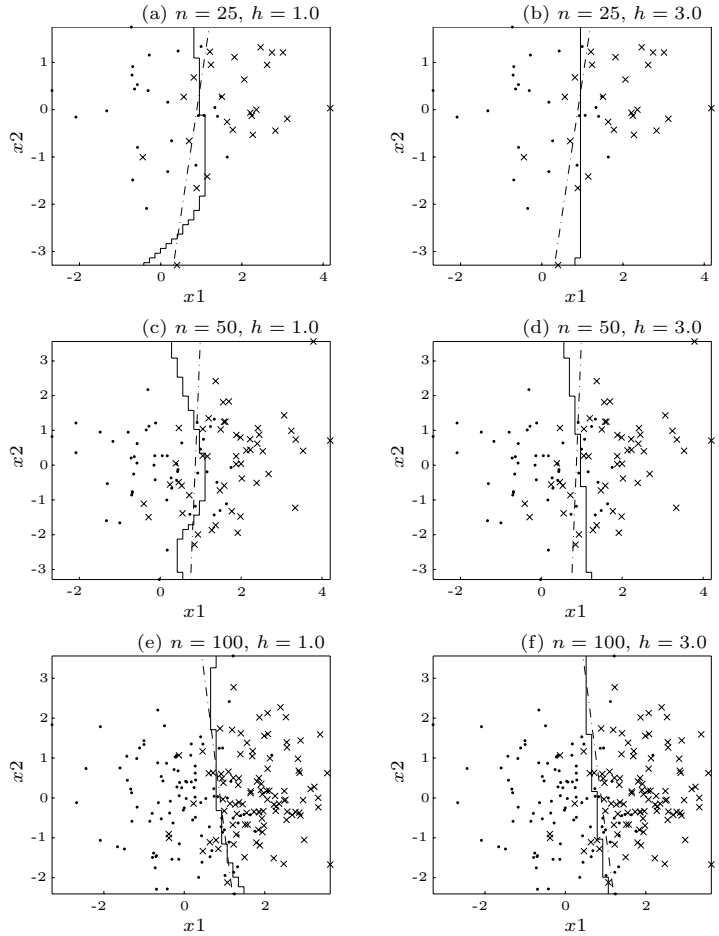

Figure 2.3. Classification boundaries for fixed bandwidth kernel classifiers and linear classifiers.

boundaries for two such classifiers (with $h = 1$ and $h = 3$) in the equal prior case for the same data set discussed above. From this figure it is quite evident that with growing sample size, the class boundaries for kernel-based classifiers converge to the separating line obtained by usual linear discriminant analysis.

For the kernel density estimation problem, there are many different techniques for choosing the bandwidth from the data (see e.g., Stone (1984), Silverman (1986), Hall, Sheather, Jones and Marron (1991), Sheather and Jones (1991) and Wand and Jones (1995)). Some good reviews of bandwidth selection methods in kernel density estimation are available in Jones, Marron and Sheather (1996a, 1996b). While those techniques are quite good for giving low $MISE$ for the density estimate, they may not be appropriate for handling classification problems. We pointed out already that other popular techniques, such as the V-fold cross-validation, also have some serious limitations. In Figure 2.4, we show the performance of such cross validatory techniques for the normal population problems with unequal priors (as in Figure 2.1). Estimated Δ-functions again turn out to be step functions with multiple minima for leave-one-out as well as for 10-fold cross-validation.

3. Data-Based Choice for Bandwidths in Kernel Discriminant Analysis

In this section, we propose and investigate a procedure for choosing bandwidths when a kernel density estimate based classifier is to be used. This proposal has been motivated by our findings reported in the preceding sections. In a J-class discrimination problem, if we use J different bandwidths $h_1, \ldots, h_J$ for the J populations, average misclassification probability is given by

$$\begin{aligned}\Delta(h_1, \ldots, h_J) &= \sum_{j=1}^{J} \pi_j \int P\{\pi_j \hat{f}_{jh_j}(\mathbf{x}) \le \pi_i \hat{f}_{ih_i}(\mathbf{x}) \text{ for some } i \ne j\}\ f_j(\mathbf{x}) d\mathbf{x} \\ &= 1 - \sum_{j=1}^{J} \pi_j \int P\{\pi_j \hat{f}_{jh_j}(\mathbf{x}) > \pi_i \hat{f}_{ih_i}(\mathbf{x}) \text{ for all } i \ne j\}\ f_j(\mathbf{x}) d\mathbf{x} \\ &= 1 - \sum_{j=1}^{J} \pi_j \int \Big[\int \prod_{i \ne j} P\{\pi_i \hat{f}_{ih_i}(\mathbf{x}) < u\}\ g_{jh_j}(u)\ du \Big]\ f_j(\mathbf{x}) d\mathbf{x},\end{aligned}$$

where $g_{jh_j}(\cdot)$ is the p.d.f. of $\pi_j \hat{f}_{jh_j}(\mathbf{x})$. Here, the probability function $P(\cdot)$ does not have a closed form expression. One possibility is to use resampling techniques like bootstrap (see e.g., Efron (1982) and Efron and Tibshirani (1993)) to estimate this probability. But in that case, to compute the misclassification probability at any data point a number of bootstrap samples have to be generated by a leave-one-out method and, for different data points, one has to use different bootstrap samples. As a result, the complexity of the algorithm increases substantially, and this increment is linear in the number of bootstrap samples.

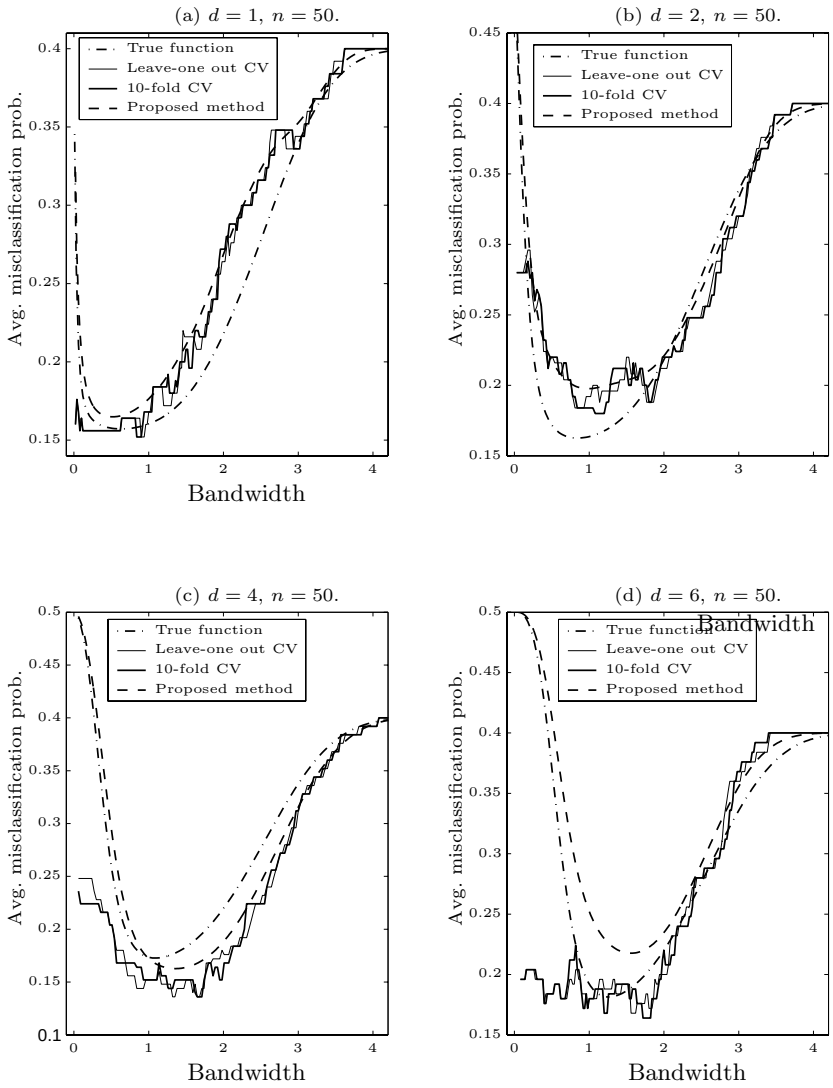

Figure 2.4. Average misclassification probabilities (unequal prior cases).

Generally, a large number of bootstrap samples is required to get a reliable estimate for the probability function, which makes the use of bootstrap approximation in practice very difficult if at all possible.

For large and moderately large samples, we can use normal approximation to the distribution of kernel density estimates and, since a kernel density estimate is a simple average of i.i.d. random variables, there is not much loss of accuracy in such approximation. Let $\mu_{jh_j}(\mathbf{x})$ and $s^2_{jh_j}(\mathbf{x})$ be the mean and the variance of $\hat{f}_{jh_j}(\mathbf{x})$ $(j = 1, \ldots, J)$. Then the average misclassification probability can be

approximated by

$$\psi(h_1,\ldots,h_J) = 1 - \sum_{j=1}^{J} \pi_j \int \Big[\int \prod_{i\neq j} \Phi\Big\{ \frac{u - \pi_i \mu_{ih_i}(\mathbf{x})}{\pi_i s_{ih_i}(\mathbf{x})} \Big\} \times \phi\Big\{ u, \pi_j \mu_{jh_j}(\mathbf{x}), \pi_j s_{jh_j}(\mathbf{x}) \Big\}\, du \Big] f_j(\mathbf{x}) d\mathbf{x},$$

where $\Phi(\cdot)$ is the c.d.f. of standard normal distribution and $\phi(\cdot, \mu, s)$ is the p.d.f. of a normal distribution with mean μ and standard deviation s.

Theorem 3.1. *Suppose that $n_1,\ldots,n_J$ $(N = \sum n_j)$ are the training sample sizes from the J populations, and $h_{n_1},\ldots,h_{n_J}$ are the bandwidths used in kernel estimates of population densities $f_1,\ldots,f_J$, respectively. Further, assume that the densities $f_1,\ldots,f_J$ have bounded third derivatives, and the kernel K is bounded and symmetric about $\mathbf{0}$ satisfying $\int \|\mathbf{y}\|^3 K^2(\mathbf{y}) d\mathbf{y} < \infty$. For every $j \in \{1,\ldots,J\}$, as $N \to \infty$ we assume that $h_{n_j} \to 0$, $h_{n_j}/h_{n_i} \to \gamma_{ji} > 0$ for all i, $n_j h_{n_j}^d \to \infty$ and $n_j/N \to \lambda_j$ such that $0 < \lambda_j < 1$. Then as $N \to \infty$, $|\Delta(h_{1_{n_1}},\ldots,h_{J_{n_J}}) - \psi(h_{1_{n_1}},\ldots,h_{J_{n_J}})| \to 0$, and both $\Delta(h_{1_{n_1}},\ldots,h_{J_{n_J}})$ and $\psi(h_{1_{n_1}},\ldots,h_{J_{n_J}})$ tend to the optimal Bayes risk.*

Thus, if one minimizes $\psi(h_1,\ldots,h_J)$ w.r.t. $h_1,\ldots,h_J$, one has a kernel density estimate based classification rule with asymptotic average misclassification probability equal to the optimal Bayes risk, under suitable regularity conditions.

3.1. Data analytic implementation

In practice, it is not possible to compute ψ as it involves unknown population parameters. Instead, we first go to its sample analogue

$$\psi_N(h_1,\ldots,h_J) = 1 - \sum_{j=1}^{J} \frac{\pi_j}{n_j} \sum_{k=1}^{n_j} \Big[\int \prod_{i\neq j} \Phi\Big\{ \frac{u - \pi_i \mu_{ih_i}(\mathbf{X}_{jk})}{\pi_i s_{ih_i}(\mathbf{X}_{jk})} \Big\} \times \phi\Big\{ u, \pi_j \mu_{jh_j}(\mathbf{X}_{jk}), \pi_j s_{jh_j}(\mathbf{X}_{jk}) \Big\}\, du \Big],$$

where $\mathbf{X}_{jk}$ is the kth observation of the jth class. Even the terms $\mu_{jh_j}(\mathbf{X}_{jk})$ and $s^2_{jh_j}(\mathbf{X}_{jk})$ that appear in the expression above can only be estimated from the available data. In our investigation, we used estimates for them based on kernel density estimates of the population densities, and such estimates were found to yield very good results. For these kernel density estimates, we used the simple least squares cross-validation method (see e.g., Hall (1983), Silverman (1986), Hall and Marron (1987) and Scott (1992)) that looks to minimize $MISE$ for choosing the bandwidths. Since in our numerical study we have used normal kernels, we got closed form expressions for the estimates $\mu^*_{jh_j}(\mathbf{X}_{jk})$ and $s^{*2}_{jh_j}(\mathbf{X}_{jk})$

of $\mu_{jh_j}(\mathbf{X}_{jk})$ and $s^2_{jh_j}(\mathbf{X}_{jk})$, respectively. This led to a further approximation of $\psi_N(h_1,\ldots,h_J)$ by $\psi^*_N(h_1,\ldots,h_J)$, where

$$\psi^*_N(h_1,\ldots,h_J) = 1 - \sum_{j=1}^{J} \frac{\pi_j}{n_j} \sum_{k=1}^{n_j} \Big[\int \prod_{i\neq j} \Phi\Big\{\frac{u-\pi_i\mu^*_{ih_i}(\mathbf{X}_{jk})}{\pi_i s^*_{ih_i}(\mathbf{X}_{jk})}\Big\} \times\phi\left\{u, \pi_j\mu^*_{jh_j}(\mathbf{X}_{jk}), \pi_j s^*_{jh_j}(\mathbf{X}_{jk})\right\} du \Big].$$

The integral appearing in the above expression can be evaluated numerically without much difficulty and to a great degree of accuracy. For computing the estimates $\mu^*_{jh_j}(\mathbf{X}_{jk})$ and $s^*_{jh_j}(\mathbf{X}_{jk})$, we used the leave-one-out strategy and did not use the $\mathbf{X}_{jk}$ in the corresponding kernel density estimate. These estimates are

$$\mu^*_{jh_j}(\mathbf{X}_{jk}) = \frac{1}{n_j-1} \sum_{\substack{l=1\\ l\neq k}}^{n_j} \phi_d\left(\mathbf{X}_{jk}, \mathbf{X}_{jl}, \{h_j^2 + h^2_{\circ j}\}\mathbf{I}_d\right),$$

$$s^{*2}_{jh_j}(\mathbf{X}_{jk}) = \frac{1}{n_j-1}\Big[\Big(\frac{1}{4\pi h_j^2}\Big)^{d/2}\Big\{\frac{1}{n_j-1}\sum_{\substack{l=1\\ l\neq k}}^{n_j} \phi_d\left(\mathbf{X}_{jk}, \mathbf{X}_{jl}, \{0.5h_j^2 + h^2_{\circ j}\}\mathbf{I}_d\right)\Big\} - \mu^{*2}_{jh_j}(\mathbf{X}_{jk})\Big],$$

where $\phi_d(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\{-(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2\}$, and $h_{\circ j}$ is the bandwidth that minimizes estimated $MISE$ of a kernel density estimate of the jth population. For computing $\mu^*_{ih_i}(\mathbf{X}_{jk})$ and $s^{*2}_{ih_i}(\mathbf{X}_{jk})$ $(i\neq j)$, almost identical formulae are used except for the fact that the sum extends over all $1\le l\le n_i$ (i.e., no observation is left out), and the factor $1/(n_j-1)$ gets replaced by $1/n_i$. Note that, unlike the step function obtained in V-fold cross-validation, $\psi^*_N(h_1,\ldots,h_J)$ is a smooth function, and we propose to minimize it over $h_1,\ldots,h_J$ to find optimal bandwidths. In all the examples considered in Figures 1.2 and 2.4, we plotted our ψ^*_N function for $h_1=h_2$ together with the corresponding 10-fold and N-fold (leave-one-out) cross-validation functions as well as the true average misclassification probability functions. Note that, since we considered only normal distribution models where two different populations were just location shifts of each other, a common choice of the bandwidth for different populations was quite justified, and it reduced the computing time significantly. It is quite transparent from the pictures that our proposed criterion function for choosing the optimum bandwidth for classification does a fairly good job in all the examples, and does visibly better than both 10-fold and N-fold cross-validation criteria.

3.2. Results from simulation experiments

In this section, we report some simulation studies that illustrate the performance of our proposed method. To reduce computational complexities, we used a common bandwidth for different populations and minimized $\psi_N^*(h)$ over the single parameter h. In general, this leads to a conservative evaluation of the performance of our method because the use of different bandwidths for different populations will lead to a lower average misclassification probability, at the cost of increased complexity. Note also that if we have different population densities satisfying location shift models (as we do in the simulations), using a common bandwidth for different populations is quite justified if the training sample sizes for different populations happen to be the same.

We start with some two-class problems with normally distributed populations that differ only in their location parameters. To make our examples simpler, we take $\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = \mathbf{I}$ and choose the location parameters $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ in such a way that they differ only in their first co-ordinates. This difference (μ) is taken to be 1, 2 and 3 in our experiments. For each of these examples, we generated 100 sets of observations taking samples of equal sizes (50 or 100) from both the classes. Since the true underlying densities are known, it is possible to compute the true optimum bandwidth minimizing the $MISE$ ($h_\circ$) and that minimizing the average misclassification probability (h_*). Both leave-one-out and 10-fold cross-validation techniques suffer from the problem of having multiple minima when estimating the Δ-function. This makes it difficult to select the optimum bandwidth based on such criteria. In those cases, however, the bandwidth which is largest among the minimizers can be considered, and we denote the bandwidths obtained from leave-one-out and 10-fold cross validation by h_+ and $h_+^{(10)}$, respectively. Averages and standard errors of the corresponding true Δ values of those 100 simulation runs are reported in Tables 3.1 (3.1A and 3.1B) and 3.2 (3.2A and 3.2B). True Δ values are reported for $h_\circ$ and h_* as well. Optimal Bayes errors are also given to facilitate comparison.

The results for normally distributed populations with equal priors for dimensions 2, 4 and 6 are presented in Table 3.1A. In all these examples, the proposed method showed an excellent performance and achieved nearly the true optimum average misclassification rates. Cross validation based methods performed better than $h_\circ$, but they could not match the performance of our proposed bandwidth selection procedure. As far as average misclassification probabilities are concerned, our proposed choice of bandwidth had a slight edge over the cross validation based techniques, but in terms of consistency it substantially outperformed both h_+ and $h_+^{(10)}$. These cross validation based techniques were found to have much higher standard errors as compared to the proposed bandwidth selection procedure.

Table 3.1A. Normal distributions with equal priors.

μ	Bayes risk	d	n	Δ in percentage				
				$h_\circ$	h_*	h_+	$h_+^{(10)}$	Proposed method
1.0	30.85	2	50	33.22	31.77	32.96 (0.242)	32.97 (0.256)	31.81 (0.009)
			100	32.41	31.36	32.18 (0.150)	32.26 (0.148)	31.40 (0.008)
		4	50	37.28	32.62	33.65 (0.198)	33.78 (0.263)	32.67 (0.009)
			100	35.95	31.82	32.40 (0.137)	32.36 (0.107)	31.85 (0.006)
		6	50	40.34	33.38	34.77 (0.321)	34.53 (0.280)	33.44 (0.018)
			100	39.24	32.25	32.83 (0.147)	32.72 (0.088)	32.26 (0.003)
2.0	15.87	2	50	16.98	16.10	17.10 (0.265)	16.79 (0.213)	16.13 (0.005)
			100	16.56	15.92	16.46 (0.149)	16.30 (0.098)	15.96 (0.009)
		4	50	20.73	16.53	17.48 (0.253)	17.40 (0.300)	16.57 (0.009)
			100	19.51	16.18	16.60 (0.091)	16.77 (0.193)	16.20 (0.006)
		6	50	25.17	16.88	17.79 (0.275)	17.43 (0.139)	16.91 (0.005)
			100	23.68	16.37	16.78 (0.076)	16.81 (0.099)	16.38 (0.003)
3.0	6.68	2	50	7.66	6.94	8.05 (0.325)	8.13 (0.324)	6.97 (0.010)
			100	7.37	6.85	7.21 (0.180)	7.46 (0.202)	6.88 (0.008)
		4	50	10.70	7.04	7.97 (0.258)	8.17 (0.339)	7.06 (0.004)
			100	9.74	6.89	7.59 (0.226)	7.54 (0.274)	6.90 (0.004)
		6	50	15.13	7.21	8.31 (0.405)	8.19 (0.385)	7.22 (0.002)
			100	13.80	6.99	7.99 (0.311)	7.64 (0.141)	7.00 (0.002)

Table 3.1B. Double exponential distributions with equal priors.

μ	Bayes risk	d	n	Δ in percentage				
				$h_\circ$	h_*	h_+	$h_+^{(10)}$	Proposed method
1.0	30.33	2	50	34.94	33.08	34.42 (0.152)	34.31 (0.177)	33.36 (0.041)
			100	33.62	31.78	32.71 (0.110)	32.66 (0.142)	31.96 (0.025)
		4	50	40.78	36.38	37.98 (0.214)	37.86 (0.212)	36.70 (0.037)
			100	39.41	34.43	35.34 (0.111)	35.25 (0.113)	34.70 (0.031)
		6	50	43.58	38.10	39.86 (0.231)	39.90 (0.249)	38.59 (0.044)
			100	42.72	35.91	37.25 (0.177)	37.36 (0.177)	36.28 (0.039)
2.0	18.39	2	50	21.35	18.90	20.13 (0.182)	20.14 (0.230)	19.01 (0.028)
			100	20.58	18.47	19.16 (0.090)	19.15 (0.097)	18.64 (0.015)
		4	50	27.73	20.46	22.08 (0.189)	22.15 (0.195)	20.68 (0.027)
			100	26.28	19.46	20.35 (0.084)	20.37 (0.089)	19.58 (0.013)
		6	50	32.46	21.46	23.67 (0.187)	23.69 (0.203)	21.88 (0.037)
			100	31.28	19.89	21.18 (0.148)	21.11 (0.140)	20.17 (0.024)
3.0	11.16	2	50	14.44	11.82	12.94 (0.191)	12.73 (0.169)	12.02 (0.021)
			100	13.85	11.69	12.34 (0.192)	12.34 (0.178)	11.78 (0.009)
		4	50	20.10	12.02	13.62 (0.242)	13.15 (0.171)	12.14 (0.011)
			100	18.92	11.80	12.52 (0.107)	12.68 (0.122)	11.87 (0.006)
		6	50	24.99	12.17	13.90 (0.192)	14.03 (0.200)	12.38 (0.014)
			100	24.06	11.92	12.87 (0.183)	12.67 (0.102)	12.03 (0.007)

Table 3.1B gives the same picture when, instead of normal, we consider double exponential distributions with independent component variables. Even in this case, where the population distributions are not spherically symmetric, the mean and the variance of the kernel density estimate have nice analytic expressions when a normal kernel is used. In all these examples, the proposed method performed quite well. Once again, cross validation based methods performed better than $h_\circ$, but had slightly higher error rates and substantially worse standard errors than our method.

Table 3.2A. Normal distributions with unequal priors.

π_1	Bayes risk	d	n	Δ in percentage				
				$h_\circ$	h_*	h_+	$h_+^{(10)}$	Proposed method
0.6	15.38	2	50	16.73	16.26	17.61 (0.239)	17.55 (0.238)	16.42 (0.030)
			100	16.30	15.94	16.82 (0.148)	16.89 (0.197)	16.10 (0.034)
		4	50	20.35	17.27	18.74 (0.252)	18.73 (0.259)	17.48 (0.027)
			100	19.14	16.63	17.38 (0.098)	17.67 (0.152)	16.70 (0.010)
		6	50	24.68	18.14	19.70 (0.331)	19.44 (0.186)	18.48 (0.033)
			100	23.22	17.22	18.10 (0.155)	18.05 (0.121)	17.38 (0.020)
0.7	13.87	2	50	15.28	15.07	16.48 (0.244)	16.62 (0.295)	15.17 (0.019)
			100	14.82	14.64	15.59 (0.143)	15.71 (0.191)	14.74 (0.020)
		4	50	18.77	16.48	18.38 (0.289)	18.60 (0.377)	16.65 (0.023)
			100	17.61	15.68	16.94 (0.212)	16.84 (0.196)	15.73 (0.006)
		6	50	22.98	17.67	19.31 (0.305)	19.57 (0.380)	18.01 (0.034)
			100	21.58	16.63	18.27 (0.299)	17.89 (0.241)	16.75 (0.015)

Table 3.2B. Double exponential distributions with unequal priors.

π_1	Bayes risk	d	n	Δ in percentage				
				$h_\circ$	h_*	h_+	$h_+^{(10)}$	Proposed method
0.6	18.02	2	50	21.29	19.72	20.57 (0.118)	20.66 (0.180)	19.90 (0.028)
			100	20.50	19.04	19.74 (0.108)	19.78 (0.115)	19.24 (0.039)
		4	50	27.39	22.31	23.60 (0.199)	23.89 (0.270)	22.56 (0.032)
			100	25.95	20.98	22.04 (0.179)	21.93 (0.155)	21.12 (0.020)
		6	50	32.27	24.51	27.15 (0.435)	26.83 (0.340)	25.26 (0.086)
			100	31.14	22.88	23.90 (0.184)	24.09 (0.204)	23.11 (0.031)
0.7	16.86	2	50	20.50	19.56	20.88 (0.232)	20.56 (0.144)	20.04 (0.128)
			100	19.59	18.58	19.30 (0.100)	19.23 (0.102)	18.71 (0.023)
		4	50	26.31	22.65	24.47 (0.305)	24.45 (0.250)	23.15 (0.102)
			100	24.99	21.38	22.69 (0.213)	22.56 (0.202)	21.61 (0.043)
		6	50	30.89	24.62	27.92 (0.430)	27.31 (0.332)	25.22 (0.072)
			100	29.94	23.46	24.66 (0.192)	24.61 (0.169)	23.78 (0.051)

Bandwidth selection is more critical when priors are not equal. To evaluate the performance of the proposed method in such situations, we considered the same examples (as above) but set $\pi_1 = 0.6$ and 0.7, respectively. For $\mu = 1$, 2 and 3, we observed similar results and therefore, instead of reporting all of them, we report the results for $\mu = 2$ only (Tables 3.2A and 3.2B). These results again show good performance of our method as a bandwidth selector for both normal and double exponential populations.

3.3. Results from the analysis of benchmark data

We now demonstrate the performance of our method using two well known data sets. In each of them, we first standardized the data by some appropriate dispersion matrix before applying the kernel density estimation technique. For each given data set, we divided it randomly 1,000 times into two parts to form a training and a test sample. In all the examples and in each random split, we took 40 observations from each of the classes to form the training sample, and the remaining observations were used as the test set. The average of test set misclassification errors over these 1,000 random splits is reported in all cases, along with their corresponding standard errors. We have plotted 10-fold and leave-one-out (N-fold) cross-validation estimates of the average misclassification probability curves (as functions of the bandwidth) for all cases in Figures 3.1 and 3.2. As before, in all cases, the estimated curves are step functions with multiple minima and are not of much help in guiding us in choosing the optimum bandwidth for the classification problem. As in simulated examples, the largest bandwidth that minimizes the estimated average misclassification probability can be used.

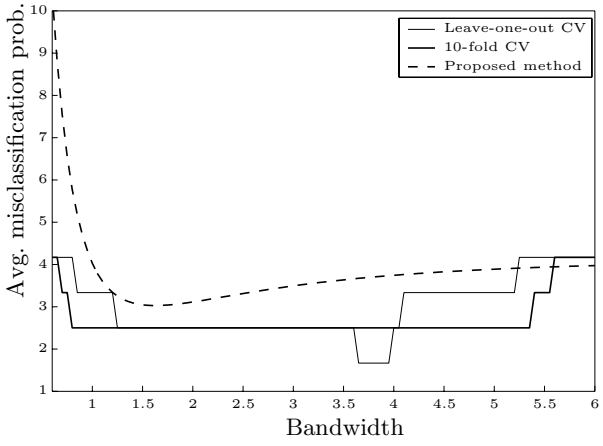

Figure 3.1. Average misclassification probabilities (Iris Data).

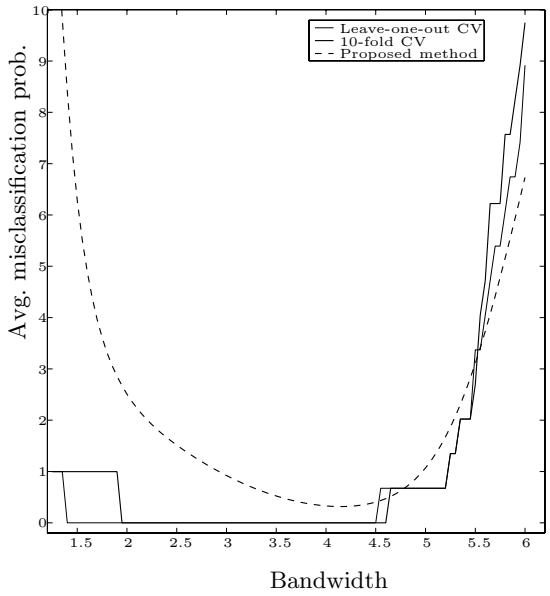

Figure 3.2. Average misclassification probabilities (Wine Data).

We begin with Fisher's (1936) Iris data, where four measurements are taken on each observation coming from one of the three classes: 'Setosa', 'Virginica' and 'Versicolor'. There are 150 observations equally distributed in those three classes. Therefore, it is reasonable to consider it as a problem where the priors are equal. The data points were standardized using the pooled dispersion matrix. Of course, it is possible to use other methods of standardization (see e.g., Coolie and MacEachern (1998)). The traditional linear discriminant analysis is known to perform well in this data set. It led to an error rate of 3.12% with a standard error of 0.09%. This example nicely demonstrates the importance of proper choice of the bandwidth parameter for kernel density estimate based discriminant analysis. When bandwidths for different population densities were estimated by the usual leave-one-out least squares cross-validation technique (see e.g., Silverman (1986) and Scott (1992)) that tries to minimize the estimated $MISE$ for the kernel density estimate, the estimated average misclassification rate turned out to be 5.36% (std. error = 0.11%). Interestingly, this error rate is much higher than that for simple linear discriminant analysis. However, for both cross validation methods with the largest minimizer of error rate estimates, and for our proposed procedure of choosing the bandwidth, we obtained much better performance. Leave one out and 10-fold cross validation techniques could achieve error rates of 3.27% (std. error = 0.11%) and 3.25% (std. error = 0.11%), respectively. Our method of bandwidth selection could further reduce the error rate. The estimate of average misclassification rate turned out to be 3.01% (std. error = 0.09%).

The other data set that we have analyzed is known as 'Wine data'. It contains the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. Chemical analysis determined the quantities of 13 constituents found in each of the three types of wines. Aeberhard, Coomans and de Vel (1994) used this data set to compare the performance of different classifiers in high dimension, and it can be obtained from http://www.uci.ics.edu. This data set contains different number of observations (59, 71 and 48 respectively) in different classes, which justifies the use of unequal priors for different populations. Proportions of observations (belonging to different classes) in the dataset were used to estimate these prior probabilities. Classical linear discriminant analysis misclassified 2.01% of the test set observations (std. error = 0.05%) but the performance of the kernel density estimate based classifier was much better. When the least squares cross-validation technique that minimizes $MISE$ was used to choose the optimum bandwidths, it led to an error rate of 0.85% (std error = 0.03%). Both cross validation methods reduced this error rate to 0.60% (std. error = 0.05%). Our proposed method brought the error rate to 0.48% with a standard error of 0.04%.

Acknowledgement

We are thankful to an associate editor and two referees for their careful reading of earlier versions of the paper and for providing us with several helpful comments.

Appendix: Proofs

Proof of Theorem 2.1. For all $j \in \{1, \ldots, J\}$, $\hat{f}_{jh}(\mathbf{x}) = n^{-1}h^{-d}\sum_{i=1}^{n} K\{(\mathbf{x} - \mathbf{X}_{j_i})/h\}$ is an average of n i.i.d. random variables with finite means and variances. Therefore, for every $\mathbf{x}$ and $h > 0$, as $n \to \infty$, $V\{\hat{f}_{jh}(\mathbf{x})\} \to 0$, and the distribution of $\hat{f}_{jh}(\mathbf{x})$ tends to be degenerate at $E\{\hat{f}_{jh}(\mathbf{x})\}$. Consequently, the asymptotic average misclassification probability of the classification rule based on kernel density estimates will be same as that of the classification rule based on the theoretical scale space functions $E\{\hat{f}_{1h}(\mathbf{x})\}, E\{\hat{f}_{2h}(\mathbf{x})\}, \ldots, E\{\hat{f}_{Jh}(\mathbf{x})\}$ associated with population densities $f_1(\mathbf{x}), \ldots, f_J(\mathbf{x})$ and the kernel K with bandwidth h.

Note that $E\{\hat{f}_{jh}(\mathbf{x})\}$ is a convolution of spherically symmetric density f_j and a spherically symmetric kernel K with bandwidth h. Hence, $E\{\hat{f}_{jh}(\mathbf{x})\}$ is also a spherically symmetric density with $\boldsymbol{\mu}_j$ as the center of symmetry.

Now, choose $\mathbf{x}_1$ and $\mathbf{x}_2$ such that $\|\mathbf{x}_1 - \boldsymbol{\mu}_j\| < \|\mathbf{x}_2 - \boldsymbol{\mu}_j\|$ (i.e., $f_j(\mathbf{x}_1) > f_j(\mathbf{x}_2)$). Consider the hyperplane $\|\mathbf{x} - \mathbf{x}_1\| = \|\mathbf{x} - \mathbf{x}_2\|$. It divides the d-dimensional space into two half-spaces (H^+ and H^-). It is clear that $\mathbf{x}_1$ and $\boldsymbol{\mu}_j$ belong to the same half-space (let us denote it by H^+) and $\mathbf{x}_2$ to the other.

For every point $\mathbf{y} \in H^-$, take $\mathbf{y}^* \in H^+$, to be the image of $\mathbf{y}$ obtained by reflecting it along the hyperplane. Then $f_j(\mathbf{y}^*) > f_j(\mathbf{y})$ and for all $h > 0$, we have $K\{(\mathbf{x}_1 - \mathbf{y}^*)/h\} - K\{(\mathbf{x}_2 - \mathbf{y}^*)/h\} = K\{(\mathbf{x}_2 - \mathbf{y})/h\} - K\{(\mathbf{x}_1 - \mathbf{y})/h\} > 0$. Therefore, whatever may be the value of h,

$$
\begin{aligned}
& h^{-d} \int_{\mathbf{y}^* \in H^+} f_j(\mathbf{y}^*) \left[K\{(\mathbf{x}_1 - \mathbf{y}^*)/h\} - K\{(\mathbf{x}_2 - \mathbf{y}^*)/h\}\right] d\mathbf{y}^* \\
& > h^{-d} \int_{\mathbf{y} \in H^-} f_j(\mathbf{y}) \left[K\{(\mathbf{x}_2 - \mathbf{y})/h\} - K\{(\mathbf{x}_1 - \mathbf{y})/h\}\right] d\mathbf{y}. \\
\Rightarrow\ & h^{-d} \int_{\mathbf{y} \in H^+} f_j(\mathbf{y}) \left[K\{(\mathbf{x}_1 - \mathbf{y})/h\} - K\{(\mathbf{x}_2 - \mathbf{y})/h\}\right] d\mathbf{y} \\
& \quad + h^{-d} \int_{\mathbf{y} \in H^-} f_j(\mathbf{y}) \left[K\{(\mathbf{x}_1 - \mathbf{y})/h\} - K\{(\mathbf{x}_2 - \mathbf{y})/h\}\right] d\mathbf{y} > 0. \\
\Rightarrow\ & E\{\hat{f}_{j_h}(\mathbf{x}_1)\} - E\{\hat{f}_{j_h}(\mathbf{x}_2)\} \\
& = h^{-d} \int_{\mathbf{y} \in R^d} f_j(\mathbf{y}) \left[K\{(\mathbf{x}_1 - \mathbf{y})/h\} - K\{(\mathbf{x}_2 - \mathbf{y})/h\}\right] d\mathbf{y} > 0.
\end{aligned}
$$

So the convolution is also a decreasing function of the distance from its center of symmetry. Now, $f_i(\mathbf{x}) > f_j(\mathbf{x}) \Leftrightarrow \|(\mathbf{x} - \boldsymbol{\mu}_i)\|^2 < \|(\mathbf{x} - \boldsymbol{\mu}_j)\|^2 \Leftrightarrow E\{\hat{f}_{i_h}(\mathbf{x})\} > E\{\hat{f}_{j_h}(\mathbf{x})\}$ since the distribution satisfies the location shift model. Hence, for all $h > 0$, theoretical scale space functions preserve the ordering of the original density functions, and the corresponding classifier based on theoretical scale space functions is the optimal Bayes classifier.

Proposition 2.1. *Suppose that $f(\mathbf{x})$ is such that $\int \|\mathbf{x}\|^6 f(\mathbf{x}) d\mathbf{x} < \infty$ and K is a density with a mode at $\mathbf{0}$ and bounded third derivatives. Then as $h \to \infty$, $E\{\hat{f}_h(\mathbf{x})\} = h^{-d}[K(\mathbf{0}) + (1/2h^2)E_f\{(\mathbf{x} - \mathbf{X})' \nabla^2 K(\mathbf{0})(\mathbf{x} - \mathbf{X})\} + O(h^{-3})]$ and $\operatorname{Var}\{\hat{f}_h(\mathbf{x})\} = (4nh^{2d+4})^{-1}[\operatorname{Var}_f\{(\mathbf{x} - \mathbf{X})' \nabla^2 K(\mathbf{0})(\mathbf{x} - \mathbf{X})\} + O(h^{-1})]$.*

Proof of Proposition 2.1. The expectation and the variance of $\hat{f}_h(\mathbf{x})$ can be written as $E\{\hat{f}_h(\mathbf{x})\} = h^{-d} E[K\{(\mathbf{x} - \mathbf{X}_1)/h\}]$ and $\operatorname{Var}\{\hat{f}_h(\mathbf{x})\} = n^{-1}h^{-2d} \operatorname{Var}_f[K\{(\mathbf{x} - \mathbf{X})/h\}]$. Using a Taylor expansion about $\mathbf{0}$, $K\{(\mathbf{x} - \mathbf{X}_1)/h\}$ can be expressed as

$$
\begin{aligned}
K\{(\mathbf{x} - \mathbf{X}_1)/h\} = {} & K(\mathbf{0}) + (1/2h^2)\{(\mathbf{x} - \mathbf{X}_1)' \nabla^2 K(\mathbf{0})(\mathbf{x} - \mathbf{X}_1)\} \\
& + (1/6h^3) \sum_{i,j,k} Y_{i,j,k}, \quad (\text{since } \nabla K(\mathbf{0}) = 0),
\end{aligned}
$$

where $Y_{i,j,k} = (x_i - X_i)(x_j - X_j)(x_k - X_k) \frac{\partial^3 K(\mathbf{t})}{\partial t_i \partial t_j \partial t_k}\big|_{\mathbf{t} = \boldsymbol{\xi}}$ for some intermediate

vector $\boldsymbol{\xi}$ between $\mathbf{0}$ and $(\mathbf{x}-\mathbf{X})/h$. Therefore

$$\begin{aligned} E_f\left[K\{(\mathbf{x}-\mathbf{X})/h)\}\right] &= K(\mathbf{0}) + (1/2h^2)E_f\{(\mathbf{x}-\mathbf{X})'\nabla^2 K(\mathbf{0})(\mathbf{x}-\mathbf{X})\} \\ &\quad + O(h^{-3}), \\ \mathrm{Var}_f\left[K\{(\mathbf{x}-\mathbf{X})/h)\}\right] &= \mathrm{Var}_f\Big[(1/2h^2)\{(\mathbf{x}-\mathbf{X})'\nabla^2 K(\mathbf{0})(\mathbf{x}-\mathbf{X})\} \\ &\quad + (1/6h^3)\sum_{i,j,k} Y_{i,j,k}\Big] \\ &= (1/4h^4)\mathrm{Var}_f\{(\mathbf{x}-\mathbf{X})'\nabla^2 K(\mathbf{0})(\mathbf{x}-\mathbf{X})\} + O(h^{-5}), \end{aligned}$$

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

Lemma 2.1. *Suppose that f_1 and f_2 are two density functions and $\hat{f}_{1h}$ and $\hat{f}_{2h}$ are their corresponding kernel density estimates. Further assume that f_1, f_2 and K satisfy the conditions of Proposition* 2.1. *Then, for any given $\mathbf{x}$, $\hat{f}_{1h}(\mathbf{x})$ and $\hat{f}_{2h}(\mathbf{x})$ have the following properties.*
(a) *If $\pi_1 = \pi_2 = 1/2$ as n, $h \to \infty$, we have $P\{\hat{f}_{1h}(\mathbf{x}) < \hat{f}_{2h}(\mathbf{x})\} \to 0$ or 1 depending on whether*

$$\begin{aligned} &\mathbf{x}'\nabla^2 K(\mathbf{0})\left\{E_{f_2}(\mathbf{X}) - E_{f_1}(\mathbf{X})\right\} \\ >\ \text{or}\ < &(1/2)\left[E_{f_2}\left\{\mathbf{X}'\nabla^2 K(\mathbf{0})\mathbf{X}\right\} - E_{f_1}\left\{\mathbf{X}'\nabla^2 K(\mathbf{0})\mathbf{X}\right\}\right]. \end{aligned}$$

(b) *If $\pi_1 > \pi_2$, we have $P\{\hat{f}_{1h}(\mathbf{x}) < \hat{f}_{2h}(\mathbf{x})\} \to 0$ as $h \to \infty$.*

Proof of Lemma 2.1. Take $Y_h(\mathbf{x}) = \pi_1\hat{f}_{1h}(\mathbf{x}) - \pi_2\hat{f}_{2h}(\mathbf{x}), \mu_h(\mathbf{x}) = E\{Y_h(\mathbf{x})\}$ and $s_h^2(\mathbf{x}) = \mathrm{Var}\{Y_h(\mathbf{x})\}$. When $\pi_1 = \pi_2 = 1/2$, it is evident from Proposition 2.1 that (i) as $h \to \infty$, the sign of $\mu_h(\mathbf{x})$ and the sign of $\mathbf{x}'\nabla^2 K(\mathbf{0})\{E_{f_2}(\mathbf{X}) - E_{f_1}(\mathbf{X})\} - (1/2)[E_{f_2}\{\mathbf{X}'\nabla^2 K(\mathbf{0})\mathbf{X}\} - E_{f_1}\{\mathbf{X}'\nabla^2 K(\mathbf{0})\mathbf{X}\}]$ will eventually be the same, and (ii) $s_h^2(\mathbf{x})/\mu_h^2(\mathbf{x}) \to 0$ as n, $h \to \infty$.

Now, by Chebychev's inequality $P\{Y_h(\mathbf{x}) \le 0\} \ge \mu_h^2(\mathbf{x})/\{\mu_h^2(\mathbf{x}) + s_h^2(\mathbf{x})\}$ when $\mu_h(\mathbf{x}) \le 0$ and $P\{Y_h(\mathbf{x}) \le 0\} \le s_h^2(\mathbf{x})/\{\mu_h^2(\mathbf{x}) + s_h^2(\mathbf{x})\}$ when $\mu_h(\mathbf{x}) > 0$. As n, $h \to \infty$, the right side tends to 1 and 0, respectively, in the first and the second cases. Therefore, $P\{Y_h(\mathbf{x}) \le 0\}$ also tends to 1 and 0 in the respective cases. On the other hand, when $\pi_1 > \pi_2$, as $h \to \infty$, $\mu_h(\mathbf{x})$ remains positive and $s_h^2(\mathbf{x})/\mu_h^2(\mathbf{x}) \to 0$. Therefore, by above inequality, $\lim_{h\to\infty} P\{Y_h(\mathbf{x}) \le 0\} = 0$.

Proof of Theorem 2.2. For $1 \le i \ne j \le J$, let A_{ij}^h be the event that $\hat{f}_{ih}(\mathbf{x}) - \hat{f}_{jh}(\mathbf{x}) > 0$. Clearly, $P(A_{ij}^h) + P(A_{ji}^h) = 1$ and $\sum_{j:\ j\ne i} P(A_{ij}^h) - (J-2) \le P\{\bigcap_{j:\ j\ne i} A_{ij}^h\} \le \min_{j\ne i} P(A_{ij}^h)$. Now, it is easy to see that, for any i, $P\{\bigcap_{j:\ j\ne i} A_{ij}^h\} \to 1$ iff $P(A_{ij}^h) \to 1$ for all $j \ne i$. The proof of the theorem then follows from the first part of Lemma 2.1.

Proof of Theorem 2.3. Since the population densities satisfy the location shift model (with location parameter $\boldsymbol{\mu}_j$ for the jth class), for $i \ne j$, $E_{f_i}(\mathbf{X}) -$

$E_{f_j}(\mathbf{X}) = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ and $E_{f_i}\{\mathbf{X}'\nabla^2 K(\mathbf{0})\mathbf{X}\} - E_{f_j}\{\mathbf{X}'\nabla^2 K(\mathbf{0})\mathbf{X}\} = \boldsymbol{\mu}_i'\nabla^2 K(\mathbf{0})\boldsymbol{\mu}_i - \boldsymbol{\mu}_j'\nabla^2 K(\mathbf{0})\boldsymbol{\mu}_j$. The proof of (a) is now immediate from Theorem 2.2. Proofs of (b) and (c) follow from Lemma 2.1, using the same logic as in the proof of Theorem 2.2.

Lemma 3.1. *Assume that the density function f has bounded third derivatives and the kernel K is symmetric about $\mathbf{0}$ with $\int \|\mathbf{y}\|^3 K^2(\mathbf{y})d\mathbf{y} < \infty$. Then, as $h \to 0$, $E\{\hat{f}_h(\mathbf{x})\} = f(\mathbf{x}) + O(h^2)$ and $\mathrm{Var}\,\{\hat{f}_h(\mathbf{x})\} = (nh^d)^{-1}\{\beta f(\mathbf{x}) + O(h^2)\}$, where $\beta = \int K^2(\mathbf{y})d\mathbf{y}$.*

Proof of Lemma 3.1.

$$\begin{aligned} E\{f_h(\mathbf{x})\} &= h^{-d}\int K\{(\mathbf{x}-\mathbf{X})/h\}\ f(\mathbf{X})\ d\mathbf{X} \\ &= \int K(\mathbf{y})\ f(\mathbf{x}-h\mathbf{y})\ d\mathbf{y} \\ &= \int K(\mathbf{y})\Big[f(\mathbf{x}) - h\{\mathbf{y}'\nabla f(\mathbf{x})\} + \frac{h^2}{2}\{\mathbf{y}'\nabla^2 f(\mathbf{x})\mathbf{y}\} + \frac{h^3}{3!}\sum_{i,j,k} Z_{i,j,k}\Big]d\mathbf{y}, \end{aligned}$$

where $Z_{i,j,k} = y_i y_j y_k\ \frac{\partial^3 f(\mathbf{t})}{\partial t_i \partial t_j \partial t_k}|_{\mathbf{t}=\boldsymbol{\xi}}$ for some intermediate vector $\boldsymbol{\xi}$ between $\mathbf{x}$ and $\mathbf{x} - h\mathbf{y}$. Therefore,

$$E\{\hat{f}_h(\mathbf{x})\} = f(\mathbf{x}) + \frac{h^2}{2}\int \{\mathbf{y}'\nabla^2 f(\mathbf{x})\mathbf{y}\}\ d\mathbf{y} + o(h^2) = f(\mathbf{x}) + O(h^2).$$

Similarly,

$$\begin{aligned} E\left[h^{-2d}K^2\{(\mathbf{x}-\mathbf{X})/h)\}\right] &= h^{-d}\int K^2(\mathbf{y})\ [f(\mathbf{x}) - h\{\mathbf{y}'\nabla f(\mathbf{x})\} \\ &\qquad + (h^2/2)\{\mathbf{y}'\nabla^2 f(\mathbf{x})\mathbf{y}\} + (h^3/3)\sum_{i,j,k} Z_{i,j,k}]\ d\mathbf{y} \\ &= h^{-d}[\beta f(\mathbf{x}) + O(h^2)], \quad \text{where } \beta = \int K^2(\mathbf{y})d\mathbf{y}. \end{aligned}$$

$$\Rightarrow \mathrm{Var}\,\{\hat{f}_h(\mathbf{x})\} = n^{-1}\mathrm{Var}\,\{h^{-d}K((\mathbf{x}-\mathbf{X})/h)\} = n^{-1}h^{-d}\{\beta f(\mathbf{x}) + O(h^2)\}.$$

Lemma 3.2. *Suppose that $n_1, \ldots, n_J$ $(N = \sum n_j)$ are the training sample sizes from J competing populations and $h_{n_1}, \ldots, h_{n_J}$ are the bandwidths used in kernel estimates of population densities $f_1, \ldots, f_j$, respectively. Further, assume that the densities $f_1, \ldots, f_j$ and the kernel K satisfy the conditions of Lemma 3.1. For every $j \in \{1, \ldots, J\}$ and $N \to \infty$, we also assume that $h_{n_j} \to 0$, $h_{n_j}/h_{n_i} \to \gamma_{ji} > 0$ for all i, $n_j h_{n_j}^d \to \infty$ and $n_j/N \to \lambda_j$ such that $0 < \lambda_j < 1$. For $\mathbf{x} \in R^d$, take $X_1, \ldots, X_J$ as independently distributed normal variates with $E(X_i) = \pi_i m_{ih_{n_i}} = E\{\pi_i \hat{f}_{ih_{n_i}}(\mathbf{x})\}$ and $\mathrm{Var}\,(X_i) = \pi_i^2 s_{ih_{n_i}}^2 = \mathrm{Var}\,\{\pi_i \hat{f}_{ih_{n_i}}(\mathbf{x})\}$.*

Then, whatever be $\mathbf{x}$*, we have*

$$\lim_{N\to\infty}\left| P(X_1 > X_i, \ \textit{for all} \ \ i \neq 1) - \prod_{i\neq 1}\Phi\left(\frac{\pi_1 m_{1h_{n_1}} - \pi_i m_{ih_{n_i}}}{\pi_i s_{ih_{n_i}}}\right)\right| = 0.$$

Proof of Lemma 3.2.

$$\begin{aligned} P\{X_i < X_1 \text{ for all } i\neq 1\} &= \int P\left\{X_i < x \text{ for all } i \neq 1\right\}\ g(x)\ dx \\ &\qquad\qquad [\ g(\cdot) \text{ being the p.d.f. of } X_1] \\ &= \int \prod_{i\neq 1}\Phi\left(\frac{x - \pi_i m_{ih_{n_i}}}{\pi_i s_{ih_{n_i}}}\right) g(x)dx \\ &= E_g\left\{\prod_{i\neq 1}\Phi\left(\frac{x - \pi_i m_{ih_{n_i}}}{\pi_i s_{ih_{n_i}}}\right)\right\}. \end{aligned}$$

Let $\theta_N(x) = \prod_{i\neq 1}\Phi[(x - \pi_i m_{ih_{n_i}})/(\pi_i s_{ih_{n_i}})]$. Using a Taylor expansion about $\pi_1 m_{1h_{n_1}}$, $\theta_N(x)$ can be expressed as $\theta_N(x) = \theta_N(\pi_1 m_{1h_{n_1}}) + (x - \pi_1 m_{1h_{n_1}})\ \theta_N^{'}(\xi)$ for some value ξ that lies between $\pi_1 m_{1h_{n_1}}$ and x. We have $(x-\pi_1 m_{1h_{n_1}})\theta_N^{'}(\xi) = \sum_{j=2}^{J}[(x-\pi_1 m_{1h_{n_1}})/(\pi_j s_{jh_{n_j}})]\beta_{jh_{n_j}}(\xi)$, where $\beta_{jh_{n_j}}(\xi) = \phi[(\xi-\pi_j m_{jh_{n_j}})/(\pi_j s_{jh_{n_j}})]$ $\prod_{k:k\neq 1,j}\Phi[(\xi - \pi_k m_{kh_{n_k}})/(\pi_k s_{kh_{n_k}})]$. Then

$$\begin{aligned} & E\left(\left|\frac{x - \pi_1 m_{1h_{n_1}}}{\pi_j s_{j_{h_{n_1}}}}\ \beta_{jh_{n_j}}(\xi)\right|\right) \\ &\leq E\left\{\left|\frac{x - \pi_1 m_{1h_{n_1}}}{\pi_j s_{jh_{n_j}}}\ \phi\left(\frac{\xi - \pi_j m_{jh_{n_j}}}{\pi_j s_{jh_{n_j}}}\right)\right|\right\} \quad (\text{since } |\Phi(\cdot)| \leq 1) \\ &\leq E^{1/2}\left(\frac{x - \pi_1 m_{1h_{n_1}}}{\pi_j s_{jh_{n_j}}}\right)^2 E^{1/2}\left\{\phi^2\left(\frac{\xi - \pi_j m_{jh_{n_j}}}{\pi_j s_{jh_{n_j}}}\right)\right\} \\ &= (\pi_1^2/\pi_j^2)(s_{1h_{n_1}}^2/s_{jh_{n_j}}^2)E^{1/2}\left\{\phi^2\left(\frac{\xi - \pi_j m_{jh_{n_j}}}{\pi_j s_{jh_{n_j}}}\right)\right\}. \end{aligned}$$

As $N \to \infty$, under the given conditions, both $s_{1h_{n_1}}^2$ and $s_{jh_{n_j}}^2$ tend to 0 (from Lemma 3.1) but $s_{1h_{n_1}}^2/s_{jh_{n_j}}^2$ tends to a constant $c_{1j} > 0$. Therefore, as $N \to \infty$, $\phi^2[(\xi - \pi_j m_{jh_{n_j}})/(\pi_j s_{jh_{n_j}})] \to 0$ which implies $E\{\phi^2[(\xi - \pi_j m_{jh_{n_j}})/(\pi_j s_{jh_{n_j}})]\}$ $\to 0$ (by the Dominated Convergence Theorem). This, in turn, implies that $E|(x - \pi_1 m_{1h_{n_1}})\ \theta_N^{'}(\xi)| \to 0$.

Proof of Theorem 3.1. For a J-class problem, $\psi(h_{n_1}, \ldots, h_{n_J})$ is given by

$$\psi(h_{n_1}, \ldots, h_{n_J}) = 1 - \sum_{j=1}^{J}\pi_j \int\Big[\int\prod_{i\neq j}\Phi\Big\{\frac{u - \pi_i m_{ih_{n_i}}(\mathbf{x})}{\pi_i s_{ih_{n_i}}(\mathbf{x})}\Big\}$$

$$\times \phi\left\{u, \pi_j m_{jh_{n_j}}(\mathbf{x}), \pi_j s_{jh_{n_j}}(\mathbf{x})\right\} du\Big] f_j(\mathbf{x}) d\mathbf{x},$$

where $\pi_j m_{jh_{n_j}}(\mathbf{x})$ and $\pi_j^2 s^2_{jh_{n_j}}(\mathbf{x})$ are the mean and the variance of $\pi_j \hat{f}_{ih_{n_i}}(\mathbf{x})$. Lemma 3.2 implies that, as $N \to \infty$, for all j

$$\left| \int \left[\prod_{i \neq j} \Phi\left\{ \frac{u - \pi_i m_{ih_{n_i}}(\mathbf{x})}{\pi_i s_{ih_{n_i}}(\mathbf{x})} \right\} \phi\left\{u, \pi_j m_{jh_{n_j}}(\mathbf{x}), \pi_j s_{jh_{n_j}}(\mathbf{x})\right\} \right] du \right.$$
$$\left. - \prod_{i \neq j} \Phi\left\{ \frac{\pi_j m_{jh_{n_j}}(\mathbf{x}) - \pi_i m_{ih_{n_i}}(\mathbf{x})}{\pi_i s_{ih_{n_i}}(\mathbf{x})} \right\} \right| \to 0.$$

Also, from Lemma 3.1, for every j, as $N \to \infty$,

$$\prod_{i \neq j} \Phi\left\{ \frac{\pi_j m_{jh_{n_j}}(\mathbf{x}) - \pi_i m_{ih_{n_i}}(\mathbf{x})}{\pi_i s_{ih_{n_i}}(\mathbf{x})} \right\} \to \begin{cases} 1, & \text{if } \pi_j f_j(\mathbf{x}) > \pi_i f_i(\mathbf{x}) \text{ for all } i \neq j, \\ 0, & \text{otherwise.} \end{cases}$$

and this means $\psi(h_{n_1}, \ldots, h_{n_J}) \to 1 - \sum_{j=1}^{J} \pi_j \int_{f_j > f_i \ \forall i \neq j} f_j(\mathbf{x}) d\mathbf{x}$ (by the Dominated Convergence Theorem), which is the optimal Bayes risk.

In view of the asymptotic orders of $E\{\hat{f}_{jh_{n_j}}(\mathbf{x})\}$ and $\text{Var}\{\hat{f}_{jh_{n_j}}(\mathbf{x})\}$ obtained in Lemma 3.1, Lindeberg's condition for the Multivariate Central Limit Theorem holds for

$$\left\{ \frac{\hat{f}_{1h_{n_1}}(\mathbf{x}) - m_{1h_{n_1}}(\mathbf{x})}{s_{1h_{n_1}}(\mathbf{x})}, \frac{\hat{f}_{2h_{n_2}}(\mathbf{x}) - m_{2h_{n_2}}(\mathbf{x})}{s_{2h_{n_2}}(\mathbf{x})}, \ldots, \frac{\hat{f}_{Jh_{n_J}}(\mathbf{x}) - m_{Jh_{n_J}}(\mathbf{x})}{s_{Jh_{n_J}}(\mathbf{x})} \right\}$$

as $N \to \infty$, for every given $\mathbf{x}$. This implies that

$$\Big| P\{\pi_j \hat{f}_{jh_j}(\mathbf{x}) > \pi_i \hat{f}_{ih_i}(\mathbf{x}) \text{ for all } i \neq j\}$$
$$- \int \prod_{i \neq j} \Phi\Big\{ \frac{u - \pi_i m_{ih_{n_i}}(\mathbf{x})}{\pi_i s_{ih_{n_i}}(\mathbf{x})} \Big\} \phi\Big\{u, \pi_j m_{jh_{n_j}}(\mathbf{x}), \pi_j s_{jh_{n_j}}(\mathbf{x})\Big\} du \Big| \to 0$$

as $N \to \infty$ using the results in Bhattacharya and Ranga Rao (1976, pp.6-23, Section 2) on uniform convergence to multivariate normal probabilities for convex sets with boundaries having zero Lebesgue measure. Finally, by the Dominated Convergence Theorem, we have $|\Delta(h_{n_1}, \ldots, h_{n_J}) - \psi(h_{n_1}, \ldots, h_{n_J})| \to 0$ as $N \to \infty$.

References

Aeberhard, S., Coomans, D. and de Vel, O. (1994). Comparative analysis of statistical pattern recognition methods in high dimensional settings. *Pattern Recognition* **27**, 1065-1077.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

Bensmail, H. and Bozdogan, H. (2002). *Model-Based Kernel Discriminant Analysis with Optimal Scaling*. In press for The Institute of Statistical Mathematics (ISM) in Japan, Springer-Verlag, Tokyo.

Bhattacharya, R. N. and Ranga Rao, R. (1976). *Normal Approximation and Asymptotic Expansions*. Wiley, New York.

Bose, S. (1996). Classification using splines. *Comput. Statist. Data Anal.* **22**, 505-525.

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

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

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

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*. Research Studies Press, Letchworth.

Dasarathy, B. V. (1991). *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society, Washington.

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

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia.

Efron, B. and Tibshirani, R. (1993). *An Introduction to Bootstrap*. Chapman and Hall, New York.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7**, 179-188.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. Academic Press, New York.

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

Hall, P. and Marron, J. S. (1987). Extent to which least squares cross validation minimizes integrated square error in nonparametric density estimation. *Probab. Theory Related Fields* **74**, 567-581.

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

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.

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

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

Hastie, T. and Tibshirani, R. (1996). Discriminant adaptive nearest neighbor classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **18**, 607-616.

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

Hills, M. (1966). Allocation rules and their error rates (with discussion). *J. Roy. Statist. Soc. Ser. B* **28**, 1-31.

James, M. (1985). *Classification Algorithms*. Wiley, New York.

Jones, M. C., Marron, J. S. and Sheather, S. J. (1996a). Progress in data-based bandwidth selection for kernel density estimation. *Comput. Statist.* **11**, 337-381.

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

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

Kooperberg, C., Bose, S. and Stone, C. J. (1997). Polychotomous regression. *J. Amer. Statist. Assoc.* **92**, 117-127.

Lachenbruch, P. A. and Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics* **10**, 1-11.

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

Loh, W.-Y. and Shih, Y.-S. (1997). Split selection methods for classification trees. *Statist. Sinica* **7**, 815-840.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, London.

McLachlan. G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.

Mosteller, F. and Wallace, D. L. (1963). Inference in an authorship problem. *J. Amer. Statist. Assoc.* **58**, 275-309.

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

Rao, C. R. (1973). *Linear Statistical Inference*. Wiley, New York.

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

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

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

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

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

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

Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203, B. T. Road, Calcutta-700108, India.

E-mail: res9812@isical.ac.in

Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203, B. T. Road, Calcutta-700108, India.

E-mail: probal@isical.ac.in

(Received September 2002; accepted July 2003)